JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

Testangst und deren Zusammenhang mit Testleistung: Effekte von Messzeitpunkt, Instruktion und funktionaler Bewertung der Angst

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Philosophie des Fachbereiches 06

der Justus-Liebig-Universität Gießen

vorgelegt von

Michael Konrad Ott

aus Frankfurt am Main

2017

Dekan:	Prof. Dr. Dr. Jürgen Hennig
1. Berichterstatter/in:	Prof. Dr. Martin Kersting
2. Berichterstatte/in:	Prof. Dr. Joachim Stiensmeier-Pelster
Tag der Disputation:	21.06.2017

Meiner Frau Anna-Katharina

Danksagung

Mein erster Dank geht an Prof. Dr. Martin Kersting für die hervorragende Betreuung dieser Arbeit. Besonders bedanken möchte ich mich für die zahllosen (und immer zeitnahen) Anregungen, Rückmeldungen und Reflektionen sowie das hohe Maß an kreativem Freiraum, das mir gewährt wurde.

Herzlich bedanken möchte ich mich bei Prof. Dr. Joachim Stiensmeier-Pelster für seine Zweitgutachtertätigkeit und für die äußerst hilfreichen Diskussionen und Anregungen, insbesondere zu Studie 1. Ebenfalls möchte ich mich bei Dr. Claudia Schöne bedanken für ihre Rückmeldungen zur Konzeption und Interpretation von Studie 1.

Ein Dank geht auch an meine Kolleginnen in der Abteilung für Psychologische Diagnostik, Dr. Anna-Sophie Ulfert, Dr. Carolin Palmer und M. Sc. Pascale Bothe sowie an meinen Mitdoktoranden Dr. Dennis Beermann für den hilfreichen Austausch, die vielen Diskussionen und Rückmeldungen. Darüber hinaus danke ich Eva Tuschen und Christin Köhler, die die Daten für Studie 2 im Rahmen ihrer Masterarbeiten erhoben haben. Ein Dank geht außerdem an die Hilfskräfte und Praktikantinnen der Abteilung, die die Durchführung von Studie 3 unterstützt haben.

Ein besonderes Dankeschön geht an die Personen, die als „Fachfremde“ Teile dieser Arbeit gegengelesen und wertvolle Rückmeldungen gegeben haben: meine Geschwister Raphaela und Benjamin Ott sowie Jonas Blum, Julia Pott, Manuel Elsässer und Kevin Günthner. Ein großer Dank geht auch an Thomas Gatzka, der parallel an seiner Dissertation gearbeitet hat, für die unzähligen Gespräche und Rückmeldungen zu meiner Arbeit sowie manches motivierende Wort.

Ich danke meinen Eltern, Marijana und Jakob, die mir mein Studium ermöglicht und mich auch während meiner Promotionsphase immer voll unterstützt haben. Von Herzen danken möchte ich meiner Frau Anna-Katharina, die sämtliche Höhen und Tiefen in der Entstehung dieser Arbeit miterlebt hat und mir mit ihrer Liebe, Kraft und Geduld immer zur Seite stand – ihr ist diese Arbeit gewidmet.

Vielen Dank!

Zusammenfassung

Die vorliegende Arbeit befasst sich mit dem negativen Zusammenhang von Testängstlichkeit (als trait) und Testangst (als state) mit Leistungen in Prüfungen und Tests. Drei in diesem Forschungsfeld bislang offene Fragen wurden behandelt.

Studie 1 basierte auf Befunden, dass Testängstlichkeit bzw. -angst stärker mit Leistung korreliert, wenn sie nicht vor, sondern nach einem Test erfasst wird. In einem Online-Experiment wurden N = 152 Studierende vor und nach einem Intelligenztest nach ihrer Testangst befragt. Die subjektive Leistung mediierte den Effekt der objektiven Leistung auf die nach dem Test berichtete Testangst. Die Ergebnisse legen nahe, dass Testangst womöglich eine Folge von (subjektiv schlechter) Leistung ist. Die objektive Leistung hatte jedoch auch einen direkten Effekt auf Testangst, was Raum lässt für die alternative Interpretation (Testangst verursacht Leistungseinbußen).

Studie 2 basierte auf der Frage, inwiefern Testangst den Effekt des „stereotype threat“ (STT) erklären kann. Einer Stichprobe von N = 168 männlichen und weiblichen Schülern wurde in einem Experiment ein numerischer Intelligenztest mit einer von drei verschiedenen Testinstruktionen vorgelegt. Durch diese Testinstruktionen sollte der STT und die evaluative Testatmosphäre manipuliert werden. Zwar zeigten sich Geschlechtsunterschiede in der Leistung in Abhängigkeit der Bedingungen, jedoch konnten diese entgegen der Erwartung nicht durch Testangst erklärt werden.

Studie 3 befasste sich mit leistungsförderlichen Effekten von Testangst („anxiety motivation“). Eine studentische Stichprobe (N = 496) bearbeitete in einem Experiment einen Intelligenztest. Eine reappraisal-Manipulation sollte eine leistungsförderliche Bewertung von Angst induzieren. In der Treatmentgruppe zeigte sich erwartungsgemäß eine schwächere Korrelation von Testangst und Testleistung als in der Kontrollgruppe, die weiteren Befunde stützten jedoch die Wirksamkeit der Manipulation nicht. Die Ergebnisse liefern darüber hinaus Hinweise, dass der negative Zusammenhang von Testängstlichkeit bzw. Testangst und Leistungskriterien durch „anxiety motivation“ moderiert wird – je eher Testängstlichkeit bzw. Testangst als förderlich empfunden wurde, desto schwächer war deren negativer Effekt.

Die Ergebnisse weisen darauf hin, dass sich zukünftige Forschung stärker damit befassen sollte, wann und wie Testangst erfasst wird. Zudem sollte genauer untersucht werden, welche gewollten und ungewollten Effekte einzelne Elemente von Testinstruktionen nach sich ziehen. Außerdem sprechen die Ergebnisse dafür, dass die (motivierende) Bewertung von Testangst stärker berücksichtigt und getrennt vom „bloßen“ Erleben von Testangst analysiert werden sollte.

Schlüsselwörter: Prüfungsangst, Testängstlichkeit, Testangst, Leistung, Messzeitpunkt, Testistruktion, stereotype threat, anxiety motivation

Abstract

The present dissertation concentrates on the negative relationship between test anxiety (trait and state) and performance in exams and tests. Three research questions which were so far unanswered in this field of study were investigated.

Study 1 was based on findings that test anxiety correlates stronger with performance when it is measured after, not before a test. In an online-experiment, $N = 152$ university students were asked to indicate their test anxiety before and after a test. Perceived performance mediated the effect of actual performance on test anxiety measured after the test. Results indicate that test anxiety might be a consequence of (subjectively low) performance. Yet, actual performance still had a direct effect on test anxiety, which leaves space for the alternative causal interpretation (test anxiety causes performance decrements).

Study 2 addressed the question, whether test anxiety explains the effects of "stereotype threat" (STT). A sample of $N = 168$ male and female school students was given a numerical intelligence test with one out of three different test instructions. These test instructions should manipulate STT as well as the evaluative atmosphere of the test. Gender differences in performance were observed depending on the different experimental conditions. However, these differences could not be explained by test anxiety.

Study 3 investigated facilitating effects of test anxiety ("anxiety motivation"). In an experiment, a sample of university students ($N = 496$) completed an intelligence test. A reappraisal-manipulation was conducted to induce a facilitating appraisal of anxiety. In line with expectations, there was a weaker correlation between test anxiety and test performance in the treatment group compared to the control group, but further results did not support the effectiveness of the manipulation. Additionaly, the results indicate that the negative relationship between test anxiety and performance criteria was moderated by "anxiety motivation" – the negative effects of test anxiety diminished when test anxiety was believed to facilitate performance.

The results suggest that future studies should take into account when and how test anxiety is being measured. Furthermore, desired and undesired effects of different elements of test instructions should be scrutinized more thoroughly. Moreover, the results propose a stronger attention towards (motivating) appraisals of test anxiety, which should be analyzed separately from the "mere" experience of test anxiety.

Key words: test anxiety, performance, measurement time, test instruction, stereotype threat, anxiety motivation

Einleitung

Prüfungssituationen begleiten Menschen ihr gesamtes Leben lang. Prüfungen und Tests regulieren in vielen Gesellschaften den Zugang zu bestimmten Bildungsabschlüssen, zu Qualifikationen, zu bestimmten beruflichen Laufbahnen oder Positionen. Da sie in erheblichem Maße über die Entwicklungswege eines Individuums und ganzer Gruppen entscheiden, sind sie für die geprüften oder getesteten Personen häufig mit Stress verbunden, dessen Ausmaß zuweilen gravierend ist. Spätestens seit der Arbeit von Mandler und Sarason (1952) beschäftigt sich die Psychologie intensiv mit der Frage, welche Rolle Angst in einer Test- oder Prüfungssituation für die darin erzielte Leistung spielt. Dies ist auch Gegenstand und Leitfrage der vorliegenden Arbeit.

Prüfungen und prüfungsähnliche Situationen (zur Eingrenzung dieser Situationsklassen siehe Abschnitt 1.1) treten im Verlauf des gesamten Lebens auf. Insbesondere aber in der ersten Lebenshälfte konzentrieren sie sich, nämlich in der Schulzeit, in der beruflichen Ausbildung und im Studium, wenn Prüfungen biographische Marksteine des individuellen Lebenslaufs bilden. Die hohe Bedeutsamkeit von Prüfungen erklärt gleichsam auch das große Interesse innerhalb der Psychologie für das Erleben und Verhalten in Prüfungssituationen. Obwohl die Verwendung unterschiedlicher Definitionen und der Einsatz verschiedenster Messverfahren präzise Aussagen über die Verbreitung von Prüfungs- bzw. Testangst erschweren (Fehm & Fydrich, 2011), veranschaulichen einige beispielhafte Untersuchungen die subjektive und gesellschaftliche Relevanz des Phänomens und der damit einhergehenden Belastungen.

Einer Untersuchung der Techniker Krankenkasse bei 1.000 Studierenden zufolge sind Prüfungen die am häufigsten genannte Quelle von Stress, noch vor der Quantität oder der Schwierigkeit des Lernstoffs und der Doppelbelastung durch Studium und Jobben: so fühlten sich 58 % der befragten Studentinnen und 46 % der Studenten durch Prüfungen „unter Druck“ bzw. „stark unter Druck“ (2015, S. 10). Auch Sorgen bezüglich schlechter Noten waren ein bedeutsamer Stressor (33 % der Studentinnen und 19 % der Studenten). Eine vom AOK-Bundesverband herausgegebene Untersuchung bei 18.214 Studierenden bestätigte diesen Befund: demnach sind Prüfungen der mit Abstand bedeutsamste Stressor unter den hochschulbezogenen Quellen von Stress (u. a. „Studien- und Semesterorganisation“ sowie „Wahl und Einstieg ins Studium“) (Herbst, Voeth, Eidhoff, Müller & Stief, 2016). Schätzungen für die Prävalenz bei Schülern[1] liefern Döpfner, Schnabel, Goletz und Ollendick (2006). Bei einer Erhebung des Phobiefragebogens für Kinder und Jugendliche (PHOKI) bei $N = 277$ Personen (Alter von 8 bis 18 Jahren) gaben ca. 14 % der Befragten an, „oft“ Angst vor schlechten Noten zu haben. Ca. 20 % gaben an, „oft“ Angst davor zu haben, eine Prüfung nicht zu bestehen (der Rest antwortete mit „gar nicht“ oder „manchmal“; siehe auch Suhr

[1] Zur besseren Lesbarkeit wird das generische Maskulinum verwendet. Sofern nicht anders indiziert, gelten sämtliche Personenbezeichnungen für beiderlei Geschlecht.

& Döpfner, 2000). In einer Literaturübersicht, die sich auf Studien mit Schülern fokussiert, kommt McDonald (2001) zu dem Schluss, dass Prüfungen zu den wichtigsten Ursachen von Angst gehören und die Häufigkeit von Prüfungsangst in den letzten Jahrzehnten tendenziell zugenommen hat. In der 19. Sozialerhebung des Deutschen Studentenwerks (Isserstedt, Middendorff, Kandulla, Borchert & Leszczensky, 2010) wurden über 16.000 Studierende zu unterschiedlichsten Aspekten ihrer Lebens- bzw. Studiensituation befragt. Danach gefragt, wo sie subjektiv Beratungs- und Informationsbedarf sehen, gaben 13 % der Befragten im Erststudium das Thema „Prüfungsangst" und 12 % das Thema „Lern-/Leistungsprobleme" an. Diese Zahlen mögen angesichts der oben dargelegten Bedeutsamkeit von Prüfungsstress im Studium niedrig erscheinen. Jedoch ist es wahrscheinlich, dass eher Personen mit einer sehr hohen Ausprägung an Prüfungs- bzw. Testängstlichkeit (und entsprechendem Leidensdruck) tatsächlich angeben, dass sie Beratung oder Information benötigen. Da das Phänomen Prüfungs- bzw. Testängstlichkeit weit verbreitet und somit vielen Menschen aus eigener Erfahrung bekannt ist, findet es auch regelmäßig Beachtung in den Medien, so z. B. in Wochen- und Tageszeitungen sowie Onlinemedien (Greiner & Padtberg-Kruse, 2015; Zoske, 2015).

Prüfungs- bzw. Testangst gehört für viele Menschen zu Prüfungen und Tests dazu. Zwei Gründe für das große wissenschaftliche und gesellschaftliche Interesse an diesem Thema liegen damit auf der Hand: Prüfungs- bzw. Testangst ist eine Form von Stress und weit verbreitet. Ein dritter Grund sind die Konsequenzen von Prüfungs- bzw. Testangst für die Ergebnisse bzw. Leistungen in Prüfungen und Tests. Mit diesen Konsequenzen befasst sich auch die vorliegende Arbeit. Mandler und Sarason (1952) führten eine der ersten, heute als „klassisch" zu bezeichnenden Studien zur Relation von Angst und Intelligenztestleistung durch und schlussfolgerten:

> *The results of the present study suggest that anxiety present in the testing situation is an important variable in test performance. It is questionable whether intelligence test scores adequately describe the underlying abilities of individuals who have high anxiety drive in the testing situation* (S. 172)

Mandler und Sarason (1952) formulierten hier die Frage, ob Prüfungs- oder Testergebnisse valide Indikatoren der Fähigkeiten von hoch Testängstlichen sind, wenn diese Personen unter bestimmten Bedingungen, nämlich in typischen Prüfungs- bzw. Testsituationen, schlechtere Leistungen erbringen (siehe Abschnitt 1.2). Diese Fragestellung war für die folgenden Jahrzehnte der Forschung prägend. Wenn Personen aufgrund ihrer Testangst nicht in der Lage sind, ihr optimales Leistungsvermögen zu zeigen, könnte dies dazu führen, dass die mit einem Test jeweils zu erfassenden Fähigkeiten, Fertigkeiten, Kompetenzen oder Kenntnisse *unterschätzt* werden. In Anbetracht der Steuerungsfunktion von Prüfungen und Tests kann dies enorme individuelle und gesellschaftliche Auswirkungen haben (Spielberger & Vagg, 1995a).

Diese Gedanken sind Ausgangspunkt für die Fragestellungen der vorliegenden Arbeit. Die Arbeit nimmt einen differentiell-psychologischen bzw. psychologisch-diagnostischen Fokus ein und basiert auf der Frage, welche Konsequenzen Prüfungs- bzw. Testangst für die Ergebnisse bzw. Leistungen in Prüfungen und Tests hat. Ein Ziel der Arbeit ist es, einen Beitrag zur Erklärung des (negativen) Zusammenhangs von Prüfungs- bzw. Testangst und Leistung zu leisten. Die vermutete Kausalität (und deren Richtung) hinter diesem Zusammenhang steht dabei ebenso im Zentrum wie Variablen, die den Zusammenhang von Prüfungs- bzw. Testangst und Leistung beeinflussen, sprich verstärken oder auch verringern. Variablen oder Maßnahmen, die zu einer Reduktion des Zusammenhangs von Prüfungs- bzw. Testangst und Leistung beitragen, können Ansatzpunkte für Interventionen darstellen und sind somit von hoher praktischer Relevanz.

Die Ausführungen zu den theoretischen Grundlagen dienen dazu, diese Grundgedanken zu elaborieren. Sie sollen schließlich zu den Fragestellungen der drei in dieser Arbeit dargestellten Studien führen. Der Theorieteil gliedert sich in drei Abschnitte. Abschnitt 1.1 dient dazu, Testängstlichkeit und Testangst als Begriffe zu definieren und zu anderen Konstrukten in Bezug zu setzen sowie notwendige Abgrenzungen vorzunehmen. Neben der Ebene des Erlebens von Testängstlichkeit bzw. Testangst wird auch den Entstehungsbedingungen von Testangst Aufmerksamkeit gewidmet, was schließlich eine integrative Betrachtung von Testangst als Prozess ermöglicht. Abschnitt 1.2 befasst sich mit Zusammenhängen von Testängstlichkeit und Testangst mit Leistung. Nachdem die grundlegende Befundlage dargestellt ist, werden die wichtigsten Theorien zur Erklärung dieser Zusammenhänge behandelt. Abschnitt 1.3 beschreibt Interventionen, also Maßnahmen zur Reduktion von Testangst und / oder zur Steigerung der Prüfungs- bzw. Testleistung. Schwerpunkt liegt dabei auf sogenannten Kurzinterventionen, also Interventionen (bzw. Manipulationen) innerhalb der Testsituation mit den genannten Zielen.

Alle drei im Rahmen dieser Arbeit beschriebenen Studien befassten sich mit der Relation von Testängstlichkeit bzw. Testangst und Leistung, weshalb der überwiegende Teil der theoretischen Ausführungen für alle drei Studien relevant ist. Nach einem Abschnitt mit methodischen Vorbemerkungen (Abschnitt 3) erfolgt eine separate Betrachtung der Studien in den Abschnitten 4, 5 und 6. In diesen werden jeweils die Hypothesen hergeleitet sowie die Methoden und Ergebnisse beschrieben und anschließend diskutiert.

Studie 1 befasste sich mit der Moderation des Zusammenhangs von Testangst und Leistung durch den Messzeitpunkt von Testangst: so hängt Testangst stärker mit Leistung zusammen, wenn sie nach einem Test erfasst wird (gegenüber einer Erfassung vor dem Test; siehe Abschnitt 1.2.1.3). Dieser Befund wurde in der Literatur immer wieder mit der Vermutung in Verbindung gebracht, dass Testangst nicht *Ursache*, sondern *Folge* von (schlechter) Leistung ist. Ziel der Studie war es, diesen Befund genauer zu untersuchen und dadurch Rückschlüsse auf den Kausalzusammenhang

beider Variablen zu ziehen. Dabei spielten auch die Fragen eine Rolle, wie sich die vor und nach einem Test gemessene Testangst unterscheidet, wovon sie abhängt und welche Implikationen dies für den (durch Praktiker oder Forscher gewählten) Messzeitpunkt von Testangst hat.

Ausgangspunkt von Studie 2 war die Verknüpfung von Erkenntnissen aus dem Forschungsgebiet zum „stereotype threat“ mit jenen aus der Testängstlichkeitsforschung. Beim „stereotype threat“ handelt es sich um das Phänomen, dass Personen, über deren Gruppe ein negatives Stereotyp existiert, in einer Aufgabe tatsächlich schlechtere Leistungen erbringen, wenn sie Gefahr laufen, eben jenes Stereotyp zu bestätigen (siehe Abschnitt 1.2.2.2). Dieser inzwischen sehr bekannte Effekt und dessen Erforschung weisen Parallelen, aber auch Unterschiede zur Testängstlichkeitsforschung auf. So hängt der Zusammenhang von Testangst und Leistung von bestimmten Eigenschaften der Testsituation, insbesondere der Testinstruktion, ab. Auch beim „stereotype threat“ wird die Art der Instruierung eines Tests mit Leistungsunterschieden zwischen bestimmten Gruppen in Verbindung gebracht. Ziel der Studie war es, die Bedeutung von Testangst beim Auftreten des „stereotype threat“ zu klären.

Ausgangspunkt von Studie 3 waren theoretische Vorstellungen über die motivierenden Aspekte von Testangst. Hierbei stand die Frage im Raum, ob Testangst immer ungünstig sein muss, oder auch positive Effekte haben kann (siehe Abschnitt 1.3.2.1). Besondere Aufmerksamkeit wurde dabei auf die kognitive Bewertung der Angst gelegt. Dieser kommt bei der Erklärung der Konsequenzen von Prüfungs- bzw. Testangst möglicherweise eine Schlüsselrolle zu. Ziel der Studie war es, die subjektive, leistungsförderliche Bewertung von Angstprozessen zu untersuchen, welche in der Literatur mit dem Begriff „anxiety motivation“ bezeichnet wurde. Dabei sollte der Frage nachgegangen werden, ob „anxiety motivation“ den Zusammenhang von Testängstlichkeit bzw. Testangst und Leistung moderiert. Darüber hinaus wurde in Studie 3 eine Kurzintervention getestet, die auf der funktionalen Bewertung von Angstprozessen basierte.

1. Theoretische Grundlagen

1.1 Eingrenzung der Thematik

Die langjährige Forschung zur Prüfungs- bzw. Testangst umfasst sowohl Entwicklungen der theoretischen Konzeption als auch der methodischen Erfassung des Konstrukts. Was sich hinter dem Konstrukt Prüfungs- bzw. Testangst verbirgt, ist keineswegs trivial. Darin enthalten sind Prozesse, die sich in ihrer Beschreibungsebene (state vs. trait) sowie in den mutmaßlich betroffenen Leistungsbereichen (einzelne oder aggregierte akademische Leistungen vs. einfache bis komplexe kognitive Leistungen) unterscheiden. Dies bringt eine gewisse Begriffsheterogenität mit sich, weshalb zunächst einige Begriffsbestimmungen notwendig sind. So findet sich im deutschen Sprachgebrauch der bereits genutzte Begriff „Prüfungsängstlichkeit“ (Hodapp, Rohrmann & Ringeisen, 2011) sowie der allgemeiner gehaltene Begriff „Leistungsängstlichkeit“ (Rost & Schermer, 2007), vereinzelt auch der Begriff „Testangst“ (Strohbeck-Kühner, 1999). Auch in englischsprachigen Publikationen liegt eine gewisse Begriffsvielfalt vor, verwendet wird beispielsweise „evaluation anxiety“ (Coy, O'Brien, Tabaczynski, Northern & Carels, 2011) oder auch „performance anxiety“ (Hopko, Hunt & Armento, 2005), wobei „test anxiety“ (Zeidner, 1998) zweifelsfrei der dominierende Begriff ist. Was sich hinter „test“ verbirgt, ist im deutschen Sprachgebrauch ebenso klärungsbedürftig wie im englischen: gemeint sind einerseits akademische Leistungsmessungen, also mündliche oder schriftliche Lernzielkontrollen in der schulischen oder universitären Ausbildung (siehe z. B. Seipp, 1991; Sparfeldt, Rost, Baumeister & Christ, 2013). Ebenso impliziert es auch psychometrische Leistungstests, die unter kontrollierten Bedingungen durchgeführt werden, also z. B. Leistungen in Tests zum Arbeitsgedächtnis (Calvo & Eysenck, 1996) oder zur Intelligenz (Meijer & Oostdam, 2007). Vor diesem Hintergrund ist eine Eingrenzung der Thematik dieser Arbeit unabdingbar. In die Betrachtung fallen:

1. Angstprozesse, die sich spezifisch auf Leistungssituationen beziehen.
2. Leistungssituationen sind Situationen, in denen leistungsmotiviertes Verhalten vorliegt, also wenn „an das eigene Handeln ein Gütestandard angelegt und die eigene Tüchtigkeit bewertet wird“ (Brunstein & Heckhausen, 2010, S. 147).
3. Leistungssituationen in dieser Vorstellung umfassen zwei Klassen: zum einen Erhebungen von Fähigkeiten, Fertigkeiten oder Wissen im akademischen Bereich, d. h. im (Aus-)Bildungssystem, zum anderen Erhebungen von Fähigkeiten, Fertigkeiten oder Wissen mit einem psychometrischen Leistungstest. In beiden Fällen wird von einer Person dabei eine definierte und in ihrer Qualität (mehr oder weniger) eindeutig beurteilbare Leistung gefordert.

4. Fokus der Leistungserhebung sind kognitive Fähigkeiten einer Person im weitesten Sinne (Lese- oder Rechenfähigkeit, Arbeitsgedächtnis, logisches Schlussfolgern, usw.). Nicht eingeschlossen sind damit Leistungen bzw. Fähigkeiten im musikalischen, sportlichen, künstlerischen oder sozialen Bereich.

Um eine Einengung auf akademische Prüfungen zu verhindern, werden fortan die Begriffe „Testangst" beziehungsweise „Testängstlichkeit" genutzt, um den jeweiligen akuten Zustand beziehungsweise die entsprechende Disposition zu kennzeichnen. Diese im Deutschen mögliche begriffliche Differenzierung wird im nächsten Abschnitt (siehe Abschnitt 1.1.1) näher erläutert. Für *beide* Betrachtungsebenen soll folgende Definition von Zeidner (1998) zugrunde gelegt werden:

> *The term "test anxety", as a scientific construct, referts to the set of phenomenological, physiological, and behavioral responses that accompany concern about possible negative consequences or failure on an exam or similar evaluative situation* (S. 17)[2]

Drei Argumente sprechen für die gemeinsame Betrachtung von Leistungen im (Aus-)Bildungswesen einerseits und solchen in psychometrischen Testverfahren andererseits. Erstens kommen in beiden Kontexten dieselben Verfahren zum Einsatz, um Testangst bzw. –ängstlichkeit zu erfassen (z. B. Lang & Lang, 2010; Musch & Bröder, 1999b). Zweitens zeigen sich negative Zusammenhänge zwischen Testangst bzw. –ängstlichkeit und Leistung sowohl unter Laborbedingungen (Hopko, Crittendon, Grant & Wilson, 2005) als auch in „realen" akademischen Prüfungen (Cassady & Johnson, 2002). Das dritte Argument lässt sich aus der oben genannten Definition ableiten: die Reaktionen, welche das Erleben von Testangst ausmachen, gehen einher mit einer bestimmten Klasse von Situationen, die von Zeidner (1998) zusammenfassend als Bewertungssituationen beschrieben werden. Ziel eines „Tests" – in einem breiten Sinne verstanden – ist es stets, die Fähigkeiten, Fertigkeiten, Wissen oder Kompetenzen einer Person in einem bestimmten Bereich festzustellen. Ob nun in einer Statistikklausur oder in einem Aufmerksamkeitstest – stets findet eine Bewertung statt. Diese Bewertung kann Gedanken an die Konsequenzen des eigenen Scheiterns oder den Vergleich mit der Leistung anderer auslösen. Wenngleich also beide Situationen wichtige Unterschiede aufweisen (siehe auch Abschnitt 1.1.2), so rechtfertigt die Annahme ähnlicher psychologischer Prozesse deren gemeinsame Betrachtung.

[2] Zeidner (1998) zitiert dabei Sieber, O'Neil und Tobias (1977), jedoch wird die Definition gewöhnlich Zeidner zugeschrieben.

1.1.1 Spezifikation des Konstruktraums

1.1.1.1 Testangst und Testängstlichkeit als Formen von Angst und Stress

Das Konstrukt Testangst bzw. –ängstlichkeit hat eine eigenständige Forschungstradition. Eng verwoben ist diese mit der Forschung zum übergeordneten Konstrukt Angst. Häufig finden sich in Bezug auf die eingesetzten Verfahren oder zugrunde gelegten Theorien Überschneidungen. Doch was ist Angst? Schmidt-Atzert, Peper und Stemmler (2014) vergleichen unterschiedliche Konzeptionen von Grund- bzw. Basisemotionen und stellen fest, dass Angst bzw. Furcht in allen vier betrachteten Konzeptionen (darunter Ekman und Cordaro (2011) sowie Izard (2009)) zu den Grund- bzw. Basisemotionen gezählt wird. Darunter versteht man „emotions that organize and motivate rapid virtually automatic yet malleable responses that are critical in meeting immediate challenges to survival or well-being“ (Izard, 2009, S. 6). Furcht lässt sich auffassen als Reaktion auf eine physische oder psychologische Bedrohung (Ekman & Cordaro, 2011). Wie bereits in der zentralen Definition von Zeidner (1998) angedeutet, ergibt sich die Bedrohung im Falle der Testangst aus einer Bewertungssituation. Wesentlich für das Verständnis des Phänomens ist die theoretische Unterscheidung zwischen dem akuten *Zustand* (Angst) und der *Neigung*, in diesen Zustand zu „geraten“ (Ängstlichkeit). Spielberger und Vagg (1995b) beschreiben den *state* als einen emotionalen Zustand, der durch Empfindungen wie Nervosität, Anspannung und Besorgnis charakterisiert ist. Auch physiologische Erregungsprozesse gehören hierzu. Dieser Zustand ist wiederum eine Reaktion auf eine Bedrohung bzw. einen Stressor (vgl. die Definition von Ekman & Cordaro, 2011). Je stärker die (subjektive) Bedrohung, desto stärker ist auch der Angstzustand (Spielberger & Vagg, 1995b). Das Ausmaß der subjektiven Bedrohung wird wiederum von Merkmalen der Situation (siehe Abschnitt 1.1.2) und individuellen Variablen bestimmt. Zu letzteren gehört unter anderem die Disposition oder Neigung, der *trait*: „Trait Anxiety […] refers to relatively stable individual differences in anxiety proneness“ (Spielberger & Vagg, 1995b, S. 6). Demnach unterscheiden sich Personen darin, inwieweit sie bestimmte Situationen als bedrohlich interpretieren (Spielberger & Vagg, 1995b). Spielberger und Vagg (1995b) fassen „test anxiety“ in diesem Kontext explizit als Persönlichkeitseigenschaft (trait) auf, die sich auf bestimmte Situationsklassen bezieht: „Test anxiety scales […] appear to reflect a specific type of trait anxiety“ (Spielberger, 1972b, S. 490). Den Zusammenhang von trait und state fassen Spielberger, Gonzalez, Taylor, Algaze und Anton (1978) so zusammen:

> *„Most test anxiety theorists seem to agree that test-anxious people are more likely: (a) to perceive examination situations as more dangerous or threatening than do people who are low in test anxiety, and (b) to experience worry cognitions and intense elevations in state anxiety in situations in which they are evaluated.”* (S. 174)

Grundlegend für das Verständnis des Zusammenspiels von Testangst und Testängstlichkeit ist das transaktionale Stressmodell von Lazarus und Kollegen. Den Kern dieses Modells zur Erklärung der Stressentstehung bildet die kognitive Bewertung (appraisal) von Reizen (siehe z. B. Lazarus & Folkman, 1987). Später wurde das Modell auf den Prozess der Emotionsentstehung übertragen (Lazarus, 1991a). Dies erforderte einige Modifikationen, da der Bewertungsprozess bei der Entstehung und Erklärung vieler verschiedener Emotionen theoretisch deutlich komplexer ist (Lazarus, 1991a). Lazarus (2006) betont, dass Stress und Emotion nicht voneinander getrennt betrachtet werden können und bezeichnet einige Emotionen als „Stressemotionen" – hierzu gehört auch Angst. Die kognitive Bewertung gilt prinzipiell als notwendige und hinreichende Bedingung für die Entstehung einer Emotion (Lazarus & Smith, 1988). Jede Emotion resultiert aus einer spezifischen Person-Umwelt-Konstellation (Lazarus, 1991b). Zwischen diesem „Ereignis" („encounter" oder auch Transaktion; Folkman, Lazarus, Dunkel-Schetter, DeLongis & Gruen, 1986; Lazarus, 1991b) und einer Emotion steht ein primärer und ein sekundärer Bewertungsprozess. Die *primäre Bewertung* beinhaltet das Urteil über die grundlegende Relevanz eines Ereignisses – ohne diese entsteht keine Emotion (Lazarus, 1991b). Im Kontext der Stressentstehung können primäre, stressbezogene Bewertungen drei Ausprägungen annehmen. Schaden bzw. Verlust (harm/loss) bedeutet, dass ein Verlust oder ein Schaden eingetreten ist, Bedrohung (threat) bedeutet, dass dies möglicherweise bevorsteht und Herausforderung (challenge), dass die Chance darauf besteht, eine Weiterentwicklung oder einen Vorteil zu erreichen. Emotionen wie Angst gehen dabei einher mit harm/loss- oder threat-Bewertungen (Folkman, 1984). Die nachfolgende, *sekundäre Bewertung* ist definiert als eine Einschätzung der Ressourcen und Möglichkeiten zur Bewältigung (Folkman, 1984). Um die Entstehung unterschiedlicher Emotionen erklären zu können, wurden die Konzepte der primären und sekundären Bewertung in Einzelbewertungen ausdifferenziert (Lazarus, 1991a). Auf diese soll nun jeweils eingegangen werden.

Lazarus (1991b) unterscheidet drei Formen der primären Bewertung. Dabei inkludiert er in sein Modell explizit motivationale Prozesse in Form von Zielen. Aus dieser Perspektive sind Emotionen „reactions to the status of goals in everyday adaptational encounters and in our lives overall." (Lazarus, 1991b, S. 820). Die *primäre Bewertung* umfasst erstens das Urteil, ob das Ereignis einen subjektiven Bezug zu einem Ziel hat (goal relevance). Ist dies der Fall, kann eine Emotion entstehen. Die Stärke der Emotion wird durch die Bedeutsamkeit des Ziels beeinflusst. Zweitens erfolgt ein Urteil, ob das Ereignis schädlich bzw. bedrohlich oder günstig ist (goal congruence). Die Valenz einer Emotion (positiv oder negativ) hängt hiervon ab. Die dritte Form der primären Bewertung ist die Art des betroffenen Ziels (goal content bzw. ego-involvement), die wiederum unterschiedliche Emotionen nach sich zieht. Lazarus (1991b) nennt als Beispiele für Zielarten den Schutz der eigenen Identität oder die Einhaltung eigener moralischer Ansprüche. Auch die *sekundäre Bewertung* gliedert sich in drei Teilbewertungen. Das erste Urteil beinhaltet zwei Attributionsurteile,

nämlich zum einen nach der Verursachung des Ereignisses (das bereits als schädlich oder günstig bewertet wurde) und ob eine Kontrollierbarkeit des Ereignisses vorlag (blame or credit). Hinzu kommt eine Bewertung, ob die eigene oder eine andere Person von dem Ereignis betroffen ist (Lazarus, 1991b). Beispielsweise würde ein schädliches Ereignis, das von der eigenen Person verursacht wurde und eine andere Person betrifft, zur Emotion Schuld führen. Die Möglichkeit zu Coping stellt das zweite Urteil dar (coping potential). Coping bezeichnet kognitive und behaviorale Aktivitäten, die der Bewältigung innerer oder äußerer Anforderungen dienen. Diese Bewältigung bezieht sich auf Anforderungen, denen Ressourcen gegenüberstehen (Folkman et al., 1986). Transaktionen, welche die eigenen Ressourcen subjektiv beanspruchen oder auch überschreiten sowie das eigene Wohlbefinden beeinträchtigen können werden als Stress bezeichnet (Lazarus & Folkman, 1984). Das dritte Urteil sind schließlich Erwartungen über die weitere, positive oder negative Entwicklung der jeweiligen Person-Umwelt-Konstellation (future expectations). Eine Übersicht über die Bewertungskomponenten findet sich in Tabelle 1.

Tabelle 1: Komponenten der primären und sekundären Bewertung nach Lazarus (1991b, 2001)

	Komponenten	Beschreibung
Primäre Bewertung	• Goal relevance	Relevanz eines Ereignisses für ein Ziel?
	• Goal congruence	Ereignis bedrohlich oder günstig?
	• Goal content*	Welches Ziel ist betroffen?
Sekundäre Bewertung	• Blame or credit	Verursachung und Kontrollierbarkeit des Ereignisses?
	• Coping potential	Bewältigungsmöglichkeiten?
	• Future expectations	Erwartung der weiteren Entwicklung?

* auch type of ego-involvement (z. B. Lebensziele, Selbstwert, moralische Wertvorstellungen)

Nach Lazarus (1991b) liegt der Entstehung von Angst eine spezifische primäre Bewertung zugrunde: „I propose that the goal content relevant to anxiety is existential, that is, centered on meanings and a sense of identity that the individual has constructed“ (Lazarus, 1991b, S. 829)[3]. Übertragen auf Testangst können diese Ziele vielschichtig sein und sich in komplexer Weise gegenseitig bedingen. So dient z. B. eine gute Note in einem Test nicht nur dem Erreichen eines Bildungsabschlusses, sondern auch der Aufrechterhaltung eines hohen Selbstwerts und eines entsprechenden Selbstkonzepts. Die zentrale Rolle von kognitiven Bewertungen für die Entstehung und das Verständnis von Testangst bzw. Testängstlichkeit wird in der Definition von Schwarzer (2000) besonders deutlich: „Leistungsangst ist die Besorgtheit und Aufgeregtheit angesichts von Leistungsanforderungen, die als selbstwertbedrohlich eingeschätzt werden." (S. 105).

In dieser Tradition wird Testangst in der vorliegenden Arbeit als Emotion verstanden und Testängstlichkeit als Neigung, Situationen (oder Transaktionen) so zu interpretieren, dass Testangst

[3] Lazarus rekurriert hierbei auf Lazarus und Averill (1972): „the environment, internal as well as external, is organized and given meaning through a system of cognitive schema. […] Anxiety appraisals entail the apprehension that one's system of interpretive schemata is not adequate for the situation." (S. 248)

entsteht, nämlich als bedrohlich. Diese Unterscheidung ist nicht immer unproblematisch, da sie dazu führen kann, dass Konzepte miteinander vermischt oder irrtümlicherweise als äquivalent interpretiert werden. So argumentiert etwa Zeidner (1998), dass die Inkonsistenz von Befunden zur Relation von Angst und Leistung eine konzeptuell-methodische Ursache haben kann, nämlich den (wechselnden) Einsatz von state- und trait-Maßen. Hinzu kommt, dass gelegentlich Maße allgemeiner Angst bzw. Ängstlichkeit verwendet werden, an anderer Stelle hingegen situationsspezifische Skalen zur Testangst bzw. Testängstlichkeit. Daher sollen noch einige Abgrenzungen bzw. Begriffsschärfungen vorgenommen werden.

In einer Übersicht hält Putwain (2008a) drei verschiedene Verständnisse des Konstrukts Testängstlichkeit fest. Dies ist erstens die Persönlichkeitseigenschaft (trait). Nach Spielberger (1972b) ist Testängstlichkeit (test anxiety) eine bestimmte, auf spezifische Situationen bezogene Variante der Ängstlichkeit (trait anxiety). Auffällig ist, dass „test anxiety" hierbei als Disposition, also als trait, aufgefasst wird. Spielberger und Vagg (1995b) fassen Angstgefühle während einer Prüfungssituation allgemein als „state anxiety" auf, ohne einen spezifischen Zustand der „state test anxiety" zu formulieren. Dies ist aus ihrer Beschreibung des state-trait-Zusammenhangs unmittelbar erkennbar: „During examinations, high test-anxious persons respond to the evaluative threat inherent in most test situations with greater elevations in S-Anxiety [gem.: state anxiety]" (Spielberger & Vagg, 1995b, S. 8). Putwain (2008a) hingegen nutzt zur Beschreibung des emotionalen Zustandes den spezifischeren Begriff „state test anxiety". Die Frage, ob bei Betrachtung der Zustände eine Trennung „unspezifischer" Angst von „spezifischer" Testangst notwendig ist, würde eine eigenständige Behandlung erfordern. Allerdings handelt es sich hier möglicherweise um ein begriffliches, aber kein phänomenologisches Problem: wenn eine Person vor, während oder nach einem Test Angst empfindet, die durch den Test hervorgerufen wird bzw. sich auf den Test bezieht, liegt offensichtlich *Testangst* vor. Anders ist die Lage beim trait: die Erfassung der dispositionellen, allgemeinen Angstneigung ist nicht identisch mit Fragen nach der Angstneigung, die sich auf *eine* Klasse von Situationen bezieht, nämlich Prüfungs- und Testsituationen.

Die dritte Konzeption ist nach Putwain (2008a) ein klinisches Verständnis von Testängstlichkeit. Test- oder Prüfungsängstlichkeit ist nicht als eigene Störungskategorie im DSM-V gelistet (American Psychiatric Association, 2013a). Die Diagnose der „Sozialen Angststörung" (im DSM-IV „Soziale Phobie"; American Psychiatric Association, 2013b) umfasst zwar explizit die Angst vor Leistungssituationen, koppelt diese allerdings an eine unmittelbare soziale Situation: „Performance fears may also manifest in work, school, or academic settings in which regular public presentations are required." (American Psychiatric Association, 2013a, S. 203). Testängstlichkeit ist zweifellos eine soziale Angst, da die individuelle Leistung häufig durch andere Personen bewertet wird (Putwain, 2008a). Mit diesem Fokus auf unmittelbar soziale Situationen (z. B. einen Vortrag halten) wird im DSM-V allerdings die Angst vor einem schriftlichen Test nicht abgedeckt, da diese

nicht (oder zumindest nicht direkt) sozialer Natur ist. Testängstlichkeit *per se* wird in der Literatur nicht als psychische Störung aufgefasst. Etabliert hat sich das Verständnis von Testängstlichkeit als einer Eigenschaft, auf der interindividuelle Ausprägungsunterschiede vorliegen (z. B. McDonald, 2001; Spielberger & Vagg, 1995a). Damit verbunden ist die Annahme, dass es, orientiert an der gesamten Merkmalsverteilung, extreme (hohe) Ausprägungsgrade von Testängstlichkeit gibt, die klinisch relevant sind (siehe z. B. Herzer, Wendt & Hamm, 2014; Hodapp et al., 2011), also auch Gegenstand von Therapie oder Beratung sein sollten.

Im Rahmen dieser Arbeit stehen die ersten beiden Konzeptionen nach Putwain (2008a) im Fokus, also die Konzeption von „test anxiety" als trait (Testängstlichkeit) und state (Testangst). Gleichzeitig soll differenziert werden zwischen Testängstlichkeit und allgemeiner Ängstlichkeit. Dass diese Konstrukte eng verwandt sind steht außer Frage. Hembree (1988) berichtet einen metaanalytischen Zusammenhang von $r = .53$ zwischen Testängstlichkeit und Ängstlichkeit (trait) ($k = 10$, $N = 961$). Eine differenzierte Betrachtung ist dennoch sinnvoll, da es dem Verständnis der Befundlage sowie der Präzision daraus gezogener Schlussfolgerungen dienlich ist. Eine Kurzübersicht dieser Definitionen ist in Tabelle 2 zu sehen.

Tabelle 2: Definitionen der Begriffe Ängstlichkeit, Testängstlichkeit und Testangst (bzw. Angst) im Kontext dieser Arbeit

		Definition	Quelle
Ängstlichkeit	trait	Disposition bzw. Neigung zum Erleben von Angst	Spielberger (1972b)
Testängstlichkeit	trait	Situationsspezifische Form der Ängstlichkeit, die sich auf Bewertungssituationen (hier: Prüfungen und Tests) bezieht	Spielberger (1972b)
Testangst / Angst	state	(Vorübergehender) emotionaler Zustand	Spielberger (1972a) Putwain (2008a)

Zur Begriffsverwendung siehe Ausführungen im Text

Es wurde darauf eingegangen, dass die Differenzierung von „Testangst" und „Angst" womöglich nur eine begriffliche ist. „Testangst" soll genutzt werden, wenn ausschließlich das Erleben in Prüfungen und Tests thematisiert wird. „Angst" wird allgemeiner verwendet, d. h. wenn es um Prozesse geht, die neben Testsituationen auch andere Situationen betreffen. Hierbei muss beachtet werden, dass in den Ausführungen eine strikte Trennung der Begriffe Testängstlichkeit und Testangst nicht immer möglich ist, auch aufgrund der zuweilen ungenauen Trennung in empirischen Untersuchungen. Wie deutlich werden wird, wurde in der Literatur häufig von der Ausprägung auf der Disposition (Testängstlichkeit) auf das Auftreten des Zustands in der Situation (Testangst) geschlossen, ohne dass beide Konstrukte erfasst wurden. Wenn also die Unterscheidung der Konzepte wichtig ist, wird jeweils indiziert welches Konzept gemeint ist (z. B. durch die Ergänzungen „trait" und „state").

Nachdem nun das grundlegende Verständnis von Testangst und Testängstlichkeit dargelegt wurde, ist es erforderlich, den Konstruktraum aus drei wichtigen Perspektiven auszuleuchten. In Abschnitt 1.1.1.2 wird die Multidimensionalität von Testangst bzw. –ängstlichkeit beschrieben, die von wichtiger Bedeutung ist. Diese Ausführungen orientieren sich an einem historischen und methodischen Leitfaden, da sich der Wandel des Konstruktverständnisses relativ gut anhand der Veröffentlichung, Neu- und Weiterentwicklung der entsprechenden Erhebungsinstrumente nachverfolgen lässt. Abschnitt 1.1.1.3 klärt die Beziehungen des Konstrukts zu anderen traits und Dispositionen. In Abschnitt 1.1.1.4 wird auf Besonderheiten der Informationsverarbeitung bei Testängstlichen eingegangen. Abschnitt 1.1.1.5 behandelt alters- und geschlechtsspezifische Ausprägungen der Testängstlichkeit. Situative Determinanten der Testangst sind schließlich Gegenstand von Abschnitt 1.1.2.

Da mittlerweile sehr umfangreiche Literatur zu Testängstlichkeit bzw. Testangst vorliegt, erheben die nachfolgenden Ausführungen keinen Anspruch auf Vollständigkeit. Sie beschränken sich auf jene Inhalte, die für ein Verständnis der Fragestellungen erforderlich sind. Einige der zitierten Studien befassten sich auch mit der Leistungsrelevanz von Testängstlichkeit und Testangst. Um die Übersichtlichkeit der Darstellung zu erhalten, wird auf dieses eigene Thema in Abschnitt 1.2 separat eingegangen. Dies hat zur Folge, dass einige der in Abschnitt 1.1 zitierten Quellen nochmals aufgeführt werden.

1.1.1.2 Historischer Wandel des Verständnisses des Konstrukts

Kennzeichnend für die Konstruktgeschichte von Testängstlichkeit bzw. Testangst ist das Wechselspiel von theoretischer Konzeption und diagnostischer Erfassung. Analysen auf Basis existierender Fragebogeninstrumente führten, wie deutlich werden wird, insbesondere in der frühen Forschungsperiode zu Modifikationen des Konstruktverständnisses. Ebenso führten theoretische Überlegungen dazu, dass neue Instrumente entwickelt oder vorhandene ergänzt wurden. Möchte man einen Beginn in dieser Historie setzen, so eignet sich die Publikation von Mandler und Sarason (1952). Mandler und Sarason (1952) entwickelten einen Fragebogen, der spezifisch das Erleben vor, während und nach einem Test (was sowohl Einzel- und Gruppenintelligenztests als auch akademische Prüfungen einschließt) erfasst. Dieser Fragebogen, der von Mandler und Cowen (1958) als Test Anxiety Questionnaire (TAQ) vorgestellt wurde, war der erste Fragebogen zur Testängstlichkeit, der eine häufige Verwendung fand (Spielberger & Vagg, 1995b). Der TAQ erfasst unterschiedliche Aspekte des Erlebens in Testsituationen wie Besorgnis, Unbehagen und körperliche Empfindungen, beispielsweise erhöhte Herzrate und Schwitzen (Mandler & Sarason, 1952).

Der TAQ wurde in den folgenden Jahren zu einer Art „Steinbruch“ für weitere Testentwicklungen. Sarason (1958b) setzte einen Fragebogen ein, der 21 Items aus dem TAQ enthielt, wobei er den

Fokus auf das dispositionelle Erleben beibehielt. Sarason (1978) rekapitulierte die weitere Skalenentwicklung, die auf diesen 21 Items basierte. Dabei wurden auch neue Items generiert, mit dem Ziel, Reliabilität und Sensitivität des Verfahrens zu erhöhen. Der neue Fragebogen, die Test Anxiety Scale (TAS)**,** bestand aus 37 Items. Die TAS etablierte sich gewissermaßen in Nachfolge des TAQ (Spielberger & Vagg, 1995b). Wie der TAQ erfasst auch die TAS die Neigung, vor, während oder nach einem Test bzw. einer Prüfung verschiedenste Symptome von Angst zu erleben. Die Bandbreite dieser kognitiven, emotionalen und körperlichen Symptome ist recht groß und beispielhaft in Tabelle 3 dargestellt:

Tabelle 3: Inhaltsbereiche und Beispielitems der Test Anxiety Scale (TAS; Sarason, 1978)

Inhaltsbereich	Beispielitem[4]
soziale Vergleiche	While taking an important exam I find myself thinking of how much brighter the other students are than I am.
sorgenvolle Gedanken	If I were to take an intelligence test, I would worry a great deal before taking it.
Gedanken, die sich nicht auf die Testsituation beziehen	During course examinations I find myself thinking of things unrelated to the actual course material.
emotionales Erleben	I get to feel very panicky when I have to take a surprise exam.
Selbstzweifel	During exams I sometimes wonder if I'll ever get through college.
körperliche Begleiterscheinungen	I sometimes feel my heart beating very fast during important tests.
subjektive Kausaltheorien von Angst und Leistung	Thoughts of doing poorly interfere with my performance on tests.
Einstellungen zu Tests und deren Interpretation	The University ought to recognize that some students are more nervous than others about tests and that this affects their performance.

Anm.: Inhaltsbereiche durch eigene Zuordnung gebildet

Die TAS sieht ein Richtig-Falsch-Antwortformat vor, wobei in der Auswertung ein Summenscore über alle Items vorgesehen ist. Ebenso wie im TAQ beziehen sich die Items der TAS sowohl auf akademische Prüfungen („exam") als auch auf psychometrische Leistungstests („intelligence test").

Gewissermaßen parallel wandelte sich die theoretische Vorstellung von Testängstlichkeit. Sassenrath (1964) setzte sich kritisch mit der eindimensionalen Interpretation des TAQ auseinander. Er demonstrierte mit einer explorativen Faktorenanalyse auf Basis von 34 der 37 Items der ursprünglichen Form des TAQ, dass bei orthogonaler Rotation insgesamt sieben Faktoren resultieren. Gorsuch (1966) ermittelte auf Basis der Korrelationsmatrix von Sassenrath (1964) mit obli-

[4] Itemformulierungen oder Instruktionstexte aus Fragebögen oder Testverfahren werden in der vorliegenden Arbeit nicht mit Seitenzahl zitiert.

quer Rotation ebenfalls sieben Faktoren, konnte aber auf zweiter Ebene zwei Faktoren identifizieren, die er „Emotionality“ und „Anxious Avoidance of Testing“ benannte. Auf dritter Ebene zeigte sich ein Generalfaktor, den er „Test Anxiety“ nannte.

Aus beiden Arbeiten ergibt sich ein Kerngedanke, der fundamental für die weitere Forschung war: Testängstlichkeit ist nicht eindimensional, sondern setzt sich aus unterschiedlichen, aber voneinander nicht unabhängigen Facetten zusammen. Ein wichtiger Meilenstein in der theoretischen Entwicklung war die Zwei-Komponenten-Theorie von Liebert und Morris (1967). Auf Basis der bis zu jenem Zeitpunkt vorliegenden Studien zum TAQ schlussfolgerten die Autoren, dass sich Testängstlichkeit auf zwei Komponenten unterteilen lässt: „worry“ und „emotionality“. *Worry* bezeichnet die kognitive Komponente, wobei Liebert und Morris (1967) darunter alle sorgenvollen, auf die eigene Leistung bezogenen Gedanken subsumierten. Diese beinhalten beispielsweise den Vergleich der eigenen Fähigkeiten mit denen anderer, oder die mit dem Scheitern in einem Test verbundenen Folgen. *Emotionality* umfasst in Reaktion auf Testsituationen stattfindende Prozesse des autonomen Nervensystems. Liebert und Morris (1967) analysierten den TAQ inhaltlich und entnahmen 10 Items, von denen jeweils die Hälfte aus ihrer Sicht worry und emotionality erfasste, wobei sie das Antwortformat der Items im Hinblick auf das akute Erleben während einer Testsituation umformulierten. Sie befragten Studierende unmittelbar vor einer Prüfung nach ihrer Leistungserwartung und führten den verkürzten TAQ durch (sie erhoben also den state). Als ersten Beleg für ihre Theorie interpretierten die Autoren, dass die Leistungserwartung mit der worry-, nicht aber mit der emotionality-Komponente zusammenhing – worry war am stärksten ausgeprägt bei den Studierenden mit einer niedrigen Leistungserwartung. Emotionality hingegen unterschied sich nicht in Abhängigkeit der Leistungserwartung. Spielberger und Vagg (1995b) bemerken zu dieser initialen Trennung von worry und emotionality[5], dass bei letzterer die körperlichen Prozesse selbst im Vordergrund stehen, weniger die damit korrespondierenden subjektiven Empfindungen.

Dieser erste Fragbogen zur Differenzierung der beiden Komponenten (worry-emotionality-questionnaire, häufig als WEQ zitiert) wurde mehrfach eingesetzt und modifiziert (z. B. Morris & Fulmer, 1976). Morris, Davis und Hutchings (1981) entwickelten den Fragebogen weiter. Wenngleich keine explizite Abgrenzung von der Definition von Liebert und Morris (1967) erfolgte, ist deren Definition von emotionality breiter und präziser: „indications of autonomic arousal and unpleasant feeling states such as nervousness and tension“ (S. 541). Diese Definition umschließt nicht nur die Wahrnehmung von Erregungsprozessen des autonomen Nervensystems (z. B. „I feel my heart beating fast“; „I am so tense that my stomach is upset“), sondern auch aversive emotionale Zustände der Anspannung („I have an uneasy, upset feeling“; „I am nervous“; „I feel panicky“). Inhalte

[5] Kleinschreibung wird für die Konstrukte benutzt, Großschreibung erfolgt bei Referenz auf die Skalen.

der worry-Items sind Zweifel an der eigenen Vorbereitung („I am afraid that I should have studied more for this test“) und Leistungsfähigkeit („I feel I may not do as well on this test as I could“) sowie Gedanken an die soziale Bewertung der eigenen Person („I feel that others will be disappointed in me“) (Morris et al., 1981).

Ein weiterer wichtiger Meilenstein war die Veröffentlichung des Test Anxiety Inventory (TAI) (Spielberger, 1980; zitiert nach Spielberger & Vagg, 1995b). Dieses basiert auf dem Zwei-Komponenten-Modell von Liebert und Morris (1967) und sollte die Interpretation zweier faktorenanalytisch gebildeter Skalen für worry und emotionality mit einem kurzen Fragebogen ermöglichen. Insgesamt besteht das TAI aus 20 Items und erfasst Testängstlichkeit als Disposition (Spielberger & Vagg, 1995b). Eine deutsche Übersetzung folgte mit dem TAI-G (Hodapp, Laux & Spielberger, 1982). Der TAI basiert auf der TAS, wobei einige Schärfungen des Gegenstandsbereichs vorgenommen wurden. So wurden etwa Items exkludiert, die sich auf Einstellungen gegenüber Prüfungen bzw. Tests bezogen, auch wurden Bezüge zu Intelligenztests in den Itemformulierungen entfernt, ebenso wie Häufigkeitsangaben (Spielberger & Vagg, 1995b). In der TAS ist in jedem Item ein bestimmtes Angstsymptom an einen bestimmten Zeitpunkt in einer bestimmten Testsituation gekoppelt (s. o.). Der TAI löst dieses Problem, indem bereits in der Instruktion die Situationsklasse definiert wird. Im TAI-G wird dementsprechend nach dem Erleben in „Klassenarbeiten, Tests oder mündlichen Prüfungen“ gefragt, wobei ein vierstufiges Antwortformat eine Angabe der Häufigkeit des jeweiligen Symptoms ermöglicht („fast nie – manchmal – oft – fast immer“). Das abgefragte Erleben wurde also auf die Zeit während eines Tests eingegrenzt (Hodapp et al., 1982; Hodapp, 1991).

Bevor die weitere Entwicklung der Facettenkonzeption von Testängstlichkeit aufgezeichnet wird, soll kurz darauf eingegangen werden, welche Belege (über faktorenanalytische Befunde hinaus) für die getrennte Interpretierbarkeit der Facetten worry und emotionality erbracht wurden.

1.1.1.2.1 Empirische Belege für die Trennung von worry und emotionality

Wie bereits erwähnt wurde, lieferten Liebert und Morris (1967) einen ersten Hinweis für die Differenzierbarkeit, indem sie demonstrierten, dass worry, nicht aber emotionality mit Leistungserwartungen vor einer Prüfung kovariiert.

Einige frühe Untersuchungen konnten zeigen, dass worry und emotionality als Abbilder des akuten Erlebens auf verschiedene experimentelle Manipulationen differentiell reagieren. Morris und Liebert (1969) legten Studierenden (N = 48) Aufgaben aus einem Intelligenztest vor und variierten dabei unter anderem die Schwierigkeit der Items (leicht vs. schwer). Unmittelbar vor und nach dem Test erfassten sie mit dem TAQ worry und emotionality (state). Bei Betrachtung der

Veränderung der Werte von der ersten zur zweiten Messung stellten sie eine Verringerung der worry-Werte bei leichten Aufgaben, und (im Trend) eine Erhöhung bei schwierigen Aufgaben fest. Dieser Effekt zeigte sich bei emotionality nicht. Die Autoren erklärten dies damit, dass worry mit konkreten Leistungserwartungen in Zusammenhang steht (die hier durch einen schwierigen Test gesenkt wurden). Emotionality hingegen „is aroused by the stressful test-taking situation per se" (Morris & Liebert, 1973, S. 322). Morris und Liebert (1973) führten ein Experiment durch, bei dem Studenten (N = 175) in drei verschiedenen Bedingungen Aufgaben zu Zahlenspannen lösen mussten. Nach einem ersten Aufgabendurchgang wurde den Probanden der ersten Gruppe die Rückmeldung gegeben, dass sie sich mehr anstrengen und schneller antworten sollen (failure-threat), einer zweiten Gruppe wurden Elektroden angelegt und Stromschläge angedroht (shock-threat). Eine dritte Gruppe erhielt keine Bedrohung (no-threat). Worry und emotionality (state) wurden nach einer Reihe weiterer Testdurchgänge mit dem TAQ erfasst. Worry war in der failure-threat-Bedingung signifikant höher als bei no-threat und shock-threat, in denen es nahezu gleich hoch ausgeprägt war. Im Vergleich zur Kontrollgruppe wurde worry durch eine negative Leistungsrückmeldung induziert, nicht aber emotionality – letztere zeigte lediglich einen tendenziellen Anstieg bei einem drohenden physischen Schmerz (shock-threat). Dass worry und emotionality auch auf die Testinstruktion unterschiedlich reagieren, zeigte beispielsweise Deffenbacher (1978). Studierende (N = 68) bearbeiteten Anagramme unter zwei unterschiedlichen Bedingungen. In der ersten Bedingung wurde der Test in der Instruktion in expliziter Weise als Intelligenztest dargestellt sowie auf die Zeitbegrenzung und die Wichtigkeit eines guten Ergebnisses verwiesen. Die zweite Bedingung war stressreduzierend, indem beispielsweise darauf hingewiesen wurde, dass der Proband nur wenige Anagramme würde lösen können. Nach dem Test wurden einige Fragen nach dem eigenen Erleben während der Anagrammaufgabe gestellt, u. a. wurde worry und emotionality erfasst. Es zeigte sich ein Haupteffekt der Stressmanipulation auf worry, nicht aber auf emotionality: worry war höher in der Bedingung mit stressinduzierender Instruktion (Deffenbacher, 1978). Auf die besondere Bedeutung von Testinstruktionen wird in Abschnitt 1.1.2 näher eingegangen.

Ein weiteres Argument für die getrennte Interpretation von worry und emotionality ist der unterschiedliche zeitliche Verlauf. In ihrer Literaturübersicht kommen Morris et al. (1981) zum Schluss, dass emotionality im Verlauf einer Testsituation abnimmt, was bei worry nicht unbedingt auftreten muss. Beispielhaft sei hierfür die Untersuchung von Morris und Fulmer (1976) beschrieben. Studierende (N = 55, Studie 1[6]) lasen einen Zeitschriftenartikel und mussten danach zu diesem ein Quiz lösen. Die erste Gruppe erhielt kontinuierliche Rückmeldungen über die Richtigkeit ihrer Antworten, die zweite Gruppe nicht. Bei allen Probanden wurden vor und nach dem Quiz

[6] Wenn ein Artikel mehrere Studien berichtet, wird im Folgenden bei der Zitation die entsprechende Studie indiziert.

worry und emotionality mit dem TAQ erfasst sowie eine Einschätzung, ob man die Leistung erzielen würde die man subjektiv anstrebt (expectancy). In beiden Bedingungen sanken die Werte für emotionality, bei worry traf dies nur auf jene Bedingung zu, in der Leistungsrückmeldungen gegeben wurden. Darüber hinaus stieg die expectancy in beiden Bedingungen an. Der Gruppenunterschied bei der Veränderung von worry konnte also nicht erklärt werden durch die expectancy. Stattdessen interpretierten die Autoren, dass die Unsicherheit über die eigene Leistung einhergeht mit kognitiven Symptomen der Angst – kontinuierliche Leistungsrückmeldungen können also ein Absinken von worry bewirken, während emotionality unabhängig davon absinkt (Morris & Fulmer, 1976).

Morris et al. (1981) nehmen an, dass worry und emotionality durch unterschiedliche Reizkonstellationen ausgelöst werden. Anders als worry wird emotionality durch Reize ausgelöst, die für sich genommen nicht evaluativ sind. Das kann der Raum sein in dem eine Prüfung stattfindet oder das Austeilen des Testmaterials. Sobald eine Person mit der Bearbeitung eines Tests beginnt, nimmt die Salienz dieser Reize ab, weshalb auch emotionality abnimmt (Morris et al., 1981). Gleichwohl konzeptuell trennbar, kovariieren worry und emotionality (als states) innerhalb von Testsituationen miteinander und zeigen nicht immer diese idealtypischen, unterschiedlichen Verläufe. Zum Beispiel zeigte sich in einem zweiten Experiment bei Morris und Fulmer (1976) in beiden Facetten durchgängig ein Absinken mit Fortschreiten der Teststituation. Ein Grund hierfür ist, dass die unterschiedlichen auslösenden Reize von worry und emotionality üblicherweise gekoppelt auftreten (Morris et al., 1981). Morris et al. (1981) liefern eine theoretische Erklärung für die Auslösung von worry und emotionality, die gleichzeitig die Interpretation von worry und emotionality als state *und* trait ermöglicht. Demnach stellt emotionality zum einen unkonditionierte emotionale Reaktionen dar, die bei allen Personen gleichermaßen, beispielsweise durch die Anwesenheit anderer Personen oder einen subjektiven Zustand der Ungewissheit vor einem Test, ausgelöst werden. Zum anderen besteht es aus konditionierten Reaktionen, die bei verschiedenen Personen individuell in Reaktion auf Prüfungs- bzw. Testsituationen erworben wurden. Worry hingegen ist unter dieser Definition ein Ausdruck kognitiver Strukturen, die ein Ergebnis der individuellen Lernhistorie sind. Diese kognitiven Strukturen oder Schemata beinhalten Annahmen über die eigene Person, die eigene Leistung oder auch die eigenen Fähigkeiten und determinieren letztlich die Art und Weise, wie eine Person eine Testsituation kognitiv bewertet (Morris et al., 1981). Dass diese Bewertung von den eigenen Erfahrungen geprägt ist, liegt auf der Hand: so wird Schüler A, auf Basis der bisherigen Erfahrungen in den beiden Fächern, ängstlich auf eine Mathematikklausur reagieren, aber gelassen auf eine Deutschklausur. Ebenso können Testsituationen Reize beinhalten, die *per se* worry auslösen, weil sie bestimmte Informationen wie die Tragweite eines Testergebnisses transportieren („Der folgende Test erfasst Ihre Intelligenz. Intelligenz ist sehr wichtig für den Erfolg im Beruf und im Leben.“). Die unterschiedliche, aber doch gekoppelte

Entwicklung und Auslösung von worry und emotionality wird schließlich durch Morris et al. (1981) auf den Punkt gebracht:

> *„Thus, emotionality and worry responses are held to be qualitatively different (affective versus cognitive), developed through different learning experiences that may or may not coincide for a given individual, and under the control of different situational stimuli that may or may not coincide in a given situation."* (S. 552)

Ein weiterer Aspekt der Differenzierung sind die unterschiedlichen Zusammenhänge mit Leistung, auf die in Abschnitt 1.2 eingegangen wird.

Worry und emotionality sind also sowohl empirisch als auch theoretisch trennbare, aber gleichzeitig miteinander in Zusammenhang stehende Phänomene. Beispielhaft sei hierfür eine Studie von Ware, Galassi und Dew (1990) genannt, in der der TAI einer konfirmatorischen Faktorenanalyse (CFA) unterzogen wurde. Ein zweidimensionales Modell mit obliquen (nicht orthogonalen) Dimensionen für worry und emotionality erzielte den besten Modell-Fit. Deffenbacher (1980) berichtet Interkorrelationen von $r = .55$ bis .76 zwischen worry und emotionality. In einer späteren Zusammenschau bisheriger Befunde attestieren Keith, Hodapp, Schermelleh-Engel und Moosbrugger (2003) eine Spanne der Interkorrelationen von $r = .55$ bis .76 (manifest) und sogar $r = .82$ bis .92 (latent). Trotz des hohen Maßes an geteilter Varianz ist eine separate Interpretation unter Berücksichtigung der aufgeführten Argumente dennoch vertretbar (Sparfeldt, Schilling, Rost, Stelzl & Peipert, 2005).

1.1.1.2.2 Weitere Entwicklungen des Facettenkonzepts

Gegen Ende der 70er und zu Beginn der 80er Jahre wurde die Konzeption von Testängstlichkeit abermals erweitert. Wesentliche Impulse hierfür lieferte die Forschung zu der Frage, warum und welche Facetten von Testängstlichkeit bzw. Testangst mit Leistungseinbußen einhergehen. Wine (1971) veröffentlichte eine Literaturübersicht zum damaligen Stand der Forschung, dessen Titel die Hauptbotschaft transportiert („Test anxiety and direction of attention"). Demnach richten hoch testängstliche Personen während einer Prüfung ihre Aufmerksamkeit nicht nur auf die Aufgabe, sondern in besonderem Maße auch auf die eigene Person – damit ist ihre Aufmerksamkeit *geteilt*. Dieser Fokus auf die eigene Person und innere Prozesse unterscheidet sie von niedrig testängstlichen Personen, die ihre Aufmerksamkeit mehr auf die eigentlichen Aufgabenanforderungen richten. Gedanken über die eigene Person haben dabei häufig selbstabwertenden Charakter (Wine, 1971).

Beispielhaft sei hierzu eine Studie von Ganzer (1968) zitiert, in der eine Stichprobe von Studentinnen ($N = 72$) instruiert wurde, Silben zu lernen. Die Stichprobe wurde geteilt in Probanden mit

hoher, mittlerer und niedriger Testängstlichkeit anhand ihres Wertes auf der TAS. Einer Hälfte der Stichprobe wurde mitgeteilt, dass sie während der Lernaufgabe durch einen Einwegspiegel unter Beobachtung stünde. Analysiert wurden unter anderem Antworten bzw. Äußerungen, die nichts mit der eigentlichen Aufgabe zu tun hatten. Die hoch testängstliche Teilstichprobe unter Beobachtung formulierte am häufigsten derartige Kommentare. Inhalte dieser Äußerungen waren vorwiegend selbstabwertend und die eigene Leistung entschuldigend (z. B. „my mind's a blank"; „I really feel stupid") (Ganzer, 1968). Die Arbeit von Wine (1971) war ein wichtiger Schritt in den Bemühungen, Testängstlichkeit bzw. Testangst und deren Wirkungen im Hinblick auf Aufmerksamkeitsprozesse zu verstehen. Nicht überraschend liegt das Hauptaugenmerk dieser Perspektive daher auf den kognitiven Prozessen, weniger auf den physiologischen Prozessen, die unter emotionality fallen würden. Kerngedanke dabei ist, dass die Selbstzweifel und andere negative, selbstbezogenen Gedanken von Testängstlichen *aufgabenirrelevant* sind – so wird eine explizite Unterscheidung von „self-relevant variables" und „task-relevant variables" vorgenommen, die Objekt der Aufmerksamkeit sein können (Wine, 1971). Die Idee, dass Angst mit einer normalen bzw. optimalen Aufgabenbearbeitung interferiert, war nicht neu – bereits Mandler und Sarason (1952) formulierten ähnliche Hypothesen. In diesem Zeitraum jedoch wurde, auch im Zuge der kognitiven Wende (Fehm & Fydrich, 2011), ein verstärkter Fokus darauf gelegt, wie diese Interferenz aussieht und welche kognitiven Prozesse sie beinhaltet.

Deffenbacher (1978) griff diese Perspektive auf und postulierte, dass worry und emotionality Prozesse sind, die die Aufmerksamkeit von Testängstlichen, welche eigentlich auf die Bearbeitung von Prüfungsfragen oder Testaufgaben gerichtet sein müsste, ablenken. Neben worry und emotionality nahm er einen dritten Prozess an, „task-generated interference": „the tendency to be susceptible to or distracted by irrelevant task parameters" (Deffenbacher & Hazaleus, 1985, S. 171). Hierfür entwickelte Deffenbacher (1978) eine Skala mit 5 Items. Inhalte dieser Skala waren (Deffenbacher, 1978; Deffenbacher & Hazaleus, 1985):

- Nachdenken über Testaufgaben oder Fragen, die nicht gelöst werden konnten
- Schwierigkeiten, die Bearbeitung von Fragen abzubrechen wenn keine Lösung gefunden wird sowie Schwierigkeiten, die Aufmerksamkeit gezielt auf eine Aufgabe zu richten
- Abschweifen des eigenen Blicks auf andere Personen oder auf die Uhr

Deffenbacher (1978) teilte eine Stichprobe von Studierenden in zwei Extremgruppen anhand ihrer Werte auf der TAS. Die Probanden bearbeiteten Anagramme entweder unter hohem oder niedrigem evaluativem Stress (siehe Abschnitt 1.1.2). Nach den Anagrammen wurden die Probanden u. a. nach der erlebten worry, emotionality und task-generated interference (state) während des Tests befragt. Die drei Facetten der Testangst korrelierten deutlich miteinander (worry und task-generated interference: r = .73, emotionality und task-generated interference: r = .50). Die

Teilgruppe mit hoher Testängstlichkeit unter hohem Stress berichtete die höchsten Werte auf worry, emotionality und task-generated interference. Darüber hinaus wurden den Probanden nach dem Test weitere Fragen gestellt. Unter anderem wurden sie explizit nach der subjektiven Beeinträchtigung durch Angst gefragt sowie danach, wie viel Zeit sie tatsächlich mit den Aufgaben verbrachten und nicht anderen Aktivitäten folgten oder an Dinge dachten, die mit der Aufgabe nichts zu tun hatten. Die Gruppe mit hoher Testängstlichkeit in der evaluativen Bedingungen berichtete auf beiden Variablen die höchsten Werte – diese Gruppe gab an, sich geschätzt nur 60 % der Bearbeitungszeit tatsächlich der Aufgabe gewidmet zu haben (die hoch testängstliche Teilstichprobe in der nicht-evaluativen Bedingung gab ca. 78 % als Schätzung an, die wenig Testängstlichen in beiden Bedingungen über 80 %) (Deffenbacher, 1978).

Auch vor dem Hintergrund der Bedeutung von (kognitiven) Interferenzprozessen schlug Sarason (1984) eine Ergänzung der Zwei-Komponenten-Theorie vor. Er legte die TAS sowie zahlreiche weitere Items zu individuellen Reaktionen auf Tests bzw. Bewertungssituationen einer Stichprobe von Studierenden vor. Er führte eine Hauptkomponentenanalyse mit orthogonaler Rotation durch und bildete daraus ein neues Verfahren, die Reactions to Tests (RTT)**.** Die RTT besteht aus vier Skalen à 10 Items, die jeweils mit einem Beispiel in Tabelle 4 veranschaulicht sind:

Tabelle 4: Skalen des Reactions to Tests (RTT; Sarason, 1984)

Skala	Itembeispiel
Tension	I feel distressed and uneasy before tests.
Worry	Before taking a test, I worry about failure.
Bodily Reactions	I get a headache during an important test.
Test-Irrelevant Thinking	During tests, I find myself thinking of things unrelated to the material being tested.

Zwei wesentliche Neuerungen kennzeichnen die RTT. Erstens sind Gedanken, die nichts mit der Prüfung bzw. dem Test zu tun haben und auch keine sorgenvollen Gedanken umfassen, nun in einer eigenen Facette berücksichtigt. Zweitens ist emotionality ausdifferenziert in körperliche Reaktionen (genauer: deren Wahrnehmung) und affektive Empfindungen der Anspannung (Benson, Moulin-Julian, Schwarzer, Seipp & El-Zahhar, 1992; Sarason, 1984). Die RTT erfuhr jedoch aus mehreren Gründen Kritik. Einer davon betraf die Methodik der Skalenbildung (Hodapp, 1991). Neben den drei Faktoren tension, worry und test-irrelevant thinking zeigten sich in der Hauptkomponentenanalyse mehrere weitere Faktoren, die auf körperliche Symptome abzielten. Aus Items dieser Faktoren wurde dann die Skala bodily reactions gebildet (Sarason, 1984). Kritisiert wurden überdies die teilweise niedrigen Skalenreliabilitäten (Hodapp, 1991), so berichtet Sarason (1984) α-Koeffizienten zwischen .68 und .81. Ebenso kritisch bewertet wurde die hohe Korrelation zwischen tension und bodily reactions von r = .69 bei Sarason (1984) (Hodapp, 1991).

Diese letztere Problematik – wenn man sie denn als Problematik auffasst, da es sich um theoretisch eng verwandte Prozesse handelt – gilt jedoch auch für den TAI (s. o.). Ein weiteres Problem der RTT ist, dass in den Items die erfragten Reaktionen an verschiedene Testphasen gekoppelt sind (siehe Tabelle 4).

Trotz der Kritik an der RTT hat sich, sicherlich auch in dessen Folge, Interferenz als eigenständig interpretierbares Element von Testängstlichkeit etabliert. Sarason, Sarason, Keefe, Hayes und Shearin (1986) arbeiteten in mehreren Studien das Konzept der kognitiven Interferenz weiter aus und entwickelten entsprechende Instrumente. Der Cognitive Interference Questionnaire (CIQ) besteht aus 22 Items, wird nach einem Test administriert und dient der Erfassung von instrusiven Gedanken, die während der Aufgabenbearbeitung aufgetreten sind. Diese Gedanken, deren Häufigkeit retrospektiv angegeben werden soll, können sich sowohl auf die Aufgabe als auch auf etwas anderes beziehen. 10 der Items weisen einen starken Bezug zu worry auf (z. B. „I thought about how poorly I was doing“), 11 weitere umfassen eine große Bandbreite unterschiedlicher Gedanken (z. B. „I thought about something that happened earlier today“) und das letzte Item stellt eine Globaleinschätzung des Abschweifens der eigenen Gedanken dar („mind wandering“). Hintergrund dieser Konzeption war, dass nach Sarason et al. (1986) nicht nur Testangst, sondern auch andere Einflüsse die Aufmerksamkeitsprozesse bei der Aufgabenbearbeitung stören können. Eine Hauptkomponentenanalyse mit orthogonaler Rotation erbrachte zwei Faktoren, „Task-Irrelevant Interference“ und „Task-Related Interference“ (Studie 1). Sarason et al. (1986) ermittelten die Zusammenhänge zwischen RTT und CIQ. Task-Related Interference (CIQ) korrelierte dabei am höchsten mit der Skala Worry zu $r = .41$ sowie mit Tension (beides RTT) zu $r = .29$, Task-Irrelevant Interference sowie das Item zu mind wandering (beides CIQ) korrelierten hingegen am stärksten mit der Skala Test-Irrelevant Thinking (RTT), $r = .34$ bzw. .31 (Studie 2A). Da der CIQ ein Instrument zur Erfassung von states ist, wurde der Thought Occurence Questionnaire (TOQ) erstellt, der die Neigung zu intrusiven Kognitionen als trait abbildet. Der TOQ enthält Items aus dem CIQ und bezieht sich auf Situationen, in denen von der Person Aufmerksamkeit bzw. Konzentration gefordert ist, wie z. B. bei der Arbeit. Dabei wird angenommen, dass kognitive Interferenz kein Phänomen ist, von dem nur Testängstliche betroffen sind. Die 28 Items des TOQ wurden in einer Faktorenanalyse auf drei Faktoren verteilt, die unterschiedliche Gedankeninhalte betreffen (Studie 3). Dies sind Gedanken an soziale Beziehungen sowie an Emotionen, die nichts mit der Aufgabe zu tun haben (Faktor 1, z. B. „I think about friends“). Faktor 2 betrifft Fluchtkognitionen (z. B. „I think about how I can’t stand it anymore“). Faktor 3 schließlich beinhaltet Items zu Sorgen, die sich auf die Aufgabe richten (z. B. „I think about my level of ability“) (Sarason et al., 1986). Der Gesamtscore der RTT und des TOQ wiesen eine substantielle Korrelation von $r = .54$ auf (Sarason et al., 1986), was wahrscheinlich an überlappenden Inhalten der Items beider Instrumente liegt. Die Trennung von aufgabenrelevanten und aufgabenirrelevanten störenden Kognitionen im CIQ

ist hilfreich für die Klassifikation möglicher Aspekte von Interferenz. Die Überschneidungen von RTT, CIQ und TOQ wurden allerdings als weiterer Kritikpunkt an der RTT aufgeführt, da die Skaleninhalte von Test-Irrelevant-Thinking im CIQ und TOQ vertreten sind (Hodapp, 1991).

Hodapp (1991) veröffentlichte eine modifizierte Version des TAI-G**,** mit der einige der Probleme der Skalenbildung gelöst werden sollten (Hodapp et al., 1982). Der TAI-G behielt das 4-stufige Antwortformat („fast nie", „manchmal", „oft", „fast immer") bei, ebenso wie den Fokus auf das Denken und Fühlen „in Prüfungssituationen (Klassenarbeiten, Tests oder mündlichen Prüfungen)" (S. 123). Ähnlich wie bei Sarason (1984) in der Entwicklung der RTT wurde für die Skalenbildung ein größerer Itempool erstellt. Für diesen wurden vorab drei thematische Bereiche definiert. Dies waren erstens Items, die kognitive Angstaspekte betrafen, wobei Hodapp (1991) als neuen Aspekt Items zur Erfassung von Zuversicht einbezog. Die Items zum (zweitens) emotionalen Erleben beinhalteten sowohl Aussagen zu körperlichen Empfindungen sowie zu affektiven Eindrücken wie Nervosität. Ein dritter Bereich umfasste nicht näher definierte „Fluchtkognitionen" und darüber hinaus Items, die sich auf „kognitive Interferenz im Sinne der Beeinträchtigung des aufgabenbezogenen Denkens" (Hodapp, 1991, S. 123)[7] beziehen. Inhalt dieser Items sollten *nicht* unspezifische Kognitionen sein, die aufgabenirrelevant sind, sondern spezifisch interferierende Kognitionen, „die die aufgabenbezogene Informationsverarbeitung beeinträchtigen" (Hodapp et al., 2011, S. 11). Der Fragebogen wurde 713 Berufs- und Gymnasialschülern zur Bearbeitung vorgelegt. Auf Basis einer Hauptkomponentenanalyse mit orthogonaler Rotation wurden schließlich vier Faktoren extrahiert. Die den Faktoren zugeordneten vier Subskalen des TAI-G sind in Tabelle 5 mit je einem Itembeispiel aufgelistet.

Tabelle 5: Skalen des TAI-G (Hodapp, 1991)

Skala	Itembeispiel
Mangel an Zuversicht	Ich vertraue auf meine Leistung.
Aufgeregtheit	Ich spüre ein komisches Gefühl im Magen.
Besorgtheit	Ich denke über meine Fähigkeit oder Begabung nach.
Interferenz	Mir schießen plötzlich Gedanken durch den Kopf, die mich blockieren.

Anm.: Besorgtheit entspricht Worry, Aufgeregtheit entspricht Emotionality

Alle Subskalen wurden unter Beibehaltung einer hohen Reliabilität so weit wie möglich gekürzt, so dass der TAI-G aus insgesamt 30 Items besteht. Der TAI-G beinhaltet somit eine Erweiterung des Zwei-Komponenten-Modells um die Facetten Mangel an Zuversicht und Interferenz. Mangel an Zuversicht wird dabei explizit als eigenständiger Faktor und nicht als Bestandteil von Besorgtheit aufgefasst. Hodapp (1989) führte eine längsschnittliche Untersuchung mit Schülern (N = 134;

[7] Die Autoren berufen sich hier auf Rost und Schermer (1989). Rost und Schermer (2007) kritisieren, dass der Skala Test-Irrelevant Thinking des RTT der „für erhöhte Leistungsängstlichkeit als typisch angesehene Belastungsaspekt" (S. 99) fehle.

7. Klasse[8]; Studie 2) durch, bei der im Vorlauf einer Prüfung u. a. die Testängstlichkeit, Zuversicht (in dieser Studie kam eine positiv gepolte Skala zum Einsatz) sowie das Anspruchsniveau erfasst wurden. Auf Basis der Ergebnisse wurde ein Pfadmodell berechnet, in dem Zuversicht ein negativer Prädiktor von worry (-.35; beides trait), des Anspruchsniveaus (-.22) und der Einschätzung, dass man das individuell angestrebte Ergebnis erreichen würde (.33), war. Zuversicht ist in diesem Kontext eine Determinante von Testängstlichkeit (Hodapp, 1989) bzw. wirkt dieser entgegen. Zuversicht bedeutet in diesem Fall auch das Erleben von Selbstwirksamkeit sensu Bandura (1977), gleichzeitig schließen sich Zuversicht und Besorgtheit nicht gegenseitig aus (Hodapp, 1989). So ist beispielsweise denkbar, dass eine Person zuversichtlich auf eine anstehende Prüfung blickt (vgl. das Item „Ich bin zuversichtlich, was meine Leistung betrifft"), gleichzeitig aber über die Konsequenzen dieser nachdenkt (vgl. das Item „Ich mache mir Gedanken, wie mein Zeugnis aussehen wird"). Die Frage, inwiefern Mangel an Zuversicht ein genuiner Bestandteil oder eher ein Korrelat von Testängstlichkeit ist, wird am Ende dieses Abschnitts erneut aufgegriffen. Interferenz ist die zweite neue Subskala des TAI-G. Anders als im CIQ und TOQ ist der Inhalt der störenden Kognitionen unspezifisch (z. B. „Ich denke an andere Dinge und werde dadurch abgelenkt"), sie werden also auch nicht kategorisiert (aufgabenrelevant vs. –irrelevant) (Hodapp, 1991). Darüber hinaus ist für die Skala Interferenz im TAI-G der *störende*, *interferierende* Charakter der Kognitionen zentral (Hodapp et al., 2011), während besagte Kognitionen gemäß der Konzeption in der RTT bzw. im CIQ und TOQ nicht zwingend störend sein müssen (vgl. z. B. das CIQ-Item „I thought about something that happened in the distant past").

Die sich verändernden Vorstellungen der Struktur des Konstrukts gingen auch mit Versuchen einher, unterschiedliche theoretische Konzeptionen bzw. deren Operationalisierung zusammen zu führen. Benson et al. (1992) bemühten sich um eine Integration der RTT von Sarason (1984) und des TAI von Spielberger et al. (1978). Auf Basis einer studentischen Stichprobe (N = 818) aus drei Ländern (USA, Deutschland, Ägypten) wurde der gemeinsame Itempool mehreren Faktorenanalysen und Itemselektionen unterzogen. Das Resultat war die Revised Test Anxiety Scale (RTA), die von Anderson und Sauser (1995) zum damaligen Zeitpunkt als „state of the art" (S. 22) bezeichnet wurde. Die RTA erfasst mit 18 (in überarbeiteter Form 20) Items die vier Testängstlichkeits-Facetten, die auch in der RTT abgebildet sind (Benson et al., 1992; Benson & El-Zahhar, 1994).

In den letzten Jahrzehnten hat sich die Vorstellung von Testängstlichkeit als einem multidimensionalen Merkmal durchgesetzt. Einige Studien befassten sich mit der Konstruktvalidität der Interpretation von Testängstlichkeit als mehrdimensional. Hodapp und Benson (1997) unternahmen

[8] Aufgrund der großen Altersheterogenität wird bei Schülerstichproben das Alter bzw. die Schulklasse berichtet; bei Studierendenstichproben wird das Alter nicht berichtet.

eine strengere Prüfung der Frage, ob die unterschiedlichen Faktoren tatsächlich allesamt *Bestandteile* (und nicht nur Korrelate) von Testängstlichkeit sind. Einer amerikanischen und einer deutschen Stichprobe von Studierenden (jeweils N = 218) wurde die RTA (Benson et al., 1992) und der TAI-G (Hodapp, 1991) (jeweils übersetzt) zur Bearbeitung vorgelegt, ebenso wie die Skala zur Erfassung der allgemeinen Selbstwirksamkeit (Jerusalem & Schwarzer, 1986). Die Autoren prüften (mit auf Basis von Faktorladungen reduzierten Itempools) Modelle mit den vier Faktoren Worry, Emotionality, Distraction und Lack of Confidence. Der Faktor Emotionality umfasste dabei die Items der gleichnamigen Skala des TAI-G sowie die RTA-Items aus den Skalen Tension und Bodily Symptoms. Kognitive Interferenz (TAI-G) und Test-Irrelevant Thinking (RTA) hingegen bildeten den Faktor Distraction. Während Worry sich aus den entsprechenden Items beider Instrumente zusammensetzte, bestand Lack of Confidence nur aus den Items des TAI-G (Hodapp & Benson, 1997). Das Modell mit vier Faktoren sowie das entsprechende Modell mit einem übergeordneten Primärfaktor (Test Anxiety) erreichten den besten Modell-Fit. Allerdings zeigte sich in letzterem Modell, dass die Ladung von Distraction auf dem übergeordneten Faktor Test Anxiety mit .40 deutlich unter den Koeffizienten von Worry (.84), Emotionality (.88) und Lack of Confidence (.73) lag. Die Autoren schlussfolgerten, dass Distraction möglicherweise kein genuiner Bestanteil von Testängstlichkeit ist, während Lack of Confidence als weiterer zentraler Bestandteil von Testängstlichkeit neben den beiden „klassischen" Komponenten interpretiert wurde: „Confidence may be conceived as the affective correlate of positive self-efficacy and outcome expectancies and may be contrasted with worry." (Hodapp & Benson, 1997, S. 238). Auch die Beziehung zu Selbstwirksamkeit konnte geklärt werden. Ein Modell mit Selbstwirksamkeit als fünftem Primärfaktor und einem Faktor höherer Ordnung wies keinen guten Fit auf, weshalb sich eine Subsumierung von Selbstwirksamkeit als Bestandteil von Testängstlichkeit nicht ableiten ließ.

Keith et al. (2003) unterzogen den TAI-G abermals konfirmatorischen Faktorenanalysen und erhoben diesen gemeinsam mit einer studienspezifischen Skala zur Erfassung der Selbstwirksamkeit (Jerusalem & Schwarzer, 1986) bei einer vorwiegend studentischen Stichprobe (N = 302). Ein Modell mit den vier Faktoren erster Ordnung Besorgtheit, Aufgeregtheit, Mangel an Zuversicht und Interferenz erzielte einen akzeptablen Modell-Fit. Ein modifiziertes Modell mit einem zusätzlichen Faktor zweiter Ordnung wies ebenfalls einen akzeptablen Fit auf. Die Ladungen der primären Faktoren auf dem sekundären Faktor lagen alle in einem ähnlichen Bereich (.65 bis .75) – die bei Hodapp und Benson (1997) festgestellte schwache Ladung von Distraction auf dem übergeordneten Faktor Testängstlichkeit wurde also hier mit dem inhaltlich äquivalenten Faktor Interferenz nicht repliziert. Ein Modell mit Selbstwirksamkeit als fünftem, primärem Faktor und einem Faktor zweiter Ordnung erreichte nur einen schwachen Fit, was, ähnlich wie bei Hodapp und Benson (1997), gegen die Auffassung von Selbstwirksamkeit als Bestandteil von Testängstlichkeit

spricht. Zusammengefasst sprechen diese Befunde für die Konstruktvalidität des vierfaktoriellen Modells des TAI-G.

In jüngerer Zeit war der TAI-G abermals Grundlage für Weiterentwicklungen. Wacker, Jaunzeme und Jaksztat (2008) replizierten auf Basis einer Stichprobe von *N* = 720 Studierenden mit einer Hauptkomponentenanalyse und obliquer Rotation die vierfaktorielle Struktur. Durch Elimination von wenig trennscharfen Items wurde der Gesamtfragebogen auf 15 Items reduziert, wobei die Faktorstruktur erhalten blieb. Die internen Konsistenzen sanken nur geringfügig. Dieser TAI-G Form X-U wurde in der vorliegenden Arbeit zur Erfassung der dispositionellen Testängstlichkeit eingesetzt (und in modifizierter Form auch zur Erfassung der Testangst als state). Den vorläufig jüngsten Schritt in der Skalenentwicklung stellt der Prüfungsangstfragebogen (PAF) von Hodapp et al. (2011) dar. Diese Weiterentwicklung des TAI-G besteht aus 20 Items. Die hierarchische Struktur mit vier Faktoren erster und einem Faktor zweiter Ordnung wurde abermals konfirmatorisch bestätigt. Darüber hinaus bietet der PAF Normwerte für Studierende (*N* = 1.350) und Grobnormen für Schüler (*N* = 340) an. Zuletzt sei noch auf die Interkorrelation der Skalen (bzw. latenten Variablen) im PAF eingegangen. Diese finden sich in Tabelle 6. Die Problematik der sehr starken Interkorrelationen zwischen Worry und Emotionality des TAI konnte im PAF (bzw. im TAI-G und TAI-G XU) gelöst werden.

Tabelle 6: Interkorrelationen der Skalen bzw. Faktoren des PAF (Hodapp et al., 2011)

	AU	BE	IN	MZ
Aufgeregtheit		.47	.51	.48
Besorgtheit	.38		.33	.20
Interferenz	.40	.25		.51
Mangel an Zuversicht	.41	.15	.41	

N = 1.350; Werte unterhalb der Diagonalen entsprechen Skaleninterkorrelationen; Werte über der Diagonalen entsprechen Korrelationen der Primärfaktoren

Eine wichtige Limitation des TAI-G bzw. PAF ist die Eingrenzung des Erlebensspektrums auf die Prüfungs- bzw. Testsituation selbst. Eine alternative Herangehensweise an das Konstrukt bietet das Differentielle Leistungsangst Inventar (DAI) von Rost und Schermer (2007). Das DAI unterscheidet vier verschiedene Bereiche von Leistungsängstlichkeit und nimmt dabei eine prozessuale Perspektive ein. Diese Bereiche sind „Angstauslösung", „Angstmanifestation", „Angststabilisierung" und „Angst-Copingstrategien", die jeweils mit mehreren Skalen erfasst werden. Auch der DAI erfasst das aktuelle Erleben von Angst („Angstmanifestation") multidimensional mit den Skalen „Physiologische Manifestation", „Emotionale Manifestation" und „Kognitive Manifestation". Kognitive Angstmanifestation wird dabei nicht gleichgesetzt mit dem Inhalt von Interferenz des TAI-G bzw. PAF. Erstere umfasst dabei „Störungen der koordinierten Informationsaufnahme, Informationsverarbeitung und des Informationsabrufs" (Sparfeldt et al., 2005, S. 226). Beide Skalen

konvergieren jedoch insofern, dass diese Gedanken wiederum störend wirken. Die höchsten Korrelationen zwischen den Skalen des PAF und DAI finden sich nach Hodapp et al. (2011), auf Basis einer Untersuchung von Stingel (2009), zwischen Aufgeregtheit (PAF) und Physischer bzw. Emotionaler Angstmanifestation (DAI) mit r = .67 bzw. r = .43 sowie zwischen Interferenz (PAF) und Kognitiver Angstmanifestation (DAI) mit r = .50.

1.1.1.2.3 Alternative Konzeptionen von Testängstlichkeit

Die Ausführungen zur theoretischen Konzeption von Testängstlichkeit und der entsprechenden Erhebungsinstrumente sind nicht erschöpfend. Einige Verfahren konzentrieren sich auf bestimmte Aspekte von Testängstlichkeit oder damit eng in Bezug stehende Phänomene. Der CIQ sowie der TOQ wurden bereits erwähnt (Sarason et al., 1986). Ein weiteres Beispiel ist die FRIEDBEN Test Anxiety Scale (FTA) (Friedman & Bendas-Jacob, 1997), bei der eine eigenständige Skala vorgesehen ist für die Angst vor der negativen sozialen Bewertung bzw. dem Verlust an Wertschätzung durch andere im Falle eines Misserfolgs. Ebenfalls jüngeren Datums ist die Cognitive Test Anxiety Scale (CTA) von Cassady und Johnson (2002), die sich allein auf die Erfassung der kognitiven Komponente von Testängstlichkeit beschränkt. (Test-)Ängstlichkeit findet sich darüber hinaus auch als Inhalt bzw. Skala in einigen Verfahren, die sich allgemeiner mit dem Erleben in Tests bzw. Leistungssituationen befassen, beispielsweise im Test Attitude Survey (TAS) (Arvey, Strickland, Drauden & Martin, 1990) oder dem Achievement Emotions Questionnaire (AEQ) (Pekrun, Goetz, Frenzel, Barchfeld & Perry, 2011).

Abschließend soll auf ein Verfahren eingegangen werden, das noch in der frühen Forschungsperiode entwickelt wurde, dessen theoretische Konzeption aber deutlich von den bisher vorgestellten Verfahren abweicht. Der Achievement Anxiety Test (AAT) von Alpert und Haber (1960) erfasst Angst mit zwei Faktoren: leistungshemmende (debilitating, AAT-) und leistungsförderliche (facilitating, AAT+) Aspekte. Die Items erfassen dabei nicht nur das subjektive Erleben von Angst in Prüfungssituationen, sondern auch deren gefühlte positive oder negative Konsequenzen auf die eigene Leistung (z. B. „Nervousness while taking an exam or test hinders me from doing well“; „I work most effectively under pressure, as when the task is very important“). Diese Konzeption geht letztlich zurück auf das Yerkes-Dodson-Gesetz (Anderson & Sauser, 1995; Yerkes & Dodson, 1908), demgemäß der Zusammenhang zwischen Leistung und Aktivierung bzw. Erregung einem umgekehrt U-förmigen Verlauf folgt: optimale Leistung wird dann gezeigt, wenn die Aktivierung moderat ist, demgegenüber ist die Leistung schlechter bei zu niedriger und zu hoher Aktivierung (Landy & Conte, 2013). Erste Hinweise für die Konstruktvalidität lieferten gegensätzliche Zusammenhänge der Skalen mit dem Fragebogen zur Testängstlichkeit von Mandler und Sarason (1952) zu r = .64 (AAT-) bzw. r = -.40 (AAT+). Bei der Kriteriumsvalidität zeigte sich ein ähnliches Muster.

So korrelierte beispielsweise AAT- zu *r* = -.29 bzw. -.35, AAT+ hingegen zu *r* = .21 bzw. .37 mit der verbalen Leistung im SAT bzw. mit einer Durchschnittsnote (Alpert & Haber, 1960). Die Korrelationen zwischen AAT- und AAT+ lagen zwischen *r* = -.34 und -.48 (mehrere studentische Stichproben). Für den Nutzen des Einsatzes beider Skalen spricht, dass sich die Vorhersage der Durchschnittsnote bei Hinzunahme der jeweils anderen Skala verbesserte (Alpert & Haber, 1960). Eine Schwäche des AAT ist jedoch die „unzulässige Konfundierung von Prädiktor (= Angst) und Kriterium (= Leistung)" (Rost & Schermer, 2007, S. 15) innerhalb der Items, wie etwa in folgendem Fall: „Nervousness while taking an exam or test hinders me from doing well". Kritisch zu werten ist auch der hohe Anspruch des AAT an die Introspektion der Befragten: „Time pressure on an exam causes me to do worse than the rest of the group under smiliar conditions". Trotz der vielversprechenden Ausgangsbefunde hat sich die Konzeption von Testängstlichkeit, wie sie im AAT vorliegt, nicht etabliert. Im Vergleich zu den Vorgängern und Nachfolgern von TAI und RTT wurde der AAT deutlich seltener eingesetzt. Anderson und Sauser (1995) stellen hierzu fest: „It is disappointing that the ripples from the Alpert-Haber (1960) splash have all but disappeared." (S. 19). Die Idee, dass Testängstlichkeit und Testangst auch funktionale Wirkungen haben können wurde jedoch immer wieder formuliert und wird in Abschnitt 1.3.2 wieder aufgegriffen. Sie war auch die Grundlage für Studie 3.

1.1.1.2.4 Fazit zur Struktur von Testängstlichkeit

In den Jahrzehnten der Forschung hat sich die Konzeption von Testängstlichkeit gewandelt. Die ein-, und später zweidimensionale Konzeption wurde schließlich durch multidimensionale Vorstellungen und entsprechende Operationalisierungen abgelöst. Gleichwohl verschiedene Facetten oder Dimensionen von Testängstlichkeit vorgeschlagen wurden, weisen die Modelle mehr oder weniger starke Konvergenzen auf (siehe z. B. die Ausführungen zu RTT und TAI-G). Wichtigstes Fazit aus Abschnitt 1.1.1.2 ist, dass Testängstlichkeit – und auch Testangst – kein uniformes Phänomen ist. Die Multidimensionalität von Testängstlichkeit erfordert, dass die Existenz der unterschiedlichen Dimensionen berücksichtigt wird. Die Betrachtung eines aggregierten Gesamtwertes der Testängstlichkeit lässt somit stets Fragen offen, da dadurch unterschiedliche Zusammenhänge einzelner Facetten zu anderen Variablen „verborgen" bleiben können. Unter bestimmten Bedingungen kann es aber sinnvoll sein, sich auf einzelne Facetten zu fokussieren.

Testängstlichkeit als Disposition ist mit einer Reihe von typischen Emotionen und Kognitionen verbunden, die in einer bestimmten Klasse von Situationen ausgelöst werden. Menschen unterscheiden sich im Ausmaß, in dem sie zu Testangst neigen. Die Unterschiede zwischen hoch und niedrig Testängstlichen erstrecken sich auch auf weitere Persönlichkeitsmerkmale, die mehr oder

weniger unmittelbar die Reaktion auf Leistungs- bzw. Bewertungssituationen beeinflussen. Auf das nomologische Netzwerk des Konstrukts soll im Folgenden näher eingegangen werden.

In Abschnitt 1.1.1.3 werden Zusammenhänge von Testängstlichkeit zu Persönlichkeitseigenschaften im engeren Sinne aufgezeigt. In Ermangelung einer etablierten Systematik zur Ordnung von Persönlichkeitsbereichen (Asendorpf & Neyer, 2012) werden die Merkmalsbereiche grob thematisch sortiert. Zunächst werden Befunde herangezogen, die Testängstlichkeit in Bezug zum Big Five Modell gesetzt haben. Danach werden die Beziehungen zu motivationalen Merkmalen und zu Aspekten des Selbstkonzepts geklärt. Über diese Persönlichkeitseigenschaften hinausgehend wird in Abschnitt 1.1.1.4 vertieft, wie sich hoch und niedrig Testängstliche in ihren kognitiven Strukturen und der Informationsverarbeitung unterscheiden. Schließlich werden in Abschnitt 1.1.1.5 Unterschiede bei der Testängstlichkeit in Abhängigkeit von demographischen Merkmalen beleuchtet. Der Fokus der Ausführungen liegt in allen drei Abschnitten auf der Metaanalyse von Hembree (1988) sowie auf ausgewählten Einzelbefunden. Sofern Befunde auf Facettenebene vorliegen und diese für ein weiteres Verständnis des Konstrukts notwendig sind, wird dies entsprechend berichtet.

1.1.1.3 Zusammenhänge und Abgrenzungen zu Persönlichkeitseigenschaften im engeren Sinne

Hodapp et al. (2011) berichten für den PAF auf Basis einer Untersuchung von Pohl (2006) Zusammenhänge mit dem NEO-FFI (Borkenau & Ostendorf, 1993) (N = 262 Realschüler, 7. bis 9. Klasse). Der mit Abstand stärkste Zusammenhang lag mit Neurotizismus vor mit r = .56 (Aufgeregtheit wies mit r = .48 die stärkste Korrelation auf, die übrigen Facetten zwischen r = .32 und .39). Kaum Zusammenhänge fanden sich zu Extraversion (r = .08, wobei Besorgtheit mit r = .18 einen schwachen, signifikanten Zusammenhang aufwies) und Offenheit (r = -.12, hier lag ein signifikanter Zusammenhang mit Mangel an Zuversicht vor, r = -.25). Zwischen Verträglichkeit und Testängstlichkeit lag ein signifikanter Zusammenhang vor mit r = -.26 (zurückgehend auf signifikante Zusammenhänge mit Interferenz von r = -.35 sowie Mangel an Zuversicht von r = -.26). Ebenso fand sich eine signifikante Korrelation mit Gewissenhaftigkeit von r = -.28 (sowohl Aufgeregtheit als auch Interferenz und Mangel an Zuversicht wiesen signifikante Zusammenhänge zu Gewissenhaftigkeit auf mit r = -.15, -.26, -.34). Besorgtheit korrelierte nicht mit Gewissenhaftigkeit.

Chamorro-Premuzic, Ahmetoglu und Furnham (2008) erhoben die RTT sowie den NEO-FFI (Costa & McCrae, 1992) an 388 Studierenden. Die stärkste bivariate Korrelation fand sich zwischen Testängstlichkeit (Gesamtwert) und Neurotizismus mit r = .45. Mit Offenheit, Verträglichkeit und Gewissenhaftigkeit fanden sich signifikante, aber schwache negative Zusammenhänge zwischen r = -.12 und -.13, während zu Extraversion kein Zusammenhang vorlag (r = .06). McIlroy und Bunting (2002) erhoben die RTA sowie die Skala Gewissenhaftigkeit der 16PF (Cattell, Eber &

Tatsuoka, 1985) bei N = 219 Studierenden. Sie berichten keine signifikanten Korrelationen für Worry (r = .07) und Test-Irrelevant Thinking (r = -.03).

Diese Befunde sprechen dafür, dass in Bezug auf die Big Five ein konsistenter und starker Zusammenhang zu Neurotizismus besteht. Dies ist nicht überraschend, beinhaltet Neurotizismus ja auch Nervosität und Ängstlichkeit (Costa & McCrae, 1992). Zumindest was die berichteten studentischen Stichproben betrifft, liegen zu den übrigen Dimensionen niedrige oder gar keine Zusammenhänge vor. Die von Hodapp et al. (2011) berichteten Befunde weisen jedoch auch darauf hin, dass sich die Testängstlichkeitsfacetten in ihren Beziehungen zu den Big Five unterscheiden können.

Testängstlichkeit beeinflusst als Disposition die Reaktion auf Leistungssituationen – eine Berücksichtigung motivationaler Variablen ist daher für das Konstruktverständnis von Bedeutung. Hembree (1988) führte eine umfassende Metaanalyse durch in die k = 562 Studien im Zeitraum von 1950 bis 1986 inkludiert wurden. Neben Zusammenhängen von Testängstlichkeit und Leistung (siehe dazu Abschnitt 1.2) sowie Interventionen wurden auch Persönlichkeitskorrelate betrachtet. Dabei berichtet er eher schwache Zusammenhänge zwischen Leistungsmotivation und Testängstlichkeit: so zeigte sich bei Grundschülern ein mittlerer Zusammenhang von r = -.16 (k = 5, N = 629), bei Schülern der High School von r = .26 (k = 7, N = 1.154) und bei College-Studierenden von r = .03 (n. s.; k = 19, N = 1.744). Allerdings wird in der Metaanalyse die genaue Operationalisierung der Leistungsmotivation nicht ausgeführt.

Eine theoretische Betrachtung der Struktur der Leistungsmotivation legt den Schluss nahe, dass die Beziehungen zur Testängstlichkeit komplex sind. Eine sinnvolle Orientierung liefert hier die Unterscheidung von Erfolgs- und Misserfolgsmotiv oder (äquivalent) von Hoffnung auf Erfolg und Furcht vor Misserfolg (Asendorpf & Neyer, 2012; Atkinson, 1957; Brunstein & Heckhausen, 2010). Atkinson interpretierte Testängstlichkeit aus der Perspektive seines Risiko-Wahl-Modells. Testängstlichkeit wird demnach mit dem Misserfolgsmotiv gleichgesetzt (Atkinson & Litwin, 1960): „The motive to avoid failure is considered a disposition to avoid failure and/or a capacity for experiencing shame and humiliation as a consequence of failure." (Atkinson, 1957, S. 360). Hagtvet und Benson (1997) führten die RTA bei N = 260 Studierenden durch und erhoben das Misserfolgsmotiv mit Items der Achievement Motivation Scale (AMS; Nygård & Gjesme, 2006). Sie postulierten, „that the motive to avoid failure should serve as a basic and general factor of test anxiety" (S. 42). Es wurde ein Strukturgleichungsmodell (SEM) gerechnet, in dem das Misserfolgsmotiv als latente, exogene und die vier Facetten der RTA als latente, engodene Variablen angelegt wurden, so dass das Misserfolgsmotiv – im Sinne des Generalfaktors – ein Prädiktor der vier Facetten war. Dieses Modell erzielte einen akzeptablen Fit, wenn zusätzlich die Residualkovarianzen der RTA-Facetten spezifiziert wurden, hingegen einen schlechteren, wenn diese nicht im Modell

geschätzt wurden. Das Misserfolgsmotiv wies substantielle (standardisierte) Pfadkoeffizienten zur Prädiktion der RTA-Facetten auf, von .34 (Bodily Symptoms) bis hin zu .65 (Worry; Tension). Die partiellen (um das Misserfolgsmotiv korrigierten) Faktorinterkorrelationen der RTA-Facetten lagen zwischen .26 und .40, mit Ausnahme von Tension und Test-Irrelevat Thinking (.07) sowie Bodily Symptoms und Test-Irrelevant Thinking (.02). Die Interkorrelationen der RTA-Facetten sanken durch die Spezifikation des Misserfolgsmotivs als endogener Variable deutlich, verschwanden aber nicht vollständig. Die Ergebnisse stützten insgesamt die Idee des Generalfaktors Furcht vor Misserfolg, wenngleich ungeklärte Varianz und Kovarianz der RTA-Facetten übrig blieb. Lang und Fries (2006) berichten für eine von ihnen entwickelte, gekürzte Version des AMS (AMS-R) bei einer Stichprobe von N = 126 Schülern (9. und 10. Klasse) Korrelationen mit dem TAI-G (Hodapp et al., 1982). Die beiden Skalen des AMS-R wiesen ein divergierendes Zusammenhangsmuster mit Testängstlichkeit auf: während Furcht vor Misserfolg zu r = .40 korrelierte, wies Hoffnung auf Erfolg mit r = .03 keinen Zusammenhang auf.

Sind also die Beziehungen zu Hoffnung auf Erfolg und Furcht vor Misserfolg recht eindeutig, ist die Klärung in Bezug auf andere Aspekte der Leistungsmotivation schwieriger. Prinzipiell wäre zu erwarten, dass – durchaus analog zu der Vorstellung von Hagtvet und Benson (1997) – die Wichtigkeit von Leistung *per se* überhaupt erst die Voraussetzung ist, damit in einer Situation Testangst entsteht. Hembree (1988) berichtet jedoch einen nicht signifikanten Zusammenhang von Testängstlichkeit mit dem Anspruchsniveau, r = -.05 (k = 8, N = 431). Ähnlich wie bei der Gewissenhaftigkeit ist zu vermuten, dass divergierende Zusammenhangsmuster der einzelnen Facetten eine Rolle spielen. Musch und Bröder (1999a) erhoben bei N = 91 Studierenden den TAI-G (Hodapp, 1991) und die Skalen „Leistungsstreben“ (z. B. „Mehr zu leisten als andere, finde ich wichtig“) und „Ausdauer und Fleiß“ (z. B. „Wenn ich mit einer schwierigen Aufgabe beschäftigt bin, bleibe ich meistens dabei“) des Leistungs-Motivations-Test LMT (Hermans, Petermann & Zielinski, 1978; zitiert nach Krohne & Hock, 2015). Sie berichten bivariate Korrelationen auf Facettenebene. So korrelierte lediglich Besorgtheit signifikant positiv mit Leistungsstreben zu r = .30, während Mangel an Zuversicht und Interferenz signifikant negativ mit Ausdauer und Fleiß korrelierten, r = -.33 bzw. -.23 (Musch & Bröder, 1999a). Der signifikante Zusammenhang von Besorgtheit und Leistungsstreben scheint den nicht signifikanten Zusammenhängen von Besorgtheit und Gewissenhaftigkeit bei Hodapp et al. (2011) zu widersprechen. Diese Diskrepanz könnte allerdings darauf zurückgehen, dass in der Skala Leistungsstreben des LMT soziale Vergleichsaspekte eine Rolle spielen (Krohne & Hock, 2015), anders als bei der Skala Gewissenhaftigkeit des NEO-FFI (siehe hierzu Borkenau & Ostendorf, 1993) .

Eine nähere Klärung dieser komplexen Befundsituation ermöglicht das Hierarchische Motivationsmodell von Elliot und Kollegen (z. B. Elliot & McGregor, 2001). Zentrale Vorannahme des Modells ist die Differenzierung von leistungsbezogenen Motiven und Zielen. Fokus der Theorie liegt

auf den Zielen (achievement goals). Leistungsmotive werden als grundlegende Neigungen verstanden, die motiviertem Verhalten zugrunde liegen. Ziele hingegen „are construed as more concrete, midlevel cognitive representations that direct individuals toward specific end states“ (Elliot & McGregor, 1999, S. 628). Sie lassen sich anhand der beiden Dimensionen Annäherung und Vermeidung unterschieden. Annäherungsmotivation ist „the energization of behavior by, or the direction of behavior toward, positive stimuli (objects, events, possibilities)“; Vermeidungsmotivation ist „the energization of behavior, or the direction of behavior away from, negative stimuli (objects, events, possibilities)“ (Elliot, 2006, S. 112). Leistungsmotive beeinflussen die Ausformung von Zielen, welche ihrerseits – unmittelbarer als die Motive – dem tatsächlichen Erleben und Verhalten zugrunde liegen (Brunstein & Heckhausen, 2010; Elliot, 2006). Die Dimension Annäherung (approach) vs. Vermeidung (avoidance) beschreibt die Valenz des Gütemaßstabs (Brunstein & Heckhausen, 2010), die entweder positiv (Erfolg erreichen) oder negativ ist (Misserfolg verhindern). Eine zweite Dimension spezifiziert die Definition oder auch Setzung des Gütemaßstabes. Dieser kann entweder absolut, intrapersonal oder normativ sein. Ein absoluter Gütemaßstab ergibt sich unmittelbar aus inhärenten Aufgabenmerkmalen. Intrapersonale Gütemaßstäbe werden in Bezug auf die eigene Person definiert, normative hingegen in Relation zu anderen Personen. Wie kann also eine Leistung konkret beurteilt werden? Ein absoluter Standard wäre erreicht, wenn eine Aufgabe bewältigt wurde. Ein intrapersonaler Standard basiert auf vergangener Leistung oder dem möglichen Leistungsmaximum – er wäre erfüllt, wenn eine Person relativ zu früheren Leistungen eine Leistungssteigerung erreicht hat, ihre Fähigkeiten erweitert oder eine neue Fähigkeit erworben hat (was letztlich einer Leistungssteigerung entspricht). Ein normativer Standard wäre erreicht, wenn die eigene Leistung jene von anderen übersteigt (Elliot & McGregor, 2001). Ein Beispiel: Ein Student muss in seiner Ausbildung die Grundlagen der deskriptiven Statistik verstehen. Hat er die wichtigsten deskriptivstatistischen Kennwerte und deren Zusammenhang verstanden, hat er nicht nur ein Lernkriterium erfüllt (absolut), sondern auch seine Kenntnisse erweitert (intrapersonal). Wäre er dabei gleichzeitig besser als seine Kommilitonen, wäre auch ein normativer Standard erfüllt. Wie im Falle des Beispiels gehen Elliot und McGregor (2001) prinzipiell davon aus, dass sich absolute und intrapersonale Standards in vielen Fällen nicht separieren lassen. Demnach werden absolute und intrapersonale Standards als „mastery“-Ziele zusammengelegt, normative Standards werden als „performance“-Ziele bezeichnet (in der deutschsprachigen Forschung findet sich die Differenzierung in Lern- und Leistungsziele, siehe z. B. Bachmann, 2009; Schöne, 2007). Durch die Kombination der Valenz- und Standard-(oder auch Kompetenz-)Dimension, ergibt sich ein 2x2-Modell mit 4 Zielen, das in Tabelle 7 veranschaulicht ist.

Tabelle 7: Taxonomie der Zielorientierung im Hierarchischen Motivationsmodell (Elliot & McGregor, 2001)

		Definition*	
		Mastery**	Performance
Valenz	Approach	Mastery-approach	Performance-approach
	Avoidance	Mastery-avoidance	Performance-avoidance

* in deutschsprachigen Publikationen findet sich häufig die Unterschiedung von Lern- und Leistungszielen
** beinhaltet absolute und intrapersonale Standards

Ein ursprüngliches trichotomes Modell ohne Differenzierung der mastery-goals (Elliot & McGregor, 1999) wurde schließlich zu dem genannten 2x2-Modell erweitert (Elliot & McGregor, 2001). Was die Ziele konkret bedeuten, lässt sich am genannten Beispiel aufzeigen. Mastery-approach (M-App) würde bedeuten, dass der Student die deskriptive Statistik beherrschen möchte, mastery-avoidance (M-Av) hingegen, dass der Student – im gewissen Sinne spiegelbildlich zu M-App – vermeiden möchte, ein fehler- oder lückenhaftes Verständnis der deskriptiven Statistik aufzubauen. Performance-approach (P-App) läge vor, wenn der Student in der deskriptiven Statistik besser sein will als seine Kommilitonen, performance-avoidance (P-Av) hingegen, wenn er nicht schlechter als diese sein möchte.

Diese vier Ziele haben spezifische, sich aber teilweise überschneidende Antezedenzen und Folgen (Elliot & McGregor, 2001). Elliot und McGregor (1999) haben Testängstlichkeit und Testangst explizit in das (frühere, trichotome) Modell eingebettet. Die theoretische Kernannahme lässt sich folgendermaßen zusammenfassen: Testängstlichkeit (trait), die eng mit der Furcht vor Misserfolg verwandt ist (s. o.), beeinflusst sowohl die Bildung von P-App- als auch P-Av-Zielen. Während die Entstehung von P-Av-Zielen aufgrund von Furcht vor Misserfolg naheliegt, gehen Elliot und Church (1997) davon aus, dass auch P-App-Ziele durch das Misserfolgsmotiv determiniert sein können. Grund hierfür ist, dass das Erreichen eines normativen Gütemaßstabes (performance) stets der Erfahrung von Misserfolg entgegen wirkt, in diesem Sinne also auch der Vermeidung von Misserfolg dienlich ist („approach in order to avoid failure"; Elliot & Church, 1997, S. 220). Beide Zielformen wirken sich wiederum auf die Leistung aus, wobei jedoch der Effekt von P-Av durch Testangst (state) mediiert wird (Elliot & McGregor, 1999). Nach dieser Annahme ist dieser Mediator vorwiegend worry, nicht emotionality, was aus der unterschiedlichen Leistungsrelevanz beider Facetten (siehe Abschnitt 1.2) abgeleitet wird (Elliot & McGregor, 1999).

Elliot und McGregor (1999) führten zur Prüfung dieser Annahmen eine Studie mit 150 Studierenden durch. Bei den Probanden (Teilnehmern eines Kurses) wurden die Leistungsziele bezüglich des Kurses erfasst. Nach einer Prüfung wenige Wochen nach Beginn des Kurses wurde die in der Prüfung erlebte Testangst (state) erhoben (u. a. worry und emotionality). In einem Pfadmodell war P-App ein signifikanter, direkter Prädiktor der Prüfungsleistung (β = .22). Der Effekt von P-

Av auf die Leistung wurde mediiert: P-Av prädizierte worry (β = .29), welches wiederum die Leistung vorhersagte (β = -.43). Emotionality hingegen wies keinen Zusammenhang zur Leistung auf. In einer zweiten Studie wurde ein weiteres Modell geprüft unter Inklusion der Testängstlichkeit (trait). In einem ähnlichen Untersuchungsdesign wurde u. a. auch die Testängstlichkeit bei 172 Studierenden eines Kurses erhoben. Wie in der ersten Studie wurde nach einer Prüfung im späteren Kursverlauf die erlebte Testangst erhoben. Dabei wurden zunächst die Ergebnisse der ersten Studie repliziert. Testängstlichkeit sagte sowohl P-App (β = .33) als auch P-Av (β = .54) vorher. Testängstlichkeit hatte einen direkten (β = .23) und indirekten, über P-Av vermittelten Effekt auf worry. Ein wichtiges Teilergebnis war überdies, dass die Relation zwischen P-App und der Prüfungsleistung nicht durch worry mediiert wurde (Elliot & McGregor, 1999). Auf Basis weiterer Studien fassen Elliot und McGregor (2001) für die 2x2-Taxonomie zusammen, dass Furcht vor Misserfolg beide Vermeidungsziele (M-Av und P-Av), aber auch P-App antezediert. Testangst (state) wiederum ist eine Folge von M-Av und P-Av.

Elliot und Pekrun (2007) nehmen theoretisch an, dass P-Av enger mit Testangst zusammenhängt als M-Av, da ein normativer Gütemaßstab (P-Av) mit einem höheren Maß an Bewertungsstress verbunden ist als ein intrapersonaler oder absoluter Gütemaßstab (M-Av). Empirisch zeigt sich aber teilweise ein umgekehrtes Bild, nämlich stärkere Zusammenhänge zu M-Av als zu P-Av (z. B. Eum & Rice, 2011; Putwain & Daniels, 2010; Putwain & Symes, 2012). Putwain und Symes (2012) berichten auf Basis einer Schülerstichprobe (M_{Alter} = 17.1; N = 275) Zusammenhänge zwischen der RTA und Zielorientierung. Insgesamt zeigten sich, mit Ausnahme von Test-Irrelevant Thinking, positive bivariate Korrelationen zwischen den RTA-Facetten und M-Av, P-App und P-Av. Worry wies dabei den stärksten Zusammenhang zu M-Av auf (r = .37), ebenso wie Tension (r = .35), während Bodily symptoms am stärksten mit P-Av korrelierte (r = .21). Ein schwach positiver Zusammenhang zwischen Tension und M-App (r = .17) könnte auf funktionale Wirkungen dieser Facette hinweisen (Putwain & Symes, 2012).

Zusammenfassend lässt sich also sagen, dass Testängstlichkeit eng mit dem Misserfolgsmotiv verknüpft ist, wobei ersteres spezifischer, also auf Prüfungs- bzw. Testsituationen bezogen ist (Elliot & McGregor, 1999). Wenngleich diese Zuordnung auf Motivebene recht klar ist, sind die Zusammenhänge auf Ebene der Zielorientierung komplexer. Letztere Beziehungen sind ambivalent und keineswegs rein negativ, da Testangst sowohl mit Annäherungs- als auch Vermeidungszielen assoziiert sein kann.

Der subjektive Wunsch, ein gutes Ergebnis in einem Test zu erzielen (oder auch ein schlechtes zu verhindern) ist eine notwendige, aber noch nicht hinreichende Voraussetzung dafür, dass eine Person Testangst verspürt. Die Vermutung liegt nahe, dass hohes Leistungsstreben selbst eine

Bedingung für die Entstehung von Testangst ist, also *per se* maladaptiv sein könnte. Arbeiten, die sich mit der Beziehung zwischen Testängstlichkeit und Perfektionismus befasst haben, tragen hier zur Klärung bei. Perfektionismus lässt sich an zwei Merkmalen festmachen, nämlich der Setzung und Verfolgung hoher Ansprüche an die eigene Leistung, wobei das eigene Verhalten einer sehr strengen und kritischen Bewertung unterzogen wird (Frost, Marten, Lahart & Rosenblate, 1990). Kernmerkmal von Perfektionismus ist die niedrige bzw. gar nicht vorhandene Toleranz gegenüber Fehlern oder Abweichungen von dem hohen Standard (vgl. z. B. das Item „It makes me uneasy to see an error in my work" der Multidimensional Perfectionism Scale MPS; Hewitt & Flett, 1991). Perfektionismus wird als multidimensionales Merkmal aufgefasst, das relativ stabil ist (Hewitt & Flett, 1991). Befunde zu Perfektionismus zeigen immer wieder, dass Perfektionismus nicht per se maladaptiv ist. Eine wichtige Unterscheidung ist die von self-oriented und socially-prescribed perfectionism (kurz: SOP vs. SPP), die auf Hewitt und Flett (1991) zurückgeht. SOP bedeutet, dass eine Person selbst den Wunsch hat und motiviert ist, sehr hohe Leistung zu erbringen (z. B. das MPS-Item „One of my goals is to be perfect in everything I do"). Demgegenüber bedeutet SPP, dass eine Person hohe Standards erreichen will, die ihr (subjektiv) von anderen Personen, also sozial, abverlangt werden (z. B. das MPS-Item „The people around me expect me to succeed at everything I do") (Hewitt & Flett, 1991). Stoeber, Feast und Hayward (2009) fassen auf Basis der vorliegenden Literatur zusammen, dass SPP eher maladaptiv, SOP hingegen ambivalent ist, also sowohl positive als auch negative Konsequenzen (und Korrelate) aufweist.

Dies lässt sich auf die Beziehungen zu Testängstlichkeit übertragen. Stoeber et al. (2009) erhoben an N = 105 Studierenden u. a. SPP und SOP mit der MPS von Hewitt und Flett (1991) sowie Testängstlichkeit mit einer ins Englische rückübersetzten Form des TAI-G (Hodapp, 1991; Hodapp & Benson, 1997). SOP und SPP, die moderat positiv miteinander korrelierten (r = .25), wiesen deutlich unterschiedliche Zusammenhänge zu Testängstlichkeit auf. Stoeber et al. (2009) berechneten Partialkorrelationen zwischen SOP bzw. SPP und Testängstlichkeit, d. h. kontrolliert um die jeweils andere Perfektionismusdimension. Der Gesamtwert des TAI-G korrelierte nicht mit SOP (r = -.02), jedoch signifikant mit SPP (r = .33). Ähnliche Ergebnisse berichten auch Weiner und Carton (2012), die bei N = 170 Studierenden u. a. die TAS sowie die Frost Multidimensional Perfectionism Scale (FMPS) (Frost et al., 1990) einsetzten. Diese unterscheidet sechs Dimensionen von Perfektionismus. Die Dimensionen „Concern Over Mistakes" (z. B. „If I fail at work/school, I am a failure as a person") und „Doubts About Actions" (z. B. „It takes me a long time to do something "right"."") wurden zu einer Skala „Evaluative Concerns Perfectionism", kurz ECP, zusammengelegt. Eine weitere Dimension ist „Personal Standards" (z. B. „I have extremely high goals", kurz PS). Die Differenzierung von ECP und PS ist deckungsgleich mit der von SPP und SOP (Weiner & Carton, 2012). Testängstlichkeit korrelierte deutlich positiv mit ECP, r = .42, hingegen schwach negativ (aber signifikant) mit PS, r = -.16 (Weiner & Carton, 2012). Dieses Zusammenhangsmuster

unterscheidet sich allerdings bei den Facetten der Testängstlichkeit. Stoeber et al. (2009) berichten hierzu Ergebnisse für den TAI-G. So korrelierte Besorgtheit positiv mit SOP (r = .29), während Interferenz (r = -.24) und Mangel an Zuversicht (r = -.39) signifikant negative Zusammenhänge hiermit aufwiesen. Dieses Muster verdeutlicht auch die Ambivalenz von SOP: Während es einerseits mit weniger Interferenz und einer höheren Zuversicht assoziiert ist, geht es gleichzeitig mit Sorgen über die eigene Leistung einher. SPP hingegen wies deutliche positive Zusammenhänge zu Interferenz (r = .46) und Mangel an Zuversicht auf (r = .35), allerdings weder signifikante Zusammenhänge zu Besorgtheit noch zu Aufgeregtheit. Die Befunde zur Interferenz verweisen darauf, dass hohe, sozial begründete Ansprüche an die eigene Person selbst eine Quelle aufgabenirrelevanter Kognitionen sein könnten. SOP hingegen ist zwar mit einer erhöhten Sorge über die eigene Leistung verknüpft, aber gleichzeitig auch mit stärkerer Zuversicht (Stoeber et al., 2009).

Eine letzte in diesem Kontext zitierte Studie betrachtete neben Perfektionismus auch Leistungsziele als Prädiktoren von Testängstlichkeit. Eum und Rice (2011) erhoben bei N = 134 Studierenden kognitive Aspekte der Testängstlichkeit mit der CTA (Cassady & Johnson, 2002) sowie den Achievement Goal Questionnaire (AGQ) zur Erfassung der Leistungsziele (Elliot & McGregor, 2001). Ferner bildeten sie Indizes für maladaptiven bzw. adaptiven Perfektionismus aus den Skalen „Discrepancy“ (z. B. „My performance rarely measures up to my standards.“) bzw. „High Standards“ (z. B. „I have a strong need to strive for excellence.“) der ebenfalls erhobenen Revised Almost Perfect Scale (APS-R; Slaney, Rice, Mobley, Trippi & Ashby, 2001). In einer multiplen Regression zur Vorhersage der (kognitiven) Testängstlichkeit waren (abgesehen vom Prädiktor Geschlecht mit β = -.25[9]) drei Prädiktoren signifikant: M-Av (β = .17), P-Av (β = .24) sowie maladaptiver Perfektionismus (β = .37). Diese vier Variablen klärten fast 50 % der Varianz des CTA-Wertes auf.

Testängstliche Personen sind also gekennzeichnet durch den Wunsch, Misserfolg zu vermeiden. Dies geht häufig einher mit unrealistischen, hohen Ansprüchen an die eigene Person, welche häufig (aus Sicht der Person) von außen an diese gestellt werden. Warum sich Testängstliche vor Leistungssituationen fürchten, kann anhand der Theorie zur Stress- bzw. Emotionsentstehung von Lazarus erklärt werden (Lazarus & Folkman, 1984). Die Einschätzung der eigenen Ressourcen (sekundäre Bewertung) zur Bewältigung einer Situation spielt in der Entstehung von Testangst eine wichtige Rolle: „For example, if in the course of secondary appraisal a student judges her or his cognitive coping resources to be adequate for dealing with a threatening midterm college exam, the degree of threat, as assessed during primary appraisal, would be diminished.” (Zeidner, 1998, S. 186). Da Personen mit hoher Testängstlichkeit eher dazu neigen, einen

[9] Gemäß der Kodierung von Geschlecht indiziert dies höhere Werte bei Frauen

Test als bedrohlich zu empfinden, kommen sie offenkundig häufig zu dem Urteil, dass ihre Ressourcen eben *nicht* genügen, um eine Anforderung erfolgreich zu bewältigen (siehe hierzu Spielberger & Vagg, 1995b; Zeidner, 1998).

Es liegt daher auf der Hand, dass Testängstlichkeit mit globalen und spezifischen Überzeugungen über die eigene Person verbunden ist, sprich mit Repräsentationen der eigenen Fähigkeiten und Fertigkeiten. Hembree (1988) berichtet in seiner Metaanalyse einen mittleren Zusammenhang zwischen Selbstwert und Testängstlichkeit von r = -.42 (k = 36, N = 8.839). Testängstliche Personen haben also einen niedrigeren Selbstwert. Testängstlichkeit ist darüber hinaus assoziiert mit Kontrollüberzeugungen. Nachvollzogen werden kann das an der Differenzierung von Kontrollüberzeugungen, die auf Rotter (1966) zurückgeht. Demnach unterscheiden sich Menschen in der Wahrnehmung der Assoziation (konkret: Kontingenz) von eigenem Verhalten und Ereignissen, wobei Rotter (1966) eine lerntheoretische Perspektive einnimmt und von der erlebten Kontingenz zwischen eigenem Verhalten (auch eigenen Eigenschaften) und Verstärkung spricht. Externale Kontrollüberzeugung liegt vor, wenn Ereignisse als weitgehend unabhängig vom eigenen Verhalten empfunden werden, also auf Zufall oder Glück oder auch Handlungen anderer Personen zurückgeführt werden. Wird ein Ereignis hingegen maßgeblich auf das eigene Verhalten oder eigene Eigenschaften zurückgeführt, liegt eine interne Kontrollüberzeugung vor (Rotter, 1966). Diese Konstrukte werden häufig unter dem Begriff locus of control (z. B. Rotter, 1975) behandelt. Hembree (1988) berichtet einen Zusammenhang von r = .22 (k = 16, N = 2.222) zwischen Testängstlichkeit und locus of control, Testängstliche weisen also tendenziell eine externale Kontrollüberzeugung auf. Dieser Befund passt gut zu den berichteten Assoziationen mit dem Selbstwert.

Auch die Beziehungen zum Selbstkonzept fügen sich in dieses Bild. Von besonderer Relevanz ist hier das Fähigkeitsselbstkonzept, also „die Gesamtheit der kognitiven Repräsentationen eigener Fähigkeiten“ (Dickhäuser, Schöne, Spinath & Stiensmeier-Pelster, 2002, S. 394). Eine Reihe beispielhafter Befunde hierzu sei skizziert. Krampen (1988) führte mit N = 346 Schülern (6. bis 10. Klasse) eine längsschnittliche Untersuchung durch. Die beiden Messzeitpunkte lagen zehn Monate auseinander. Erhoben wurde unter anderem das Selbstkonzept eigener mathematischer Fähigkeiten sowie die Testängstlichkeit mit der Skala „Prüfungsangst“ des Angstfragebogens für Schüler (AFS; Wieczerkowski, Nickel, Janowski, Fittkau & Rauer, 1975). Zu beiden Messzeitpunkten zeigte sich ein deutlicher negativer Zusammenhang zwischen mathematischem Selbstkonzept und Testängstlichkeit, r = -.47 bzw. -.50. Überdies sagte das Selbstkonzept zu t1 die Testängtlichkeit zu t2 mit r = -.27 vorher. Rohrmann, Bechtoldt, Schnell und Hodapp (2010) erhoben den PAF, den Fragebogen zu Kompetenz- und Kontrollüberzeugungen (FKK) von Krampen (1991) sowie die Skala zur Erfassung der Selbstwirksamkeit (Jerusalem & Schwarzer, 1999) bei N = 349 Studierenden. Der FKK erfasst in vier Primärskalen das Selbstkonzept eigener Fähigkeiten (FKK-SK, z.

B. „Auch in schwierigen Situationen fallen mir immer viele Handlungsalternativen ein.“), Internalität (FKK-I, z. B. „Wenn ich bekomme, was ich will, so ist das immer eine Folge meiner Anstrengung und meines persönlichen Einsatzes.“), Soziale Externalität (FKK-P, z. B. „Mein Wohlbefinden hängt in starkem Masse vom Verhalten anderer Menschen ab.“) und Fatalistische Externalität (FKK-C; z. B. „Vieles von dem, was in meinem Leben passiert, hängt vom Zufall ab.“) (Krampen, 1991). Der Gesamtwert des PAF korrelierte zu r = -.39 mit FKK-SK, darüber hinaus negativ mit FKK-I zu r = -.18 und positiv mit FKK-P zu r = .32 sowie positiv mit FKK-C zu r = .29. Testängstlichkeit geht also mit einem niedrigen Fähigkeitsselbstkonzept einher sowie niedrigen internen, aber ausgeprägten externen Kontrollüberzeugungen. Der Gesamtwert der Testängstlichkeit korrelierte überdies deutlich negativ mit der Selbstwirksamkeit zu r = -.40. Hierzu passt auch ein Befund von Ringeisen, Buchwald und Hodapp (2010), die einen Zusammenhang zwischen dem Gesamtwert des TAI-G und dem dispositionellen Optimismus (erfasst mit dem Life Orientation Test LOT-R, Glaesmer, Hoyer, Klotsche & Herzberg, 2008) von r = -.40 ermittelten (deutsche Teilstichprobe, N = 183 Studierende).

Die dargestellten Befunde legen den Schluss nahe, dass Testängstlichkeit als situationsspezifische Form der Ängstlichkeit (Spielberger, 1972b) eine maladaptive Eigenschaft ist. Testängstlichkeit ist stark assoziiert mit Neurotizismus und Testängstliche haben überdies einen niedrigeren Selbstwert, ein schwächeres Selbstkonzept ihrer eigenen Fähigkeiten und eine niedrigere Selbstwirksamkeit als wenig Testängstliche. Zu dieser tendenziell ungünstigen Sicht auf die eigene Person passen auch die Neigung zu ungünstigen Seiten des Perfektionismus sowie die ausgeprägte Furcht vor Misserfolg. Bislang wurden vorwiegend korrelative Befunde berichtet, die die Frage nach kausalen Beziehungen offen lassen. So ist es denkbar, dass Testängstlichkeit bzw. Testangst das Resultat einer negativen – aber objektiv zutreffenden – Einschätzung der eigenen (In)Kompetenz ist. Diese Frage soll im Kontext der Defizitperspektive (Abschnitt 1.2.1.2) vertieft werden, welche auch für Studie 1 von besonderer Relevanz ist.

Es ist naheliegend und empirisch gut abgesichert, dass sich diese interindividuellen Unterschiede in Persönlichkeitseigenschaften auch im konkreten Erleben und Verhalten in Testsituationen niederschlagen. Die Forschung hat gezeigt, dass – neben den „genuinen“ kognitiven Inhalten von Testangst – eine Reihe weiterer kognitiver Prozesse und Strukturen typisch für Testängstliche sind. Teilweise handelt es sich um Besonderheiten bzw. interindividuelle Unterschiede in der basalen Informationsverarbeitung und –interpretation, wie z. B. der Aufmerksamkeitssteuerung. Diese Besonderheiten müssen nicht per se auf Leistungssituationen beschränkt sein. Teilweise betreffen diese Besonderheiten aber das konkrete und bewusste Testerleben, also z. B. Leistungserwartungen. Gemeinsam ist diesen Prozessen, dass Sie eine wichtige Rolle bei der Erklärung von

Leistungseffekten spielen (siehe Abschnitt 1.2.1). Auf diese Prozesse und Strukturen soll daher nun eingegangen werden. Einige Theorien und Befunde stammen dabei nicht aus der engeren Forschung zu Testängstlichkeit, sondern aus Untersuchungen zu allgemeiner Ängstlichkeit sowie aus der klinischen Perspektive auf Angststörungen.

1.1.1.4 Kognitive Prozesse bei Testängstlichen

Schwarzer (2000) stellt zwei Hauptunterschiede zwischen hoch und niedrig ängstlichen Personen heraus. Zum einen richten hoch Ängstliche ihre Aufmerksamkeit nicht allein auf Anforderungen der jeweiligen Aufgabe, sondern auch auf ihre eigene Person, womit ihre Aufmerksamkeit gewissermaßen geteilt ist. Dieser fundamentale Gedanke wurde bereits von Wine (1971) formuliert. Die „Selbstvoreingenommenheit“ (Schwarzer, 2000) Testängstlicher kennzeichnet sich darüber hinaus durch bestimmte negative Kognitionen wie die Zweifel an den eigenen Fähigkeiten oder Gedanken an Konsequenzen eines Misserfolges. Jenseits dieser kognitiven Inhalte von Testangst zeigen sich Auffälligkeiten in der Informationsverarbeitung bei Testängstlichen bzw. allgemein bei Ängstlichen. Dies betrifft unter anderem die vorhandenen kognitiven Schemata (Zeidner, 1998). Kognitive Schemata beeinflussen die Informationsverarbeitung und können für eine verzerrte (biased) Informationsverarbeitung verantwortlich sein (Beck & Clark, 1997). Bei Personen mit Angststörungen liegt beispielsweise insofern ein bias vor, dass selektiv bedrohliches Reizmaterial verarbeitet wird (Beck & Clark, 1997), wobei unterschiedliche Annahmen darüber existieren, welche Stufen des Informationsverarbeitungsprozesses von diesem bias betroffen sind (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg & van IJzendoorn, 2007).

Bar-Haim et al. (2007) führten die erste systematische Metaanalyse zum Aufmerksamkeitsbias bei Ängstlichen durch, in die insgesamt 172 zwischen 1986 und 2005 publizierte Studien inkludiert wurden. Die Autoren kodierten die Studien unter anderem nach den verschiedenen experimentellen Paradigmen zum Nachweis eines Aufmerksamkeitsbias. Diese Paradigmen sollen kurz beschrieben werden. Zum einen ist dies das emotionale Stroop, eine Modifikation des Stroop-Paradigmas. Der Stroop-Effekt zeigt sich bei der Inkongruenz von Farbe und Bedeutung eines Wortes (z. B. das Wort „blau“ in roter Farbe) beim Benennen der Farbe des Wortes. Verglichen mit einer kongruenten Reizkonstellation (z. B. das Wort „blau“ in blauer Farbe) erzeugt dies Interferenz, welche über verlängerte Reaktionszeiten operationalisiert wird (Stroop, 1935). Auch beim emotionalen Stroop müssen Probanden die Farbe eines Wortes benennen, wobei Reaktionszeiten auf emotionale und neutrale Begriffe verglichen werden. Bedrohliche Wörter sind beispielsweise „disease“, „cancer“ oder auch „failure“ (siehe hierzu eine Literaturübersicht von Williams, Mathews & MacLeod, 1996). Verzögerte Reaktionen bei emotionalen Wörtern gegenüber neutralen dienen hierbei als Operationalisierung der Interferenz bzw. eines entsprechenden attentionalen

bias. Dieser bias lässt sich bei bestimmten Personen zeigen, z. B. bei dispositionell hoch Ängstlichen oder auch klinischen Probanden mit einer Posttraumatischen Belastungsstörung oder einer spezifischen oder sozialen Phobie (Williams et al., 1996). Es finden sich auch Designs, in denen statt Wörtern Bilder als Material genutzt werden, z. B. Gesichter mit ärgerlichem Ausdruck (Bar-Haim et al., 2007).

Ein zweites verbreitetes Paradigma ist das dot-probe. Hierbei werden in zahlreichen Durchgängen für kurze Zeit (500 ms) Wortpaare präsentiert, die dann verschwinden. In einigen der Durchgänge erscheint anstelle *eines* der Worte ein Punkt (dot probe). Ist dies der Fall, muss der Proband so schnell wie möglich reagieren, z. B. in Form eines Knopfdrucks (MacLeod, Mathews & Tata, 1986). Die von MacLeod et al. (1986) initial aufgestellte und bestätigte Hypothese lautet, dass die Aufmerksamkeit ängstlicher Probanden auf bedrohliche Reize fokussiert, mit der Folge, dass sie *schneller* auf dot probes reagieren, die im Anschluss auf solche Reize erscheinen als auf dot probes, die anstelle eines neutralen Reizes erscheinen.

Ein drittes Design ist das modifizierte spatial cueing-Paradigma. Dieses geht auf Posner (1980) zurück: Probanden müssen einen Reiz entdecken und erhalten vor dessen Erscheinen einen Hinweisreiz (cue). Dieser zeigt in 80 % der Durchgänge (valid cue trial) die spätere Position des Zielreizes korrekt an, in 20 % zeigt der cue die Position des Zielreizes falsch an (invalid cue trial). Es lassen sich gegenüber neutralen trials kürzere (bei validen trials) sowie längere (bei invaliden trials) Reaktionszeiten feststellen, die jeweils auf Kosten bzw. Gewinne bei der Steuerung der Aufmerksamkeit zurückgeführt werden (Bar-Haim et al., 2007; Posner, 1980). Fox, Russo, Bowles und Dutton (2001) führten eine Studie mit einem emotionalen spatial cueing Paradigma durch. Bei einer Verwendung bedrohlicher Reize wird dabei folgende Hypothese aufgestellt: Der attentionale bias von Ängstlichen führt dazu, dass diese in validen Durchgängen schnellere Reaktionen zeigen als eine Kontrollgruppe. Demgegenüber reagieren sie (noch) langsamer als die Kontrollgruppe bei einem invaliden Durchgang. Zwar konnten Fox et al. (2001) den Effekt bei validen trials nicht zeigen, jedoch bei invaliden. Das spricht dafür, dass die Lösung der Aufmerksamkeit von bedrohlichen Reizen bei Ängstlichen zusätzliche Kosten (in Form von längeren Reaktionszeiten) erzeugt. In diesem Paradigma liegt also ein bias vor, wenn die Reaktionszeitdifferenz zwischen invaliden und validen cues bei bedrohlichen Reizen *größer* ist als bei neutralen (Bar-Haim et al., 2007).

Bar-Haim et al. (2007) berichten über alle Paradigmen hinweg bei hoch ängstlichen Probanden eine Effektstärke von d = .45 (k = 112, N = 2.263) für den auf bedrohliche Reize bezogenen bias (within-group, d. h. Vergleich von bedrohlichen vs. neutralen Bedingungen innerhalb einer Gruppe). Bei den analysierten Kontrollgruppen fand sich kein bias, d = -.007 (k = 87, N = 1.768). Auch beim Vergleich von Ängstlichen und entsprechenden Kontrollgruppen (between-group)

zeigte sich der bias, d = .41 (k = 125, N = 2.906). Beim within-group-Vergleich zeigten sich biases bei ängstlichen Stichproben sowohl beim emotionalen Stroop (d = .49, k = 70), beim dot-probe (d = .37, k = 35) und beim emotionalen spatial cueing (d = .43, k = 7), nicht aber bei Kontrollgruppen (für alle drei Paradigmen). Auch bei between-group Analysen zeigten sich ähnliche Effektstärken für das emotionale Stroop und das dot-probe, während beim emotionalen spatial cueing kein Effekt vorlag (was allerdings auf Besonderheiten der dort inkludierten Primärstudien sowie deren geringe Anzahl zurückgeführt wurde). Interessanterweise unterschied sich die Höhe des bias nicht in Abhängigkeit des Reizmaterials (Wörter vs. Bildmaterial). Bemerkenswerterweise zeigte sich auch kein signifikanter Unterschied in der Höhe des bias bei Betrachtung von Stichproben, die eine diagnostizierte Angststörung vorwiesen (k = 62) gegenüber Stichproben, die aufgrund von Fragebogenmaßen (k = 50) als hoch ängstlich identifiziert wurden.

Diese Befunde decken sich mit einer zu Beginn der 1990er Jahre von M. W. Eysenck aufgestellten Theorie der Hypervigilanz**.** Hypervigilanz zeichnet Personen mit hoher trait-Ängstlichkeit aus und beinhaltet mehrere Aspekte. Generell neigen diese Personen dazu, ihre Aufmerksamkeit auf aufgabenirrelevante Reize zu richten, sind also per se stark ablenkbar (allgemeine Hypervigilanz). Ein Ausdruck von Hypervigilanz ist auch das ausgeprägte Absuchen der Umgebung („environmental scanning"; Eysenck, 1992, S. 43). Spezifische Hypervigilianz äußert sich darin, dass hoch ängstliche Personen ihre Aufmerksamkeit stärker (also selektiv) auf bedrohliche als auf neutrale Reize richten. Darüber hinaus ist Hypervigilanz durch einen breiten Fokus der Aufmerksamkeit gekennzeichnet in der Phase, *bevor* ein salienter Reiz entdeckt wird, wobei eine Einengung der Aufmerksamkeit folgt, sobald dies geschehen ist. Saliente Reize sind beispielsweise bedrohliche Stimuli (Eysenck, 1992). Eysenck (1992) nimmt an, dass diese Phänomene bei hoch trait-Ängstlichen insbesondere dann auftreten, wenn auch eine hohe state-Angst vorliegt. Wichtig ist dabei, dass sich diese Beschreibungen auf die „normale", d. h. auf die nicht-klinische Population beziehen, wenngleich Hypervigilanz einen Vulnerabilitätsfaktor für die Entstehung einer generalisierten Angststörung darstellt. Eine Demonstration dieses Effekts bei Testängstlichen leisteten Calvo, Eysenck und Estevez (1994). Eine Stichprobe von N = 32 spanischen Studierenden wurde auf Basis des Wertes im TAI in Extremgruppen (je die Hälfte niedrige vs. hohe Ausprägung) separiert. Die Probanden wurden mit mehrdeutigen Sätzen geprimt, die entweder nicht bedrohliche, potenziell physisch bedrohliche (z. B. „The old woman was crossing the motorway on foot when a lorry approached her at high speed.") oder potenziell evaluativ-bedrohliche (ego-threat; z. B. „Many students observed in the public lists that Emile's marks were the lowest in that course.") Situationen schilderten. Anschließend wurde ein Satzanfang (z. B. im Falle des ego-threat: „He would be considered more …") präsentiert, auf den dann entweder ein Wort oder ein Nichtwort folgte. Die Probanden mussten entscheiden, ob ein Wort oder Nichtwort vorlag. Die Nichtwörter unterschieden sich nur durch einen Buchstaben von den Wörtern und sollten als ambige Information fungieren.

Die Wörter bzw. Nichtwörter bestätigten oder widerlegten den bedrohlichen Charakter der geschilderten Situation, z. B:

- „Silly“, im Spanischen „TONTO“ (Wort) vs. „TUNTO“ (Nichtwort) als bestätigendes Wort
- „Lazy“, im Spanischen „VAGO“ (Wort) vs. „VEGO“ (Nichtwort) als widerlegendes Wort

Dieses bestätigende oder widerlegende (Nicht-)Wort war letztlich eine Operationalisierung von wahrscheinlich erwarteten (im Falle des bestätigenden Wortes) Konsequenzen, z. B. dass Emile als dumm erachtet wird aufgrund seiner schlechten Note. Annahme der Autoren war, dass Testängstliche bei ambiger, potenziell bewertend-bedrohlicher Information einen bias vorweisen, der auch die weitere Verarbeitung steuert. Tatsächlich zeigten hoch Testängstliche eine schnellere Reaktion bei Wörtern, die den ego-threat bestätigten. Hingegen zeigten sie langsamere Reaktionszeiten bei Wörtern, die den ego-threat widerlegten, sowie auch langsamere Reaktionen bei Nichtwörtern, die den ego-threat bestätigten. Somit wurde gezeigt, dass der bias in der Interpretation von Informationen bei Testängstlichen auftritt und sich spezifisch auf ego-threat, nicht auf physische Bedrohung bezieht, da sich bei letzterer kein vergleichbares Befundmuster zeigte (Calvo, Eysenck & Estevez, 1994). Zusammenfassend lässt sich feststellen, dass sich der bias bei Ängstlichen sowohl bezüglich der Selektion und Interpretation von Informationen als auch bei Gedächtnisprozessen finden lässt (Eysenck, 1997). In einer Literaturübersicht zum attentionalen bias gegenüber bedrohlichen Reizen stellen Cisler, Bacon und Williams (2009) fest, dass sich der attentionale bias bei Ängstlichen sowohl in der bevorzugten Aufmerksamkeit für bedrohliche Reize als auch in der erschwerten attentionalen Ablösung (disengagement) von diesen Reizen äußert.

Die geschilderten Befunde machen deutlich, dass sich die basale Informationsverarbeitung bei hoch Ängstlichen bzw. Testängstlichen gegenüber niedrig Ängstlichen unterscheidet. Sarason (1972) bringt dies folgendermaßen auf den Punkt: „Whereas the less test-anxious person plunges into a task when he thinks he is being evaluated, the high test-anxious plunges inward. He either (1) neglects or misinterprets informational cues that may be readily available to him or (2) experiences attentional blocks.“ (S. 393). Beide von Schwarzer (2000) formulierten Hauptunterschiede hoch und niedrig Testängstlicher, nämlich die Teilung der Aufmerksamkeit und der Selbstfokus, kommen hier zum Ausdruck. Testangst lässt sich in diesem Sinne als ein zweifach ungünstiges Phänomen interpretieren: Die Aufmerksamkeit wird nicht nur auf die eigene Person (und damit teilweise von der Aufgabe weg) gelenkt, die Inhalte der Aufmerksamkeit sind überdies negativ. Eine Studie von Sarason und Stoops (1978) veranschaulicht, wie sich dies auf das subjektive Testerleben auswirkt. In mehreren Experimenten wurde gezeigt, dass hoch Testängstliche unter evaluativen Instruktionen sowohl die Wartezeit auf einen Test als auch die Dauer des Tests

selbst länger einschätzen als hoch Testängstliche in neutraler sowie mittel und niedrig Testängstliche unter evaluativer und neutraler Instruktion. Diese Ergebnisse deuten darauf hin, dass hoch Testängstliche die evaluative, also für sie aversive Situation tatsächlich auch als länger andauernd wahrnehmen. Ferner berichtete die hoch testängstliche Subgruppe unter der evaluativen Instruktion am meisten aufgabenirrelevante Gedanken (z. B. über die Meinung des Versuchsleiters von der eigenen Person oder über die Leistung anderer bei der Aufgabe) (Sarason & Stoops, 1978).

Die Ausführungen in Abschnitt 1.1.1.3 und 1.1.1.4 haben einerseits die Zusammenhänge zu anderen Dispositionen und darüber hinaus Besonderheiten bei der Informationsverarbeitung Ängstlicher bzw. Testängstlicher aufgezeigt. Unzweifelhaft stehen beide Bereiche miteinander in Beziehung. Besondere Bedeutung bekommen die Befunde und Theorien aus dem Bereich der Informationsverarbeitung bei der Frage nach Leistungsunterschieden zwischen hoch und niedrig Testängstlichen (siehe Abschnitt 1.2).

Zur Betrachtung von Testängstlichkeit als situationsspezifischer Disposition gehören darüber hinaus demographische Merkmale, die bislang nicht berücksichtigt wurden. Insofern demographische Merkmale mit unterschiedlichen Ausprägungen der Testängstlichkeit einhergehen, sind sie auch potenziell relevant für die Analyse von Testangstprozessen während der Testbearbeitung. Daher soll nun knapp auf diese Merkmale eingegangen werden.

1.1.1.5 Alters- und geschlechtsspezifische Ausprägungen von Testängstlichkeit

Die Ausprägung von Testängstlichkeit unterscheidet sich bei verschiedenen Altersgruppen. Während des Grundschulalters (Schuljahre 1 bis 6) lässt sich längsschnittlich ein Anstieg der Testängstlichkeitswerte feststellen (Hill & Sarason, 1966). Manley und Rosemier (1972) berichten querschnittlich niedrigere Werte bei Schülern der junior high school (7. bis 9. Klasse) als bei solchen der senior high school (10. bis 12. Klasse). Chapell et al. (2005) berichten querschnittlich auf Basis einer großen Stichprobe (N = 5.414) niedrigere Testängstlichkeitswerte für graduate students gegenüber undergraduates, wenngleich der Effekt klein war. Hembree (1988) berichtet in seiner Metaanalyse ebenfalls einen tendenziell umgekehrt U-förmigen Verlauf, d. h. einen Anstieg im Grundschulalter, ein hohes, stabiles Niveau im Jugendalter und einen anschließenden Niveaurückgang ab etwa der 10. Klasse (k = 2, N = 22.282). Nach einem Anstieg in der Kindheit und Höhepunkt im Jugendalter scheint sich also die Ausprägung der Testängstlichkeit im Hochschulalter zu verringern. Hierfür werden in der Literatur mehrere Ursachen aufgeführt. Eine wichtige Erklärung sind steigende Leistungsanforderungen, welche mit häufigeren Misserfolgserfahrungen ein-

hergehen (Zeidner, 1998). Überdies werden realistischere Selbsteinschätzungen und die Zunahme von sozialen Vergleichsprozessen als Erklärungen genannt (Wigfield & Eccles, 1989). Die niedrigeren Ausprägungen bei Studierenden in querschnittlichen Untersuchungen könnten auch auf Selektionseffekte zurückgehen und nicht zwingend auf intraindividuelle Entwicklungsprozesse (Hembree, 1988). Beispielsweise ist denkbar, dass innerhalb einer Kohorte Personen mit extrem hoher Testängstlichkeit (Aus-)Bildungswege mit häufigen Prüfungen und hohen Leistungsanforderungen meiden und sich nicht mehr in den entsprechenden Stichproben finden. Dass es auch während des Studiums in Form von Abbrüchen zu Selektionseffekten kommt, ist vor dem Hintergrund der Belastungen durch Prüfungsangst wahrscheinlich (vgl. die in der Einleitung zitierten Befragungen unter Studierenden).

Sehr gut belegt sind Geschlechtsunterschiede in der Ausprägung der Testängstlichkeit. Über alle untersuchten Altersgruppen hinweg berichtet Hembree (1988) höhere Ausprägungen der Testängstlichkeit bei Mädchen bzw. Frauen gegenüber Jungen bzw. Männern, wobei sich im mittleren Schulalter (5. bis 10. Klasse) mit d = .43 (k = 73, N = 13.244) ein größerer Unterschied fand als bei älteren Altersgruppen, d = .27 (11. und 12. Klasse sowie postsecondary; k = 39, N = 10.615). Dieses Muster ist nicht überraschend, angesichts des robusten Befunds der höheren Ausprägung von Frauen auf Neurotizismus (Costa, Paul, Jr., Terracciano & McCrae, 2001; Schmitt, Realo, Voracek & Allik, 2008). Einzelne Untersuchungen zeigen, dass sich sowohl die Facettenstruktur (für das Zweifaktorenmodell des TAI siehe Everson, Millsap & Rodriguez, 1991) sowie die Zusammenhänge mit anderen Konstrukten (z. B. Selbstwirksamkeit, Selbstkonzept in Mathematik; Benson, Bandalos & Hutchinson, 1994) zwischen den Geschlechtern nicht unterscheiden. Inwiefern dies auch für die Zusammenhänge mit Leistung zutrifft wird in Abschnitt 1.2 betrachtet.

Bislang wurde dargestellt, was Testängstlichkeit ist und mit welchen anderen Dispositionen bzw. Merkmalen sie in Zusammenhang steht. Testängstlichkeit als situationsspezifische Form der Ängstlichkeit richtet sich immer auf bestimmte Situationsklassen. In Abschnitt 1.1 wurde bereits dargelegt, auf welchen Situationen der Fokus in dieser Arbeit liegt. Aus der Betrachtung des Phänomens als trait *und* state resultiert, dass die Analyse des Phänomens ohne eine Betrachtung situativer Determinanten nicht gelingen kann. Mit den wichtigsten situativen Determinanten von Testängstlichkeit und Testangst befasst sich nun Abschnitt 1.1.2. Diese Ausführungen sind eine wichtige Grundlage insbesondere für die Fragestellung von Studie 2, da der „stereotype threat" letztlich aus bestimmten Variationen in der Darbietung bzw. Instruktion eines Tests resultiert, sich also aus „Situationsmerkmalen" ergibt. Auf Basis dieser Ausführungen soll Testangst als ein Prozess verstanden werden, der durch Testängstlichkeit, aber auch durch ein komplexes Wirken

von situativen Elementen bestimmt wird. Die weitgehend separate Betrachtung von Dispositionen in Abschnitt 1.1.1.3 und 1.1.1.4 sowie situativen Determinanten in Abschnitt 1.1.2 bildet die Grundlage für eine integrative Betrachtung der Entstehung und Wirkung von Testangst in Abschnitt 1.1.3.

1.1.2 Situative Determinanten von Testängstlichkeit und Testangst

Im Sinne der bereits beschriebenen Theorie der Stress- bzw. Emotionsentstehung von Lazarus gehen sekundäre Bewertungen eines Reizes als Bedrohung mit der Emotion Angst einher. Eine entscheidende Vorannahme dieses Abschnitts leitet sich aus der Definition von Schwarzer (2000) ab, nach der Leistungsangst die Konsequenz einer subjektiven Bedrohung des Selbstwerts ist. Natürlich werden unterschiedliche Prüfungen bzw. Tests in unterschiedlichem Maße als selbstwertbedrohlich aufgefasst und lösen daher in unterschiedlichem Maße Testangst aus. Auf die wichtigsten Situationsparameter soll nun eingegangen werden.

Zu den zentralen Situationsparametern dürfte die subjektive Aufgabenwichtigkeit gehören. Die in Abschnitt 1.1.1.1 dargelegte Theorie der transaktionalen Stress- bzw. Emotionsentstehung von Lazarus (1991b) postuliert, dass Transaktionen dann Emotionen auslösen, wenn sie Relevanz für ein Ziel haben, und dass die Stärke der Emotion von der Wichtigkeit dieses Ziels abhängt. Diese Urteile sind Inhalt der primären Bewertung. Analog zur Definition von Schwarzer (2000) müssen also Prüfungen und Tests potenziell umso mehr Angst erzeugen, je stärker sie den Selbstwert bedrohen. Die Kontroll-Wert-Theorie zur Erklärung von Leistungsemotionen von Pekrun (2006) ermöglicht eine Präzisierung dieser Wechselwirkungen. Demgemäß ist die Entstehung von Leistungsemotionen maßgeblich determiniert durch die Bewertung von Aktivitäten und Ergebnissen, die mit Lernen bzw. mit Leistung zu tun haben. Dieses Urteil wird anhand der beiden Dimensionen Kontrolle und Wert vorgenommen, d. h. dass die subjektive Kontrolle über entsprechende Aktivitäten (z. B. Lernen) und Ergebnisse (z. B. eine gute Note) sowie der jeweils subjektive Wert bewertet werden (Frenzel, Götz & Pekrun, 2009). In diesem Kontext zentral ist die Differenzierung des Wertes in eine kategoriale Komponente (also ob Aktivität bzw. Ergebnis aus Sicht der Person positiv oder negativ ist) und eine dimensionale Komponente (die subjektive Wichtigkeit von Aktivität oder Ereignis) (Frenzel et al., 2009). Merkmale der Situation wirken sich nicht nur auf die Kontrolle, sondern auch auf den Wert aus. Frenzel et al. (2009) nennen als Beispiel für letzteres das Gewicht, mit dem eine Teilnote in ein Zeugnis eingeht. Ein weiteres Beispiel wäre die Frage, ob es sich bei einer Prüfung im Studium um einen Erstversuch handelt oder um einen Wiederholungsversuch, bei dem ein Durchfallen das Ende des Studiums bedeutet. Neben Merkmalen der Situation bestimmen auch generalisierte Kontroll- und Wertüberzeugungen die Bewertung, welche sich aus dem Fähigkeitsselbstkonzept oder auch aus Leistungszielen ergeben können (Frenzel

et al., 2009). Dabei handelt es sich stärker um Merkmale der Person. Beispielsweise dürfte für einen Schüler, der sich für mathematisch begabt hält und ein Informatikstudium anstrebt, ein schlechtes Ergebnis in einer Mathematikprüfung besonders bedrohlich sein (in diesem Fall wäre sowohl das Selbstkonzept als auch die eigene Ausbildungsplanung von einem Misserfolg tangiert). Ähnlich bedrohlich könnte für einen Studenten eine lediglich durchschnittliche Studienabschlussnote sein in einem Fach, in dem nur die besten Absolventen attraktive Stellen bekommen können (in diesem Fall wäre ein Leistungsvermeidungsziel entscheidend für die Bewertung des Wertes).

Einige beispielhafte Untersuchungen veranschaulichen diese Zusammenhänge. Wigfield und Meece (1988) berichten für eine altersheterogene Stichprobe (6. bis 12. Klasse; N = 564) einen positiven Zusammenhang zwischen der kognitiven Komponente von Mathematikängstlichkeit und der subjektiven Wichtigkeit von Mathematik. Frenzel, Pekrun und Goetz (2007) erhoben bei einer großen Stichprobe (N = 2.053) von Schülern der 5. Klasse u. a. die Ängstlichkeit in Mathematik, die subjektive Kompetenz in Mathematik, aber auch den Wert der Domäne selbst („Mathematics is my favorite subject“) sowie den Wert guter Leistung in der Domäne („It is very important for me to get good grades in mathematics“). Auch nach Kontrolle von Geschlecht und Mathematiknoten zeigte sich ein positiver Zusammenhang des Werts der Leistung mit der Ängstlichkeit. Kompetenzüberzeugungen und der Wert der Domäne hingegen standen in einem negativen Zusammenhang mit mathematikbezogener Ängstlichkeit. Nie, Lau und Liau (2011) untersuchten Zusammenhänge zwischen bereichsspezifischer Testängstlichkeit in Mathematik (N = 1978) und Englisch (N = 1670) (beides 9. Klasse, singapurische Stichprobe) und der Wichtigkeit des jeweiligen Faches („I think learning English is important“). Es zeigte sich, dass besagter Zusammenhang durch Selbstwirksamkeit moderiert wird – positive Zusammenhänge zwischen Testängstlichkeit und Wichtigkeit zeigten sich in beiden Fächern bei niedriger Selbstwirksamkeit, nicht aber bei hoher Selbstwirksamkeit.

Hembree (1988) berichtet in seiner Metaanalyse einige Effekte für Testangst[10] in unterschiedlichen Testbedingungen. Demnach spielt die subjektive Schwierigkeit einer Prüfung bzw. eines Tests eine Rolle. So zeigte sich ein Unterschied in der Testangst bei schwer gegenüber leicht empfundenen Tests bzw. Prüfungen von d = .35 (k = 3, N = 590) (Hembree, 1988). Dieser Befund ist vor dem Hintergrund der Theorie von Lazarus gut zu erklären, nach der Stress eine Situation darstellt, in der die eigenen Ressourcen subjektiv beansprucht werden (Lazarus & Folkman, 1984), was umso stärker der Fall ist, je schwieriger ein Test wahrgenommen wird. Auch Itemrückmeldungen spielen eine Rolle. So zeigten sich niedrigere Werte der Testangst, wenn pro Item oder

[10] Es wird davon ausgegangen, dass in diesen Studien Testangst (state) erfasst wurde, da sich die Effekte auf Unterschiede zwischen Testbedingungen beziehen. Präzise Informationen hierzu finden sich jedoch in dem Artikel nicht.

Aufgabe eine Rückmeldung erfolgt als wenn keine Rückmeldung erfolgt, d = -.64 (k = 4, N = 239) (Hembree, 1988).

Ein weiteres wichtiges situatives Merkmal ist der getestete Bereich. Die Testängstlichkeit einer Person muss sich nicht auf alle Leistungsbereiche (oder Fächer, Disziplinen, je nach Kontext) gleichermaßen richten, sie kann mitunter auch nur auf einen sehr spezifischen Bereich beschränkt sein (siehe z. B. Sparfeldt et al., 2005). So gibt es auch einige Verfahren, die explizit die Ängstlichkeit in bestimmten Fächern adressieren (siehe z. B. Hopko, 2003; Krinzinger et al., 2007; Plake & Parker, 1982; Schnell, Tibubos, Rohrmann & Hodapp, 2013). Auch kann sich Testängstlichkeit unterscheiden in Abhängigkeit der Prüfungsform, also ob eine Prüfung schriftlich oder mündlich erfolgt (siehe z. B. Sparfeldt et al., 2013). Jedoch sind Divergenzen zwischen schriftlichen und mündlichen Prüfungen bislang kaum erforscht, was problematisch ist, da Prüfungen (sowohl in der Schule als auch an Hochschulen) sehr oft in mündlicher Form stattfinden (Sparfeldt et al., 2013), die Forschung sich jedoch überwiegend auf schriftliche Prüfungen und schriftliche Tests konzentriert.

Ein weiteres wichtiges Merkmal der Testsituation bezieht sich auf die Art der Instruierung. Da dieser Aspekt für alle im Rahmen dieser Arbeit behandelten Fragestellungen bzw. Studien von großer Bedeutung ist, wird ausführlicher darauf eingegangen.

Testatmosphäre

Unter Testatmosphäre (Zeidner, 1998) fällt die Art und Weise, wie ein Test bzw. eine Testsituation gegenüber Probanden dargestellt wird. Starken Einfluss auf die Forschung hatte das experimentelle Untersuchungsparadigma, das auf I. G. Sarason und Kollegen zurückgeht. Diese Forschung fokussierte vorrangig auf die Frage, welche Leistungsunterschiede bei dispositionell hoch und niedrig Testängstlichen in Abhängigkeit der Testatmosphäre auftreten (siehe hierzu Abschnitt 1.2). In diesen Experimenten wird ein Test oder eine Aufgabe auf unterschiedliche Weise angekündigt, wobei sich evaluative, also bewertende und leistungsorientierte Instruktionen von neutralen, nicht bewertenden Instruktionen unterscheiden lassen. I. G. Sarason und Kollegen setzten dieses Paradigma in zahlreichen Studien in unterschiedlichen Varianten ein, z. B. high vs. low motivation (Sarason, 1956), Standardinstruktionen vs. ermutigende bzw. beruhigende Instruktionen (Sarason, 1958a) oder bedrohliche vs. nicht bedrohliche Instruktionen (Sarason, 1961). Ein typisches Beispiel für eine evaluative Instruktion von Sarason und Stoops (1978) sei dargestellt:

> *As I said, the test you are about to take is part of an intelligence test. This test has been found to predict such things as course grades, success in later life, and to some extent, the kind of personality you possess. Of course, your own intelligence*

> *will primarily determine whether you do well or poorly on the test. At a later date you will have an opportunity to compare your IQ score with those of the other people in this study. You will then be able to determine how your abilities and capacities compare with other people like you.* (S. 103 f.)

Aus diesem Beispiel sowie weiteren Studien lassen sich die zentralen Kernmerkmale einer evaluativen Instruktion ableiten, die alle darauf abzielen, Stress bzw. Angst zu induzieren. Wichtiger Bestandteil ist die Aussage, dass das Ergebnis im nachfolgenden Test ein aussagekräftiger Indikator für die individuellen Fähigkeiten ist, wobei oft der Begriff „Intelligenz" explizit genannt wird (z. B. Sarason, 1973). Letzteres ist häufig verbunden mit Hinweisen auf die Vorhersagekraft von Intelligenz für den akademischen oder beruflichen Erfolg (alle diese Merkmale sind enthalten im Beispiel von Sarason & Stoops, 1978). Es finden sich Varianten evaluativer Instruktionen, die explizite Leistungserwartungen kommunizieren („most college students should be able successfully to complete the task"; Sarason, 1961, S. 166). Häufig enthalten ist auch ein Verweis auf soziale Vergleiche, also dass das individuelle Ergebnis in Relation zu einer Referenzgruppe gesetzt wird (z. B. Darke, 1988a, 1988b). Weiteres Mittel zur Induktion von Bewertungsdruck ist die Information, dass ein Testresultat in schulische bzw. akademische Noten eingeht (z. B. French, 1962; Prins & Hanewald, 1997). Es finden sich noch stärkere Instruktionselemente wie die Information, dass die Eltern über schlechte Ergebnisse in Kenntnis gesetzt werden (Meijer & Oostdam, 2007). Deffenbacher und Hazaleus (1985) kontrastierten eine konventionelle mit einer beruhigenden Testinstruktion. Sie argumentierten am Beispiel des Wonderlic Personnel Test, kurz WPT (Wonderlic Inc., 1996), dass erstere Form von Instruktionen stressinduzierende, aber auch -reduzierende Elemente enthält. Folgender Abschnitt aus der Testinstruktion des WPT verdeutlicht dies:

> *Dieser Test besteht aus 50 Fragen. Es ist unwahrscheinlich, dass Sie alle Fragen schaffen werden, aber tun Sie Ihr Bestes. Nachdem der Prüfer das Startzeichen gegeben hat, haben Sie genau 12 Minuten Zeit, um so viele Fragen zu beantworten, wie Sie können. Um unnötige Fehler zu vermeiden, sollten Sie nicht zu schnell arbeiten. Die Fragen werden fortwährend schwieriger, also springen Sie nicht hin und her. Verbrauchen Sie nicht zu viel Zeit an den einzelnen Aufgaben. Der Prüfer wird nach Beginn des Tests keine Fragen mehr beantworten.*

Die Instruktion enthält neben Hinweisen auf die Zeitbegrenzung und ansteigende Aufgabenschwierigkeit auch die Anweisung, schnell und dennoch fehlerlos zu arbeiten und dass eine hohe Anzahl gelöster Aufgaben das Ziel ist. Gleichzeitig wirkt die Information, dass kaum alle Fragen zu lösen sind, stressreduzierend.

In vielen experimentellen Untersuchungen ist die Administrierung evaluativer Instruktionen ein Mittel zur Schaffung einer Situation, die einer echten Prüfungen ähnelt – im Sinne von Pekruns

Kontroll-Wert-Theorie dient dies dazu, den Wert eines guten Testergebnisses zu steigern. In zahlreichen Untersuchungen von I. G. Sarason und anderen wurden derartige Bedingungen kontrastiert mit Kontrollgruppen, wobei auch hier verschiedene Varianten zu finden sind. Kennzeichen derartiger Bedingungen sind beispielweise, dass eine individuelle Fähigkeitsdiagnostik durch den Test nicht erwähnt oder auch explizit negiert wird, manchmal wird auch die Erforschung von Eigenschaften der Aufgabe bzw. des Tests als Ziel der Testung beschrieben (z. B. Sarason, 1956; Sarason, 1973; Sarason & Stoops, 1978). Neben dem *Fehlen* von expliziten evaluativen Hinweisen finden sich auch gezielt beruhigende Instruktionen, die den Probanden anweisen, sich keine Sorgen über die eigene Leistung zu machen sowie Hinweise, dass Fehler in der Bearbeitung wahrscheinlich und nicht gravierend sind (z. B. Sarason, 1958a). Das Stressniveau kann auch herabgesetzt werden, indem die Ergebnisse der Gruppe (nicht die einzelner Probanden) in den Vordergrund gestellt werden oder durch eine bestimmte Gestaltung des Testnamens (z. B. „Green Test, Form II, Experimental“ statt „Wonderlic Personnel Test“; Deffenbacher & Hazaleus, 1985).

Dieses verbreitete Paradigma weist einige interpretative Probleme auf. Erstens sind die dargestellten Manipulationen und deren Variationen sehr heterogen. So variieren bei evaluativen Instruktionen Anzahl und Intensität der „Stressreize“. Auch das Kommunizieren unrealistisch hoher Leistungsanforderungen (wie z. B. bei Sarason, 1961) birgt nicht nur das Potenzial zur Stressinduktion, sondern kann (zusätzlich) Misserfolgsgefühle hervorrufen. Noch heterogener sind nichtevaluative Instruktionen. So finden sich sowohl „neutrale“ als auch „beruhigende“ Instruktionen: das bloße Fehlen evaluativer Hinweisreize wird gelegentlich ergänzt um offenkundig stressreduzierende Hinweisreize. Diese Heterogenität erschwert die Vergleichbarkeit der Befunde. Ein weiteres Problem ist, dass hoch und niedrig testängstliche Probanden häufig aufgrund von extrem niedrigen oder hohen Ausprägungen auf der jeweils eingesetzten Skala ausgewählt wurden, womit Extremgruppenvergleiche vorliegen (z. B. Deffenbacher, 1978). Drittens lässt sich feststellen, dass insbesondere in den Arbeiten von I. G. Sarason und Kollegen (z. B. Sarason, 1956; Sarason, 1958a; Sarason, 1961; Sarason, 1973; Sarason & Stoops, 1978) oftmals keine Erfassung der in der Situation vorliegenden Testangst (state) erfolgte. Zeidner (1998) kritisiert hierzu, dass in vielen Studien Gruppenunterschiede in Abhängigkeit der Testbedingung auf die Testängstlichkeit (trait) zurückgeführt wurden, das Gros der Studien jedoch weder state-Messungen der Angst bzw. Testangst noch andere Manipulationschecks enthielt. Dieser Kritikpunkt verdeutlicht, dass auch die umfangreiche Forschung zur Testängstlichkeit Fragen offen lässt, da systematisch bestimmte Aspekte unzureichend berücksichtigt wurden.

Einige Studien, deren Design auf dem von I. G. Sarason geprägten Paradigma basieren, prüften jedoch direkt Effekte der Testatmosphäre auf das Testerleben. In der bereits zitierten Untersuchung von Deffenbacher (1978) wurde in einem 2x2-Design (hoch vs. niedrig Testängstlich x evaluative vs. beruhigende Instruktion) nach einer Anagrammaufgabe das Erleben während des

Tests erfragt. Die hoch Testängstlichen in der evaluativen Bedingung zeigten auf mehreren Variablen die extremsten Ausprägungen: so fanden sich hier die höchsten Werte bei Angst sowie Interferenz durch Angst und die niedrigsten Werte bei Kompetenzempfinden. Auch zeigten sich die höchsten Werte bei worry, emotionality und task-generated interference (wenngleich der Haupteffekt der Bedingung beim Kompetenzempfinden und emotionality nicht signifikant war, siehe auch Abschnitt 1.1.1.2.1). Deffenbacher und Hazaleus (1985) konnten dieses deutliche Befundmuster bei denselben abhängigen Variablen mit einem ähnlichen Design jedoch nicht replizieren (N = 129, studentische Stichprobe). Zwar ergab sich ein Haupteffekt der Ausprägung der Testängstlichkeit, nicht aber der Manipulation. Die Autoren vermuteten, dass die gewöhnliche Testinstruktion des WPT, die hier als evaluative Instruktion administriert wurde, nicht in ausreichendem Maß Stress induziert. Einige Studien konnten wiederum Hinweise darauf finden, dass sich derartige Variationen der Testatmosphäre auch in situativen Maßen von Angst bzw. Testangst niederschlagen (Coy et al., 2011; Englert, Bertrams & Dickhäuser, 2011; Leininger & Skeel, 2012; Meijer & Oostdam, 2011).

Es lässt sich also festhalten, dass es eine Reihe von Situationsparametern gibt, die das Ausmaß an Testangst in einer Test- bzw. Prüfungssituation beeinflussen. Zentrale Merkmale sind die Aufgabenwichtigkeit, die empfundene Schwierigkeit, der getestete Bereich und die Testatmosphäre. Diese und weitere Situationsmerkmale spielen nicht nur eine Rolle für das absolute Niveau von Testangst in der Testsituation, sondern auch für die Relation von Testängstlichkeit bzw. Testangst und Leistung. Sie sind für die vorliegende Arbeit wichtig, da sich alle drei Studien mit der Relation von Testangst und Leistung befassten. In Abschnitt 1.2 werden diese Parameter daher wieder aufgegriffen.

Die Ursachen von Testangst, deren Erleben und Auswirkungen lassen sich als ein komplexer Prozess auffassen, der zunächst ausgeht von relativ stabilen Dispositionen, die mit Besonderheiten in der Informationsverarbeitung assoziiert sind. Diese Prozesse treten verstärkt dann auf, wenn bestimmte Situationsmerkmale vorliegen. Der folgende Abschnitt soll die Perspektive auf Disposition und Situation verknüpfen und ein prozessuales Verständnis von Testängstlichkeit und Testangst ermöglichen.

1.1.3 Integrative Betrachtung: Entstehung und Wirkung von Testangst als Prozess

Die Entstehung und Wirkung von Testangst ist ein komplexer Prozess. Ziel dieses Abschnitts ist es, eine Brücke zu schlagen von der Entstehung und Erscheinungsweise von Testangst hin zu den

Konsequenzen von Testangst im Leistungsbereich. Wie Personen eine Situation (oder allgemeiner: einen Reiz) beurteilen, hängt von stabilen Dispositionen ab, ebenso von Besonderheiten in kognitiven Strukturen und der Informationsverarbeitung. Aus der Perspektive von Testängstlichkeit als situationsspezifischem trait wird theoretisch angenommen, dass dispositionell hoch Testängstliche in Bewertungssituationen häufiger und in stärkerem Maße Angst (state) erleben (Spielberger, 1972b). Zeidner (1994) führte eine Studie zur Untersuchung der Interaktion von Disposition und Situation bei Ängstlichkeit durch. Bei einer Stichprobe von *N* = 198 Studierenden wurde unter anderem die dispositionelle Ängstlichkeit in sozialen Bewertungssituationen erhoben. Bei den Probanden wurde vier Wochen vor der Prüfungsphase (t1) sowie während der Prüfungsphase, unmittelbar vor einer Prüfung, die aktuell erlebte Angst (state) erhoben (t2), wobei angenommen wurde, dass es sich jeweils um eine neutrale und evaluative Situation handelte. Bei Analyse der state-Angst von hoch und niedrig Ängstlichen (trait) zu t1 und t2 zeigten sich Haupteffekte des Messzeitpunkts (höhere state-Angst zu t2) und der Ängstlichkeitsgruppe (höhere state-Angst bei hoch Ängstlichen). Überdies wurde der erwartete Interaktionseffekt festgestellt: hoch Ängstliche zeigten einen stärkeren Anstieg als niedrig Ängstliche.

Eine explizite Integration interindividueller Unterschiede, akuter Bewertungsprozesse und deren Konsequenzen im Leistungsbereich ermöglicht die self-regulatory executive function theory (S-REF Theorie) (Wells & Matthews, 1996). Die S-REF Theorie verknüpft Annahmen aus der Theorie kognitiver Schemata (Beck & Clark, 1997) und Erkenntnisse aus dem Informationsverarbeitungsansatz (z. B. zum Aufmerksamkeitsbias) zur Erklärung der Entstehung von Distress. Anwenden lässt sich die Theorie sowohl auf das Erscheinungsbild affektiver Störungen als auch auf nichtklinisches Stresserleben (Wells & Matthews, 1996). Die S-REF Theorie basiert auf dem bereits dargestellten Stressmodell von Lazarus sowie auf Theorien der Selbstregulation (Carver & Scheier, 2001) (Zeidner & Matthews, 2007).

Die Entstehung von Distress (im konkreten Fall: Testangst) lässt sich dabei erklären durch eine komplexe Wechselwirkung von drei Teilsystemen. Dies sind erstens automatische Verarbeitungsprozesse, zweitens kontrollierte bzw. absichtsvolle Verarbeitungsprozesse, die Aufmerksamkeitsressourcen erfordern (das „executive system“ oder auch S-REF), und drittens im Langzeitgedächtnis gespeicherte Überzeugungen über die eigene Person (Matthews, Hillyard & Campbell, 1999; Wells & Matthews, 1996). Selbstbezogene Überzeugungen lassen sich dabei unterscheiden in deklarative und prozedurale Überzeugungen. Deklarative Überzeugungen sind z. B. Inhalte des Fähigkeitsselbstkonzepts. Prozedurale Überzeugungen („plans“) üben eine Steuerungsfunktion auf das zweite, absichtsvolle Verarbeitungssystem aus, konkret auf Bewertungsprozesse, Metakognitionen, selektive Aufmerksamkeit sowie den Abruf von Gedächtnisinhalten (Wells & Matthews, 1996). Persönlichkeitseigenschaften erklären in dieser Theorie interindividuelle Unterschiede im selbstbezogenen Wissen bzw. sind teilweise mit diesem gleichgesetzt. Demnach

wird unterschieden zwischen „broad traits“ wie den Big Five und „self-referent traits“, zu denen z. B. Selbstwirksamkeit und Selbstwert, aber auch dispositionelle Besorgtheit („trait worry“) und metakognitive Überzeugungen gehören. „Self-referent traits“ sind teilweise „typical content of a person’s thoughts“ (S. 173) (z. B. zum Wert der eigenen Person), repräsentieren aber auch auch dispositionelle Besonderheiten in der Informationsverarbeitung (z. B. das Denken über die eigenen Sorgen) (Matthews, Schwean, Campbell, Saklofske & Mohamed, 2000). Bezogen auf alle Teilsysteme gilt: „The system as a whole may operate in different configurations. The one most relevant to negative emotion is the self-regulatory executive function (S-REF), in which processing is driven by self-regulatory goals derived from discrepancies between actual and preferred self-status.“ (Matthews & Wells, 1999, S. 183). Diese Regulation zielt darauf ab, die entsprechende Diskrepanz zu verringern (Matthews & Wells, 1999; Wells & Matthews, 1996).

Im Folgenden wird kurz beschrieben, wie das Modell durch Matthews und Kollegen auf die Entstehung und Wirkung von Testangst (als akut erlebter Emotion) übertragen wurde. Ausgangspunkt für S-REF sind interne oder externe Reize, die im ersten, automatischen Verarbeitungssystem zur Bildung von Intrusionen, konkret bedrohliche Gedanken, führen (Matthews et al., 1999; Zeidner & Matthews, 2007). Intrusionen wiederum lösen kontrollierte (exekutive) Verarbeitungsprozesse aus (S-REF) (Wells & Matthews, 1996; Zeidner & Matthews, 2007). Analog zur kognitiven Interferenz, wie sie von Sarason (1984) beschrieben wurde, und dem Aufmerksamkeitsbias bei Ängstlichen (Bar-Haim et al., 2007), richtet sich bei S-REF die Aufmerksamkeit auf die eigene Person, „biasing the system toward registration of self-relevant intrusions and toward accessing self-discrepancies“ (Matthews et al., 2000, S. 176). Im Sinne des Wechselspiels der Teilsysteme beeinflussen selbstbezogene Überzeugungen das S-REF. Sorgenvolle Kognitionen (allgemein: Distress) werden auf diese Weise durch den Abruf negativer selbstbezogener Überzeugungen erzeugt. Beispielsweise ruft eine dispositionell testängstliche Person in einer Bewertungssituation negative Informationen aus dem eigenen Fähigkeitsselbstkonzept ab. S-REF beinhaltet nun nicht nur die Bewertung der Situation, sondern auch die Suche nach Bewältigungsstrategien (im Sinne der Reduktion der o. g. Diskrepanz) (Zeidner & Matthews, 2007).

Metakognitionen, insbesondere bezüglich der Bedeutung der eigenen Sorgen und des Umgangs damit, spielen für S-REF eine entscheidende Rolle. Matthews et al. (1999) erhoben bei einer studentischen Stichprobe ($N = 84$) unter anderem die RTT und verschiedene Skalen zu dispositioneller Coping-Neigung und zu metakognitiven Überzeugungen. Grundsätzlich lässt sich differenzieren zwischen emotionsorientiertem Coping (Regulation der Emotion selbst), problemorientiertem Coping (Beeinflussung der auslösenden Bedingungen der Emotion bzw. des Stresses) und vermeidungsorientiertem Coping (Vermeidung der Situation, die mit Stress verbunden ist) (Zeidner, 1998). Bezogen auf die Angst vor einer Prüfung wäre problemorientiertes Coping z. B. eine

gezielte und sorgfältige Vorbereitung, emotionsorientiertes Coping hingegen der Versuch, die eigene Angst zu reduzieren, z. B. durch die Abwertung der Wichtigkeit der Prüfung. Vermeidungsorientiertes Coping wäre z. B. die Ablenkung vom Stressor durch Musik und Filme (Zeidner, 1998). Die Skalen der RTT korrelierten deutlich positiv mit emotionsorientiertem Coping (insb. Tension und Worry zu r = .60 und .59). Worry und Test-Irrelevant Thinking korrelierten negativ mit problemorientiertem Coping zu r = -.31 und -.46. Test-Irrelevant Thinking korrelierte überdies positiv mit vermeidungsorientiertem Coping, r = .27. Von besonderer Relevanz sind auch die Zusammenhänge zu Metakognition, die mit dem Meta-Cognitions Questionnaire (MCQ; Cartwright-Hatton & Wells, 1997) erfasst wurden. Testängstlichkeit ging in allen Facetten einher mit der subjektiven Unkontrollierbarkeit und Bedrohlichkeit von Sorgen (z. B. „I find it difficult to control my thoughts"; „My worrying could make me go mad") sowie einem geringen Vertrauen in die Zuverlässigkeit der eigenen Kognition (z. B. „I have little confidence in my memory for words and names"). Stöber und Esser (2001) konnten letzteres in einer experimentellen Studie (N = 56, studentische Stichprobe) bestätigen. Dispositionell hoch Testängstliche gaben an, Informationen lieber external (z. B. über eine Notiz) als internal (d. h. durch Einprägen im Gedächtnis) abzuspeichern und schätzten zudem die Wahrscheinlichkeit, sich korrekt erinnern zu können, bei externaler Speicherung höher ein. Wenn die manipulierte Wichtigkeit der Information hoch war, präferierten zwar auch niedrig Testängstliche die externale Speicherung, hoch Testängstliche aber bevorzugten diese unabhängig von der Wichtigkeit der Information.

Gemäß der S-REF Theorie tragen metakognitive Überzeugungen zur Aufrechterhaltung von negativem, auf die eigene Person fokussiertem Denken bei, wobei die Wahl maladaptiver Copingstrategien ein Ergebnis von S-REF ist (Zeidner & Matthews, 2007). Maladaptives Coping äußert sich auf verschiedene Weise. Dies kann z. B. emotionsorientiertes Coping oder Vermeidung (vgl. auch Leistungsvermeidungsziele; Elliot & McGregor, 2001) sein. Ebenso gehört hierzu self-handicapping, das definiert ist als „any action or choice of performance setting that enhances the opportunity to externalize (or excuse) failure and to internalize (reasonably accept credit for) success." (Berglas & Jones, 1978, S. 406). Konsequenzen von S-REF sind also nicht nur das Erleben (und die Aufrechterhaltung) von Testangst, sondern eine langfristige Hemmung der Fähigkeitsentwicklung einer Person, da beispielsweise Vermeidungstendenzen den Erwerb neuer Kompetenzen blockieren – ein extremes Beispiel wäre eine hoch testängstliche Person, die mehrere Prüfungen nicht antritt und schließlich ihr Studium abbricht (Zeidner & Matthews, 2007). Das komplexe Zusammenspiel der Prozesse lässt sich mit folgendem Zitat von Zeidner und Matthews (2007) zusammenfassen:

> *The source of anxiety is dysfunctional self-knowledge (both declarative and procedural), but its expression as maladaptive situational coping, and perpetuation over time, require the dynamic perspective of the transactional model of stress*

> *and emotion* […] *Self-referent processing driven by metacognitive goals initiates dysfunctional coping strategies (emotion focus, avoidance, self-handicapping) that draw attentional ressources, working memory, and effort away from task at hand, leading to impairments if the task is demanding.* (S. 157)

Die S-REF Theorie erklärt also nicht nur Entstehung und Aufrechterhaltung von Testangst sowie regulatorische Bewältigungsversuche, sondern auch deren Konsequenzen auf Leistung. Zeidner und Matthews (2007) machen deutlich, dass sowohl die selbstbezogenen Kognitionen und deren (kognitiver) Ressourcenbedarf, als auch die maladaptiven Bewältigungsstrategien und deren ungünstiger Einfluss auf das Fähigkeitspotenzial einer Person zu Leistungsunterschieden zwischen hoch und niedrig Testängstlichen beitragen. In dieser Konzeption lässt sich eine Verknüpfung zweier verschiedener theoretischer Erklärungsmodelle für den negativen Leistungseffekt von Testängstlichkeit bzw. Testangst, nämlich der Interferenz- und Defizitperspektive, erkennen. Diese werden in Abschnitt 1.2 behandelt.

1.2 Testängstlichkeit, Testangst und Leistung

Der negative Zusammenhang von Testängstlichkeit, Testangst und Leristung ist Kern der vorliegenden Arbeit und Untersuchungsgegenstand aller drei Studien. In diesem Abschnitt sollen zunächst die zentralen Erkenntnisse zum Zusammenhang von Testängstlichkeit und Testangst einerseits und Leistung andererseits betrachtet werden. Aufgrund der Fülle an Studien hierzu liegt der Fokus insbesondere auf den Metaanalysen von Hembree (1988), Seipp (1991), Ackerman und Heggestad (1997) sowie auf einzelnen Studien späteren Datums. Auch auf eine Metaanalyse von Richardson, Abraham und Bond (2012) wird eingegangen. Der Schwerpunkt wird hierbei auf Zusammenhänge mit Leistungen in kognitiven Fähigkeitstests sowie mit Schul- bzw. Studiennoten gelegt.

Hembree (1988) schloss insgesamt k = 562 Studien aus dem Zeitraum von 1950 bis 1986 in seine Analysen ein. Aus der Fülle an Kriterien werden hier Zusammenhänge zu vier verschiedenen Leistungsmaßen berichtet: Intelligenztests („IQ“), Kursnoten („course grade“), Durchschnittsnoten (grade point average, GPA) und andere Leistungsmaße („aptitude and achievement tests[11]). Schwerpunkt der hier dargestellten Ergebnisse liegt auf Jugendlichen und jungen Erwachsenen. Insgesamt ermittelte er einen mittleren Zusammenhang mit dem IQ von r = -.23 und mit anderen Leistungsmaßen von r = -.29. Für letztere wurde überdies auch nach fachspezifischen Leistungen differenziert. So waren negative Zusammenhänge beispielsweise bezüglich Lesen & Englisch (r = -.24; k = 67, N = 10.761, 3. Klasse bis postsecondary) sowie Mathematik (r = -.22; k = 46, N = 6.534, 4. Klasse bis postsecondary) vorhanden. Schwächere Zusammenhänge fanden sich zu GPA (siehe Tabelle 8).

[11] In Anlehnung an das Glossar des National Council on Measurement in Education [NCME] (2015) handelt es sich bei „achievement tests“ um Verfahren zur Überprüfung von Wissen oder Fertigkeiten in einem Schulfachbereich (z. B. „mathematics“, „science“ oder „writing“). Unter „aptitude test“ fallen Fähigkeitstests, die auf einen bestimmten Bereich bezogen sind (z. B. „scholastic“, „verbal“, „mechanical“).

Tabelle 8: Metaanalytisch geschätzte Zusammenhänge von Testängstlichkeit und verschiedenen Leistungsmaßen (Hembree, 1988)

Kriterium	*k*	*N*	Altersgruppe[1]	Mittlerer Effekt
Testängstlichkeit (nicht differenziert)				
IQ	61	8.438	3-P	-.23
Andere Leistungsmaße	44	6.390	4-6, 8-12, P	-.29
GPA				
High School	8	1.164	9-12	-.12
College: TAQ[2]	9	1.499	P	-.12
College: AAT-[2]	9	1.423	P	-.29
Worry				
Andere Leistungsmaße	13	1.112	9-12, P	-.31
Kursnoten	13	272	12, P	-.26
Emotionality				
Andere Leistungsmaße	3	1.112	9-12, P	-.15
Kursnoten	3	272	12, P	-.19

[1] Klasse, P = postsecondary [2] Erfassung der Testängstlichkeit mit dem TAQ bzw. hemmender Angst aus dem AAT

Wesentliche Erkenntnis aus dieser Metaanalyse ist, dass worry stärker mit Leistungsmaßen zusammenhängt als emotionality, sowohl was Kursnoten als auch andere Leistungsmaße betrifft. Überdies ist erkennbar, dass sich negative Zusammenhänge mit einer breiten Spanne von Leistungsmaßen ergeben. Dies schließt sowohl Intelligenzmaße, akademische Leistungsmaße und Noten sowie basalere kognitive Leistungen ein, z. B. Problemlösen (r = -.20, k = 7, N = 1.225, 5. und 6. Klasse sowie postsecondary) und Gedächtnisaufgaben (r = -.28, k = 4, N = 172, postsecondary).

Seipp (1991) führte eine weitere Metaanalyse durch und inkludierte für den Zeitraum von 1975 bis 1988 126 Studien mit k = 156 Effektstärken[12]. Während Hembree (1988) nur Studien mit englischsprachigen Stichproben einschloss, bezog Seipp (1991) auch nicht-englischsprachige Stichproben ein. Fokus der Analyse lag auf Studien, die akademische Leistungsmaße enthielten (standardisierte Verfahren wie der Achievement College Test oder der Stanford Achievement Test sowie Noten und GPA, was Schul- und Studiennoten beinhaltete). Nichtakademische Leistungsmaße wie Intelligenztests wurden nicht berücksichtigt. Auch Seipp (1991) fand einen negativen Zusammenhang zwischen Ängstlichkeit und Leistung, wenngleich mit r = -.212 kleiner als bei Hembree (1988). Überdies wurden in der Metaanalyse eine Reihe bedeutsamer Moderatoren des Angst-Leistungs-Zusammenhangs festgestellt (siehe Tabelle 9).

[12] In dieser Metaanalyse wurde die Anzahl der Effektstärken mit k bezeichnet.

Tabelle 9: Metaanalytisch ermittelte Moderatoren der Relation von Ängstlichkeit und akademischer Leistung (Seipp, 1991)

Moderatoren		*k*	*N*	Mittlerer Effekt
Gesamt		156	36.626	-.212
Geschlecht	Weiblich	38	58.356	-.215
	Männlich	40	10.921	-.182
Facette	Worry	38	6.885	-.219
	Emotionality	37	5.182	-.147
Disposition vs. Zustand	State	29	3.867	-.210
	Trait	137	30.670	-.210
Art der Ängstlichkeit	Allgemeine Ängstlichkeit	53	11.680	-.163
	Testängstlichkeit	114	28.424	-.233
Messzeitpunkt der Angst	Vor der Leistungsmessung	31	4.498	-.211
	Nach der Leistungsmessung	9	2.056	-.283
	Unabhängig[1]	35	10.906	-.212

[1] d. h. ohne für den Probanden erkennbare Verbindung zwischen Angst- und Leistungsmessung

Trotz des absoluten Unterschieds in der Ausprägung der Testängstlichkeit bei den Geschlechtern (siehe Abschnitt 1.1.1.5) fand sich kein Geschlechtsunterschied im Angst-Leistungs-Effekt. Auch beim Vergleich der Effekte von Studien aus den USA, der BRD und anderen Ländern zeigten sich keine bedeutsamen Unterschiede. Der stärkere Leistungszusammenhang von worry gegenüber emotionality wurde repliziert. Entgegen der Erwartung ergab sich kein höherer Effekt für state-Angst gegenüber trait-Ängstlichkeit. Erwartungsgemäß hingegen korrelierte Testängstlichkeit höher mit Leistung als allgemeine Ängstlichkeit. Schließlich stellte sich noch der Messzeitpunkt als Moderator heraus: höhere Zusammenhänge ergaben sich bei einer Erfassung der Testängstlichkeit (bzw. Testangst, nicht näher definiert) nach einer Leistungsmessung gegenüber einer vorher stattfindenden Erfassung. Dieser Befund war für die Fragestellung von Studie 1 maßgeblich (siehe hierzu Abschnitt 1.2.1.3). Bemerkenswert ist auch das 95 %-Konfidenzintervall bei den inkludierten Effekten von worry ($-.40 \leq r \leq -.04$). Dieses lässt darauf schließen, dass auch der robuste Leistungseffekt von worry erheblichen „Schwankungen" unterliegt, die möglicherweise auf weitere Moderatoren zurückgehen. Der gefundene mittlere Zusammenhang ist absolut gesehen niedrig bis moderat, weist aber praktische Bedeutsamkeit auf – der Zusammenhang von $r = -.212$ (entspricht $d = -.434$)[13] bedeutet, dass hoch Ängstliche im Mittel fast eine halbe Standardabweichung schlechter abschneiden als niedrig Ängstliche (Seipp, 1991).

Ackerman und Heggestad (1997) publizierten eine umfangreiche Metaanalyse zu den Zusammenhängen von Intelligenz, Persönlichkeit und Interessensbereichen. Diese Analyse inkludierte auch

[13] Seipp (1991) berichtet ein *d* von -.043, was offenkundig ein Tippfehler ist: der Wert -.434 wurde vom Verfasser dieser Arbeit ermittelt (zur Effektstärkenberechnung siehe Abschnitt 3) und stimmt überein mit der inhaltlichen Interpretation des Leistungsunterschieds in Einheiten der *SD* durch Seipp (1991).

Studien, die Testängstlichkeit erfassten. Die Autoren orientierten sich bei der Kodierung der verschiedenen Intelligenzmaße grob an der Übersichtsarbeit von Carroll (1993). Die Ergebnisse sind in Tabelle 10 zusammengefasst.

Tabelle 10: Metaanalytisch geschätzte Zusammenhänge von Testängstlichkeit mit verschiedenen Intelligenzfähigkeiten (Ackerman & Heggestad, 1997)

Kriterium	Beispielhafte Operationalisierung	*k*	*N*	Effekt[1]
General Intelligence (G)		21	3.027	-.33
Crystallized Intelligence (Gc)	Leseverständnis	21	4.714	-.24
Ideational Fluency	Wortflüssigkeit	2	607	-.01
Knowledge and Achievement	Fach-/bereichsspezifisches Wissen	5	1.183	-.16
Learning and Memory	Freier Abruf	3	216	-.22
Speed	Erfassung von Reaktionszeiten	1	141	-.16
Visual Perception	Erkennen räuml. Beziehungen	4	755	-.23
Fluid Intelligence (Gf)	Induktives Schlussfolgern	4	784	-.25
Math-Numerical	Zahlengebundenes Schlussfolgern	16	3.943	-.27

[1] mittlerer Effekt; berichtet ist jeweils $\hat{\rho}$ (doppelt minderungskorrigiert); Kriteriumsbegriffe direkt übernommen, beispielhafte Operationalisierung anhand eigener, freier Übersetzung

Testängstlichkeit wies demnach zu einer großen Bandbreite intellektueller Fähigkeiten negative Zusammenhänge auf, wobei der Effekt für die allgemeine Intelligenz am stärksten ausgeprägt war.

Richardson et al. (2012) führten eine Metaanalyse zu den Kriteriumsvaliditäten zahlreicher nichtkognitiver Konstrukte bezüglich Studiennoten (GPA) auf Basis des Zeitraums von 1997 bis 2010 durch. Bezogen auf Testängstlichkeit fand sich ein mittlerer Zusammenhang von r = -.24 (k = 29, N = 13.497; korrigiert für Stichprobenfehler), was etwa in der Höhe der von Hembree (1988) und Seipp (1991) berichteten Werte liegt. Nach worry und emotionality differenzierte Effekte wurden nicht berichtet.

Von den zahlreichen Einzelstudien, die vor und nach den zitierten Metaanalysen den Angst-Leistungs-Zusammenhang untersucht haben, sollen nur einige wenige Arbeiten zitiert werden. Diese konnten die unterschiedliche Leistungsrelevanz der Facetten von Testängstlichkeit replizieren. Putwain (2008b) berichtet für N = 558 Schüler (11. Klasse, d. h. Alter 15-16 Jahre) mit der Gesamtnote im General Certificate of Secondary Education (GCSE) einen Zusammenhang von r = -.27 für Worry und r = -.13 für Emotionality. Chapell et al. (2005) erhoben bei einer großen studentischen Stichprobe in mehreren US-Bundesstaaten (N = 4.000 undergraduate bzw. 1.414 graduate students) den TAI sowie GPA. Bei den undergraduates fand sich ein stärkerer Effekt (Worry: r = -.21, Emotionality: r = -.08) als bei den graduates (Worry: r = -.12, Emotionality: r = -.06). Deutlich geringer ist die Anzahl an Studien, die mehrdimensionale Messungen von Testängstlichkeit mit Leistung in Bezug gesetzt haben. Der Befund, dass kognitive Aspekte von Testangst am deutlichsten zu Leistungsmaßen in Relation stehen, zeigt sich jedoch auch in diesen Arbeiten. Deffenbacher und Hazaleus (1985) berichten für eine studentische Stichprobe (N = 129) zwischen

der Leistung im Wonderlic Personnel Test[14] und der nach dem Test (retrospektiv) erfassten Testangst einen Zusammenhang zu Worry von r = -.32, zu Emotionality von r = -.18 und zu Task-generated interference von r = -.24. Putwain und Symes (2012) konstatieren, dass es bezüglich des mehrdimensionalen Modells der RTA noch kein eindeutiges Bild über die Leistungszusammmenhänge der einzelnen Facetten gibt, dennoch scheint sich ein negativer Zusammenhang von Worry auch bei der vierdimensionalen Konzeption der RTA zu bestätigen (so bei Keogh, Bond, French, Richards & Davis, 2004; McIlroy & Bunting, 2002; Putwain, Connors & Symes, 2010). Auch bezüglich der vierdimensionalen Konzeption im PAF bzw. TAI-G zeigen sich negative Leistungsassoziationen vorwiegend für die kognitiven Angstfacetten, wohingegen Aufgeregtheit kaum mit Leistung korreliert (z. B. Hodapp et al., 2011; Musch & Bröder, 1999a).

1.2.1 Theoretische Erklärungen

Korrelative Zusammenhänge liefern noch keine Aussage über tatsächliche Kausaleffekte. Nach Cook und Campbell (1979)[15] gibt es drei notwendige Bedingungen für Kausalität. Die erste Bedingung, dass zwei Variablen kovariieren, ist im Falle von Testängstlichkeit bzw. Testangst und Leistung erfüllt. Die zweite Bedingung für Kausalität lautet, dass die verursachende Variable (Testängstlichkeit bzw. Testangst) der verursachten Variable (Leistung) zeitlich vorausgeht. Trotz der langjährigen Forschung zu Testängstlichkeit ist diese Frage nicht abschließend geklärt. Ebenso plausibel wie die negative kausale Wirkung von Testängstlichkeit bzw. Testangst auf Leistung ist auch der umgekehrte Effekt, sprich die (Rück-)Wirkung von Leistung auf Testängstlichkeit bzw. Testangst (Zeidner, 1998). Beispielsweise kann das Erleben von Testangst nach einem Test unmittelbare Reaktion auf ein (subjektiv) schlechtes Abschneiden in dem Test sein, wenn Gedanken an die Konsequenzen eines Scheiterns auftreten (dieser Idee wurde in Studie 1 vertieft nachgegangen). Ebenso könnte sich Testängstlichkeit langfristig als stabile Disposition herausbilden, wenn in Leistungssituationen Misserfolgserfahrungen kumulieren. In diesem Fall ist Testängstlichkeit bzw. Testangst nicht selbst Ursache schlechter Leistung, sondern lediglich deren Folge und damit Ausdruck der niedrigeren Leistungsfähigkeit einer Person (siehe Abschnitte 1.2.1.1 und 1.2.1.2). Beide Kausalrichtungen müssen sich aber nicht zwangsläufig ausschließen (siehe Abschnitt 1.2.1.4). Entsprechend der dritten Bedingung für Kausalität muss der Einfluss der verursachenden auf die verursachte Variable nicht auf eine dritte Variable zurückführbar sein. Aufgrund der Multideterminiertheit kognitiver bzw. intellektueller sowie akademischer Leistungen ist die vollständige Erfüllung dieser Bedingung unrealistisch. Adäquater ist die Annahme, dass

[14] Häufig auch nur als „Wonderlic“ bezeichnet.
[15] Wiederum abgeleitet vom Kausalitätsverständnis von John Stuart Mill.

Testängstlichkeit bzw. Testangst eine bedeutsame, aber nicht die einzige Variable ist, die zu Leistungsunterschieden beiträgt.

In den Abschnitten 1.2.1.1 und 1.2.1.2 sollen diese beiden grundlegenden Erklärungsansätze für die Angst-Leistungs-Relation näher beschrieben werden. Theorien und Befunde zur kausalen Wirkung von Testängstlichkeit und Testangst auf Leistung werden dabei unter dem Sammelbegriff Interferenzperspektive subsumiert. Demgegenüber werden Theorien und Befunde mit umgekehrter kausaler Interpretation unter dem Sammelbegriff Defizitperspektive behandelt.

1.2.1.1 Interferenzperspektive

In Abschnitt 1.1.1.2.2 wurde „Interferenz“ als Komponente von Testängstlichkeit bzw. Testangst eingeführt. Wenn im Folgenden von „Interferenzperspektive“ gesprochen wird, meint dies jedoch nicht ausschließlich jene kognitiven Prozesse, die in der Facette „Interferenz“ subsumiert sind. Vielmehr geht es um die allgemeine Hypothese, dass Testängstlichkeit bzw. Testangst einen (kausalen) negativen Effekt auf kognitive Leistung hat, da sie störend – oder eben „interferierend“ – auf die kognitiven Prozesse während der Aufgabenbearbeitung einwirkt. In dieser Betrachtung spielen natürlich auch jene Komponenten von Testangst eine Rolle, die in Abschnitt 1.1.1.2.2 als „Interferenz“ beschrieben wurden, jedoch nicht ausschließlich.

Es gibt unterschiedliche theoretische Vorstellungen darüber, wie diese interferierende Wirkung von Testangst stattfindet. Eine Gemeinsamkeit der Ansätze ist jedoch, dass der Fokus auf kognitiven Prozessen liegt. Entsprechend dem Informationsverarbeitungsparadigma (Asendorpf & Neyer, 2012) werden Leistungsunterschiede zwischen hoch und niedrig Testängstlichen auf Divergenzen in der Informationsverarbeitung zurückgeführt.

1.2.1.1.1 Kognitive Interferenz

Hinweise auf kausale Erklärungen für Leistungsunterschiede zwischen hoch und niedrig Testängstlichen liefern insbesondere experimentelle Untersuchungen, in denen die Testatmosphäre manipuliert wurde (siehe Abschnitt 1.1.2). Bereits in den 50er Jahren beobachtete I. G. Sarason, dass sich hoch und niedrig Testängstliche in ihrer Leistung in Abhängigkeit davon unterscheiden, wie ein Leistungstest oder eine Prüfung instruiert bzw. präsentiert wird. In einer der ersten Studien hierzu verglich Sarason (1956) die Leistung hoch, mittel und niedrig ängstlicher Personen (N = 180, studentische Stichprobe) in einer Gedächtnisaufgabe bei hoch und niedrig evaluativer Instruktion. Es zeigte sich, dass hoch Ängstliche nach hoch evaluativer Instruktion schlechtere Leistungen erzielten als bei niedrig evaluativer Instruktion. Darüber hinaus schnitten mittel und niedrig Ängstliche in der evaluativen Bedingung besser ab als bei niedrig evaluativer Instruktion.

Zu ähnlichen Ergebnissen kam Sarason (1958a) beim Vergleich der Leistung von hoch und niedrig Testängstlichen (N = 64, studentische Stichprobe) bei einer konventionellen gegenüber einer beruhigenden Testinstruktion. Sarason konzentrierte sich in diesen und weiteren Studien darauf, wie die (trait-)Testängstlichkeit mit der Variation von Testinstruktionen im Hinblick auf Leistungseffekte interagiert. Typischerweise schneiden hoch Testängstliche unter einer evaluativen Instruktion schlechter ab als niedrig Testängstliche, während sich bei niedrig Testängstlichen umgekehrte Effekte finden lassen (z. B. Sarason, 1961; Sarason, 1973).

Eine Erklärung für diese Effekte liefert die Theorie der kognitiven Interferenz (Sarason, Sarason & Pierce, 1990). Demgemäß lenken hoch Testängstliche ihre Aufmerksamkeit in einer Testsituation nicht nur auf die Anforderungen der Aufgabe, sondern auch auf die eigene Person bzw. eigene (sorgenvolle) Gedanken und Empfindungen, weshalb ihre Aufmerksamkeit insgesamt *geteilt* ist. Niedrig Testängstlichen gelingt es in stärkerem Maße, ihre Aufmerksamkeit auf die Aufgabe zu richten, was letztlich die Leistungsunterschiede erklärt (Wine, 1971; Wine, 1980). Wine (1980) stellt den Bezug zur früher entwickelten Zwei-Komponenten-Theorie von Liebert und Morris (1967) her – dass worry stärker als emotionality mit Leistung kovariiert, stützt die Hypothese, dass die Gedanken an einen möglichen Misserfolg zu Leistungseinbußen führen. Prinzipiell müssen sich diese „self-relevant variables", die Gegenstand der Aufmerksamkeit sind (im Gegensatz zu den „task-relevant variables"; Wine, 1971), nicht auf Gedanken an ein Scheitern und dessen Konsequenzen beschränken. So können auch die beschriebenen Interferenzprozesse, also Gedanken an aufgabenirrelevante Themen, die *keine* Sorgen zum Inhalt haben (vgl. z. B. die Skala Task-Irrelevant Thinking der RTT; Sarason, 1984), für eine Teilung der Aufmerksamkeit sorgen. Hierzu passen die bereits aufgeführten Befunde, dass Testängstliche in stärkerem Maße Gedanken berichten, die nichts mit der Aufgabenbearbeitung zu tun haben (z. B. Deffenbacher, 1978; Sarason & Stoops, 1978).

Deutliches Indiz für einen kausalen Zusammenhang ist die Kopplung der Leistungsunterschiede an bestimmte Situationsmerkmale: „The evidence also indicates that for the most part individuals at different test-anxiety levels show either smaller or no differences in performance and cognitive interference in nontest situations" (Sarason et al., 1990, S. 6). Der Befund, dass niedrig Testängstliche unter nichtevaluativer Instruktion teilweise schlechter abschneiden, wird dabei auf Motivationsverluste zurückgeführt (Zeidner, 1998). Die Theorie der Aufmerksamkeitsteilung ist eng verknüpft mit der Forschung zur Testatmosphäre. Sarason und Stoops (1978) demonstrierten die Abhängigkeit von kognitiver Interferenz und Leistung von der Testatmosphäre in einem Experiment (N = 60 Studentinnen, Studie 3): hoch testängstliche Probanden schnitten unter evaluativer Instruktion in einer Anagrammaufgabe am schlechtesten ab (verglichen mit niedrig und mittel

Testängstlichen bei neutraler und evaluativer sowie mit hoch Testängstlichen bei neutraler Instruktion). Überdies berichteten sie das höchste Maß an kognitiver Interferenz[16] (siehe Abbildung 1).

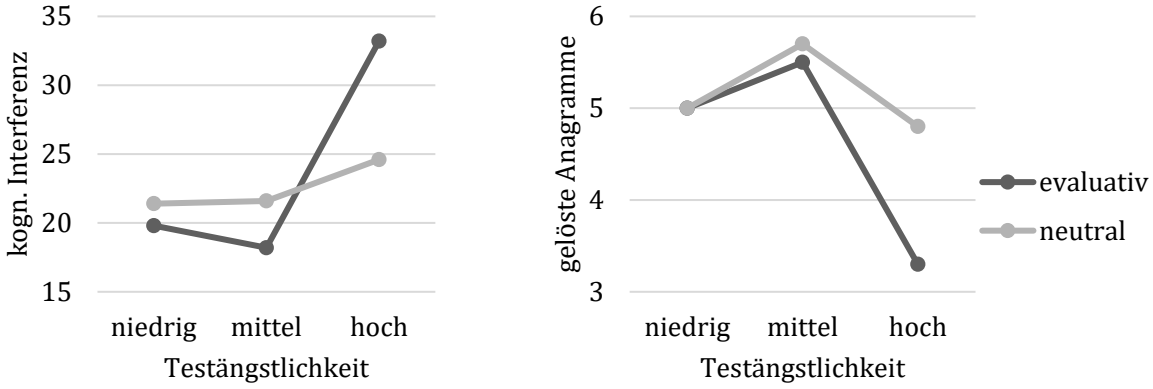

Abbildung 1: Kognitive Interferenz und gelöste Anagramme in Abhängigkeit von Testängstlichkeit und evaluativer gegenüber neutraler Instruktion (Sarason & Stoops, 1978)

Auffällig ist, dass sich die hoch Testängstlichen insbesondere bei evaluativer Instruktion von den anderen Gruppen unterscheiden. Sarasons kognitive Interferenztheorie wird von Zeidner (1998) in einem Schaubild veranschaulicht (siehe Abbildung 2):

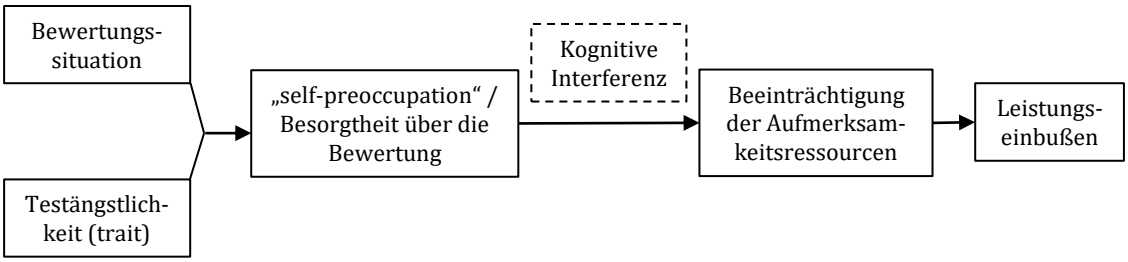

Abbildung 2: Kognitives Interferenzmodell von Sarason (Darstellung nach Zeidner, 1998, S. 67)

Nach Sarason und Sarason (1990) beeinträchtigt kognitive Interferenz Leistung auf verschiedene Weise. Zentral ist hierbei die mentale Beschäftigung mit inneren Prozessen, so beispielsweise in Form sorgenvoller Gedanken über die eigene Leistung („self-preoccupation“). Hinzu kommen „Thoughts about off-task matters and a general wandering of attention from the task” (S. 482; vgl. auch die Skalen Worry und Test-Irrelevant Thinking der RTT bzw. Besorgtheit und Interferenz des TAI-G). Sarason und Sarason (1990) subsumieren beide Aspekte unter „interfering thoughts“ (S. 482).

Die Theorie der kognitiven Interferenz wird gestützt durch zahlreiche Befunde, die sich auf das Testerleben von hoch und niedrig Testängstlichen fokussiert haben. Deffenbacher (1978) stellte in einer bereits zitierten Arbeit fest, dass hoch Testängstliche unter evaluativen Testbedingungen

[16] Erfasst wurde diese mit einer frühen Version des CIQ. Diese erfasst nach Sarason (1978) beispielsweise Zweifel an der eigenen Leistungsfähigkeit, aber auch gänzlich aufgabenirrelevante Gedanken.

deutlich weniger Zeit mit der eigentlichen Aufgabenbearbeitung verbringen als niedrig Testängstliche. Auch auf den Befund von Ganzer (1968), dass hoch Testängstliche in einer Testsituation unter Beobachtung häufiger als Vergleichsgruppen selbstabwertende Kommentare äußern, wurde bereits eingegangen. Blankstein, Toner und Flett (1989) ließen eine studentische Stichprobe (N = 47) unter evaluativer Instruktion eine Anagrammaufgabe bearbeiten. Nach dem Test sollten die Probanden frei und spontan Ihre Gedanken und Gefühle während des Tests auflisten. Auf Basis einer Kodierung der Protokolleinträge zu den geäußerten Gedanken wurde ersichtlich, dass dispositionell hoch Testängstliche im Verhältnis häufiger als mittel und niedrig Testängstliche negative, selbstbezogene Gedanken auflisteten (z. B. „My vocabulary is really poor“) und zugleich deutlich weniger positive, aufgabenbezogene Gedanken angaben (z. B. „Since a lot of words have this letter combination I'll try it“).

Eng verbunden mit der kognitiven Interferenztheorie sind ressourcentheoretische Erklärungen für Leistungsunterschiede zwischen hoch und niedrig Testängstlichen. Wesentlich hierfür ist die Annahme, dass das Informationsverarbeitungssystem ressourcenlimitiert ist, wie es beispielsweise für das Arbeitsgedächtnis in der Theorie von Baddeley angenommen wird (z. B. Baddeley, 2012). Gedanken an ein mögliches Scheitern oder an andere, aufgabenirrelevante Themen beanspruchen kognitive Ressourcen, welche folglich für die Aufgabenanforderungen nicht mehr zur Verfügung stehen (Zeidner, 1998). Leistungsunterschiede zwischen hoch und niedrig Testängstlichen nehmen unter Bewertungsdruck mit zunehmender Aufgabenkomplexität zu (Sarason, 1980), zudem zeigte sich metaanalytisch ein größerer Leistungsunterschied bei subjektiv schweren gegenüber leichten Tests (d = -.45 vs. -.07; k = 13 bzw. 12; N = 874 bzw. 849; Hembree, 1988). Dies kann aus einer Ressourcenperspektive gut erklärt werden, weil komplexere Aufgaben mehr kognitive Ressourcen fordern (Eysenck, 1982).

Die empirische Fundierung der Theorie der kognitiven Interferenz weist einen bedeutenden Schwachpunkt auf: in vielen Fällen werden Leistungsunterschiede in Abhängigkeit der Ausprägung der Testängstlichkeit (trait) kausal auf im Test erlebte Testangst (state) zurückgeführt, ohne dass akute störende Gedanken oder Sorgen direkt erfasst werden (Zeidner, 1998). Ein Beispiel für eine neuere Untersuchung, in der die Annahmen der kognitiven Interferenztheorie direkt überprüft wurden, ist die Arbeit von Coy et al. (2011). Eine Stichprobe von N = 88 Studierenden bearbeitete entweder unter evaluativer oder beruhigender Instruktion eine Reihe von Arbeitsgedächtnisaufgaben (zur phonologischen Schleife, zum visuell-räumlichen Notizblock und zur zentralen Exekutive; hier abgekürzt mit PHS, VRN und ZEX). Vor (nach der Instruktion) und nach den Aufgaben bearbeiteten die Probanden die RTA als Maß für die Testangst (reformuliert als state-Variante für das Erleben vor bzw. retrospektiv während des Tests). Ebenfalls nach den Aufgaben wurde die erlebte kognitive Interferenz mit dem CIQ erhoben. In den Ergebnissen zeigte sich, dass die Werte bei Testangst (vor und nach dem Test) und kognitiver Interferenz in der evaluativen

Bedingung signifikant höher waren als in der beruhigenden. Darüber hinaus wurden in der evaluativen Bedingung signifikant schlechtere Leistungen in den Aufgaben zur PHS und marginal signifikant (p = .094) schlechtere Leistungen in der Aufgabe zur ZEX festgestellt, jedoch kein Leistungsunterschied im VRN. Entscheidendes Ergebnis war, dass der Effekt der Manipulation auf die Leistung in der PHS durch die kognitive Interferenz mediiert wurde, welche selbst einen negativen Effekt auf die Leistung hatte. Da die PHS für die temporäre Speicherung und Aufrechterhaltung von verbaler Information zuständig ist (Baddeley, 2012), war der hier gefundene Mediationseffekt hypothesenkonform – so wird in der Literatur angenommen, dass sorgenvolle Kognitionen mit Aktivität in der ZEX und der PHS, nicht aber im VRN, verbunden sind bzw. an diesen Positionen Arbeitsgedächtnisressourcen beanspruchen (Eysenck, Derakshan, Santos & Calvo, 2007; Rapee, 1993). Dass sich dieser Effekt bei der ZEX nicht zeigte, lässt sich möglicherweise auf unterschiedliche Teilfunktionen in diesem System zurückführen (siehe dazu die Ausführungen zur Processing Efficiency Theory, Abschnitt 1.2.1.1.3) (Coy et al., 2011).

1.2.1.1.2 Aufmerksamkeitslenkung

Die bereits dargestellten Besonderheiten Testängstlicher in der Informationsverarbeitung (siehe Abschnitt 1.1.1.4) lassen sich mit der Theorie der kognitiven Interferenz und der theoretischen Konzeption kognitiver Ressourcen gut vereinbaren. Die beschriebene Hypervigilanz Ängstlicher (Eysenck, 1992) kann zum einen erklären, warum dispositionell Ängstliche mit höherer Wahrscheinlichkeit in einer Situation Testangst erleben, und zwar weil sie ihre Umgebung nach bedrohlichen Reizen absuchen und die Aufmerksamkeit stärker auf bedrohliche Reize richten. Die bereits zitierte Metaanalyse von Bar-Haim et al. (2007) konnte zeigen, dass es über verschiedene Untersuchungsparadigmen hinweg einen Aufmerksamkeitsbias Ängstlicher gegenüber bedrohlichen Reizen gibt. Zweitens ist es wahrscheinlich, dass sich auch während einer Testbearbeitung dieser bias bzw. die Hypervigilanz negativ auf kognitive Leistungen auswirken, z. B. indem ein testängstlicher Student häufiger als wenig testängstliche Studenten an nicht gelöste Aufgaben oder die immer knapper werdende Zeit denkt (letzteres ist beispielsweise Ausdruck der „task-generated inferference“ nach Deffenbacher, 1986) (Zeidner, 1998). Auch die mit der Hypervigilanz verbundene Einengung der Aufmerksamkeit auf bedrohliche Reize (Eysenck, 1992) kann zur Folge haben, dass die Aufmerksamkeit von Aufgabenanforderungen wegbewegt wird, was zu Leistungsverschlechterungen führt (Zeidner & Matthews, 2010). In diesem Sinne ist Hypervigilanz in der S-REF Theorie integriert und dort als ein Merkmal der maladaptiven Situationsbewältigung aufgeführt (siehe z. B. Zeidner & Matthews, 2007). Diese Interpretation wird auch durch den bereits zitierten Befund gestützt, dass sich der Aufmerksamkeitsbias nicht nur in der bevorzugten Aufmerksamkeit auf, sondern auch der erschwerten Ablösung von bedrohlichen Reizen ausdrückt (Cisler et al., 2009).

Obwohl die kognitive Interferenztheorie sowie die Theorien zum Aufmerksamkeitsbias und zur Hypervigilanz sich jeweils auf zahlreiche empirische Belege stützen können, gibt es noch offene Fragen. So finden sich Artikel, die die theoretischen Annahmen empirisch nicht bestätigen konnten. Beispielsweise fanden Deffenbacher und Hazaleus (1985) bei einer Manipulation der Testinstruktion (evaluative bzw. konventionelle Testinstruktion) zwar die erwarteten Haupteffekte der dispositionellen Testängstlichkeit (niedrig vs. hoch) auf Leistung und Testangst, aber keinen Haupteffekt der Bedingung. Eine wichtige Einschränkung der Theorie ist die Frage nach der kausalen Richtung. Morris et al. (1981) bestätigen in ihrer Literaturübersicht, dass worry und nicht emotionality mit Leistungseinbußen einhergeht, konstatieren aber, dass dies kein kausaler Effekt sein muss: „worry may reflect concern about accurately perceived past and present performance difficulties, rather than acting as a cause of poor performance." (S. 544). Dieser wichtige Gedanken wird in Abschnitt 1.2.1.2 im Kontext der Defizitperspektive wieder aufgegriffen und wurde in Studie 1 untersucht. Ein methodisches Problem vieler Studien in diesem Kontext liegt überdies darin, dass die Stichproben auf Basis der Testängstlichkeitswerte in Subgruppen zerlegt wurden – wie etwa durch eine Drittelung in niedrige, moderate und hohe Ausprägung (z. B. Sarason & Stoops, 1978), durch Mediansplits (z. B. Sarason, 1973) oder auch die Bildung von Extremgruppen (z. B. Deffenbacher & Deitz, 1978). Dies ist problematisch, da es sich erstens um eine willkürliche Aufteilung der Stichprobe handelt. Zweitens wird dadurch die Vergleichbarkeit der Studien erschwert, da Unterschiede zwischen hoch und niedrig Testängstlichen je nach Vorgehen (Mediansplit oder z. B. Extremgruppenbildung) kleiner oder größer ausfallen können. Drittens wird durch eine Dichotomisierung kontinuierlicher Variablen die Teststärke reduziert (Irwin & McClelland, 2003). Aus diesen Gründen wurden in den vorliegenden Studien keine Subgruppenbildungen anhand der Testängstlichkeit vorgenommen.

1.2.1.1.3 Verarbeitungseffizienz

Die kognitive Interferenztheorie liefert wichtige Erkenntnisse und trägt zur Erklärung der Leistungsunterschiede bei. Derakshan und Eysenck (2009) kritisieren jedoch, dass die Befundmuster in einigen Fällen nicht der theoretischen Vorhersage folgen. In der bereits zitierten Untersuchung von Blankstein et al. (1989) etwa zeigten sich deutliche Unterschiede in den Protokollen der frei berichteten Gedanken und auch in der kognitiven Interferenz, die Leistung unterschied sich aber nicht in Abhängigkeit von der dispositionellen Testängstlichkeit. Auch das von Seipp (1991) berichtete 95 %-Konfidenzintervall des Leistungszusammenhangs von worry ($-.40 \leq r \leq -.04$) verdeutlicht, dass nicht signifikante Zusammenhänge mit Leistung zum empirischen Bild gehören. Derakshan und Eysenck (2009) gehen davon aus, dass die kognitive Interferenztheorie zur Erklärung der Angst-Leistungs-Relation nicht komplex genug ist und dass sie überdies keine Aussagen trifft, wie das Informationsverarbeitungssystem konkret beeinträchtigt wird.

Eysenck und Calvo (1992) entwickelten die Processing Efficiency Theory (PET), die sich insbesondere mit dem Effekt von state-Angst (weniger von trait-Ängstlichkeit) auf die Leistung befasst. Grundlegend wird dabei zwischen der Effektivität und Effizienz kognitiver Leistungen unterschieden. Während Effektivität die Güte der Leistung bezeichnet, stellt Effizienz das Verhältnis aus Effektivität und Anstrengung dar, wobei diese Anstrengung mit investierten kognitiven Ressourcen gleichgesetzt wird (Eysenck & Calvo, 1992). Sorgenvolle Kognitionen (worry) beanspruchen Ressourcen im Arbeitsgedächtnis. Anders als in der kognitiven Interferenztheorie wird worry aber auch eine motivationale Wirkung zugesprochen. Diese motivationale Funktion besteht nicht nur darin, den aversiven Angstzustand zu reduzieren, sondern auch in dem Ziel, antizipierte Konsequenzen eines Misserfolgs zu vermeiden. Dies geschieht entweder durch den Einsatz zusätzlicher Ressourcen und / oder durch den Einsatz von (besseren) Verarbeitungsstrategien (Eysenck & Calvo, 1992). Diese Kompensationen sind möglich, sofern insgesamt genug Ressourcen vorhanden sind. Dabei verringert sich die Effizienz, nicht aber die Effektivität – dies geschieht erst wenn insgesamt keine Ressourcen verfügbar sind um eine Kompensation zu ermöglichen (Eysenck et al., 2007).

Die PET wurde schließlich zur Attentional Control Theory (ACT) weiterentwickelt (Eysenck et al., 2007). Ausgangspunkt der ACT ist der Gedanke, dass die Kontrolle der Aufmerksamkeit über zwei verschiedene Systeme stattfindet. Das erste System dient der zielgerichteten Verarbeitung und wird beispielsweise durch Erfahrungen oder Erwartungen gesteuert (top-down). Das zweite System hingegen ist reizgetrieben (bottom-up) und bewirkt die Lenkung der Aufmerksamkeit auf bedeutsame Reize in der Umgebung (Corbetta & Shulman, 2002). Die ACT nimmt an, dass Angst zu einem stärkeren Gewicht der bottom-up- und einer schwächeren Bedeutung der top-down-Verarbeitung führt. Diese Annahme wird durch die vielfache empirische Evidenz für den Aufmerksamkeitsbias bei Ängstlichen gestützt (siehe z. B. Bar-Haim et al., 2007). Gemäß der PET wirkt sich Angst insbesondere auf die zentrale Exekutive aus, wobei sich Eysenck et al. (2007) an der Differenzierung exekutiver Funktionen nach Miyake et al. (2000) orientieren: „(a) shifting between tasks or mental sets, (b) updating and monitoring of working memory representations, and (c) inhibition of dominant or prepotent responses." (S. 54). Die negativen Effekte von Angst werden in erster Linie bei shifting und inhibition vermutet. Das bedeutet, dass im Zustand der Angst sowohl der Wechsel zwischen aufgabenrelevanten Reizen (shifting) als auch ein Verhindern der Aufmerksamkeitslenkung auf aufgabenirrelevante Reize und Reaktionen (inhibition) beeinträchtigt sind (Eysenck et al., 2007). In Bezug auf eine Zahlenreihenaufgabe wäre shifting beispielsweise der attentionale Wechsel von der Rechenregel zwischen erster und zweiter Zahl zur Regel zwischen zweiter und dritter Zahl, inhibition wäre zum Beispiel das Ausblenden von aufkommenden Gedanken an das Abschneiden anderer Personen.

Wie die PET differenziert auch die ACT Effizienz und Effektivität. Die Operationalisierung von Effizienz wurde auf unterschiedlichen Wegen vorgenommen, beispielsweise über das Verhältnis von Leistung und Bearbeitungszeit. Eysenck et al. (2007) verfassten eine umfassende Literaturübersicht zur PET bzw. zur ACT und berichten, dass das empirische Fundament für Leseaufgaben am solidesten ist. Typische Kompensationsmechanismen sind in derartigen Untersuchungen das Zurückspringen zu bereits gelesenem Textmaterial und die verbale oder nonverbale Wiederholung während des Lesens. Zwei Studien seien beispielhaft beschrieben, die bezüglich der Stichprobengrößen nicht allzu belastbar, jedoch für das Verständnis der PET bzw. ACT hilfreich sind. Calvo, Eysenck, Ramos und Jiménez (1994) etwa ließen eine studentische Stichprobe (N = 36), die in eine Gruppe hoch und niedrig Testängstliche geteilt wurde, in einer evaluativen Untersuchungssituation an einem Bildschirm kurze wissenschaftliche Texte lesen (Studie 3). Dabei wurde jeweils immer nur ein Satz angezeigt und die Probanden konnten selbstständig zwischen den Sätzen vor und zurück springen. Insgesamt zeigte sich in einem Verständnistest kein Leistungsunterschied zwischen hoch und niedrig Testängstlichen. Die hoch Testängstlichen lasen jedoch länger und sprangen auch häufiger zu bereits gelesenen Sätzen zurück. Hadwin, Brogan und Stevenson (2005) beobachteten bei Schülern (N = 30; M_{Alter} = 10.25 Jahre) bei einer Zahlenspannenaufgabe keine Leistungsunterschiede zwischen hoch und niedrig Ängstlichen (bezogen auf state-Angst), jedoch eine längere Bearbeitungszeit bei Ängstlichen. Überdies schätzten die ängstlichen Probanden die erlebte mentale Anstrengung als höher ein.

Die weiteren Annahmen der ACT sollen nicht weiter vertieft werden. Unverkennbar sind jedoch die Parallelen zur Theorie der Hypervigilanz und zur kognitiven Interferenztheorie von Sarason. Der wichtigste Fortschritt der ACT ist die Differenzierung von Effizienz und Effektivität der Leistung, auf Basis derer erklärt werden kann, warum sich nicht immer negative Zusammenhänge von Testängstlichkeit bzw. Testangst und Leistung zeigen. So kann die zusätzliche Investition kognitiver Ressourcen bei einer hoch ängstlichen Person dazu führen, dass diese im Leistungsresultat ebenso gut abschneidet wie eine niedrig ängstliche Person. Somit wäre die bei beiden Personen vergleichbare Effektivität verbunden mit einer niedrigeren Effizienz bei der hoch ängstlichen Person (siehe hierzu Derakshan & Eysenck, 2009; Eysenck et al., 2007). Der negative Effekt von Angst auf Leistung geht darauf zurück, dass einerseits bedrohliche Reize bevorzugt der Verarbeitung zugeführt werden (bottom-up) und andererseits die Aufmerksamkeitskontrolle (top-down) eingeschränkt ist (Eysenck et al., 2007). Die folgende Formulierung verdeutlicht dabei, dass die ACT nicht im fundamentalen Gegensatz zur kognitiven Interferenztheorie steht: „More generally, attentional control theory accounts for distraction effects in anxiety, and the distracting stimuli can either be *external* (as in most research) or *internal* (e.g., worry).“. (Eysenck et al., 2007, S. 348).

Ein Beispiel für eine Verbindung des „klassischen“ Experimentalparadigmas von Sarason mit der Betrachtung der Verarbeitungseffizienz ist die Arbeit von Calvo, Ramos und Estevez (1992). Eine

studentische Stichprobe (N = 36) wurde in hoch und niedrig Ängstliche (separiert anhand von Werten auf TAI und STAI) geteilt und bearbeitete in drei getrennten Untersuchungen (Studie 1 bis 3) verschiedene Aufgaben. Bei den hoch Ängstlichen wurden schlechtere Werte in einem allgemeinen und in spezifischen Wortverständnistests ermittelt (Studie 1; die spezifischen Wortverständnistests bezogen sich auf Begriffe, welche in Texten enthalten waren, die in Studie 3 zu lesen waren). Bei der erfassten Lesespanne zeigte sich eine schlechtere Leistung der hoch gegenüber den niedrig Ängstlichen unter einer evaluativen, nicht aber in einer nicht-evaluativen Instruktionsbedingung (in der Buchstabenspanne zeigten sich keine Leistungsunterschiede) (Studie 2). In Studie 3 lasen die Probanden eine Reihe von wissenschaftlichen und erzählerischen Texten. Im Anschluss waren Verständnisaufgaben zu lösen. Es zeigte sich insgesamt eine niedrigere Effizienz (Leistung / Lesezeit) bei hoch Ängstlichen, wobei sich die niedrigste Effizient bei Fragen zu den wissenschaftlichen Texten in der Gruppe hoch Ängstlicher ergab, wenn keine Textzusammenfassung als Hilfe gegeben wurde. Mit einer Reihe von Kovarianzanalysen konnten die Autoren darüber hinaus zeigen, dass der Haupteffekt der Testängstlichkeit auf die Effizienz geringer wurde, wenn die Lesespanne (Studie 2) oder das spezifische Wortverständnis (Studie 1) kontrolliert wurden. Der besagte Haupteffekt wurde schließlich nicht signifikant, wenn das allgemeine Wortverständnis (Studie 1) kontrolliert wurde. Die Ergebnisse zeigen zum einen, dass sich Leistungsunterschiede eher unter evaluativen Bedingungen zeigen (Studie 2). Zum anderen wird aber auch deutlich, dass bei der Erklärung von Leistungsunterschieden nicht nur aktuell verfügbare kognitive Ressourcen berücksichtigt werden müssen, sondern auch Defizite im Wissen oder den Fähigkeiten einer Person (Studie 3).

Eine zentrale Limitation der Interferenzperspektive stellt die Frage dar, welche Bedeutung (vor einem Test existierende) interindividuelle Unterschiede in bestimmten Fertigkeiten oder Wissen für Leistungsunterschiede spielen (siehe z. B. Culler & Holahan, 1980; siehe auch Zeidner, 1998). Wird mit einem Test die Beherrschung von Kursinhalten geprüft, ist die Leistung maßgeblich davon abhängig, in welchem Maße die Kursinhalte verstanden wurden und abgerufen werden können. Im Sinne der Konstruktvalidität sollte selbiges für jegliche mit Leistungstests erfasste Fähigkeit gelten. Covington und Omelich (1987) demonstrierten, dass Interferenz und die Qualität des Lernverhaltens nicht voneinander isoliert sind. Eine studentische Stichprobe (N = 189) bearbeitete an zwei aufeinanderfolgenden Tagen dasselbe Wissensquiz zu Kursinhalten. Beim zweiten Quiz wurde eine nicht-evaluative Instruktion administriert. Vorab wurden bei den Teilnehmern die Lerneffektivität (z. B. Lernzeit, Wiederholungen beim Lernen, gezielte Vorbereitung auf die Prüfungsanforderungen) und die Testängstlichkeit erhoben. Nach Vermutung der Autoren lässt sich eine „Blockade“ oder auch ein Problem beim Abruf von Wissen durch Angst daran erkennen, dass bei einer Wiederholungstestung unter nicht-evaluativer Bedingung insbesondere bei hoch

Testängstlichen ein Leistungsgewinn auftritt. Das bedeutet, dass eigentlich beherrschtes Kursmaterial in einer evaluativen Testsituation nicht verfügbar ist – Beispiel hierfür ist die typische Situation, in der einer geprüften Person die Lösung für eine Frage unmittelbar *nach* der Prüfung einfällt. Demgegenüber sollten niedrig Testängstliche geringere Leistungssteigerungen vorweisen, da sie auch unter Bewertungsdruck ihre optimale Leistung zeigen (können). Teilweise konnte diese Hypothese (bezüglich der leichten Quizitems) bestätigt werden: die hoch Testängstlichen mit hoher Lerneffektivität zeigten eine stärkere Leistungssteigerung als die niedrig Testängstlichen mit hoher Lerneffektivität. Die hoch Testängstlichen mit niedriger Lerneffektivität verbesserten ihre Leistung nicht. Der Effekt ging nicht auf unterschiedliches Lernverhalten in der Zeit zwischen den beiden Quizzen zurück. Die Autoren interpretierten das Ergebnis dahingehend, dass hoch Testängstliche unter einer weniger stressreichen Testsituation ihr Wissen besser abrufen können, was aber nur geschieht, wenn tatsächlich Wissen vorhanden ist: „Simply put, learning must be present in order for it to be interfered with." (Covington & Omelich, 1987, S. 393).

Diese Überlegungen machen deutlich, dass die Interferenzperspektive alleine nicht ausreicht, um die Angst-Leistungs-Relation hinreichend zu erklären. Nach der Interferenzperspektive führt Angst zu Veränderungen im Informationsverarbeitungssystem, die wiederum Leistungsverschlechterungen nach sich ziehen. Die konträre Alternativhypothese wäre, dass sich Leistung auf Angst auswirkt. In Abschnitt 1.1.1.3 wurde bereits dargelegt, dass Testängstlichkeit einhergeht mit negativen Überzeugungen über die eigene Person und die eigene Kompetenz (z. B. Krampen, 1988; Ringeisen et al., 2010; Rohrmann et al., 2010). Grundgedanke der Defizitperspektive ist, dass die schlechtere Leistung von hoch Testängstlichen *nicht* auf ihre erlebte Testangst, sondern auf konkrete Defizite zurückgeht, die hoch Testängstliche aufweisen.

1.2.1.2 Defizitperspektive

Wesentlicher Ausgangspunkt der Defizitperspektive ist der Gedanke, dass hoch Testängstliche in einigen Merkmalen Defizite aufweisen, welche wiederum ihrerseits Ursache für die Angst *und* für die schlechtere Leistung sind (Zeidner, 1998). In einer Übersicht zu dieser Forschungsrichtung unterscheidet Tobias (1985) Defizite im Bereich der „study skills" von jenen in „test-taking skills". Defizite in study skills (Lernfertigkeiten) führen demnach zu einer unzureichenden Enkodierung und / oder Speicherung von zu lernenden Inhalten. Als Beispiel sei ein Student genannt, der sich auf eine wichtige Prüfung nicht gut vorbereitet hat (z. B. aufgrund seiner unzureichenden study skills) und somit guten Grund hat, ängstlich zu sein. Test-taking skills beziehen sich wiederum auf den optimalen Umgang mit dem Testmaterial: „any discrete tactic, rule, or procedure that increases the probability of successful interpretation and solution of common test questions" (Bruch, 1981, S. 43).

Einige Befunde von Hembree (1988) haben Bezug zur Defizitperspektive. So fanden sich für hoch im Vergleich zu niedrig Testängstlichen (oberes vs. unteres Drittel der Verteilung) stärkere Enkodierungsprobleme, d = .74 (k = 3, N = 264), aber auch mehr Lernzeit in Stunden pro Woche, d = .53 (k = 3, N = 216). Ferner fand sich ein negativer Zusammenhang von r = -.27 zwischen Testängstlichkeit und study skills (k = 4, N = 399).

Benjamin, McKeachie, Lin und Holinger (1981) befragten N = 146 Studierende nach einer Kursprüfung zu Problemen beim Lernen während des Kurses und der Prüfung selbst sowie nach der Testangst während der Prüfung. Die hoch Testängstlichen berichteten nicht nur die stärksten Abrufprobleme während der Prüfung, sondern gaben auch am meisten Probleme beim Lernen im Verlauf des Kurses und in Vorbereitung auf die Prüfung an. Die Ergebnisse einer zweiten Studie von Benjamin et al. (1981) mit N = 48 Studierenden veranschaulichen die konkreten Defizite in den study skills Testängstlicher. So gaben diese beispielsweise häufiger an, sich Begriffe einzuprägen, ohne diese zu verstehen.

Culler und Holahan (1980) stellten bei einer studentischen Stichprobe (N = 96) fest, dass hoch Testängstliche nicht nur mehr Lernzeit in Stunden berichteten, auch korrelierte nur bei hoch Testängstlichen die Lernzeit signifikant mit GPA, r = .30 vs. -.10. Eine Interpretation der Autoren war, dass hoch Testängstliche mehr Lernzeit aufwenden, um defizitäre study skills auszugleichen. Diese Befunde legen nahe, dass hoch Testängstliche (auch) deswegen in Prüfungen schlechtere Leistungen erbringen, weil sie die getesteten Inhalte, bedingt durch ihre niedrigen study skills, weniger gut beherrschen. Die Wahrnehmung dieser unzureichenden Vorbereitung führt schließlich zum Erleben von Testangst während des Tests (Tobias, 1985).

Darüber hinaus konnten verschiedene Studien Unterschiede in den test-taking skills feststellen. Bruch (1981) stellte bei einer studentischen Stichprobe (N = 118) schlechtere test-taking skills bei hoch und moderat Testängstlichen gegenüber niedrig Testängstlichen fest. Paulman und Kennelly (1984) konnten bei einer studentischen Stichprobe (N = 64) Haupteffekte von Testängstlichkeit und test-taking skills auf kognitive Leistungen nachweisen. Die Autoren vermuten, dass es hoch testängstliche Personen gibt, die über gute Fertigkeiten bezüglich Lernen und effektiver Testbearbeitung verfügen, womit sie den negativen Effekt der Testangst ausgleichen können.

Belege für diese Vermutung fanden Naveh-Benjamin, McKeachie und Lin (1987). Sie verglichen bei einer studentischen Stichprobe (N = 65; Studie 2) hoch und niedrig Testängstliche mit jeweils gut oder schlecht ausgeprägten study skills. Als Leistungsmaße wurden die Kursprüfung (evaluative Situation) sowie eine Aufgabe zum Verständnis der Kursinhalte (nonevaluative Situation) herangezogen. In der nonevaluativen Situation erbrachten die hoch Testängstlichen mit guten study skills eine vergleichbare Leistung wie die niedrig Testängstlichen. In der evaluativen Situation hingegen schnitten alle hoch Testängstlichen schlecht ab – unabhängig von ihren study skills.

Die Autoren folgerten daraus, dass hoch Testängstliche mit guten study skills (lediglich) Schwierigkeiten beim Abruf der Kursinhalte hatten. Demgegenüber hatten hoch Testängstliche mit schlechten study skills zusätzlich auch Probleme beim vorherigen Erwerb des Materials.

Implikationen im Sinne der Defizitperspektive haben auch Untersuchungen dazu, wie Testängstlichkeit durch Interventionen reduziert werden kann und insbesondere inwiefern sich dies auch in einer verbesserten Leistung niederschlägt. Kirkland und Hollandsworth (1980) beispielsweise zeigten in ihrer Interventionsstudie eine Überlegenheit eines skill-orientierten gegenüber angstreduzierenden Trainings. Dies nahmen sie zum Anlass, den praktischen Nutzen des Konstrukts Testängstlichkeit per se in Frage zu stellen. Nach der Metaanalyse von Hembree (1988) finden sich jedoch positive Effekte auf verschiedene Leistungsmaße sowohl für skill-orientierte und angstreduzierende Maßnahmen als auch für Kombinationen beider Herangehensweisen (siehe hierzu Abschnitt 1.3). Auf Basis dieser Befundlage kann also nicht geklärt werden, ob nun die Interferenz- oder die Defizitperspektive gültig ist.

Einige Untersuchungen haben versucht, die relative Bedeutung von study skills gegenüber Testängstlichkeit in der Prädiktion von Leistung zu eruieren. Smith, Arnkoff und Wright (1990) führten hierzu eine Studie mit N = 178 Studierenden durch. Sie erhoben im Laufe eines Kurses zahlreiche Variablen, unter anderem Bewertungssorgen im akademischen Bereich, Testangst direkt vor der Prüfung, Gedanken während der Prüfung im Rückblick sowie akademische Fertigkeiten, d. h. Lerngewohnheiten und die vor dem Test eingeschätzte Qualität der eigenen Prüfungsvorbereitung. Als Maß für die individuellen Fähigkeiten wurde der SAT-Wert herangezogen. Die Autoren schlussfolgerten aus den Ergebnissen, dass (neben den Fähigkeiten als bedeutsamem Prädiktor) kognitiv-attentionale Variablen (also Bewertungssorgen, negative Gedanken und Sorgen während der Prüfung) relativ gesehen ein stärkeres Gewicht in der Vorhersage haben als akademische Fertigkeiten und weitere, ebenfalls erhobene selbstwirksamkeitsbezogene Variablen. Ein Teilergebnis war, dass die akademischen Fertigkeiten bivariat bedeutsam mit der Note (r = .35) korrelierten. In der hierarchischen Regression zur Vorhersage der Note klärten Fähigkeiten und kognitiv-attentionale Variablen signifikant Varianz auf (11 bzw. 19 %), die akademischen Fertigkeiten jedoch nicht (3 %). Musch und Bröder (1999b) erhoben bei N = 66 Studierenden eines Statistikkurses Lerngewohnheiten, die Mathematikfähigkeiten (operationalisiert über die letzte Mathematiknote) und die während der Statistikprüfung erlebte Testangst. In einer hierarchischen Regression zeigte sich, dass die Mathematikfähigkeiten (17 %) und Testangst (5 %, $p < .06$), nicht aber die Lerngewohnheiten (3 %), Varianz in der Klausurnote aufklärten. Eine mögliche Erklärung dieses Befunds liegt darin, dass Lerngewohnheiten und die Mathematikfähigkeiten nicht voneinander unabhängig sind, was auch deren Korrelation von r = -.30 nahelegt (Musch & Bröder, 1999b).

Die Befunde von Smith et al. (1990) sowie Musch und Bröder (1999b) machen insgesamt deutlich, dass es sinnvoll ist, Interferenz- *und* Defizitperspektive zu berücksichtigen. Statt nach einer Entscheidung für oder gegen einen Ansatz zu suchen sollte es vielmehr das Ziel sein, die spezifischen Geltungsbereiche, Bedingungen und Wechselwirkungen der in der jeweiligen Perspektive postulierten Prozesse zu verstehen (siehe auch Zeidner, 1998). Dies war auch ein Bestandteil der Fragstellung von Studie 1. In einer globalen Abgrenzung der beiden Perspektiven stellt Zeidner (1998) heraus, dass aus Sicht der Defizitperspektive Schwierigkeiten in der Enkodierung und mentalen Strukturierung von zu lernenden Inhalten Ursache von Leistungsunterschieden sind, wohingegen aus Sicht der Interferenzperspektive Schwierigkeiten lediglich beim Abruf existieren. Der wichtigste Unterschied betrifft die Rolle der Testangst – diese ist im Defizitmodell lediglich ein Korrelat unzureichender skills sowie eine Folge von subjektiv wahrgenommenen Defiziten. Anders als in der Interferenzperspektive spielt Testangst dabei keine kausale Rolle zur Erklärung von Leistungsunterschieden (Zeidner, 1998)[17].

Bislang wurde von Defiziten bei den „skills" gesprochen. Sofern die unzureichende Beherrschung eines Prüfungsstoffs aus den mangelhaften study skills einer Person resultiert, ist das Defizit in den study skills gleichbedeutend mit einem Defizit im Wissen oder der Kompetenz in dem entsprechenden Bereich. Die bisher behandelten Studien umfassten skills (Fertigkeiten), welche sich auf Lernverhalten und Bearbeitungsstrategien in Prüfungen und Tests beziehen. Nun können manche Leistungsunterschiede nicht ohne weiteres mit interindividuell unterschiedlich guten study skills erklärt werden, so zum Beispiel bei Intelligenztests oder in Aufgaben zu basalen kognitiven Leistungen, da für derartige Tests üblicherweise keine Lernvorbereitungen stattfinden (Reeve & Bonaccio, 2008). Hingegen ist es möglich, dass Leistungsunterschiede hier durch niedrige test-taking skills zustande kommen. Eine dritte Quelle von Defiziten führt zu einer grundlegenderen Interpretation des Defizitbegriffs. Leistungsunterschiede könnten demnach auf latente Fähigkeitsunterschiede zurückgeführt werden, die überdies auch das Ausmaß an Testangst determinieren (Reeve & Bonaccio, 2008). Diese Interpretation soll in dieser Arbeit explizit mit in die Defizitperspektive integriert werden (hierauf wird im Anschluss eingegangen). Trotz der Konfundierung von Fähigkeiten und Fertigkeiten (siehe z. B. Musch & Bröder, 1999b) ist es plausibel, dass ein niedriges Fähigkeitsniveau *direkten* Einfluss auf die Leistung *und* Testangst hat, insbesondere in Leistungssituationen, in denen study skills und gezielte Vorbereitung eine untergeordnete Rolle spielen. Die Kernannahmen der Defizitperspektive und der für diese Arbeit gesetzten Schwerpunkte darin sind in Abbildung 3 veranschaulicht.

[17] Ein weiteres, von Zeidner (1998) vorgeschlagenes Defizitmodell, wird hier nicht behandelt.

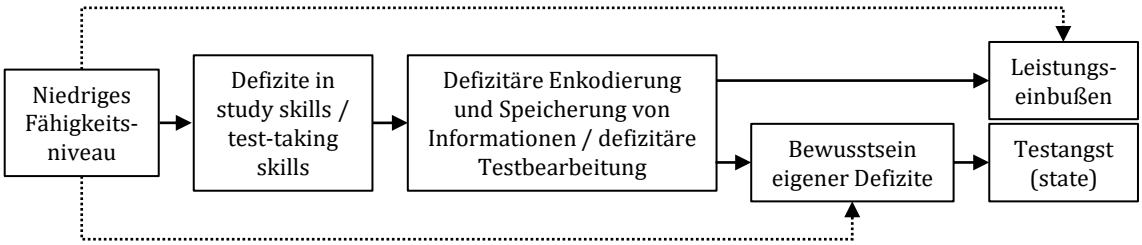

Abbildung 3: Variante des Defizitmodells auf Basis von Tobias (1985), Covington und Omelich (1988) sowie Zeidner (1998, S. 71)

Einige Studien haben versucht, Interferenz- und Defizitperspektive mit SEM zu kontrastieren. Diese Studien befassten sich nicht mit study oder test-taking skills, sondern mit latenten Fähigkeitsunterschieden, die im Rahmen dieser Arbeit als dritte Quelle von Defiziten verstanden werden. Der Grundgedanke dieser Ansätze wurde bereits zu Beginn dieser Arbeit formuliert. Nach Hembree (1988) resultiert aus dem negativen Zusammenhang von Testängstlichkeit und Leistung, dass „the IQs, aptitudes, and progress of test-anxious students are consistently misinterpreted and undervalued […] The validity of the entire testing process is challenged." (S. 75). Aus einer messtheoretischen Sicht liegt ein Messfehler vor, wenn Personen mit *derselben* Ausprägung auf einem latenten Merkmal *nicht* dieselben Item- oder Testwerte vorweisen und sich gleichzeitig in anderen Merkmalen unterscheiden (Sommer & Arendasy, 2014). Ein solches Konstrukt, das von der Fähigkeit zu unterscheiden ist, wäre beispielsweise Testängstlichkeit. Ein Messfehler bedeutet also, dass die beobachteten Werte (z. B. Testwerte) sowohl durch die (zugehörige) latente Variable als auch durch weitere Variablen determiniert sind (Reeve & Bonaccio, 2008). Reeve und Bonaccio (2008) argumentieren, dass sich Defizit- und Interferenzperspektive anhand der Existenz des bias differenzieren lassen. Ausgangspunkt ist, dass die latente Variable Intelligenz Pfade zu deren Indikatoren aufweist. Bei Gültigkeit der Interferenzperspektive müsste es signifikante Pfade von Testangst auf diese Indikatoren (in diesem Fall Itemparzellen der Intelligenzskalen) geben (General Bias Model), wobei in diesem Fall Pfade der Testangst auf die latente Variable Intelligenz von nachrangiger Bedeutung seien. Unter Annahme der Defizitperspektive sollte Testangst mit der latenten Variable Intelligenz korrelieren, aber keine signifikanten Pfade auf die Indikatoren letzterer aufweisen (No Bias Model). Zur Überprüfung dieser Annahme wurden bei N = 185 Studierenden mehrere Intelligenztests durchgeführt, wobei latente Variablen für verbales, numerisches, räumliches Denken, fluide Intelligenz, cognitive speededness und den g-Faktor modelliert wurden. Im Anschluss an die Tests wurde die Testangst erfasst. Insgesamt zeigte das No Bias Model einen etwas besseren fit und die Pfade von Testangst auf die Itemparzellen waren nicht signifikant. Reeve und Bonaccio (2008) folgerten daraus, dass durch Testangst kein Messfehler erzeugt wird.

Auch Sommer und Arendasy (2014) stellten bessere fit-Indizes für No Bias gegenüber General Bias Modellen fest, sowohl bei Testängstlichkeit (trait) und Testangst (state). In einer weiteren

Studie replizierten Sommer und Arendasy (2015) diesen Befund in Bezug auf die Zusammenhänge von Testängstlichkeit und Intelligenz bei einer großen Stichprobe von Bewerbern für ein Medizinstudium (N = 1.768). Reeve und Bonaccio (2008) weisen jedoch einschränkend darauf hin, dass ihre Befunde Raum lassen für beide Kausalrichtungen – so könne Testangst zu einer niedrigeren (latenten) Fähigkeit führen, aber auch das Bewusstsein der eigenen Unfähigkeit zu Testangst.

Eine bereits angesprochene Annahme der Defizitperspektive ist, dass individuelle Defizite kausale Ursachen schlechter Leistung sind und das Bewusstsein der eigenen Defizite Ursache der Angst ist, wobei die Angst selbst keine eigentliche Leistungsrelevanz aufweist (Covington & Omelich, 1988; Zeidner, 1998). Außer Frage steht, dass die subjektiven Fähigkeiten eine wichtige Determinante der erlebten Testangst sind. So wurde bereits auf den quer- und längsschnittlichen Zusammenhang des Fähigkeitsselbstkonzepts mit Testängstlichkeit eingegangen (Krampen, 1991; Ringeisen et al., 2010; Rohrmann et al., 2010). Hong und Karstensson (2002) konnten bei einer studentischen Stichprobe (N = 298) zeigen, dass der Zusammenhang zwischen Mathematikfähigkeit (operationalisiert durch einen Test) und der Ängstlichkeit bezogen auf einen Statistikkurs durch die wahrgenommene Schwierigkeit des Kurses mediiert wird.

Dieser Gedanke führt zu einer wichtigen Limitation der Defizitperspektive. Angenommen, ein Fähigkeits- oder Fertigkeitsdefizit ist die kausale Ursache für schlechte Leistung *und* erhöhte Testangst, wobei Testangst *nicht* kausal mit (niedriger) Leistung verknüpft ist. In diesem Fall ist ein schlechtes Leistungsresultat Ausdruck einer niedrigen Fähigkeit und hohe Testangst ein Resultat der erlebten Defizite (siehe Abbildung 3). Dies setzt allerdings voraus, dass eine Person ihre (niedrigen) Fähigkeiten valide einschätzt (siehe auch Strohbeck-Kühner, 1999). Freund und Kasten (2012) führten eine Metaanalyse zum Zusammenhang von selbsteingeschätzten und gemessenen kognitiven Fähigkeiten durch. Sie ermittelten insgesamt einen moderaten Zusammenhang von r = .33 (k = 154; N = 22.256). Dieser Zusammenhang erhöhte sich zwar bei numerischen (gegenüber allgemeinen kognitiven) Fähigkeiten und bei Selbsteinschätzungen relativ zu einer Vergleichsgruppe (gegenüber einer absoluten Einschätzung). Das Ergebnis macht jedoch deutlich, dass die Selbsteinschätzung von Fähigkeiten weit von einer perfekten Validität entfernt ist. Dass subjektive Defizite (d. h. subjektiv niedrige Ressourcen) zum Erleben von Testangst führen, ist auf Basis der Theorie von Lazarus zu erwarten. Leistungsunterschiede werden jedoch in der Defizitperspektive letztlich auf objektive Fähigkeits- bzw. Fertigkeitsunterschiede zurückgeführt. Wie Freund und Kasten (2012) zeigen, sind aber subjektive und objektive Defizite nicht deckungsgleich. Einige Untersuchungen konnten darüber hinaus Zusammenhänge zwischen kognitiven Verzerrungen und Testängstlichkeit feststellen (Putwain et al., 2010; Wong, 2008), was die Validität der Selbsteinschätzungen von Testängstlichen in Frage stellt. Ausgehend von den Befunden von Freund und Kasten (2012) dürfte die Validität einer Selbsteinschätzung (und damit eines

empfundenen Defizits) stark von der Art des Prüfung bzw. des Tests abhängen. Beispielsweise sollte eine Defiziteinschätzung umso zutreffender sein, je mehr deklaratives Wissen in einer Prüfung abgefragt wird („Ich weiß was ich nicht weiß"). Vor diesem Hintergrund und den gefundenen Unterschieden in lernbezogenen Fertigkeiten und Metakognitionen (Abschnitt 1.1.3) ist aber noch eine komplexere Situation denkbar: so könnten Testängstliche *per se* schlechter in der Lage sein, ihre Stärken und Schwächen adäquat einzuschätzen.

Defizit- und Interferenzperspektive wurden häufig als gegensätzlich gesehen (Zeidner, 1998) – sie müssen sich jedoch nicht per se widersprechen. Tobias (1985) schlug beispielsweise vor, beide Ansätze aus einer ressourcentheoretischen Sicht zu vereinbaren. Demnach bindet Testangst kognitive Ressourcen, während study skills und test-taking skills den gegenteiligen Effekt haben. Sie senken den Ressourcenaufwand einer kognitiven Aufgabe, „improving performance by enabling demanding tasks to fit into available processing capacity." (Tobias, 1985, S. 139). Die beste Leistung sollten also Personen zeigen, die sowohl gute skills als auch eine niedrige Testängstlichkeit vorweisen. Neuere Ansätze wie die S-REF Theorie (siehe Abschnitt 1.1.3) sind in der Lage, beide Perspektiven miteinander zu verbinden, worauf am Ende von Abschnitt 1.2 noch eingegangen wird.

In der Literatur wird immer wieder die Idee rezipiert, dass die Moderation des Angst-Leistungs-Zusammenhangs durch den Messzeitpunkt Hinweise auf die tatsächliche (oder primäre) Kausalrichtung liefert. Dieser in der vorliegenden Arbeit bislang nicht thematisierte Ansatz, den Geltungsbereich von Interferenz- gegenüber Defizitperspektive zu prüfen, ist Gegenstand des folgenden Abschnitts. Diese Thematik ist auch der Kern der Fragestellung von Studie 1, die sich mit eben diesem Effekt und dessen Implikationen beschäftigt hat.

1.2.1.3 Messzeitpunkt als Moderator und Rückschlüsse auf Kausalprozesse

Seipp (1991) verglich in ihrer Metaanalyse die Zusammenhänge von Angst und Leistung in Abhängigkeit des Messzeitpunkts der Angst (siehe Abschnitt 1.2). Es fand sich ein stärkerer Effekt bei Messung der Angst nach dem Test gegenüber vor dem Test und einer vom Test unabhängigen Messung (die letzteren beiden Effekte unterschieden sich in der Höhe nicht voneinander). Die Autorin interpretierte das Ergebnis dahingehend, dass Angst womöglich nicht negativ auf Leistung wirkt, sondern im umgekehrten Sinne Leistung Angst determiniert. Dieser Moderationseffekt ist ein immer wieder zitierter Befund, dessen Ursachen jedoch kaum systematisch erforscht wurden. Während es zahllose Studien gibt, die Angst oder Testangst als state erfasst haben, ist die Anzahl an Studien, die Testangst-Messungen sowohl vor als auch nach einem Test vorgenommen und die Effekte kontrastiert haben, deutlich geringer. Auf einige dieser Studien soll nun eingegangen werden.

Klinger (1984) erhob bei einer studentischen Stichprobe (N = 82) vor und nach einer Prüfung die aktuell erlebte state-Angst. Für die pretest-Messung ergab sich kein signifikanter Zusammenhang mit Leistung, jedoch für die posttest-Messung, r = .18 vs. .46 (inverse Polung). Der Autor folgerte, „that the affect-performance relation may best be thought of as a performance-affect relation" (S. 1385). Zeidner (1991) erhob bei einer Stichprobe von N = 378 israelischen Studienplatzbewerbern (M_{Alter} = 23.45) unmittelbar vor oder nach einer Zulassungstestbatterie die Testängstlichkeit (trait). In den Ergebnissen zeigte sich ein deutlich geringerer Zusammenhang zwischen Testängstlichkeit und Leistung bei der Messung vor gegenüber nach dem Test, r = -.11 vs. -.40, bei Besorgtheit r = -.17 vs. -.39. Bezüglich der absoluten Ausprägung unterschieden sich weder die Leistung noch die Testängstlichkeit zwischen den beiden Bedingungen. Zwar war die Streuung der nach dem Test gemessenen Testängstlichkeitswerte etwas höher, allerdings in einem so geringen Maß, dass dies kaum die Ergebnisse erklären konnte. Auch in diesem Fall interpretierte der Autor die Ergebnisse im Sinne der Defizitperspektive: „the major effect of time of measurement appears to be in 'realigning' the levels of test anxiety to accord with actual levels of test performance." (Zeidner, 1991, S. 107). Zu einer ähnlichen Schlussfolgerung gelangte Strohbeck-Kühner (1999). Er untersuchte N = 181 Personen, welche die Medizinisch-Psychologische Fahreignungsuntersuchung (MPU) absolvierten. Über verschiedene Tests aus dem sensorischen und motorischen Leistungsbereich fanden sich keine signifikanten Korrelationen für die vor den Testungen erfasste Testangst, jedoch über mehrere verschiedene Verfahren signifikant negative Zusammenhänge mit der nach den Tests erfassten Testangst. Strohbeck-Kühner (1999) interpretierte die Ergebnisse in dem Sinne, dass die erlebte Leistung die nach dem Test berichtete Testangst beeinflusst. Gleichwohl betonte er, dass dies nur der Fall sein kann, wenn die Probanden ihre Leistung tatsächlich korrekt beurteilen – die jeweils erfasste subjektive Leistung korrelierte überwiegend positiv mit der tatsächlichen Leistung, und zwar in etwa der Höhe, die Freund und Kasten (2012) metaanalytisch ermittelten.

Diese Arbeiten sind mit einer reinen Interferenzperspektive nur schwer zu vereinbaren, jedoch erschweren einige methodische Aspekte eine eindeutige Interpretation. So erhob Zeidner (1991) nicht etwa state-, sondern trait-Maße, die sich – zumindest was die Instruktion anbelangte – nicht direkt auf die Testsituation bezogen. Auch wurde die subjektive Leistung, die Kern der Ergebnisinterpretation ist, nicht erhoben. Strohbeck-Kühner (1999) erfasste die subjektive Leistung, prüfte aber nicht die eigentlich unterstellte Mediation des Zusammenhangs von objektiver Leistung und Testangst (nach dem Test) durch subjektive Leistung. Sommer und Arendasy (2014) führten in einer bereits zitierten Studie mehrere kognitive Fähigkeitstests bei N = 411 Studierenden durch und erfassten jeweils vor (prä) und nach (post) den Tests die state-Angst. Es fanden sich durchgängig keine signifikanten Zusammenhänge zu Leistung für die prä-Maße der state-

Angst, jedoch signifikant negative für die post-Maße. Auch Sommer und Arendasy (2014) interpretierten ihre Ergebnisse im Sinne der Defizitperspektive, sie erfassten allerdings die subjektive Leistung nicht.

Der Unterschied in der Angst-Leistungs-Relation zwischen prä- und post-Maßen der Testangst hängt eng mit einer weiteren Frage zusammen, die bislang noch nicht betrachtet wurde. Eine Bewertungssituation lässt sich in unterschiedliche Phasen differenzieren, in denen sich auch das Erleben von Emotionen allgemein und Angst im Spezifischen unterscheidet (Folkman & Lazarus, 1985). Dies äußert sich beispielsweise in den Interkorrelationen von Testängstlichkeit (trait) und Testangst (state) vor bzw. nach einem Test. Beispielsweise berichtet Klinger (1984) eine eher schwache Korrelation eines state-Angst-Maßes vor mit jenem nach einer Prüfung von r = .18. Ringeisen und Buchwald (2010) erhoben bei N = 82 Schülern (M_{Alter} = 18.06) drei Wochen vor einer Abschlussprüfung die Testängstlichkeit. Zu diesem Zeitpunkt (t1) sowie direkt nach der Prüfung (t2) wurden verschiedene state-Emotionen bezüglich der Prüfung erhoben, unter anderem jene, die sich auf eine bedrohliche Bewertung der Prüfung bezogen („anxious", „worried"; subsumiert unter threat appraisals). Die Testängstlichkeit korrelierte stark mit threat appraisals zu t1, aber deutlich schwächer mit threat appraisals zu t2, r = .63 vs. .27. Bei statistischer Kontrolle der threat appraisals zu t1 korrelierte Testängstlichkeit nicht mehr signifikant mit den threat appraisals zu t2, r = -.14. Auch zwischen Testängstlichkeit und den anderen Emotionen (challenge, harm & benefit appraisals) fanden sich Zusammenhänge zu t1, jedoch keine oder schwache zu t2. Zwischen den jeweiligen Emotionen zu t1 und t2 fanden sich überwiegend hohe Zusammenhänge. Eine Erklärung für diesen Befund liegt in der Interaktion von Person und Situation: die Prüfung und deren Wahrnehmung beeinflussen das aktuelle Erleben umso stärker – relativ zur Disposition Testängstlichkeit – je geringer der zeitliche Abstand zur Prüfung ist (Ringeisen & Buchwald, 2010). Dieser Befund macht aber auch deutlich, wie stark das Erleben von Testangst vom Messzeitpunkt abhängt. Zusätzliche Interpretationsschwierigkeiten werden durch die unterschiedliche Formulierung der post-Maße von Angst bzw. Testangst erzeugt. So kann nach dem Test erfasste Testangst entweder das Erleben während des Tests retrospektiv abbilden (z. B. Sommer & Arendasy, 2014) oder sich auf das aktuelle Erleben nach dem Test beziehen (z. B. Klinger, 1984; Strohbeck-Kühner, 1999). Es ist zu vermuten, dass post-Maße, die sich auf das aktuelle Erleben beziehen, letztlich die Angst vor der Leistungsrückmeldung widerspiegeln (z. B. dem Ergebnis der MPU bei Strohbeck-Kühner, 1999). Diese sollte unter dem Eindruck der subjektiven Leistung stehen. Letzteres gilt auch für die andere Form der post-Maße, wobei diese vermutlich stärker das Erleben während des Tests bzw. der Prüfung abbilden.

Diese Befunde werfen eine zentrale Frage auf, welche die Wissenschaftler mit der bisherigen Forschung bislang nicht beantwortet haben. Dies ist die Frage, wann Testangst erfasst werden sollte, oder präziser, welche konkreten Prozesse in Abhängigkeit des Messzeitpunkts erfasst werden

und welche Konsequenzen dies für die Interpretation der Ergebnisse hat. Zwar bietet die Defizitperspektive eine Erklärung dafür, warum sich für die Testangst post- vs. prä stärkere Leistungseffekte finden. Eine direkte Überprüfung dieser Annahme wurde jedoch noch nicht vorgenommen. Wann Testangst erfasst wird spielt nicht nur eine Rolle dafür, *was* jeweils erfasst wird. Eng hiermit zusammen hängt die Frage, ob nicht auch *die Erfassung selbst* Auswirkungen auf das Erleben und Verhalten in der Situation hat. Hierzu liegen nur wenige Untersuchungen vor.

Galassi, Frierson und Sharer (1981) erhoben bei einer studentischen Stichprobe (N = 234) entweder während oder nach einer Prüfung einige Maße zum Erleben während des Tests (u. a. positive und negative Gedanken, Wahrnehmung körperlicher Prozesse, Angst). Die Manipulation des Messzeitpunkts hatte keine Effekte auf die genannten Variablen sowie die Leistung. Ein kritischeres Bild zeichnen die Befunde von Brodish und Devine (2009). Diese ließen eine weibliche, studentische Stichprobe (N = 101) einen Test mit mathematischen Aufgaben bearbeiten. Die eine Hälfte der Stichprobe wurde direkt vor dem Test (nach einem Beispielitem) nach ihrer Testangst befragt, während die andere Hälfte nach dem Test befragt wurde (auf diese Studie wird in Abschnitt 1.2.2.2 nochmals eingegangen). Es zeigte sich ein deutlicher Leistungsunterschied zuungunsten der Gruppe, die vor dem Test nach ihrer Testangst befragt wurde. Ursache hierfür könnte sein, dass die Frage nach der erlebten Testangst die entsprechenden (leistungsmindernden) Testangstprozesse verursacht hat (Brodish & Devine, 2009). Prinzipiell ist es denkbar, dass die Erfassung von Testangst vor einem Test den Bewertungscharakter einer Situation verstärkt, eine Problematik, die ebenfalls in Studie 1 untersucht wurde. So könnte die Frage danach, ob man nervös ist, im Sinne eines informativen sozialen Einflusses (Deutsch & Gerard, 1955; Hewstone & Martin, 2014) so interpretiert werden, dass es in der sozialen Situation angemessen ist, Testangst zu verspüren: man stelle sich einen Schüler vor, der erst testängstlich wird, nachdem er von einem Mitschüler auf seine offensichtliche, scheinbar unangebrachte Entspanntheit vor einer Prüfung angesprochen wird.

Wie deutlich gemacht wurde, wurden Unterschiede in der Angst-Leistungs-Relation in Abhängigkeit des Messzeitpunkts im Sinne der Defizitperspektive interpretiert. Die subjektive Leistung spielt in diesem Kontext eine Schlüsselrolle: das (schlechte) subjektive Abschneiden in einem Test führt zu (hoher) Testangst. Diese Mediationsannahme, also dass subjektive Leistung den Effekt von objektiver Leistung auf Testangst mediiert, setzt eine enge Beziehung zwischen objektiver und subjektiver Leistung voraus. Wie die Befunde zu kognitiven Verzerrungen und der moderaten Relation von tatsächlicher und selbsteingeschätzter Intelligenz zeigen, kann dies jedoch nur mit deutlichen Vorbehalten angenommen werden. Es liegt nahe, dass die Erklärung für die Moderation der Angst-Leistungs-Relation durch den Messzeitpunkt in den Besonderheiten der nach dem

Test erfassten Testangst liegt. Hilfe bei der Erklärung dieses Effekts liefert eine Betrachtung aus Sicht der Selbstwertregulation.

1.2.1.3.1 Testangst und Selbstwertregulation

Die Selbstwerttheorie der Leistungsmotivation (Covington, 1984a, 1984b) ermöglicht es, das Phänomen Testangst im Kontext der Selbstwertregulation zu verstehen. Grundgedanke dieser Theorie ist die Annahme eines Selbstwertmotivs, sprich des menschlichen Bedürfnisses, eine positive Vorstellung der eigenen Person zu bilden und zu wahren. Aufgrund der hohen Bedeutung von Leistung bzw. Leistungsfähigkeit für den „sozialen Wert" einer Person gibt es eine Reihe von Mechanismen, die dazu dienen, den Selbstwert gegen unvermeidbare Misserfolge und damit verbundene Bedrohungen zu schützen. Das bereits angesprochene self-handicapping, die Vermeidung von Leistungssituationen (siehe Abschnitt 1.1.3), aber auch die Reduktion der eigenen Anstrengung (Covington, 1984a) sind hierbei mögliche Strategien. Dabei geht es nicht immer um die Vermeidung von Misserfolg selbst, sondern auch um „avoiding the implications of failure" (Covington, 1984a, S. 82). Mechanismen wie diese ermöglichen also, einen Misserfolg auf *andere* Ursachen als die eigenen Fähigkeiten zu attribuieren (Covington, 1984a).

Im Folgenden soll kurz dargelegt werden, wie der Bericht von Testangst in einem Fragebogen im Sinne eines Regulationsprozesses verstanden werden kann. Diese Überlegung erfordert eine abermalige Auseinandersetzung mit dem Konzept des self-handicapping. Leary und Shepperd (1986) unterscheiden auf Basis einer Literaturübersicht zwei Varianten von self-handicapping. Dies ist einerseits die Herbeiführung von Bedingungen, die eine optimale Leistung erschweren, aber gleichzeitig eine plausible Attributionsmöglichkeit für tatsächliches Scheitern bieten. Dies ist nahe an der bereits zitierten Definition von Berglas und Jones (1978). Behaviorales self-handicapping kann also den Selbstwert schützen, da ein Versagen nicht auf die eigenen Fähigkeiten zurückgeführt werden muss. Eine zweite Variante von self-handicapping erfolgt durch den Selbstbericht hindernder Umstände. Das bedeutet, dass Personen nicht aktiv Leistungshindernisse aufbauen, sondern verstärkt von internen Prozessen oder Befindlichkeiten berichten, welche leistungsbeeinträchtigend wirken oder wirkten, wie z. B. Angst oder körperliches Unwohlsein (Leary & Shepperd, 1986).

Eine Untersuchung von Smith, Snyder und Handelsman (1982) veranschaulicht, wie der Bericht von Testangst derartige Funktionen annehmen kann. Eine weibliche studentische Stichprobe (N = 117), die in hoch und niedrig dispositionell testängstliche Probanden geteilt wurde, bearbeitete sehr schwere Testaufgaben. Nach dem Test sollten die Probanden ihre state-Angst in Bezug auf die Testbearbeitung angeben und erhielten hierfür unterschiedliche Instruktionen. Einer ersten

Gruppe wurde gesagt, dass Angst Leistungsbeeinträchtigungen hervorrufe (Angst schadet[18]). Einer zweiten Gruppe wurde das Gegenteil mitgeteilt, nämlich dass Angst *nicht* der Leistung schade (Angst schadet nicht). Eine dritte Gruppe schließlich erhielt keine dieser Informationen (Kontroll). Bei den hoch (nicht aber bei den niedrig) Testängstlichen fanden sich Unterschiede in Abhängigkeit von der Instruktion. Nur diese Teilstichprobe wird hier beschrieben: die dritte Gruppe (Kontroll) berichtete ein höheres Angstniveau als jene in der ersten Gruppe (Angst schadet). Die zweite Gruppe (Angst schadet nicht) berichtete das niedrigste Angstniveau. Die Ergebnisse weisen darauf hin, dass hoch Testängstliche nach einem Test weniger state-Angst berichten, wenn ihnen die Möglichkeit „genommen" wird, ein schlechtes Testergebnis auf die eigene Angst zu attribuieren. Dass die erste Gruppe nicht ein noch höheres Maß an Angst berichtete erklärten die Autoren damit, dass diese „funktionale" Form des Berichts von Angst durch den expliziten Hinweis eher gehemmt worden sein könnte. Interessant ist der Befund, dass die zweite Gruppe im Selbstbericht angab, sich am wenigsten im Test angestrengt zu haben. Die Autoren folgerten, dass bei hoch Testängstlichen der Bericht von Angst eine selbstwertschützende Funktion hat. Die Angabe einer geringen Anstrengung stellte eine weitere Strategie dar, die gewählt wurde, wenn (Test)Angst nicht als regulative „Option" zur Verfügung stand (Smith et al., 1982).

Hirt, Deppe und Gordon (1991) berichten ähnliche Ergebnisse bei einer studentischen Stichprobe (N = 230), die hohe oder niedrige Ausprägungen auf einer Skala zum dispositionellen self-handicapping aufwies. Die Probanden erhielten unterschiedliche Instruktionen zu einem Test und mussten unmittelbar danach einen Stressfragebogen bearbeiten, in dem das Stresserleben der letzten zwei Wochen abgefragt wurde. Ähnlich wie bei Smith et al. (1982) wurde den Probanden in der Instruktion des Tests mitgeteilt, ob Stress sich negativ auf die Ergebnisse im Test auswirkt oder nicht. Es wurde ein erwarteter Interaktionseffekt festgestellt: Probanden mit hoher Neigung zum self-handicapping berichteten deutlich mehr Stress als jene mit niedriger, wenn ihnen Stress als relevante Einflussgröße dargestellt wurde. Wurde in der Instruktion der Effekt von Stress negiert, zeigte sich kein Unterschied im berichteten Stress in Abhängigkeit von der Neigung zum self-handicapping (Hirt et al., 1991). Snyder, Smith, Augelli und Ingram (1985) fanden ähnliche Effekte bei sozial ängstlichen Probanden in Bezug auf die in einer Testsituation berichtete soziale Angst und Schüchternheit.

Auf Basis der Selbstwerttheorie der Leistungsmotivation argumentierte Thompson (1996), dass ihren Selbstwert schützende Personen („self-worth protective") in Leistungssituationen auch die eigene Anstrengung reduzieren, um somit eine Attribution eines schlechten Ergebnisses auf die eigenen Fähigkeiten zu vereiteln (Thompson, 1996). Thompson und Dinnel (2003) entwickelten die Self-Worth Protection Scale (SWPS) zur Erfassung der entsprechenden Disposition zum

[18] Eigene Bezeichnung zur Abkürzung

Selbstwertschutz (die Datengrundlage bildeten zwei studentische Stichproben, N = 243 sowie 411). Neben einer Skala zu Zweifeln an den eigenen Fähigkeiten (ability doubts) erfasst diese mit einer zweiten Skala die Bedeutsamkeit von Leistung für den eigenen Selbstwert (importance of ability). Eine dritte Skala erfasst die Tendenz, eher Aufgaben niedriger Schwierigkeit zu wählen bzw. solche, in denen ein Scheitern unwahrscheinlich ist (avoidance orientation). Die signifikant positiven Korrelationen dieser Skalen mit Furcht vor Misserfolg (r = .41, .23 bzw. .32; zweite Stichprobe) sind Indikatoren für die Nähe dieses Konstrukts zum Konstrukt Testängstlichkeit.

Enge theoretische Verbindung zum Selbstwertschutz weist das Konzept der Selbstwertkontingenz auf. Crocker und Wolfe (2001) definieren diese folgendermaßen: „A contingency of self-worth is a domain or category of outcomes on which a person has staked his or her self-esteem, so that person's view of his or her value or worth depends on perceived successes or failures or adherence to self-standards in that domain" (S. 594). Deci und Ryan (1995) kontrastieren einen wahren, stabilen von einem kontingenten Selbstwert, welcher stark davon abhängig ist, inwieweit soziale oder intraindividuelle Erwartungen erfüllt werden. Hat eine Person einen hohen kontingenten Selbstwert in Bezug auf ihre Leistungsfähigkeit, dann steht und fällt ihr positives Selbstbild mit dem Erfolg oder Misserfolg in Leistungssituationen. Ein „wahrer" (und hoher) Selbstwert ist demgegenüber weniger volatil und in diesem Sinne sicherer und stabiler (Deci & Ryan, 1995). Personen mit einem in diesem Sinne kontingenten Selbstwert beschäftigen sich in ausgeprägtem Maße damit, wie andere Personen sie bewerten und ob sie ihre jeweiligen Erwartungen und Standards erfüllen oder nicht (Kernis, 2003). Aus diesem Grunde ist ein hoher, kontingenter Selbstwert durch eine gewisse Fragilität gekennzeichnet (Deci & Ryan, 1995; Kernis, 2003).

Crocker und Knight (2005) gehen hingegen nicht davon aus, dass sich Personen in dem Ausmaß ihrer Selbstwertkontingenz unterscheiden. Sie nehmen an, dass die Kontingenz des Selbstwerts die Regel ist, sich Personen jedoch darin unterscheiden, in welchem Bereich sie kontingent sind: so dürfte dies für manche Personen die physische Attraktivität sein, für andere die berufliche Leistungsfähigkeit und für wiederum andere Personen die Beherrschung einer bestimmten Sportart. Ein von Crocker, Luhtanen, Cooper und Bouvrette (2003) entwickelter Fragebogen zur Erfassung der Selbstwertkontingenz enthält dementsprechend sieben verschiedene Bereiche, in denen der Selbstwert kontingent sein kann, u. a. die Wertschätzung durch andere, die körperliche Attraktivität, der Wettbewerb mit anderen Personen, akademische Kompetenzen oder auch die Einhaltung moralischer Standards. Bislang bezogen sich die Ausführungen auf einen hohen, kontingenten Selbstwert. Es stellt sich jedoch die Frage, wie die Situation in Bezug auf testängstliche Personen ist, welche einen eher niedrigen Selbstwert haben. Theoretisch wird davon ausgegangen, dass es neben dem vergleichsweise stabilen, dispositionellen Selbwert (trait) auch einen situativen Selbstwert (state) gibt, der um den stabilen Selbstwert oszilliert (Crocker & Wolfe, 2001). Crocker und Knight (2005) betonen nun, dass – unabhängig von der Höhe des (dispositionellen oder auch

trait) Selbstwerts – eine Senkung bzw. Steigerung des (momentanen) Selbstwerts mit negativen bzw. positiven Emotionen verbunden ist. Ein in diesem Sinne fragiler und schwankender Selbstwert ist mit dem Erleben von Stress und Angst verknüpft (Crocker & Knight, 2005; Crocker & Wolfe, 2001; Deci & Ryan, 1995). Personen mit einem kontingenten Selbstwert sind „anxiously focused on one's own agenda, whether that agenda is being feminine, famous, fashionable, fabulously wealthy, or far out" (Deci & Ryan, 1995, S. 32), wodurch Selbstwertkontingenz starke Parallelen zur Testängstlichkeit aufweist.

Lawrence und Williams (2013) untersuchten in einem experimentellen Design das Zusammenspiel von akademischer Selbstwertkontingenz, state-Testangst und Leistung. Eine studentische Stichprobe (N = 91) bearbeitete in einer evaluativen und nonevaluativen Bedingung einen Test. Es zeigte sich, dass in der evaluativen Bedingung ein indirekter (negativer) Effekt der Selbstwertkontingenz auf die Leistung auftrat, welcher durch die nach der Testinstruktion erfasste state-Testangst mediiert wurde. Höhere Selbstwertkontingenz ging also mit einer schlechteren Leistung einher, was durch erhöhte Testangst während der Testung erklärt werden konnte.

Deci und Ryan (1995) nehmen nun in ihrer Konzeption der Selbstwertkontingenz an, dass Personen mit hohem, kontingenten Selbstwert danach streben, ihren hohen Selbstwert aufrecht zu erhalten. Diese Personen „will use whatever means are available to match the standards" (Deci & Ryan, 1995, S. 32). Aus diesem Grund liegt es nahe, dass besonders Personen mit hoher Selbstwertkontingenz self-handicapping aufweisen, um eine Schädigung ihres Selbstwerts zu verhindern (Zuckerman & Tsai, 2005).

1.2.1.3.2 Implikationen und offene Fragen

Die Moderation der Angst-Leistungs-Relation durch den Messzeitpunkt der Testangst hat unmittelbare Relevanz für die Kausalfrage in der Angst-Leistungs-Relation. Wie bereits ausgeführt wurde, wurde im Sinne der Defizitperspektive die höhere Korrelation von state-Testangst post (gegenüber prä) mit Leistung so interpretiert, dass (schlechte) Leistung zu (hoher) Testangst führt und nicht umgekehrt. Nicht die objektive, sondern die subjektive Leistung nimmt dabei die Schlüsselrolle in der Determination der erlebten Testangst ein (siehe Abbildung 3): „A person who has failed a test, but believes he/she has done well will not become upset." (Sarason et al., 1990, S. 5). Somit müssten also Zusammenhänge von objektiver Leistung und state-Testangst post durch die subjektive Leistung mediiert werden.

Zentral ist in diesem Kontext die Überlegung, dass state-Testangst nach einem Test unter dem Eindruck des Tests steht. Theorien zur Selbstwertregulation lassen nun vermuten, dass dabei der Bericht von Testangst Ausdruck von Regulationsprozessen sein kann. Testangst würde in diesem

Sinne als eine Form von self-handicapping interpretiert werden: die nach dem Test berichtete Testangst ermöglicht eine von den eigenen Fähigkeiten *unabhängige* Attribution eines subjektiv schlechten Ergebnisses und somit den Schutz des Selbstwerts (im Sinne des Gedanken: „Ich hatte einfach einen Blackout – wäre ich nicht so nervös gewesen, hätte ich besser abgeschnitten."). Wie gezeigt wurde, legen eine Reihe von Untersuchungen nahe, dass derartige Prozesse tatsächlich ablaufen, wenn Testangst erhoben wird. Bisher unbeantwortet ist jedoch die Frage, wann und in welchem Ausmaß dies geschieht. Im Extremfall könnte argumentiert werden, dass die retrospektive Erfassung von Testangst nach einem Test inhaltlich wertlos ist, da nicht das Erleben während des Tests, sondern die individuellen Regulationsneigungen und -bemühungen einer Person erfasst werden. Wenn der Bericht von Testangst Ausdruck von Regulationsprozessen ist, sollte Testangst durch die subjektive Leistung determiniert sein, denn nur bei einem subjektiv schlechten Ergebnis würde man demnach hohe Testangst berichten. Damit können aus der Beziehung zwischen subjektiver Leistung und nach dem Test berichteter Testangst Erkenntnisse über etwaige regulative Prozesse gewonnen werden.

Nimmt man nun aus einer Regulationsperspektive an, dass der Bericht von Testangst nach einem Test dem Selbstwertschutz dient, dann sollte das umso mehr der Fall sein, je höher die Selbstwertkontingenz ist, also je empfindlicher eine Person (genauer: ihr Selbstwert) für einen Misserfolg ist. Der Zusammenhang von subjektiver Leistung und state-Testangst post sollte also durch die Selbstwertkontingenz moderiert werden. Zwar haben hoch Testängstliche einen eher niedrigen Selbstwert, doch erlaubt die Konzeption der Selbstwertkontingenz nach Crocker und Kollegen (z. B. Crocker & Knight, 2005) die Annahme, dass auch Personen mit niedrigem Selbstwert in einem regulativen Sinne anstreben, dass ihr situativer Selbstwert ansteigt bzw. nicht abnimmt.

Die Betrachtung des Berichts von Testangst nach einem Test als Ausdruck von self-handicapping ist an eine Voraussetzung geknüpft: eine Person verspürt die subjektive „Notwendigkeit", die Attribution eines Ergebnisses auf die eigenen Fähigkeiten zu unterbinden. Dies ist nur dann überhaupt erforderlich, wenn eine Person grundsätzlich davon ausgeht, dass eine Prüfung oder ein Test etwas über die eigenen Fähigkeiten aussagt. Die subjektive Leistung sollte also umso stärker mit der Testangst nach dem Test zusammenhängen, je valider der Test eingeschätzt wird. Aus regulativer Sicht lässt sich dieser Gedanke fortsetzen: die subjektive Abwertung der Aussagekraft eines Tests ermöglicht bei einem schlechten Testergebnis den Schutz des Selbstwerts – in einem (subjektiv) unsinnigen Test zu versagen gibt wenig Anlass zu Besorgnis und Selbstzweifeln.

Darüber hinaus stellt Testangst natürlich nur *eine* mögliche Variante von self-handicapping beim Selbstbericht dar. In Anlehnung an die Ergebnisse von Smith et al. (1982) sollte der Zusammenhang von subjektiver Leistung und state-Testangst post umso geringer sein, je stärker eine Person andere Erklärungen für eine (schlechte) Leistung aufführt, wie z. B. mangelnde Anstrengung.

Diese Überlegungen stellen Forscher und Praktiker vor die Frage, wann nun Testangst erfasst werden sollte. Aufgrund der wahrscheinlichen Konfundierung mit Regulationsprozessen nach dem Test bietet sich (vermeintlich) die Lösung an, Testangst vor einem Test anstatt danach zu erfassen. Dabei stellt sich jedoch die Frage, welche Effekte die Erfassung von Testangst vor dem Test auf die Leistung und das Testerleben hat.

Diese Konstellation an offenen Fragen und Vermutungen sollte in einem Design untersucht werden, in dem state-Testangst sowohl vor als auch nach einem Test erfasst wird. Zur Prüfung, ob die Erfassung der state-Testangst vor dem Test sich auf die Leistung oder andere abhängige Variablen auswirkt, sollte das Design zudem eine Kontrollbedingung vorsehen, in der die Testangst nicht vor dem Test erfasst wird. Zu diesem Zweck wurde Studie 1 durchgeführt.

1.2.1.4 Fazit zur Interferenz- und Defizitperspektive

Vor dem Hintergrund der empirischen Evidenz für beide Perspektiven sollte das Ziel der Forschung zur Testängstlichkeit nicht primär darin bestehen, die Entscheidung zwischen einem der beiden Modelle zu treffen. Zielführender ist es, die Geltungsbereiche beider Perspektiven auszuloten und festzustellen, unter welchen Bedingungen welche Prozesse maßgeblich sind (siehe auch Zeidner, 1998). Eine Integration beider Ansätze ermöglicht beispielsweise die bereits ausführlich dargelegte S-REF Theorie (siehe Abschnitt 1.1.3). Demnach führt die selbstbezogene Informationsverarbeitung zu Interferenz bei der Aufgabenbearbeitung, während maladaptive Bewältigungsstrategien, wie z. B. Vermeidungsverhalten, langfristig dazu führen, dass die Entwicklung von Fähigkeiten und Fertigkeiten stagniert und sich somit Defizite herausbilden oder verschärfen (Zeidner & Matthews, 2007). Interferenz- und Defizitprozesse schließen sich also nicht aus.

1.2.2 Determinanten der Relation von Testängstlichkeit, Testangst und Leistung

Die Forschung hat sich intensiv damit befasst, welche Variablen die Beziehung von Angst und Leistung beeinflussen, sprich moderieren. Dieser Forschungsstrang überschneidet sich deutlich mit der Forschung zu den situativen Determinanten von Testängstlichkeit bzw. Testangst, die in Abschnitt 1.1.2 beschrieben wurden. Hier soll nun der Fokus auf Befunde gelegt werden, die sich spziell auf die Relation von Testängstlichkeit, Testangst und Leistung beziehen. Die Ausführungen sind folgendermaßen gegliedert: zunächst werden Merkmale der Instruktion, der Aufgabenschwierigkeit und der Aufgabendarbietung sowie deren Bedeutung dargestellt. Im Anschluss wird auf den „stereotype threat“ eingegangen, der eine besondere Kombination von Aufgabeninstruktion und –inhalt darstellt und Kern der Fragestellung von Studie 2 bildete.

1.2.2.1 Merkmale von Aufgaben: Instruktion, Schwierigkeit und Darbietung

Bereits in Abschnitt 1.1.2 wurde darauf eingegangen, dass Tests auf unterschiedliche Weise instruiert werden können. Die Variable, welche durch die Instruktion direkt beeinflusst wird, ist die Testatmosphäre (Zeidner, 1998). Insbesondere I. G. Sarason und Kollegen haben sich intensiv mit dieser Thematik befasst und die Leistung dispositionell hoch und niedrig Testängstlicher unter verschiedenen Testinstruktionen verglichen (z. B. Sarason, 1956; Sarason, 1958a; Sarason & Stoops, 1978). Zeidner (1998) bezeichnet es als robustes Befundmuster, dass evaluative Instruktionen, verglichen mit nonevaluativen, neutralen Instruktionen, mit verbesserten Leistungen bei niedrig Testängstlichen, jedoch schlechteren Leistungen bei hoch Testängstlichen einhergehen. Beispielhaft sind hierzu die Ergebnisse von Sarason et al. (1986) dargestellt (siehe Abbildung 4). Hierbei wurde die Leistung von hoch, mittel und niedrig testängstlichen studentischen Probanden (N = 302) in einem Wissenstest unter neutraler und evaluativer Instruktion verglichen.

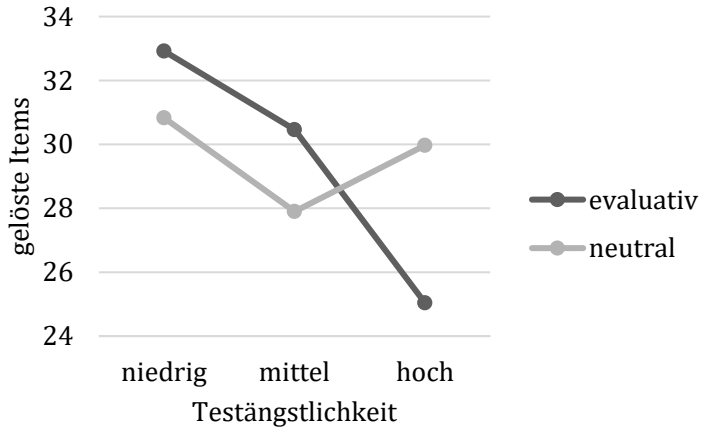

Abbildung 4: Interaktion von Testängstlichkeit (auf Basis der Worry-Skala des RTT) und Testinstruktion bei der abhängigen Variable gelöste Items (Sarason et al., 1986, Studie 2B).

Die gängige Erklärung für diesen Interaktionseffekt lautet, dass niedrig Testängstliche bei nichtevaluativer Instruktionen weniger Motivation für die Aufgabenbearbeitung verspüren, wohingegen hoch Testängstliche vom Fehlen des Bewertungsdrucks profitieren (Deffenbacher, 1978; Wine, 1971). Nach der Metaanalyse von Hembree (1988) ist das genannte Muster am eindeutigsten bei lebensälteren Stichprobengruppen: hoch testängstliche Schüler bzw. Studenten (postsecondary) waren schlechter bei evaluativen Instruktionen (ego-involving), d = -.26 (k = 17, N = 640), niedrig testängstliche Schüler (10. bis 12. Klasse, postsecondary) waren besser bei evaluativen Instruktionen, d = .29 (k = 22, N = 974)[19]. Wie bereits in Abschnitt 1.2.1.1 dargelegt wurde, ist dieser Interaktionseffekt eine wichtige Grundlage für die Annahme von Interferenzprozessen.

[19] Bei der Betrachtung dieser Interaktion muss indes berücksichtigt werden, dass Hembree (1988) bei niedrig testängstlichen Stichproben jüngeren Alters (4-6. sowie 8. Klasse, k = 12, N = 980) zwar auch den erwarteten Effekt fand, dieser jedoch sehr heterogen war. Bei hoch Testängstlichen jüngeren Alters (4. bis 6., 8. und 10. bis 12. Klasse, k = 16, N = 1.120) fand sich der Effekt nicht.

Zeidner (1998) kritisiert allerdings, dass in den meisten Studien beobachtete Unterschiede in Abhängigkeit der dispositionellen Testängstlichkeit mit dem (Angst-)Erleben während der Testsituation erklärt werden, ohne dass besagte Prozesse direkt erfasst wurden. Dies gilt insbesondere für die Arbeiten von I. G. Sarason, wobei es aber auch Ausnahmen gibt (z. B. Sarason, 1981; Sarason & Stoops, 1978).

Ein weiteres wichtiges Charakteristikum von Aufgaben ist die Aufgabenschwierigkeit oder Aufgabenkomplexität (Zeidner, 1998). Einen Zugang zum Effekt der Schwierigkeit liefert die Betrachtung der subjektiven Schwierigkeit. Hembree (1988) berichtet einen Leistungsunterschied zwischen hoch und niedrig Testängstlichen bei subjektiv schwierigen Aufgaben von d = -.45 zuungunsten der hoch Testängstlichen (k = 13, N = 874). Kein Unterschied zeigt sich bei als leicht empfundenen Tests, d = -.07 (k = 12, N = 849). Ein Beispiel für das Zusammenspiel von subjektiver Aufgabenschwierigkeit, Testangst und Leistung ist die Untersuchung von Hong (1998). Bei einer studentischen Stichprobe (N = 208) wurde unmittelbar vor einer Statistikprüfung die Testangst und die erwartete Schwierigkeit der Prüfung erfasst. Die subjektive Schwierigkeit hatte keinen direkten, jedoch einen indirekten Effekt auf die Leistung, mediiert durch die Besorgtheit (state). Über die subjektive Schwierigkeit hinausgehend definieren Eysenck und Calvo (1992) die Schwierigkeit ressourcentheoretisch. Die Schwierigkeit einer Aufgabe lässt sich demnach aus der Menge an kognitiven Ressourcen ableiten, welche diese erfordert, womit sich auch der Interaktionseffekt erklären lässt: „Given that anxiety reduces working memory capacity, anxiety will interact with task demands, with the detrimental effect of anxiety increasing directly with the demands that tasks place on the capacity of working memory“ (Eysenck, 1982, 122 f.). Darüber hinaus erleben Personen bei schwierigen Aufgaben eher Misserfolg, was wiederum die in der Situation erlebte Angst verstärkt und mit zu dem Effekt beiträgt (Eysenck, 1982). Allerdings muss bei dieser Betrachtung berücksichtigt werden, dass die Definition von Aufgaben- bzw. Testschwierigkeit (zumal die subjektiv eingeschätzte) vielschichtig ist: neben der eigentlichen Aufgabenkomplexität bestimmen unter anderem auch die individuellen Fähigkeiten die Schwierigkeit (Heinrich & Spielberger, 1982). Darüber hinaus spielen Variablen wie der getestete Inhaltsbereich (siehe z. B. Sparfeldt et al., 2005) und das individuelle Fähigkeitsselbstkonzept sicherlich auch eine Rolle.

Neben den beschriebenen Parametern Aufgabeninstruktion und –schwierigkeit wurden noch zahlreiche weitere Merkmale der formalen Darbietungsweise von Aufgaben bzw. Items untersucht, welche sich auf das Angsterleben und / oder die erbrachte Leistung auswirken. Hierzu gehören unter anderem die Entscheidungskontrolle über die Testaufgaben oder deren Darbietungsweise (Keinan & Zeidner, 1987; Wise, Roos, Plake & Nebelsick-Gullett, 1994), Zeitdruck (Morris & Liebert, 1969), aber auch onlinebasiertes Testen (Stowell & Bennett, 2010) sowie adaptives Testen gegenüber festgelegter Itemreihenfolge (Ortner & Caspers, 2011). Es ist gleichwohl zu beachten, dass es sich dabei nicht notwendigerweise um Variationen handelt, die ausschließlich bei

hoch Testängstlichen Auswirkungen haben. Beispielsweise berichten Kellogg, Hopko und Ashcraft (1999), dass Zeitdruck sowohl bei hoch und niedrig mathematikängstlichen Studierenden (N = 30) zu schlechteren Leistungen in einem Mathematiktest führt. Die Vermutung liegt nahe, dass die meisten der behandelten Aufgabenparameter letztlich über denselben Mechanismus das Erleben und Verhalten im Test beeinflussen. Sowohl Testatmosphäre, Aufgabenkomplexität als auch unterschiedliche Formen der Aufgabendarbietung bestimmen das „Stresspotenzial" eines Tests. Hierbei spielt sicherlich auch der getestete Bereich bzw. die getestete Domäne oder auch die Prüfungsform eine Rolle. Im Sinne des transaktionalen Stressmodells von Lazarus sollte sich also das Vorhandensein oder aber die Abwesenheit von Reizen, die den Leistungscharakter der Situation indizieren, auf die primäre Bewertung auswirken. Weitere Variationen (z. B. Entscheidungskontrolle über die Testaufgaben) wirken sich auf die sekundäre Bewertung aus. Manche Variationen (z. B. Zeitdruck, Domäne) dürften sowohl die primäre als auch die sekundäre Bewertung beeinflussen. In einem metaphorischen Sinne lassen sich diese Manipulationen an den Aufgaben bzw. am Test als „Stellschrauben" auffassen – je nach Position werden verschieden Personen in Abhängigkeit dispositioneller Merkmale unterschiedlich reagieren. Die sich daraus ergebende große Vielfalt an Kombinations- und Interaktionsmöglichkeiten soll an dieser Stelle nicht weiter vertieft werden. Stattdessen soll der Fokus auf eine spezifische Kombination dieser Parameter gerichtet werden.

Ein großer Teil der bis dato behandelten Befunde basiert auf folgendem experimentellen Paradigma: bestimmte Aspekte des Tests, z. B. der Bewertungscharakter der Instruktion, werden manipuliert, um im Anschluss zu prüfen, ob in Abhängigkeit von Gruppenzugehörigkeiten Effekte auftreten. Besagte Gruppen werden häufig aus den jeweiligen Merkmalsverteilungen abgeleitet (z. B. via Mediansplits). Die An- und Abwesenheit von Gruppenunterschieden in der zentralen abhängigen Variable (Leistung) ist dabei ein Indiz für kausale Erklärungen, wie sie in der Interferenzperspektive formuliert sind. Dieses Prinzip weist starke Analogien zu anderen Forschungsbereichen auf. Eine besondere Parallele ergibt sich zu einem in neuerer Zeit intensiv beforschten Phänomen, welches in Grundzügen eine Kombination von manipulierten Testinstruktionen mit einer spezifischen Aufgabendomäne darstellt, nämlich dem sog. „stereotype threat" (STT). Grundgedanke ist auch hier, dass sich bei bestimmten Instruktionen Leistungsunterschiede zwischen bestimmten Gruppen (z. B. Frauen und Männern) in einem bestimmten Leistungsbereich (z. B. numerisches Denken) zeigen bzw. vergrößern. Kern der folgenden Überlegungen ist die Frage, inwiefern Testangst zur Erklärung dieses Phänomens beiträgt, ob also der STT in Form von Interferenzprozessen verstanden werden kann. Um die Überschaubarkeit der Ausführungen zu diesem wachsenden Forschungsbereich zu gewährleisten, wird nach einer Begriffsbestimmung auf die

wichtigsten Befunde zum STT eingegangen, mit einem Fokus auf metaanalytisch gewonnenen Erkenntnissen. In einem dritten Schritt wird schließlich die Verbindung zwischen STT und Testangst hergestellt, welche auch das zentrale Thema von Studie 2 war.

1.2.2.2 Kombination von Aufgabendomäne und –instruktion: „stereotype threat"

Ein Stereotyp ist eine „kognitive Struktur, die unser Wissen, unsere Überzeugungen und Erwartungen über eine soziale Gruppe von Menschen enthält." (Pendry, 2014, S. 111). Stereotype beinhalten folglich auch Annahmen über die Fähigkeiten einer Gruppe, wobei Stereotype sowohl die allgemeine Höhe, aber auch deren bereichsspezifische Ausprägung zum Inhalt haben können (z. B. „Informatiker sind mathematisch begabt, aber sozial wenig kompetent."). Wie sich negative Stereotype über eine bestimmte Gruppe auf das Erleben und Verhalten von Personen, die Mitglied in dieser Gruppe sind, auswirken, ist Kernfrage des STT. Wesentlicher Ausgangspunkt dieses Forschungsstranges war eine vielfach zitierte, „klassische" Studie von Steele und Aronson (1995), die sich mit dem Stereotyp befasste, dass Afroamerikaner über eine niedrigere akademische Begabung verfügen als Weiße. In mehreren Experimenten konnten die Autoren demonstrieren, dass Leistungsunterschiede in einem verbalen Fähigkeitstests zwischen afroamerikanischen und weißen Probanden auftraten, wenn bestimmte Instruktionen vorgelegt wurden: afroamerikanische Probanden erbrachten unter einer evaluativen Bedingung schlechtere Leistungen als weiße Probanden, jedoch vergleichbare Leistungen unter nicht-evaluativer Instruktion (Studie 1 & 2 bei Steele und Aronson; N = 114 bzw. 20). Die Autoren konnten darüber hinaus zeigen, dass eine evaluative, verglichen mit einer neutralen Instruktion, bei afroamerikanischen Probanden mit einer stärkeren Verfügbarkeit von Stereotypen über Afroamerikaner einhergeht (Studie 3, N = 68). Auch bei neutraler, nicht diagnostischer Instruktion ergaben sich Gruppenunterschiede zuungunsten der afroamerikanischen Probanden, wenn in einem Fragebogen vor dem Test die eigene Ethnie aktiviert wurde (durch eine Frage nach der Ethnie; Studie 4; N = 44). Die ungünstigen Auswirkungen negativer Stereotype über eine Gruppe auf Mitglieder dieser Gruppe werden als „stereotype threat" (STT) bezeichnet: „It is the social-psychological threat that arises when one is in a situation or doing something for which a negative stereotype about one's group applies." (Steele, 1997, S. 614). Die in derartigen Situationen virulente Bedrohung besteht daraus, dass man ein negatives Stereotyp als zutreffend bestätigen könnte, was sich dann wiederum in einer tatsächlichen Leistungsbeeinträchtigung niederschlagen kann. Bearbeiten beispielsweise eine Frau und ein Mann einen Mathematiktest, so existiert lediglich für die Frau die Gefahr, mit einer schlechten Leistung das Stereotyp „Frauen sind schlecht in Mathematik" zu bestätigen (Steele & Aronson, 1995).

Der STT wurde bei verschiedenen Gruppen untersucht, wobei jeweils Domänen im Fokus stehen, bezüglich derer negative Stereotype über die entsprechende Gruppe existieren. Dies betrifft beispielsweise lebensältere Menschen und deren kognitive und körperliche Leistungsfähigkeit (Lamont, Swift & Abrams, 2015) oder auch Einwanderer und deren akademische Leistungen (Appel, Weber & Kronberger, 2015). Die meisten Befunde scheint es jedoch in Bezug auf die kognitiven Fähigkeiten von Frauen und ethnischen Minderheiten zu geben. Die Forschung zum STT bei Frauen bezieht sich auf das Stereotyp, dass diese in der Domäne Mathematik niedrigere Fähigkeiten besitzen als Männer. Dieser Fall des STT steht bei der vorliegenden Arbeit im Fokus.

Die Erforschung des STT ist im Kontext von gefundenen Unterschieden zwischen sozialen Gruppen bzw. den Geschlechtern zu sehen. Diese Unterschiede besitzen eine hohe gesellschaftliche Relevanz, was der Grund dafür sein dürfte, dass sich in den letzten zwei Jahrzehnten eine starke Forschungsaktivität zum STT entfaltet hat. In einer Metaanalyse über insgesamt 259 Studien stellten Hyde, Fennema und Lamon (1990) fest, dass es zwar über alle betrachteten Studien hinweg keinen nennenswerten Geschlechtsunterschied in der Mathematikleistung gab. Bei Betrachtung unterschiedlicher Altersgruppen traten jedoch mit zunehmendem Alter wachsende Geschlechtsunterschiede auf (d = .29, .41 bzw. .59 in den Altersgruppen 15-18 Jahre, 19-25 Jahre bzw. 26 Jahre und älter). In den USA zeigt sich seit Jahrzehnten ein deutlicher Leistungsvorsprung von Männern in den Mathematikergebnissen des SAT (Halpern et al., 2007). Geschlechtsbezogene Unterschiede erfahren in den letzten Jahren auch in Deutschland zunehmendes Interesse, was mit der starken Unterrepräsentation von Frauen in Studiengängen mit einem substanziellen Anteil an Mathematik, den MINT-Fächern (Leszczensky, Cordes, Kerst, Meister & Wespel, 2013), verbunden ist. Steele und Aronson (1995) stellen die Vermutung auf, dass Bedingungen, welche den STT auslösen, in vielen Prüfungs- und Testsituationen in Bildungssystemen inhärent sind, womit der STT womöglich zur Erklärung von gefundenen Geschlechtsunterschieden beiträgt (siehe z. B. Halpern et al., 2007). Der STT kann nach der theoretischen Vorstellung auch langfristig dazu führen, dass Mitglieder der stereotypisierten Gruppe sich „deidentifizieren“, also die Bedeutung der jeweiligen Domäne für ihr Selbstkonzept herabsetzen, um so den Selbstwert vor der Bedrohung zu bewahren (Steele, 1997).

Steele (1997) formulierte einige grundlegende theoretische Annahmen zu Auslösung und Auswirkung des STT. Einer dieser Aspekte wurde bereits deutlich, nämlich dass der STT in allen denkbaren Gruppen auftreten kann, insofern es verbreitete, negative Stereotype über diese Gruppe gibt (Mitglieder dieser Gruppe sind im Folgenden als „targets“ bezeichnet, Nichtmitglieder als „nontargets“). Darüber hinaus ist nach Steele (1997) weder die subjektive Überzeugung über die allgemeine Gültigkeit eines Stereotyps, noch die Annahme der Gültigkeit des Stereotyps bei der eigenen Person notwendige Voraussetzung für das Auftreten des STT. Es ist also unerheblich, ob eine Frau tatsächlich denkt, dass Frauen allgemein in Mathematik weniger begabt sind als Männer.

Auch muss sie sich selbst nicht für weniger begabt in Mathematik halten – selbstverständlich aber muss sie das Stereotyp kennen (Steele & Aronson, 1995).

Nguyen und Ryan (2008) führten eine Metaanalyse zum STT bei Frauen und ethnischen Minderheiten durch. Sie gliederten dabei auch die große Vielfalt an Untersuchungsdesigns, mit denen der STT empirisch untersucht wurde. Demnach lassen sich drei Klassen von Auslösebedingungen unterscheiden, mit denen eine „Bedrohung" empirisch induziert wird. Eine *offenkundige Aktivierung* besteht daraus, dass Probanden explizit vor einem Test mitgeteilt wird, dass targets in dem anstehenden Test und / oder der getesteten Domäne allgemein schwächer abschneiden als nontargets. In der *moderat expliziten Aktivierung* wird zwar auf Gruppenunterschiede zwischen targets und nontargets hingewiesen, jedoch wird nicht ausgeführt, welche Gruppe nun überlegen bzw. unterlegen ist. Offenkundige sowie moderat explizite Aktivierungen funktionieren nach Nguyen und Ryan (2008) auf einem bewussten Wege. Die dritte Form ist die *indirekte und subtile Aktivierung*, bei der keine Hinweise auf Gruppenunterschiede gegeben werden, aber „the context of tests, test takers' subgroup membership, or test taking experience is manipulated" (S. 1316). So kann beispielsweise das eigene Geschlecht durch eine entsprechende Frage in einem vorgeschalteten Fragebogen aktiviert werden. Auch eine Darstellung der getesteten Domäne und der diagnostischen Aussagekraft des Ergebnisses gehören hierzu (Nguyen & Ryan, 2008). Diese dritte Variante weist starke Parallelen zum Paradigma von I. G. Sarason zur Manipulation der Testatmosphäre auf (siehe Abschnitt 1.1.2.). Nguyen und Ryan (2008) betonen, dass sich offenkundige und moderat explizite Aktivierungen in der praktischen Anwendung von Tests eher selten finden dürften, die subtile „Erzeugung" des STT hingegen durchaus.

Nguyen und Ryan (2008) inkludierten nur experimentelle Studien. Für den STT bezüglich der Mathematikleistung bei Frauen (STT- vs. Kontrollbedingungen) ermittelten die Autoren einen kleinen Effekt von d = -.21 (k = 72, N = 4.935). Ein größerer Effekt zeigte sich bei subtiler Aktivierung mit d = -.24 (k = 32, N = 2.564) gegenüber moderater mit d = -.18 (k = 20, N = 1.138) und offenkundiger Aktivierung mit d = -.17 (k = 22, N = 1.279)[20]. Die theoretische Annahme, dass der STT umso deutlicher wird, je bedeutsamer die jeweilige Domäne für das Selbstkonzept einer Person ist („domain identification", kurz DI; Steele, 1997), wurde nicht bestätigt. Stattdessen zeigte sich ein stärkerer Effekt bei Frauen mit mittlerer DI (d = -.52, k = 6, N = 212) gegenüber jenen mit hoher DI (d = -.29, k = 9, N = 380). Interessant ist darüber hinaus der Vergleich von targets und nontargets. Auch unter Kontrollbedingungen zeigten Frauen schlechtere Mathematikleistungen als Männer, d = -.26 (k = 13, N = 1.803), jedoch vergrößerte sich dieser Unterschied, wenn beide Geschlechter unter einer STT-Bedingung getestet wurden, d = -.39 (k = 39, N = 3.330). Diese Befunde

[20] Es ist nicht auszuschließen, dass die von Nguyen und Ryan (2008) ermittelten Effektstärken verzerrt sind, da auch Studien von Diederik Stapel inkludiert wurden. Die zitierte Metaanalyse wurde vor Bekanntwerden der Datenfälschungen durch Diederik Stapel publiziert.

legen also nahe, dass Geschlechtsunterschiede in der Mathematikleistung, wenn nicht initial hervorgerufen, so zumindest durch den STT verstärkt werden könnten.

1.2.2.2.1 Mechanismen zur Erklärung des STT

Nguyen und Ryan (2008) schreiben zur Erklärung des STT: „One important limitation of the literature base is the lack of sufficient, successful studies on mediating mechanisms. That is, we do not know exactly what psychological processes stereotype-activating cues trigger" (S. 1330). Die zunehmende Forschung zum STT muss also noch die Frage beantworten, welche Prozesse den Effekt erklären. Schmader, Johns und Forbes (2008) postulieren in einer Literaturübersicht, dass das Auftreten des STT durch widersprüchliche Beziehungen zwischen drei verschiedenen, aktivierten Konzepten verursacht wird, die hier am Beispiel des STT bei Frauen beschrieben sind. Diese sind das Selbstkonzept, das Konzept der jeweiligen Fähigkeitsdomäne und das Konzept der eigenen Gruppe. Aufgrund des negativen Stereotyps über die mathematische Kompetenz von Frauen kommt es dazu, dass diese drei Konzepte („Ich bin eine gute, intelligente Person" – „Ich bin gut in Mathe" – „Ich bin eine Frau") miteinader in Konflikt geraten (Beispiele nach Rydell, McConnell & Beilock, 2009). Dies stellt eine Situation kognitiver Dissonanz sensu Festinger (1957) dar (Schmader & Beilock, 2012).

Schmader et al. (2008) nehmen an, dass die Leistungsbeeinträchtigungen beim STT auf drei verschiedene Prozesse zurückführbar sind. Dies sind erstens physiologische Stressreaktionen. Zweitens führt das angesprochene Ungleichgewicht dazu, dass die Aufmerksamkeit verstärkt auf die eigene Person sowie die eigene Leistung gerichtet wird. Dies ist eng mit dem Bestreben verbunden, das entsprechende Stereotyp *nicht* zu bestätigen, wodurch auch ein Gleichgewicht wiederhergestellt würde (Schmader & Beilock, 2012). Mit dieser Veränderung der Aufmerksamkeitsprozesse geht auch eine bevorzugte Verarbeitung bedrohlicher Reize einher (Schmader et al., 2008) (ähnlich dem Aufmersamkeitsbias, siehe Abschnitt 1.1.1.4). Drittens wird angenommen, dass targets unter dem STT dazu neigen, aversive Gedanken und Emotionen, die mit dem STT verbunden sind, zu unterdrücken (Schmader et al., 2008). Diese Suppression ist insofern regulatorisch notwendig, da beispielsweise Zweifel an den eigenen Fähigkeiten subjektiv als Bestätigung des negativen Stereotyps interpretiert werden könnten – eine Bewertung, die targets zu vermeiden und auszublenden suchen (Schmader & Beilock, 2012). Alle drei Prozesse wirken sich negativ auf die verfügbaren Arbeitsgedächtnisressourcen aus, was schließlich zu einer Leistungsminderung führt (Schmader et al., 2008). Demzufolge scheint das Arbeitsgedächtnis bzw. dessen verfügbare Ressourcen eine entscheidende Rolle bei der Erklärung von STT-Effekten zu spielen (Schmader & Johns, 2003).

Dieses Modell weist Parallelen zu der in Abschnitt 1.1.3 dargelegten S-REF Theorie auf. Das für die selbstbezogene Informationsverarbeitung zentrale Missverhältnis zwischen dem angestrebten und dem aktuell erlebten „self-status“ (Matthews & Wells, 1999) ähnelt dem Ungleichgewicht der aktivierten Konzepte bei Schmader et al. (2008). Beide Theorien sagen vorher, dass Personen[21] beim STT bzw. in einer Bewertungssituation ihre Aufmerksamkeit verstärkt auf die eigene Person und das eigene Erleben richten (vgl. hierzu auch die „self-preoccupation“ nach Sarason & Sarason, 1990). In beiden Modellen spielen Metakognitionen eine wichtige Rolle. Diese Ähnlichkeiten werfen auf einer spezifischeren Ebene die Frage auf, welche Parallelen es zwischen STT-Phänomenen und Testangst gibt.

1.2.2.2.2 Gemeinsamkeiten und Unterscheide von STT und Testangst

Zunächst ist offenkundig, dass das integrative Prozessmodell von Schmader und Kollegen ein Interferenzmodell ist. Aus dieser Perspektive weist der STT starke Parallelen zur Testangst auf, insofern es sich um einen „leistungshemmenden“ Prozess handelt, der für Leistungsunterschiede zwischen Gruppen verantwortlich ist. Angst bzw. Testangst wurde wiederholt als Erklärung von STT-Effekten aufgeführt (Pennington, Heim, Levy, Larkin & Pavlova, 2016). Die Parallelen und Unterschiede zwischen dem STT und Testangst sollen im Folgenden dargelegt werden.

Beide Phänomene haben eine starke soziale Komponente. Beim STT ist diese offensichtlich, da Stereotype Vorstellungen über die Eigenschaften einer sozialen Gruppe sind (Pendry, 2014). Auch Testangst bzw. Testängstlichkeit weist soziale Komponenten auf, was sich in deren Operationalisierungen niederschlägt. Putwain (2008a) verglich einschlägige Instrumente zur Erfassung von Testängstlichkeit bzw. Testangst nach den darin erfassten sorgenvollen Kognitionen. Sieben der neun betrachteten Verfahren enthielten Items, die Sorgen bezüglich der sozialen Bewertung thematisieren. Darüber hinaus weist das Erleben von Personen, die sich in einer STT-Situation befinden, Ähnlichkeiten zum Erleben von testängstlichen Personen auf. In Abschnitt 1.1.1.3 wurde beschrieben, dass Testängstlichkeit mit ungünstigen Annahmen über die eigenen Fähigkeiten verbunden ist. Einige Untersuchungen konnten zeigen, dass derartige Empfindungen auch bei targets in STT-Situationen auftreten. In der bereits zitierten Untersuchung von Steele und Aronson (1995) produzierten die afroamerikanischen Probanden in der evaluativen bzw. diagnostischen Bedingung in einer Wortvervollständigungsaufgabe deutlich mehr auf Selbstzweifel bezogene Wörter als weiße Probanden in derselben Bedingung und afroamerikanische Probanden in den Kontrollbedingungen (Studie 3). Die Befunde von Stangor, Carr und Kiang (1998) legen nahe, dass

[21] Beim STT die targets, im Kontext der S-REF Theorie dispositionell Testängstliche: „Such individuals may be prone to react to such situations [gem.: Bewertungssituationen] with the maladaptive metacognitive and coping processes described by the S-REF model” (Matthews, Hillyard & Campbell, 1999, S. 114)

sich der STT auch ungünstig auf Leistungserwartungen von targets auswirken kann. Weitere Studien konnten zeigen, dass der STT mit negativen Emotionen und Kognitionen verbunden ist. So gaben weibliche Probanden beim STT retrospektiv eine negativere Stimmung während des Tests (Keller & Dauenheimer, 2003) sowie im freien Bericht mehr negative Gedanken in Bezug auf die Domäne Mathematik an (z. B. „I am not good at math"; Cadinu, Maass, Rosabianca & Kiesner, 2005).

Schmader et al. (2008) postulieren, dass es drei wesentliche Unterschiede zwischen dem STT und Testangst gibt. Der erste Unterschied besteht darin, dass beim STT die Zugehörigkeit zu einer bestimmten Gruppe salient wird. Die negativen Ansichten über diese Gruppe müssen nicht zwingend geteilt werden (Steele, 1997), womit eine Person auch dann dem STT „zum Opfer fallen kann", wenn sie ihre Kompetenz als hoch einschätzt. Testängstliche hingegen schätzen ihre Fähigkeiten dispositionell eher schlecht ein.

Ein zweiter Unterschied erwächst aus der motivationalen Betrachtung. Es gibt empirische Hinweise darauf, dass ein mit einer offenkundigen Manipulation induzierter STT zu Reaktanzeffekten führen kann (Kray, Thompson & Galinsky, 2001). Diese Reaktion würde auch durch das beschriebene Modell von Schmader et al. (2008) vorhergesagt: ein Gleichgewicht kann wiederhergestellt werden, indem die negative Annahme über die Gruppe durch eine eigene, gute Leistung widerlegt wird („Frauen sind *nicht* schlecht in Mathe"). Schmader et al. (2008) argumentieren, dass Testangst demgegenüber eher ungünstige motivationale Effekte hat, so etwa im Sinne von Motivationsverlusten (siehe z. B. Hancock, 2001). Bereits dargestellt wurde, dass Testängstlichkeit eng mit Furcht vor Misserfolg sowie mit Vermeidungszielen verknüpft ist (siehe Abschnitt 1.1.1.3). Allerdings ist diese von Schmader et al. (2008) vertretene Abgrenzung aus meiner Sicht unscharf, da deren Modell ebenso annimmt, dass Personen eine Bestätigung des Stereotyps vermeiden wollen. Tatsächlich gibt es Befunde, nach denen sich targets beim STT eher Leistungsvermeidungsziele setzen (Smith, 2006). Diese ambivalente Konstellation von Annäherungs- („Ich will zeigen, dass Frauen gut in Mathe sind") und Vermeidungszielen („Ich will auf keinen Fall beweisen, dass Frauen schlecht in Mathe sind") erinnert an die Beziehungen von Testangst zu beiden Zielen.

Den dritten Unterschied sehen Schmader et al. (2008) darin, dass die beim STT auftretende Angst nicht in demselben Maß durch die Probanden berichtet wird wie das Erleben von Testangst. Hintergrund dieser Annahme sind Befunde, dass targets beim STT zur Unterdrückung ihres Angsterlebens neigen, was Ausdruck von Emotionsregulationsaktivitäten ist (Johns, Inzlicht & Schmader, 2008). In diesem Kontext führen Schmader et al. (2008) (unter Verweis auf Johns et al., 2008) folgende Abgrenzung auf: „stereotype threat can be induced through subtle cues that simultaneously impair performance but leave individuals unaware of (or unwilling) to acknowledge their resulting feelings of anxiety" (S. 350). Auf die plausible Problematik des Selbstberichts beim STT

wird an späterer Stelle erneut eingegangen. Demgegenüber ist jedoch zu beachten, dass STT-Manipulationen zuweilen inhaltsgleich mit Manipulationen sind, die Bewertungsstress evozieren sollen. Die von Nguyen und Ryan (2008) vorgeschlagene Kategorie der indirekten bzw. subtilen STT-Aktivierung beinhaltet konkret zwei Reizklassen: einerseits das Priming der Gruppenzugehörigkeit (Geschlecht) und andererseits die Betonung der diagnostischen Aussagekraft eines Tests. Betrachtet man letzteren Fall, so unterscheidet sich eine solche STT-Instruktion von einer in der Testängstlichkeitsforschung gängigen evaluativen Instruktion einzig und allein darin, dass der indizierte Leistungsbereich jener ist, bei dem die targets dem Stereotyp zufolge niedrige Fähigkeiten aufweisen (z. B. „mathematische Fähigkeiten“). Der theoretisch angenommenen Verschiedenheit liegen also in vielen Fällen beinahe identische Untersuchungsdesigns zugrunde (siehe hierzu Abschnitt 2.2).

1.2.2.2.3 Befunde zur Relevanz von Testangst für die Erklärung von STT-Effekten

Eine „Erklärung“ des STT (als unabhängiger Variable) kann statistisch so verstanden werden, dass ein (zu ermittelnder) Mediator den Effekt der Manipulation auf die Leistung (als abhängige Variable) vermittelt. Dieser Gedanke ist in Abbildung 5 dargestellt.

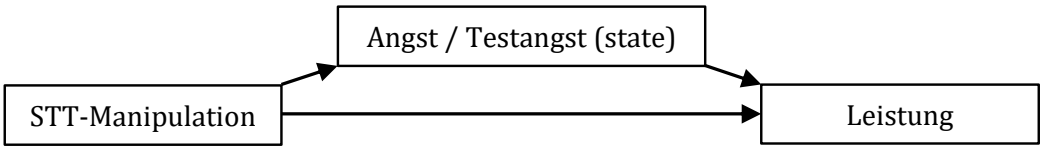

Abbildung 5: Schematische Darstellung einer vermuteten Mediation des STT-Effekts durch Angst bzw. Testangst (state)

Eine Reihe von Untersuchungen hat Angst bzw. Testangst als möglichen Erklärungsmechanismus für STT-Effekte untersucht. Pennington et al. (2016) trugen in einer Literaturübersicht Untersuchungen zusammen, die Mediationseffekte von der STT-Manipulation auf die abhängige Variable Leistung geprüft haben bzw. Hinweise darauf lieferten (z. B. i. S. eines Effekts der Manipulation auf den Mediator state-Angst). Die Befundlage ist insgesamt äußerst heterogen. Eine Reihe von Studien berichten nur tendenzielle (Spencer, Steele & Quinn, 1999) oder keine Hinweise auf Mediationseffekte durch Angst bzw. Testangst (Aronson et al., 1999; Hess, Auman, Colcombe & Rahhal, 2003; Keller & Dauenheimer, 2003; Schmader & Johns, 2003). Auch Beilock, Rydell und McConnell (2007) fanden keinen Effekt einer STT-Manipulation auf die erlebte state-Angst (Experiment 3A & B). Ein interessanter Teilbefund ist jedoch, dass sich bei freier Abfrage der Gedanken während der Aufgabenbearbeitung durchaus Effekte zeigten: relativ gesehen berichteten targets deutlich häufiger Sorgen darüber, durch ihre Leistung das Stereotyp zu bestätigen und auch häufiger Gedanken über das eigene Leistungsverhalten, was auch testangstbezogene Gedanken beinhaltete (z. B. „I wish I was better at subtracting numbers in my head“).

Einige Studien liefern jedoch Hinweise auf Mediationseffekte. In einer Untersuchung von Osborne (2001) wurden Leistungsunterschiede zwischen Ethnien (weiße vs. hispanische sowie Afroamerikaner) sowie zwischen Männern und Frauen durch state-Angst teilmediiert. Allerdings entstammen die Daten einer größeren Bildungsstudie, sodass keine Manipulation vorlag. Die Argumentation von Osborne (2001), dass bei den Leistungsmessungen ein STT vorlag, da dieser in standardisierten Testungen die Regel sei, ist fragwürdig. Chung, Ehrhart, Holcombe Ehrhart, Hattrup und Solamon (2010) konnten eine sequentielle Mediation zwischen dem subjektiv erlebten STT[22] und Leistung durch state-Angst und aufgabenspezifische Selbstwirksamkeit feststellen. Jedoch wurde auch hier der STT nicht manipuliert. Bei Harrison, Stevens, Monty und Coakley (2006) war der STT in einem experimentellen Design begleitet von erhöhter state-Testangst bei den targets (nicht bei den nontargets). Weitere Studien fanden in experimentellen Designs Hinweise auf mehrstufige Mediationen. Brodish und Devine (2009) berichten eine sequenzielle Mediation des Effekts der STT-Manipulation auf Leistung über Leistungsvermeidungsziele und state-Testangst. Mrazek et al. (2011) erhoben sowohl das Abschweifen der Gedanken (mind-wandering) während der Testung als auch die state-Angst. Der Effekt der STT-Manipulation auf die Leistung wurde durch mind-wandering mediiert, der Effekt der Manipulation auf mind-wandering über die state-Angst.

Bezüglich der Erklärung von STT-Effekten kommen Pennington et al. (2016) zu folgender, äußerst ernüchternder Schlussfolgerung: „Two decades of research have demonstrated the harmful effects that stereotype threat can exert on a wide range of populations in a broad array of performance domains. However, findings with regards to the mediators that unterpin these effects are equivocal" (S. 19). Dies gilt auch für den Mediator Angst bzw. Testangst. Die Heterogenität der Befunde kann verschiedene Ursachen haben. Eine davon ist das Vorhandensein von Moderatoren (Pennington et al., 2016). Einen interessanten Beitrag hierzu leisteten Gerstenberg, Imhoff und Schmitt (2012). In drei Experimenten zeigten sich die erwarteten Leistungseffekte unter dem STT nur bei Frauen mit einem instabilen mathematischen Selbstkonzept (d. h. hohes explizites bei gleichzeitig niedrigem implizitem Selbstkonzept). Darüber hinaus wurde der STT-Effekt bei dieser Subgruppe durch state-Testangst mediiert (Studie 3). Die Heterogenität der Befundlage lässt sich auch methodisch erklären: Die Erfassung der Angst bzw. Testangst erfolgte häufig nur in unzureichender Differenzierung. So ist nicht davon auszugehen, dass die Facette Aufgeregtheit Varianz in der Leistung aufklärt. Denkbar ist auch, dass durch die Aggregation von Angstfacetten die ablaufenden Prozesse nicht mehr sensitiv genug abgebildet werden (z. B. waren bei Keller und Dauenheimer (2003), Schmader und Johns (2003) sowie Beilock et al. (2007) kognitive und affektive Angstaspekte vermengt). Brodish und Devine (2009) sowie Gerstenberg et al. (2012) erfassten die Besorgtheit und konnten Mediationseffekte finden. Die Facette Mangel an Zuversicht könnte ebenso ein vermittelnder Prozess sein, in Anbetracht von Mediationseffekten durch das

[22] z. B. „I often feel that people's evaluations of my behavior are based on the ethnic group to which I belong."

eng verwandte Konstrukt Selbstwirksamkeit (siehe z. B. Chung et al., 2010). Ähnliches gilt auch für die Facette Interferenz, was sich wiederum aus Mediationseffekten zu mind-wandering ableiten lässt (siehe z. B. Mrazek et al., 2011). Schließlich könnte auch die Art der Aktivierung des STT (subtil oder mehr oder weniger explizit) zur heterogenen Befundlage beitragen (Pennington et al., 2016). Dass dies für die stattfindenden Prozesse (sprich Mediatoren) eine Rolle spielt, legen neben der Moderation der Effektgrößen durch die Aktivierungsform (Nguyen & Ryan, 2008) auch einzelne Studien (siehe Pennington et al., 2016) nahe. In Erweiterung dieses Gedanken ist es möglich, dass der STT in unterschiedlichen target-Gruppen (z. B. Frauen, ältere Menschen, etc.) über unterschiedliche Prozesse vermittelt wird (Pennington et al., 2016).

1.2.2.2.4 Implikationen und offene Fragen

Die bisherigen Ausführungen zum STT sowie zur Bedeutung von Testangst beim STT werfen eine Reihe von Fragen auf. Hierzu gehört insbesondere die Heterogenität der Befunde. Dabei ist auch zu klären, ob das Erleben von Angst bzw. Testangst beim STT überhaupt eine nennenswerte Bedeutung hat. Ein Ansatz zur Klärung könnte sich aus einer Betrachtung der gängigen Untersuchungsdesigns in beiden Forschungsbereichen ableiten. Wie bereits dargelegt wurde, gehört die experimentelle Induktion von Bewertungsstress zum typischen Untersuchungsdesign in der Testängstlichkeitsforschung (siehe Abschnitt 1.1.2). Die subtile Aktivierung nach Nguyen und Ryan (2008) weist deutliche Parallelen hierzu auf. Letztere besteht zuweilen lediglich aus einer evaluativen Instruktion, bei der der mutmaßlich erfasste Leistungsbereich der ist, in dem die jeweilige Gruppe gemäß des Stereotyps unbegabt ist (z. B. „mathematische Fähigkeiten“ bei Frauen). Vereinfacht lässt sich also sagen, dass derartige Manipulationen gleichsam evaluative Instruktionen sind, allerdings in der Form, dass zur Situation eine stereotype Bedeutung „addiert“ wird. Die fragliche Rolle von Testangst beim STT sollte am ehesten in Designs aufklärbar sein, in denen evaluative Instruktionen mit und ohne stereotyper Bedeutung kontrastiert werden. Aus diesem sowie aus Gründen der praktischen Relevanz sollte der Bereich der subtilen Aktivierung des STT vorrangig betrachtet werden: „If effects are not found with more subtle cues, then one might question the applicability of this line of research to employment testing contexts.“ (Nguyen & Ryan, 2008, S. 1315). Die offenkundigen Formen der Aktivierung treten vermutlich seltener in der praktischen Anwendung von Tests auf (Nguyen & Ryan, 2008)[23]. Darüber hinaus ist denkbar, dass die Fokussierung auf einzelne Facetten anstelle der Aggregation von Facetten bei der Klärung der Bedeutung von Testangst helfen kann.

[23] Es ist aber wahrscheinlich, dass auch explizite Aktivierungen in der Praxis vorkommen, z. B. durch unterschiedliches Lehrerverhalten gegenüber Mädchen und Jungen. Auf derartige Situationen soll hier aber kein Fokus gelegt werden.

Interessant sind in diesem Kontext auch die Reaktionen von nontargets. So ist das negative Stereotyp über eine Gruppe („Frauen sind schlecht in Mathe) stets gekoppelt an ein positives Stereotyp über eine andere Gruppe („Männer sind gut in Mathe“) (Smith & Johnson, 2006; Walton & Cohen, 2003). Wie sich die Aktivierung eines positiven Stereotyps über die *eigene* Gruppe („stereotype boost“) bzw. eines negativen Stereotyps über eine *andere* Gruppe („stereotype lift“) auswirkt, wird zunehmend beforscht (siehe Shih, Pittinsky & Ho, 2011). Diese Thematik soll an dieser Stelle nicht weiter vertieft werden, jedoch ist für den STT aus einem anderen Grund die Analyse der nontargets wichtig. Theoretische Grundlage von STT-Effekten ist die Annahme, dass nontargets anders auf STT-Manipulationen reagieren als targets, was Rückschlüsse auf die stattfindenden Prozesse erlauben sollte. Für den möglichen Mediator Testangst bedeutet dies, dass bei targets und nontargets ein Unterschied in der Mediation auftreten müsste, z. B. dergestalt, dass bei targets eine Mediation auftritt und bei nontargets nicht, oder dass der Effekt bei targets stärker ist als bei nontargets. Es genügt daher nicht, einen Mediationseffekt bei den targets zu prüfen, selbiges sollte auch für die nontargets vorgenommen werden. Die Forschung zur Testängstlichkeit und Testangst legt nahe, dass sich die Relation von Angst und Leistung in Abhängigkeit von der Instruktion (evaluativ vs. nonevaluativ) unterscheidet. Die Forschung zum STT legt wiederum nahe, dass sich targets und nontargets in Abhängigkeit von der Instruktion (Stereotypaktivierung vs. neutrale Instruktion) mehr oder weniger voneinander unterscheiden. Aufschluss über die tatsächliche Rolle von Testangst beim STT können also Designs liefern, die diese verschiedenen Bedingungen systematisch kontrastieren:

a) Eine evaluative Bedingung, die stereotyp formuliert ist
b) Eine evaluative Bedingung, die stereotypneutral formuliert ist
c) Eine nonevaluative Bedingung, die ebenfalls stereotypneutral formuliert ist

Die Kombination von b) und c) stellt das typische Untersuchungsparadigma nach Sarason und Kollegen dar. Bei einer subtilen Aktivierung des Stereotyps kann allerdings – je nachdem welcher Leistungsbereich angesprochen wird – b) (ungewollt) gleichbedeutend mit a) sein, womit im Design unbeabsichtigt ein STT-Effekt induziert würde. Aus diesem Grund ist die Trennung der Bedingungen a) und b) notwendig, die sich nicht im evaluativen Charakter, sondern im stereotypen Gehalt der Instruktion unterscheiden. Der Vergleich mit c) erlaubt ein Urteil, wie die Reaktionen ohne evaluativen Charakter der Instruktion aussehen. In allen Bedingungen muss unterschieden werden zwischen der Angst, das Stereotyp zu bestätigen und allgemeiner Angst bzw. Testangst (siehe auch Beilock et al., 2007). So ist es eine offene Frage, welchen Stellenwert die Angst, ein negatives Stereotyp zu bestätigen gegenüber „allgemeinen“ Angst- bzw. Testangstprozessen hat und wie diese zueinander in Bezug stehen. Diesbezüglich bringt das genannte Design einen wich-

tigen Vorteil: da der Bewertungscharakter und der stereotype Inhalt der Instruktion nicht konfundiert sind, kann spezifischer nachvollzogen werden, welche Manipulation sich auf welche „Form“ der Angst auswirkt. Zur Untersuchung dieser Fragen wurde Studie 2 durchgeführt.

1.2.3 Fazit

Testangst und Testängstlichkeit gehen mit schlechteren Leistungen in vielen verschiedenen Formen von Prüfungen und Tests einher. Zahlreiche Befunde weisen darauf hin, dass es kausale Effekte der Testangst auf die Leistung gibt, Testangst also Ressourcen bindet, die für eine optimale Leistung fehlen. Ebenso gibt es vielfache empirische Evidenz dafür, dass Testangst ein Ausdruck von erlebten und / oder tatsächlichen Defiziten und somit nicht nur Ursache, sondern auch Folge von (schlechter) Leistung ist. Interferenz- und Defizitperspektive sollten nicht als gegensätzliche, sondern sich ergänzende Perspektiven verstanden werden, deren Wechselwirkung und spezifischer Geltungsbereich von der Forschung zu klären ist. Die Höhe der Angst-Leistungs-Relation hängt von einer Reihe von Variablen ab, was wiederum Rückschlüsse auf die jeweilige Bedeutung von Interferenz- gegenüber Defiziterklärungen zulässt. Ein in diesem Kontext wichtiger Moderator ist der Messzeitpunkt von Testangst. Die Ursache für diese Moderation ist bislang noch nicht geklärt und Gegenstand von Studie 1. Darüber hinaus hängt die Ausprägung der Angst-Leistungs-Relation stark ab von situativen Variablen, sprich Aufgabeninstruktion, -komplexität und darbietung. Diese wirken sich in den meisten Fällen auf das „Stresspotenzial“ einer Prüfung oder eines Tests aus. Mit dem STT wurde eine spezifische Kombination dieser Parameter näher betrachtet, bei der eine bestimmte Instruktionsform mit einem bestimmten Aufgabenbereich verknüpft ist. Die Rolle von Testangst bei der Erklärung des STT (siehe Abbildung 5) wurde mehrfach diskutiert, aber bislang nicht zureichend geklärt. Ein Beitrag hierzu sollte in Studie 2 geleistet werden.

Da Testangst von den meisten Menschen als aversiv erlebt wird und leistungsmindernde Effekte aufweist, gibt es eine umfangreiche Forschungsaktivität zu Interventionen. Diese Interventionen haben zum Ziel, Testangst zu reduzieren, die hoch Testängstlichen zu besseren Leistungen zu befähigen oder beides. Im letzten Abschnitt der theoretischen Grundlagen soll auf dieses Thema näher eingegangen werden, welches auch die Basis für Studie 3 war.

1.3 Interventionen

Die folgenden Betrachtungen sollen einen Eindruck davon geben, nach welchen Prinzipien oder Mechanismen auf Testängstlichkeit bezogene Interventionen funktionieren. Unter Interventionen fallen dabei in diesem Kontext alle Formen von Therapien, Trainings oder Coachings, die das Individuum zu einem besseren Umgang mit Prüfungs- und Testsituationen befähigen sollen. Diese werden im Folgenden von „Kurzinterventionen" unterschieden. Letztere haben oft ähnliche Ziele wie Interventionen, unterscheiden sich aber darin, dass sie (lediglich) Veränderungen der Testsituation selbst darstellen und dadurch meistens deutlich kürzer sind (siehe Abschnitt 1.3.1).

Hembree (1988) analysierte metaanalytische Effekte von Interventionen auf Testängstlichkeit, Testangst und Leistung[24]. Sowohl behaviorale (u. a. Systematische Desensibilisierung und Entspannungstraining) als auch kognitiv-behaviorale Interventionen (u. a. Kognitive Umstrukturierung, Aufmerksamkeitstraining, Angstbewältigungstraining) erbringen demnach positive Effekte im Sinne einer Reduktion der Testängstlichkeit. Systematische Desensibilisierung war sowohl bei jüngeren (5. bis 12. Klasse) (d = -.54, k = 13, N = 870) als auch älteren Probanden (postsecondary) (d = -.59 bzw. -1.08, je nach Methode; k = 64, N = 1.850) effektiv, sowie auch Entspannungstrainings (d = -.68, k = 32, N = 953). Kognitiv-behaviorale Interventionen waren bei College-Studierenden (d = -.87, k = 43, N = 1.311) effektiver als bei jüngeren Schülern (d = -.53, k = 5, N = 486). Ein wichtiger Teilbefund war, dass study skills-Interventionen alleine keinen signifikanten Effekt hatten (d = -.14, k = 6, N = 163), jedoch in Kombination mit behavioralen oder kognitiv-behavioralen Maßnahmen. Interventionen zu test-taking skills hingegen erbrachten eine Senkung der Testängstlichkeit (d = -.55, k = 6, N = 350). Behaviorale als auch kognitiv-behaviorale Interventionen reduzierten sowohl Besorgtheit als auch Aufgeregtheit in einem ähnlichen Maß (d = -.60 bis -.82). Hembree (1988) differenzierte überdies die Effekte auf Ängstlichkeit (trait) und Angst (state) – beide wurden durch Interventionen reduziert. Eine neuere Metaanalyse von Ergene (2003) befasste sich ebenfalls mit Interventionseffekten auf Testängstlichkeit[25]. Inkludiert wurden insgesamt 56 Studien (N = 2.428, M_{Alter} = 18.86). Die größten Effekte zeigten sich bei einer Kombination von kognitiven und skill-orientierten Interventionen (d = 1.22, k = 3) sowie bei der Kombination behavioraler und skill-orientierter Interventionen (d = 1.10, k = 5). Bei Betrachtung spezifischer Interventionstechniken waren kognitive Umstrukturierung (d = 1.11, k = 4), Angstbewältigungstraining (d = .97, k = 2) und Systematische Desensibilisierung (d = .90, k = 15) besonders erfolgreich. Study-skills-Interventionen erbrachten alleine nur einen kleinen Effekt (d = .28, k = 10).

[24] Der zeitliche Umfang dieser reichte von einer bis 12 Stunden (Md = 6).
[25] Über die Differenzierung von trait und state wurden hier keine Aussagen gemacht, ebenso fehlte die Angabe des N pro Effektgröße. Die unzureichende Dokumentation der Ergebnisse in dieser Metaanalyse konstatieren auch Fehm und Fydrich (2011).

Hembree (1988) betrachtete auch Effekte auf Leistung. Systematische Desensibilisierung (d = .32, k = 38, N = 1.274) sowie kognitiv-behaviorale Methoden (d = .52; k = 32, N = 1.132) führten demnach zu Verbesserungen bei Testleistungen (für deren Kodierung siehe Abschnitt 1.2), nicht aber Entspannungstrainings (d = .13, k = 21, N = 3.109). Auch in Bezug auf GPA erbrachten systematische Desensibilisierung (d = .40, k = 20, N = 657) und kognitiv-behaviorale Methoden (d = .72, k = 8, N = 190) positive Effekte.

Die Befunde von Hembree (1988) sowie Ergene (2003) zeigen, dass Testängstlichkeit durch Interventionen abgemildert werden kann. Dabei sind kognitive und behaviorale Maßnahmen erfolgreich, während ein reiner Fokus auf skills schwächere Wirkungen zeigt. Dieser Befund verweist auch darauf, dass eine Erklärung der Angst-Leistungs-Relation allein durch die Defizitperspektive unzureichend ist – andernfalls müsste eine Beschränkung auf skill-Ansätze mindestens ebenso große Effekte erzielen wie kognitive und behaviorale Interventionen. Hoch Testängstliche können von diesen Maßnahmen also profitieren. In den letzten Jahren befasste sich eine wachsende Zahl von Studien mit dem Gedanken, wie mit noch einfacheren Mitteln Testangst reduziert bzw. Testängstlichen zu einer besseren Leistung verholfen werden kann.

1.3.1 Verschiedene Ansätze von Kurzinterventionen

Der Begriff Kurzintervention eignet sich um eine bestimmte Klasse von Maßnahmen zu beschreiben, die in erster Linie Eingriffe in die Testsituation darstellen. Von diesen experimentellen Manipulationen wird eine positive Wirkung auf Testängstliche vermutet. Dass derartige Modifikationen der Test- bzw. Prüfungssituation eine positive Wirkung auf Testängstliche erzielen könnten, leitet sich aus den zahlreichen Befunden ab, nach denen das Ausmaß der erlebten Testangst (siehe Abschnitt 1.1.2) sowie die Angst-Leistungs-Relation (siehe Abschnitt 1.2.2) von situativen Charakteristika wie der Instruktion oder dem Aufgabenbereich abhängen. Wie sich zeigen wird, haben einige Kurzinterventionen starke Ähnlichkeit mit den oben erwähnten Interventionen oder zumindest die gleichen theoretischen Ursprünge. Kurzinterventionen im hier verstandenen Sinne haben nicht (zumindest nicht primär) zum Ziel, die Testängstlichkeit dauerhaft zu reduzieren. Stattdessen geht es um eine Reduktion der erlebten Testangst und / oder eine Steigerung der Leistung von Testängstlichen in einem bestimmten Test oder einer Prüfung[26]. In diesem Abschnitt werden eine Reihe von verschiedenen Kurzinterventionen vorgestellt sowie deren Potenziale und Limitationen bewertet. Im Anschluss daran wird die kognitive Umbewertung als eine Form der Kurzintervention diskutiert, welche die angesprochenen Limitationen umgehen könnte.

[26] Eine Definition von Yeager und Walton (2011) im Rahmen einer Übersichtsarbeit zu ähnlichen Manipulationen lautet: „“small” social-psychological interventions – typically brief exercises that do not teach academic content but instead target students‘ thoughts, feelings and beliefs in and about school“ (S. 268)

Eine Form der Kurzintervention ist das expressive Schreiben (Park, Ramirez & Beilock, 2014; Ramirez & Beilock, 2011). Expressives Schreiben bezeichnet eine therapeutische Methode, bei der Personen über ein emotional bedeutsames Thema schreiben, beispielsweise über Traumata, Konflikte oder stressvolle Ereignisse (Pennebaker & Chung, 2011). Diese Technik, bei der eine Person üblicherweise in mehreren Durchgängen über das besagte Thema schreibt, weist positive Effekte auf verschiedene Indikatoren des Gesundheitszustands und des Wohlbefindens auf, aber auch auf Leistungsmaße wie Noten (Pennebaker, 1997). Eine mögliche Erklärung für leistungssteigernde Wirkungen liefern Befunde, nach denen expressives Schreiben positive Effekte auf die Arbeitsgedächtniskapazität hat (Klein & Boals, 2001; Yogo & Fujihara, 2008), welche für die Erklärung des Zusammenhangs von Testangst und Leistung eine wichtige Rolle spielt (siehe Abschnitt 1.2.1.1).

Einige Studien haben sich mit der Anwendung von expressivem Schreiben als einmaliger, kurzer Intervention direkt vor einem Test befasst. Ramirez und Beilock (2011) führten hierzu mehrere Experimente durch. In zwei Laborstudien (N = 20 bzw. 47, studentische Stichprobe[27]) bearbeiteten Probanden Mathematikaufgaben unter neutraler (pretest) und evaluativer Instruktion (posttest), wobei jeweils ein Teil der Stichprobe vor dem posttest expressives Schreiben bezüglich der anstehenden Aufgaben durchführte[28]. Während die Experimentalgruppen bessere Leistungen zeigten als im pretest, verschlechterte sich die Leistung der Kontrollgruppen. Tatsächlich produzierten die Probanden in der Experimentalgruppe beim Schreiben mehr Wörter und Aussagen, die Angst bzw. Sorgen bezüglich der eigenen Leistung im Test zum Inhalt hatten (verglichen mit einer Kontrollgruppe, die über ein anderes Ereignis schreiben sollte). In zwei weiteren Studien (N = 51 bzw. 55, 9. Klasse) wurde geprüft, ob Testängstliche, die ja stärker zu Sorgen neigen, von expressivem Schreiben besonders profitieren. Vor einer Abschlussprüfung führte eine Gruppe expressives Schreiben durch, die andere Gruppe nicht. Vorab wurde die Testängstlichkeit erfasst. Die Testängstlichkeit korrelierte signifikant mit der Prüfungsleistung in den Kontrollgruppen (r = -.45 bzw. -.48 in Studie 3 bzw. 4), aber nicht signifikant in den Experimentalgruppen (r = -.07 bzw. -.19). Eine Unterteilung der Stichprobe in hoch und niedrig Testängstliche via Mediansplit ergab, dass expressives Schreiben nur einen Effekt auf die Schüler mit hoher, nicht jedoch auf jene mit niedriger Testängstlichkeit hatte (Ramirez & Beilock, 2011). In einer neueren Studie berichten Park et al. (2014) ebenfalls von positiven Effekten durch eine einmalige Schreibintervention bei hoch mathematikängstlichen Probanden, während sich die Leistung niedrig mathematikängstlicher Probanden, die expressiv schrieben, nicht von jenen in der Kontrollgruppe unterschied. Expressives Schreiben bietet einen vielversprechenden Ansatz für Kurzinterventionen. Potenzial hat

[27] Die Stichprobe in Studie 2 wurde nicht näher beschrieben.

[28] Ein Ausschnitt aus einer solchen Instruktion: „Please take the next 7 minutes to write as openly as possible about your thoughts and feelings regarding the math problems you are about to perform on the Excel spread sheet. In your writing, I want you to really let yourself go and explore your emotions and thoughts as you are getting ready to start the second set of math problems. […]" (Park et al., 2014, S. 106)

insbesondere der Befund, dass sich die Leistung niedrig Ängstlicher durch die Intervention nicht verschlechterte (siehe Abschnitt 1.2.2.1). Eine offene Frage ist jedoch, wie viel Informationen zu einem anstehenden Test einem Probanden vorliegen müssen, damit er etwas über diesen Test schreiben kann. Um die eigenen Emotionen zu einem Test zu beschreiben sollte ein Mindestmaß an Informationen vorhanden sein. Bei Prüfungen im schulischen bzw. universitären Bereich dürfte dies kein Problem sein, da wichtige Eigenschaften der Prüfung (Dauer, Format der Fragen, Inhalt) in ihren Grundzügen bekannt sind. Bei anderen Tests (z. B. Leistungstests) ist es erforderlich, Beispielaufgaben vorzulegen (so wie bei Park et al., 2014). Inhalt und Schwierigkeit der Beispielitems dürften hier eine wichtige Rolle spielen.

Eine weitere Kurzintervention ist die Schaffung von Transparenz. Mavilidi, Hoogerheide und Paas (2014) erlaubten Schülern ($N = 117$, $M_{Alter} = 11.59$) vor einem Mathematiktest, sich für eine Minute das Aufgabenmaterial anzusehen oder nicht. Die Schüler (sowohl niedrig, mittel als auch hoch Testängstliche) zeigten in der transparenten Bedingung bessere Leistungen. Bei dieser Maßnahme ist die praktische Umsetzbarkeit jedoch nicht immer möglich. Zudem muss angenommen werden, dass das vorherige Betrachten des Testmaterials de facto eine Verlängerung der Bearbeitungszeit bedeutet, welche unmittelbar zu einer Leistungsverbesserung führt.

Eine weitere Methode von Kurzinterventionen ist das Priming von Eigenschaften oder Stereotypen. Wichtige theoretische Grundlage für die Effekte von Priming ist die ideomotorische Theorie, die auf William James zurückgeht: „For James, imagining or thinking about a behavioral response had the same kind of priming effect on the likelihood of engaging in that response." (Bargh, Chen & Burrows, 1996, S. 231). Das bedeutet, dass die Aktivierung eines Konzepts (z. B. eines Stereotyps oder einer Eigenschaft) die Wahrscheinlichkeit erhöht, dass Verhaltensweisen, die mit diesem Konzept verbunden sind, auch gezeigt werden (Wheeler & Petty, 2001). Das wohl bekannteste Beispiel für diesen Effekt ist der Befund von Bargh et al. (1996): Probanden, die mit einem Altersstereotyp geprimt wurden, liefen den Gang vor einem Untersuchungsraum nach dem Experiment langsamer entlang als eine Kontrollgruppe. Wheeler und Petty (2001) unterscheiden in einer Literaturübersicht self-stereotypes von other-stereotypes. Bei self-stereotypes (hierzu gehört z. B. der STT) werden Stereotype über eine bestimmte Gruppe bei eben jener Gruppe aktiviert. Bei other-stereotypes Situationen werden Stereotype über eine Gruppe bei Personen aktiviert, die nicht Mitglied dieser Gruppe sind. In den meisten Fällen zeigt sich bei beiden Formen eine Annäherung des Verhaltens der Probanden an das jeweilige Stereotyp (Wheeler & Petty, 2001).

Beispielhaft sei hierzu eine Studie von Hansen und Wänke (2009) geschildert. Probanden sollten hier in der Experimentalgruppe typische Eigenschaften und Verhaltensweisen eines Professors beschreiben. Verglichen mit Probanden, die Beschreibungen von als weniger intelligent geltenden

Personen (Sekretäre bzw. weibliche Putzkräfte) abgeben mussten, berichteten erstere eine höhere aufgabenspezifische Selbstwirksamkeit für Wissensfragen (Studie 1 und 2, N = 39 bzw. 40, studentische Stichprobe) und erzielten auch bessere Leistungen (Studie 2). Einen ähnlichen Ansatz verfolgten Lang und Lang (2010). In zwei Studien wurde ein Kompetenzpriming durchgeführt, bei dem sich die Experimentalgruppe eine erfolgreiche, kompetente Person vorstellen, mehrere Eigenschaften dieser Person auflisten und beschreiben sollte, wie sich diese Person bei der Lösung schwieriger Probleme fühlt. In zwei Studien (N = 219 bzw. 232 Schüler, M_{Alter} = 16.54 bzw. 15.27) wurde dieses Treatment bei einer Experimentalgruppe vor der Bearbeitung einer Intelligenztestaufgabe administriert. In beiden Studien wurde in der Kontrollgruppe eine signifikante, negative Korrelation zwischen Testängstlichkeit und Leistung registriert (r = -.35 bzw. -.28), nicht jedoch in der Experimentalgruppe (r = .15 bzw. -.06). Dabei verbesserte sich die Leistung der hoch Testängstlichen in der Experimentalgruppe, während die niedrig Testängstlichen durch die Manipulation schlechtere Leistungen erbrachten. Die Autoren erklärten diese Befunde damit, dass Selbstwirksamkeit und Motivation bzw. Anstrengung nicht linear positiv, sondern diskontinuierlich zusammenhängen (Vancouver, More & Yoder, 2008). Niedrig Testängstliche, die ohnehin schon eine höhere Selbstwirksamkeit aufweisen, würden durch die Intervention eine so hohe Erfolgserwartung aufbauen, dass sie nur noch sehr wenig Anstrengung aufbringen (Lang & Lang, 2010). Dieses Phänomen ist in der Testängstlichkeitsforschung aus der Variation von evaluativer und nonevaluativer Instruktion bekannt (siehe Abschnitt 1.2.2.1) und weist der praktischen Anwendung Schranken auf. So schlagen Lang und Lang (2010) vor, dass Probanden bezüglich ihrer Testängstlichkeit gescreent werden, wobei dann nur die hoch Ängstlichen eine Intervention erfahren sollten. Unklar ist dabei, ab welchem Wert ein cut-off gesetzt werden sollte.

Generell sind Studien zu Priming häufig eindrucksvoll, da sich mit minimalen Manipulationen erstaunliche Effekte zeigen. Auch Nelson und Knight (2010) berichten mit einer ähnlichen Intervention positive Effekte auf Testangst und Leistung. Die Debatte, die durch fehlgeschlagene Replikationen von Primingeffekten (Doyen, Klein, Pichon, Cleeremans & Lauwereyns, 2012; Shanks et al., 2013) ausgelöst wurde (siehe z. B. Abbott, 2013; Bargh, 2012), macht jedoch deutlich, dass noch Forschung zu den Rahmenbedingungen und Erklärungen für diese Effekte nötig ist. Eine Bewertung der Möglichkeiten und Grenzen der Primingmethode ist daher derzeit noch schwierig. Dabei stellt sich insbesondere die Frage, ob Eigenschaften beliebig „geprimt“ werden können, oder ob nicht das dispositionelle Kompetenzempfinden eine größere Bedeutung hat als das manipulierte (siehe z. B. van Yperen, 2007).

Auch im Kontext des STT wurden einige Kurzinterventionen untersucht, wie z. B. die Titulierung eines Tests als Herausforderung (Alter, Aronson, Darley, Rodriguez & Ruble, 2010) oder der induzierte Fokus auf das Erzielen richtiger gegenüber dem Vermeiden falscher Antworten (Keller, 2007). Allerdings bedeuten diese Interventionen einen starken Eingriff in die Testadministration,

da sie das Testziel (z. B. die diagnostische Bedeutung des Ergebnisses) oder den Bewertungsmodus verändern. Beides ist bei standardisierten Verfahren in der Praxis kaum umsetzbar.

Generell stellen Kurzinterventionen eine Veränderung der objektiven Testsituation dar, ob durch eine Variation der Testinstruktion oder die Vorgabe einer vorgeschalteten Aufgabe wie dem expressiven Schreiben. Diese Modifikationen sind aus drei Gründen nicht unbegrenzt realisierbar. Erstens sind Tests und Prüfungen in der Regel standardisiert, d. h. Abweichungen von der vorgegebenen Instruktion sind prinzipiell nicht vorgesehen. Vorgeschaltete Aufgaben ermöglichen es zwar, dass eine vorgesehene Testinstruktion unangetastet bleibt. Jedoch läge streng genommen eine Verletzung der Standardisierung vor, wenn nicht allen Probanden dieselbe vorgeschaltete Aufgabe vorgelegt würde. Zweitens sind einige Manipulationen in einer high-stakes Testung psychologisch wenig glaubhaft. Beispielsweise lässt sich die Bedrohlichkeit einer wichtigen Abschlussprüfung oder einer Assessment Center Übung im Rahmen eines Auswahlverfahrens schwerlich reduzieren, indem deren diagnostische Relevanz abgewertet wird. Lediglich in Laborsituationen lassen sich der Bewertungscharakter und damit die Bedrohlichkeit eines Tests kontrolliert und sinnvoll manipulieren. Drittens gibt es Hinweise darauf, dass unterschiedliche Personen (z. B. in Abhängigkeit der Ausprägung der Testängstlichkeit) in unterschiedlichen Situationen optimale Leistung zeigen. Eine Kurzintervention könnte somit für einige Probanden vorteilhaft, für andere nachteilig sein. Ein vorheriges Screening der Probanden wäre mit unmittelbarer Auswertung und Zuweisung der entsprechenden Kurzintervention zwar mit technischer Unterstützung möglich. Dies würde aber erfordern, dass relevante Merkmale wie Testängstlichkeit vorab erfasst werden, was möglicherweise wiederum Einfluss auf das Testerleben haben könnte (siehe Abschnitt 1.2.1.3).

Die Vielfalt an Kurzinterventionen wurde an dieser Stelle nur angeschnitten. Ein häufiges (implizites oder explizites) Ziel einer Kurzintervention ist die Reduktion der erlebten Testangst, z. B. über die Stärkung der situationsspezifischen Selbstwirksamkeit. Im Sinne von Lazarus können Kurzinterventionen sowohl eine Modifikation der primären („Ist die Situation bedrohlich, erfordert sie eine Bewältigungsreaktion?“) als auch der sekundären Bewertung („Welche Bewältigungsressourcen stehen mir zur Verfügung?“) anstreben. Im nächsten Abschnitt soll eine Variante von Kurzinterventionen dargestellt werden, die das Potenzial besitzt, alle der drei Einschränkungen zu umgehen. Diese Kurzintervention wurde auch in Studie 3 getestet. Hierzu ist eine genauere Betrachtung des Emotionsentstehungsprozesses notwendig.

1.3.2 Kognitive Umbewertung von Testangst

Einen Beitrag zum theoretischen Verständnis von Kurzinterventionen und deren (angestrebter) Wirkung liefert das modale Emotionsmodell von Gross und Thompson (2007). Das modale Modell

sieht fünf Stufen in der Emotionsentstehung vor. An deren Beginn steht eine interne oder externe, psychologisch bedeutsame Situation (siehe Abbildung 6). Aufmerksamkeit für diese Situation resultiert in einer Situationsbewertung, welche letztlich zu einer emotionalen Reaktion führt[29]. Durch die Art der Reaktion wird wiederum der (psychologische) Situationscharakter verändert, weshalb es innerhalb des Modells Rückkopplungsprozesse gibt. Die Reaktion kann aber auch auf andere Schritte in der Emotionsentstehung rückwirken. Jeder Stufe der Emotionsentstehung ist eine bestimmte Klasse von Regulationsmechanismen zugeordnet (Gross, 1998; Gross & Thompson, 2007) (siehe Abbildung 6). Emotionsregulation ist dabei ein zielgerichteter Prozess, der folgendermaßen definiert wird: „the processes by which individuals influence which emotions they have, when they have them, and how they experience and expresse these emotions" (Gross, 1998, S. 275).

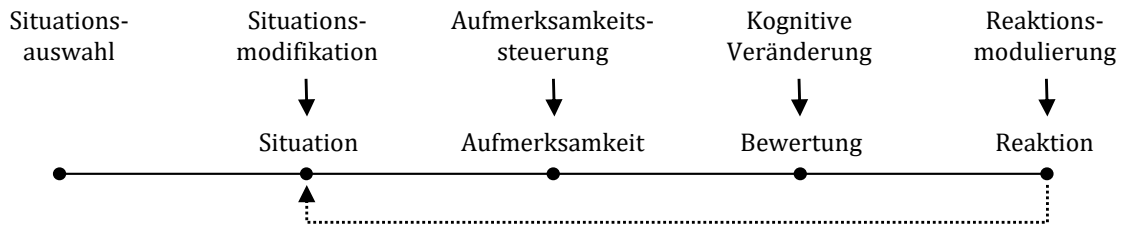

Abbildung 6: Modales Modell der Emotion nach Gross und Thompson (2007). Dargestellt sind die Schritte in der Emotionsentstehung sowie die jeweiligen Regulationsmechanismen.

Kurzinterventionen als Veränderungen der Testsituation können – ebenso wie die in Abschnitt 1.1.2 beschriebenen situativen Determinanten von Testangst – auf verschiedenen Stufen der Emotionsentstehung Auswirkungen haben und somit verschiedene Regulationsmechanismen anstoßen. Veränderungen der Testatmosphäre würden beispielsweise die Situation an sich modifizieren. Auch wäre es möglich, durch die Titulierung eines Tests als Herausforderung die Aufmerksamkeit von bedrohlichen Reizen weg zu lenken. Einige der unter Abschnitt 1.3 behandelten Interventionen (z. B. Entspannungstrainings) können eine Reaktionsmodulierung bewirken. Kognitive Veränderung findet verglichen mit der Reaktionsmodulierung früher im Emotionsentstehungsprozess statt und betrifft die Bewertung, die für die Art der Reaktion entscheidend ist (Gross, 1998). Diese Bewertung kann sich auch darauf beziehen, wie körperliche Erregungsprozesse interpretiert werden, z. B. als hemmend oder aber als anspornend (Gross & Thompson, 2007). Zu den verschiedenen Arten kognitiver Veränderung (u. a. soziale Abwärtsvergleiche) gehört das sog. reappraisal: „This involves cognitively transforming the situation so as to alter its emotional impact." (Gross, 1998, S. 284). Reappraisal kann darin bestehen, „im Schlechten das Gute" zu suchen, also Misserfolgen oder anderen negativen Ereignissen Positives abzugewinnen

[29] Zur Bedeutsamkeit einer Situation vgl. auch die primäre Bewertung im transaktionalen Stressmodell von Lazarus (siehe Abschnitt 1.1.1.1). Ebenso wie in der Theorie von Lazarus spielt die Bewertung (appraisal) hier eine zentrale Rolle.

(Augustine & Hemenover, 2009). Augustine und Hemenover (2009) führten eine Metaanalyse zur Effektivität von verschiedenen Emotionsregulationsstrategien durch (k = 75). Dabei wurde die Effektivität als „hedonic shift" operationalisiert, also als Anstieg des positiven oder Reduktion des negativen Affekts. Reappraisal gehört demnach zu den effektivsten Strategien (k = 2, d = .65). Emotionsregulation dient, aus funktionaler Sicht betrachtet, einem bestimmten Ziel (Gross & Thompson, 2007). Dieses Ziel muss aber nicht zwingend die Reduktion eines negativen Affekts sein. So gibt es Hinweise darauf, dass Personen in bestimmten Kontexten dazu motiviert sind, einen negativen Affekt zu erleben bzw. aufrecht zu erhalten, wenn dieser als funktional für die Erreichung eines Ziels empfunden wird, wie z. B. die Emotion Ärger in einer Verhandlungssituation (Tamir & Ford, 2012).

Ausgehend von den Theorien zur Emotionsentstehung und –regulation stellt sich nun die Frage, ob die (aversive) Emotion Testangst in einer funktionalen Weise kognitiv umbewertet werden kann. Diese Umbewertung könnte zweierlei Formen annehmen: Testangst könnte als *nicht* schädlich für die Leistung interpretiert werden, oder aber (darüber hinaus) sogar als leistungsförderlich. Einige Studien haben in unterschiedlichen Kontexten versucht, ein reappraisal zu induzieren. Im Sinne des Regulationsgedankens muss dies *nicht* darauf abzielen, die Entstehung von Testangst zu verhindern, sondern darauf, deren Bedeutung und Wirkung zu verändern. Auf diese Studien soll nun eingegangen werden.

Johns et al. (2008) untersuchten die Wirkung von reappraisal-Manipulationen im Kontext des STT. Dabei wurden Probanden instruiert, dass auf Basis von Forschungsergebnissen erwiesen sei, dass Angst nicht hinderlich, sondern förderlich für die Leistung sein kann (Studie 3 und 4). Bei unterschiedlichen targets (Frauen und hispanischen Amerikanern) wurde dadurch ein STT-Effekt auf Leistung verringert bzw. aufgehoben. Interessant ist der Befund, dass das reappraisal keinen Effekt auf die Ausprägung der (nach den Aufgaben erfassten) state-Angst hatte – das reappraisal reduzierte also nicht das Niveau an Angst.

Jamieson und Kollegen führten eine Reihe von Untersuchungen durch, die sich insbesondere mit dem reappraisal von physiologischer Erregung (arousal) befassten. Grundlage ist die Annahme, dass die Interpretation von arousal die Emotionsbildung bestimmt (siehe z. B. Schachter & Singer, 1962). Jamieson, Mendes und Nock (2012) argumentieren, dass arousal normalerweise negativ bewertet wird und entsprechend negative Konsequenzen auf das affektive Erleben und die Leistung hat. Da arousal in Anforderungssituationen funktional sein kann, zielt ihr Ansatz von reappraisal-Manipulationen *nicht* darauf ab, das Entstehen von arousal zu verhindern oder arousal zu senken: „arousal-reappraisal manipulations break the association between stress-based arousal and negative appraisals." (S. 52).

Jamieson, Nock und Mendes (2012) legten einer studentischen Stichprobe (N = 49) unterschiedliche Zusammenfassungen von Zeitschriftenartikeln vor. Die erste Gruppe las einen Text der mitteilte, dass die gewöhnlichen Stressreaktionen des menschlichen Körpers bei der Bewältigung von Stress und beim Erbringen von Leistung helfen. Die zweite Gruppe las einen Text der empfahl, als Bewältigungsmaßnahme die Aufmerksamkeit von bedrohlichen Reizen abzuwenden. Eine dritte Gruppe las keinen Text. Nach einer sozialen Stresssituation zeigte die reappraisal-Gruppe die niedrigste Interferenz beim emotionalen Stroop und adaptivere physiologische Reaktionsmuster (höheres Herzzeitvolumen und niedrigerer peripherer Gefäßwiderstand) als die beiden anderen Gruppen.

Ebenfalls positive Effekte einer reappraisal-Manipulation berichten Jamieson, Nock und Mendes (2013) für Probanden ohne und mit einer sozialen Angststörung. In beiden genannten Untersuchungen führte das reappraisal überdies dazu, dass die Probanden ihre Ressourcen zur Bewältigung der Aufgabe höher einschätzten. Auch Beltzer, Nock, Peters und Jamieson (2014) stellten diesen Effekt fest. Wie Jamieson und Nock et al. (2012) und Jamieson et al. (2013) setzten sie den Trier Social Stress Test (TSST; Kirschbaum, Pirke & Hellhammer, 1993) ein, der eine soziale Stress- bzw. Leistungssituation simuliert. Beltzer et al. (2014) ließen die Leistung in der darin enthaltenen Selbstpräsentation von unabhängigen Beurteilern einschätzen. Dabei ergaben sich tendenziell bessere Leistungen in der reappraisal-Gruppe. Jamieson, Mendes, Blackstock und Schmader (2010) übertrugen dieses Paradigma auf den Kontext von Prüfungsleistungen. Eine studentische Stichprobe (N = 60) bearbeitete einige Wochen vor der GRE in einem Laborsetting eine Übungs-GRE. Eine Hälfte der Probanden erhielt dabei eine reappraisal-Instruktion mit der Information, dass arousal und Angst nicht schädlich seien, sondern leistungssteigernd wirken können. In der Übungs-GRE zeigte sich in der Experimentalgruppe eine bessere Leistung als in der Kontrollgruppe (in GRE Math, nicht in GRE Verbal). Ein Teil der Stichprobe (N = 28) trat den GRE im Laufe der folgenden drei Monate an und wurde danach nach dem Erleben während der Prüfung befragt. Die Probanden, die im Labor in der reappraisal-Bedingung waren, schnitten im GRE Math besser ab. Sie berichteten überdies in höherem Maße, dass das arousal beim Test geholfen habe und in geringerem Maße Zweifel an der eigenen Leistung sowie weniger Sorgen darüber, ängstlich zu sein (Jamieson et al., 2010). Erstaunlich an diesem Befund ist, dass sich mehrere Wochen nach der Manipulation Leistungseffekte beobachten ließen. Allerdings kann ein Hawthorne-Effekt nicht gänzlich ausgeschlossen werden, da alle Probanden vorab informiert wurden, dass es um den Zusammenhang von arousal und Leistung sowie Angst gehe.

In einer weiteren Studie von Jamieson, Peters, Greenwood und Altose (2016) wurde bei leistungsschwachen Studierenden (N = 93) eine reappraisal-Manipulation mit einer Placebo-Intervention verglichen im Hinblick auf den Effekt auf die Leistung in einer Mathematikprüfung (Probanden wurden angewiesen, negative Gedanken auszublenden). Im Vergleich zu einer ersten Prüfung vor

der Intervention stieg die Leistung der reappraisal-Gruppe in einer zweiten Prüfung, bei der die Intervention administriert wurde, an. Überdies berichtete die reappraisal-Gruppe bei der zweiten Prüfung eine niedrigere Mathematikängstlichkeit und ein höheres Maß an Bewältigungsressourcen als bei der ersten Prüfung. Dies zeigte sich nicht bei der Placebo-Gruppe. Der Effekt der Manipulation auf die Leistung wurde über die Einschätzung der eigenen Bewältigungsressourcen teilmediiert.

Ein Vorteil von reappraisal als Kurzintervention ist, dass es keinen theoretischen Grund dafür gibt, dass – wie z. B. beim Priming – niedrig Testängstliche durch ein reappraisal schlechter werden. Personen, die keine Anspannung oder Angst verspüren, müssen diese auch nicht funktional umbewerten. Bei Jamieson et al. (2013) etwa reduzierte sich der Aufmerksamkeitsbias durch das reappraisal bei Probanden ohne und mit sozialer Phobie in ähnlichem Ausmaß, wenngleich ein Niveauunterschied zwischen diesen beiden Gruppen erhalten blieb. Prinzipiell ist das reappraisal eine „adaptive" Kurzintervention, welche – zumindest theoretisch – nur dann eine Wirkung entfaltet, wenn auch eine Wirkung nötig ist, sprich wenn eine Person Angst erlebt.

Allerdings wirft eine genauere Betrachtung der Manipulationen die Frage auf, welche Prozesse eigentlich konkret umbewertet werden. So legten vier der sechs zitierten Studien (Beltzer et al., 2014; Jamieson et al., 2013; Jamieson et al., 2016; Jamieson, Nock et al., 2012) einen Fokus auf die funktionale Uminterpretation von eher unspezifischen körperlichen Erregungsprozessen in Stresssituationen (z. B. „The increase in arousal you may feel during stress is not harmful. Instead, these responses evolved to help our ancestors survive by delivering oxygen to where it is needed in the body."; Jamieson et al., 2013, S. 369). Jamieson et al. (2010) hingegen vermengten in ihrem Treatment die Umbewertung von arousal und Angsterleben: „However, recent research suggests that arousal doesn't hurt performance on these tests and can even help performance. . . people who feel anxious during a test might actually do better." (Jamieson et al., 2010, S. 209). Johns et al. (2008) formulierten ihre Manipulation im Sinne der Umbewertung von Angst. Es dürfte jedoch für die Manipulation keinen erheblichen Unterschied ausmachen, ob „arousal" oder „Angst" umbewertet werden soll – beides ist phänomenologisch schwer voneinander zu trennen. So ist die Wahrnehmung von arousal als Facette Aufgeregtheit wichtiger Bestandteil von Testängstlichkeit bzw. Testangst. Die genannten reappraisal-Interventionen richten sich maßgeblich auf die Umbewertung von arousal beziehungsweise von emotionalen Angstkomponenten. Es wurde bereits darauf eingegangen, dass es in erster Linie die kognitiven Komponenten von Testangst sind, die negativ mit Leistung einhergehen (siehe Abschnitt 1.2). Aufgrund des engen Zusammenhangs kognitiver und emotionaler Elemente von Testangst liegt die Vermutung nahe, dass bei dieser Form von Kurzintervention auch die Bewertung bzw. Bedeutung – nicht unbedingt die Intensität – von kognitiven Angstprozessen (Besorgtheit) verändert wird.

Mit Blick auf die breite Evidenz für den negativen Zusammenhang von Testängstlichkeit bzw. Testangst und Leistung ist es folgerichtig, dass Kurzinterventionen häufig darauf abzielen, die Entstehung von Testangst zu verhindern bzw. das Niveau an Testangst zu reduzieren. Die Befunde zu den Effekten von reappraisal hierbei sind uneinheitlich. Johns et al. (2008) stellten keine Auswirkungen auf das absolute Niveau an selbstberichteter Angst fest. Jamieson et al. (2010) registrierten ein geringeres Maß an Leistungszweifeln durch das reappraisal. Beltzer et al. (2014) ließen den Ausdruck von Angst und Scham während des TSST durch Probanden einschätzen und berichten positive Effekte auf beide Komponenten durch reappraisal. Die reappraisal-Gruppe in der Studie von Jamieson et al. (2016) gab ein niedrigeres Niveau an Mathematikängstlichkeit nach der Intervention an. Dieses empirische Bild macht deutlich, dass die mutmaßlich ablaufenden Prozesse beim reappraisal komplex sind. Denkbar ist, dass die Umbewertung von (unspezifischem) arousal die Entstehung von Angst verhindern kann. Es ist aber fraglich, ob die Wahrnehmung von körperlichen Erregungsprozessen in einer Stresssituation, entsprechend der Theorie von Gross und Thompson, nicht *unmittelbar* die Bewertung nach sich zieht, was dann in der Emotion Angst resultiert.

Das reappraisal von arousal baut wesentlich auf Befunden auf, die für positive Effekte von arousal (bzw.: bestimmten Formen von arousal) auf Leistung sprechen (siehe z. B. Dienstbier, 1989). Auf Ansätze zur Konzeptualisierung der positiven Aspekte von Testangst soll nun eingegangen werden.

1.3.2.1 Funktionale Aspekte von Testangst

Bereits in Abschnitt 1.1.1.2.3 wurde auf die Unterscheidung von hemmenden und erleichternden Aspekten von Testängstlichkeit nach Alpert und Haber (1960) im AAT eingegangen. Hembree (1988) inkludierte auch Studien, in denen der AAT genutzt wurde. Dabei zeigten sich positive Zusammenhänge zwischen AAT+ mit IQ zu $r = .30$ ($k = 3$, $N = 315$), mit anderen Leistungsmaßen zu $r = .29$ ($k = 23$, $N = 2.624$) und mit GPA zu $r = .32$ ($k = 9$, $N = 1.664$). Trotz der beachtlichen empirischen Befundlage zum AAT kommt dieser nur noch selten zum Einsatz. Bei aller berechtigter Kritik an den Items des AAT stellen Anderson und Sauser (1995) fest: „Whatever the controversies, the facilitating-debilitating properties of anxiety deserve more research and measurement attention“ (S. 19).

Zu den verschiedenen Ansätzen zur Erklärung von hemmenden und erleichternden Effekten von Testangst gehört die Auffassung, dass die Bewertung der körperlichen Erregungsprozesse entscheidet, in welche „Richtung“ sich Testangst auswirkt: „Accordingly, feelings of arousal may actually occur in both high- and low-test-anxious subjects, but they may be interpreted differen-

tially by different types of individuals or groups; this self-labeling of arousal as motivating or debilitating, respectively, may either facilitate or disturb behavior on cognitive tasks.“ (Zeidner, 1998, S. 208). Eine Differenzierung zwischen Intensität und Richtung der Angst findet sich insbesondere in der sportpsychologischen Literatur (z. B. Jones & Hanton, 2001; Jones & Uphill, 2004; Perry & Williams, 1998). Ob Angst förderlich oder hinderlich ist, hängt aus dieser Perspektive von der individuellen Interpretation der körperlichen und kognitiven Angstkomponenten ab (Jones, 1995).

Aus diesen Überlegungen folgt, dass Testängstlichkeit bzw. Testangst nicht immer negativ sein muss, sondern auch mit adaptiven bzw. positiven Eigenschaften und Verhaltensweisen in Verbindung steht. Hinweise darauf finden sich auch bei Betrachtung des nomologischen Netzwerks von Testängstlichkeit. In Abschnitt 1.1.1.3 wurde darauf eingegangen, dass Testängstlichkeit nicht nur mit Leistungsvermeidungs-, sondern auch mit Leistungsannäherungszielen korreliert (Elliot & McGregor, 1999). Auch korreliert Besorgtheit positiv mit Leistungsstreben (Musch & Bröder, 1999a) und (teilweise) positiv mit einem selbstgesetzten, hohen Leistungsstandard, also der eher adaptiven Komponente von Perfektionismus (Stoeber et al., 2009). Davey, Hampton, Farrell und Davidson (1992) argumentierten gar, dass die Konstrukte Ängstlichkeit und Besorgtheit (operationalisiert als Ausmaß an Sorgen in verschiedenen Lebensbereichen) voneinander trennbar sind. Besorgtheit ging in ihrer Studie (unter Kontrolle der Ängstlichkeit) eher mit adaptiven Bewältigungsmustern einher. So wies Besorgtheit (in diesem allgemeinen Sinne) positive Beziehungen zu problemorientiertem Coping auf, ausgedrückt z. B. durch die Suche nach Informationen über die jeweilige (Stress-)Situation oder Aktivitäten zur unmittelbaren Problemlösung. Aus Sicht von Davey et al. (1992) sind Sorgen ein bedeutsames Element in der Problembewältigung und –lösung und nicht zwingend an das Erleben von Angst gekoppelt.

Eine in diesem Kontext zentrale Studie stammt von Strack und Esteves (2014). Ausgangspunkt ist die Vermutung, dass appraisal und emotionale Reaktion nicht ausschließlich in einer unidirektionalen Kausalbeziehung zueinander stehen. In den bisherigen Ausführungen wurde von dem Fall ausgegangen, dass die Bewertung der Situation (die primäre Bewertung sensu Lazarus) die emotionale Reaktion bestimmt. In gewissermaßen „umgekehrter“ Richtung können appraisals jedoch aus Emotionen resultieren und sind zuweilen auch Bestandteil von Emotionen (Frijda & Zeelenberg, 2001). In ähnlicher Form findet sich dieser Gedanke im Modell der Emotionsentstehung nach Gross und Thompson (2007) in den Rückkopplungsprozessen. Daraus folgt, dass die Bewertung einer Emotion sich in der Bewertung der Situation niederschlagen kann. Je nachdem, ob die Emotion „Angst“ als hinderlich oder förderlich aufgefasst wird, wird eine Situation sensu Lazarus als Herausforderung (challenge) oder Bedrohung (threat) interpretiert (vgl. Abschnitt 1.1.1.1) (Strack & Esteves, 2014). Strack und Esteves (2014) erhoben bei einer studentischen Stichprobe

(*N* = 103) in den Tagen vor einer Prüfung täglich die erlebte Angst, wie bedrohlich und wie herausfordernd die anstehende Prüfung empfunden und inwiefern die etwaige erlebte Angst als förderlich wahrgenommen wurde. Es zeigte sich, dass zum einen die Relation von Angst und threat appraisal durch die Interpretation der Angst als förderlich moderiert wurde: der Zusammenhang von Angst und threat appraisal war stärker, wenn Angst nur in niedrigem Maße bzw. nicht als hilfreich erlebt wurde. Zum anderen wurde die Relation von Angst und challenge appraisal moderiert: wenn Angst als förderlich empfunden wurde, zeigte sich ein positiver Effekt der Angst auf die challenge appraisal, jedoch ein negativer Effekt wenn dies nur geringfügig der Fall war. Darüber hinaus korrelierte die Interpretation der Angst als erleichternd negativ mit der ebenfalls erfassten emotionalen Erschöpfung in den Tagen vor der Prüfung und positiv mit der Leistung in der Prüfung. Der Interpretation von Strack und Esteves (2014) zufolge wirkte sich die Bewertung der Emotion auf die Bewertung der Situation aus.

Strack, Lopes und Esteves (2014) schlugen vor, die motivierende Wirkung von Angst („anxiety motivation", kurz: AM) in eine informative („anxiety motivation information", kurz AM-Info) und eine energetisierende Funktion („anxiety motivation energy", kurz AM-Energie) zu differenzieren. Auf dieses Konstrukt und deren Herleitung durch Strack et al. (2014) soll kurz eingegangen werden. Angst als Reaktion auf eine (physische oder psychische) Bedrohung hat zweifellos eine wichtige informative Funktion. Aus regulationstheoretischer Perspektive ist negativer Affekt allgemein und Angst im Speziellen das Resultat der subjektiven Diskrepanz zwischen einem angestrebten und einem aktuell erlebten Zustand oder Ziel (Carver & Scheier, 1988, 1990), eine Annahme, die auch Bestandteil der S-REF Theorie ist (siehe Abschnitt 1.1.3). In dem Sinne, dass Selbstregulation eine Reduktion dieser Diskrepanz anstrebt, sind mit der informativen Funktion von Angst auch motivierende Auswirkungen verknüpft (in ähnliche Richtung argumentiert auch die feelings-as-information Theorie von Schwarz, 2012). Beispielsweise könnte eine Studentin in der Prüfungsvorbereitung feststellen, dass sie ein Prüfungsthema noch nicht so gut beherrscht wie sie es für notwendig empfindet, so dass die Motivation entsteht, sich abermals mit dem jeweiligen Thema auseinander zu setzen.

Die „energetisierende" Funktion von Angst basiert auf den Betrachtungen von Furcht bzw. Angst als Basisemotion, die prinzipiell eine (über)lebensnotwendige Bedeutung hat (siehe Abschnitt 1.1.1.1): „fear performs its basic function of motivating escape and alleviating fear-eliciting conditions" (Izard & Ackerman, 2000, S. 260). Bereits zu Beginn des 20. Jahrhunderts wurde die grundlegende Relevanz von Furcht und deren körperlicher Begleiterscheinungen für das Überleben angenommen (Cannon, 1929), wie auch die Beziehungen von arousal und Leistung erforscht wurden (Yerkes & Dodson, 1908). Die Annahme der „energetisierenden" Funktionen von Angst ist also keineswegs neu. Ein neuer Aspekt hingegen ist die Vorstellung, dass auch die *Bewertung* der Angst und deren motivationaler Bedeutung interindividuell divergiert. Eine Grundlage hierfür bildet die

Theorie der emotionalen Intelligenz, welche definiert ist als „the ability to carry out accurate reasoning about emotions and the ability to use emotions and emotional knowledge to enhance thought" (Mayer, Roberts & Barsade, 2008, S. 511). Das Four-Branch Modell der emotionalen Intelligenz postuliert als zweite Komponente die Nutzung von Emotionen zur Verbesserung des Denkens (Mayer, Salovey, Caruso & Sitarenios, 2001). Diese Fähigkeit umfasst das Utilisieren von Emotionen mit dem Ziel, kognitive Prozesse wie Problemlösung oder Urteilsbildung zu unterstützen. Das schließt auch die Generierung von Emotionen ein, die dabei helfen, bestimmten Situationsanforderungen gerecht zu werden (Papadogiannis, Logan & Sitarenios, 2009). Erforderlich hierfür ist Wissen darüber, wie sich Emotionen auf Denkprozesse auswirken und wie erstere dementsprechend reguliert werden sollten (Mayer et al., 2008). Parrot (2002) sieht das Verständnis und das Nutzen der funktionalen Aspekte negativer Emotionen als einen Ausdruck emotionaler Intelligenz und schildert Beispiele für die Emotion Angst. So kann eine Person dazu motiviert sein, ein gewisses Angstniveau zu halten, wenn es bei einer bestimmten Aufgabe entscheidend ist, vorsichtig zu sein. In ähnlicher Weise ist es denkbar, dass sich eine Person von einer übermäßig entspannten und heiteren Stimmung in eine ängstlichere Stimmung versetzt, wenn sie darin aus eigener Erfahrung zielstrebiger und besser lernen kann. Das „Nutzen" der Emotion Ärger in einer Verhandlungssituation ist ein anderes, bereits genanntes Beispiel (Tamir & Ford, 2012). Ein weiteres Beispiel sind Studierende, denen die Erstellung einer Hausarbeit nur unter Zeitdruck gelingt und die daher erst wenige Tage vor der Abgabefrist beginnen.

Positive Effekte von Angst finden sich auch in bereits behandelten Theorien. Die in Abschnitt 1.2.1.1.3 diskutierte Processing Efficiency Theory (PET) postuliert, dass Besorgtheit die Motivation erzeugt, einen Misserfolg bzw. dessen negative Konsequenzen abzuwenden. Diese Motivation regt die Nutzung zusätzlicher Verarbeitungsressourcen und Aktivierung von Verarbeitungsstrategien an, was beides zu einer Erhöhung der verfügbaren Arbeitsgedächtniskapazität beitragen kann. Diese erhöhte Anstrengung bewirkt eine Reduktion der Effizienz der Leistung, kann aber deren Effektivität sicherstellen (Eysenck & Calvo, 1992). Eine Studentin könnte also bei der Vorbereitung für eine Klausur – als Reaktion auf das Erleben von Prüfungsangst – bislang nicht genutzte Zeitressourcen für die Prüfungsvorbereitung investieren und gleichzeitig die eigenen Lernaktivitäten effizienter planen. Ausgehend von diesen Aspekten vermuten Strack et al. (2014), dass die interindividuell unterschiedliche Bewertung von Angst dazu führen kann, dass Angst selbst nicht als aversiv, sondern als anregend, motivierend, kurz als positiv wahrgenommen wird.

Strack et al. (2014) erhoben bei einer Stichprobe aus Lehrern und Ärzten (N = 86; Studie 1) im Abstand von einem Jahr die Ängstlichkeit, die emotionale Erschöpfung als Komponente von Burnout (Maslach, Schaufeli & Leiter, 2001) sowie eine von den Autoren entwickelte Skala zu AM. Diese bezog sich auf die Bewertung von erlebter Angst im Arbeitskontext („Feeling somewhat anxious about a task or a situation at work…") und differenzierte die informierende (z. B. „… reminds me

that I need to work“, ein Item zu AM-Info) und energetisierende (z. B. „… makes me more active in problem-solving“, ein Item zu AM-Energie) Wirkung von Angst.

Die Ängstlichkeit (t1) hatte einen positiven, AM-Energie und AM-Info (t1) einen negativen Effekt auf die emotionale Erschöpfung (t2). In zwei separaten Moderationsanalysen wurde gezeigt, dass sowohl AM-Info als auch AM-Energie den Effekt der Ängstlichkeit (t1) auf die emotionale Erschöpfung (t2) moderieren: Ängstlichkeit hatte nur einen positiven Effekt, wenn AM-Info respektive AM-Energie niedrig ausgeprägt war. Bei hoher AM-Info respektive AM-Energie fand sich kein signifikanter Effekt der Ängstlichkeit. Diese Befunde legen nahe, dass die motivierende Bewertung von Angst negative Auswirkungen derselbigen abschwächt. Eine Inklusion von AM-Info und AM-Energie in das Modell erbrachte ein komplexes Befundbild: nur AM-Energie zeigte einen signifikanten Haupteffekt auf die emotionale Erschöpfung und nur AM-Info moderierte den Effekt der Ängstlichkeit. In einer zweiten, experimentellen Studie bearbeitete eine studentische Stichprobe (N = 91) Intelligenztestaufgaben. Dabei erhielt eine Gruppe eine AM-Instruktion, die den geschilderten reappraisal-Interventionen ähnelte und die Probanden ermutigte, die mit dem Erleben von Angst verbundene Energie in die Bearbeitung der Aufgaben zu überführen. Die übrigen Probanden bearbeiteten den Test entweder mit der Instruktion, sich beim Empfinden von Angst auf den Test zu konzentrieren (task focus) oder ohne weitere Instruktion (Kontrollgruppe). Alle Probanden wurden unter einen vergleichsweise hohen Stress gesetzt, u. a. durch ein Gewinnspiel für die besten Leistungen, die Bestrafung von falschen Antworten in der Bewertung und die deutliche Verkürzung der eigentlich vorgesehenen Bearbeitungszeit. Auch erhielten die Probanden während der Bearbeitung fingierte Rückmeldungen, die den individuellen Leistungsrückstand gegenüber den Besten signalisierten (dies lässt vermuten, dass nicht nur Stress, sondern auch Misserfolgsgefühle induziert wurden). Die AM-Gruppe berichtete nach dem Test das niedrigste Niveau an emotionaler Erschöpfung und auch das höchste Maß an AM-Energie (state-Variante adaptiert aus Studie 1) bezogen auf den Test (die erlebte Angst wurde dabei nicht erfasst). Die zitierten Befunde von Strack und Kollegen legen nahe, dass AM möglicherweise eine Eigenschaft ist, in der sich Personen voneinander unterscheiden. Die schwachen bzw. moderaten Zusammenhänge von Ängstlichkeit zu AM-Info von r = .12 bzw. zu AM-Energie von r = -.37 legen außerdem nahe, dass es sich bei Ängstlichkeit und der Neigung, Angst als motivierend zu bewerten, nicht um identische Konstrukte handelt (Studie 1; Strack et al., 2014).

1.3.2.2 Implikationen und offene Fragen

Reappraisal als Kurzintervention erbrachte in einigen Studien positive Effekte auf Leistungsmaße, auf die subjektiv verfügbaren Bewältigungsressourcen in einer Leistungs- bzw. Stresssituation und teilweise auf das Angsterleben. Reappraisal-Manipulationen sind sehr einfach umzusetzen

und orientieren sich dabei am Erleben der jeweiligen Person – nur wenn (hemmende) Angst auftritt besteht der „Bedarf“ (oder das „Bedürfnis“), diese funktional um zu bewerten. Ferner gibt es Hinweise darauf, dass sich Personen darin unterscheiden, inwiefern sie habituell Angstsymptome als motivierend auffassen, was mit positiven Konsequenzen auf das Belastungserleben verbunden ist.

Die Betrachtung von reappraisal und der funktionalen Wirkungen von Angst bzw. Testangst negiert keineswegs die vielfache empirische Evidenz zu den negativen Begleiterscheinungen und Auswirkungen von Testängstlichkeit bzw. Testangst, sondern ergänzt diese vielmehr. Die Unterscheidung zwischen dem bloßen *Auftreten* und der *Bewertung* von Angstprozessen könnte erklären, warum Testangst nicht immer negativ mit Leistung einhergeht bzw. die Höhe dieses Zusammenhangs erheblich schwankt (siehe hierzu Seipp, 1991). In Analogie zur Differenzierung hemmender und erleichternder Angst nach Alpert und Haber (1960) ist es denkbar, dass Angst nur dann negative Effekte auf Leistung hat, wenn sie *nicht* als motivierend, sondern als blockierend aufgefasst wird. Sowohl die Studien zu reappraisal-Manipulationen als auch die von Strack und Kollegen vorgeschlagene Konzeption der anxiety motivation weisen in diese Richtung.

Die zitierten Untersuchungen lassen jedoch einige Fragen offen, die sich auf drei Sachverhalte konzentrieren. Erstens ist unklar, inwiefern AM als Disposition aufgefasst werden kann. Strack et al. (2014) weisen darauf hin, dass die Relation von AM zu anderen Konstrukten noch zu klären ist. Insbesondere dürfte eine Betrachtung der Relationen zu den Facetten von Testängstlichkeit und weiteren Persönlichkeitseigenschaften Aufschluss darüber geben, inwieweit es sich um ein eigenständiges Konstrukt handelt. Zweitens gibt es noch Forschungsbedarf in Bezug auf reappraisal-Manipulationen. Die zitierten Studien, welche entsprechende Manipulationen im Kontext kognitiver Leistungen bzw. von Tests und Prüfungen untersuchten, befassten sich nicht mit den Effekten der Manipulation auf die Relation von Angst und Leistung (Jamieson et al., 2010; Jamieson et al., 2016; Johns et al., 2008). Strack et al. (2014) applizierten zwar eine entsprechende Manipulation bei einer Intelligenztestaufgabe, analysierten aber keine Leistungseffekte. Da andere Kurzinterventionen wie das expressive Schreiben (Ramirez & Beilock, 2011) oder das Kompetenzpriming (Lang & Lang, 2010) eine Verringerung der Angst-Leistungs-Relation nachweisen konnten, ist zu klären, ob dies auch für reappraisal-Manipulationen gilt. Dies leitet zur dritten Frage über. Strack und Esteves (2014) berichten eine Moderation des Effekts von Ängstlichkeit auf threat und challenge appraisal durch die funktionale Interpretation von Angst. Analog hierzu ist es theoretisch plausibel, dass auch die Relation von Angst und Leistung durch die funktionale Interpretation von Angst – die überdies experimentell induziert werden könnte – moderiert wird. So weist Testangst möglicherweise *nur dann* einen negativen Effekt auf Leistung auf, wenn sie als hinderlich empfunden wird. Offen ist, welcher Effekt sich zeigt, wenn Angst als förderlich empfunden wird. Denkbar

ist einerseits ein nicht signifikanter Effekt. Andererseits könnten auch positive Effekte von Testangst auf Leistung vorliegen. In Anbetracht der positiven Leistungseffekte der im AAT erfassten erleichternden Ängstlichkeit (Hembree, 1988) ist dies nicht abwegig. Diese Moderationshypothese lässt sich sowohl auf das situative als auch das dispositionelle Erleben von Testangst übertragen. Inwiefern also das situative Erleben von AM und – in Verbindung mit der ersten Frage – die dispositionelle Neigung zu AM als Moderatoren fungieren, ist zu klären.

Zur Prüfung dieser Fragen ist ein Design notwendig, in dem zum einen dispositionale Tendenzen zu AM erfasst und zu anderen Persönlichkeitseigenschaften, insbesondere Testängstlichkeit, in Relation gesetzt werden. Zum anderen ist zum Vergleich mit anderen Kurzinterventionen ein experimentelles Design mit einer reappraisal-Manipulation erforderlich. Dabei muss die während des Tests erlebte, leistungsförderliche Wirkung von Angst erfasst werden. Diesen Fragen wurden in Studie 3 untersucht.

2. Fragestellungen

Kern der drei in dieser Arbeit vorgestellten Studien war die Relation von Testängstlichkeit bzw. Testangst und kognitiver Leistung. Drei Bereiche von Fragestellungen wurden jeweils in einer gesonderten Studie untersucht. Bei den folgenden Erläuterungen der jeweiligen Fragestellungen werden auch die Neuerungen gegenüber der bisherigen Forschung geschildert.

2.1 Fragestellung 1: Variation des Messzeitpunkts von Testangst

Die erste Studie basierte auf den in Abschnitt 1.2.1.3.2 geschilderten Fragen. Ausgangspunkt war dabei die Moderation der Relation von Testängstlichkeit bzw. Testangst und Leistung durch den Messzeitpunkt. Die umfangreiche Forschung zur Testängstlichkeit hat sich mit der Erklärung dieses Effekts bislang nur unzureichend beschäftigt. Aus diesem Grund ist es für Forschung und praktische Anwendung nicht klar, wann Testängstlichkeit bzw. Testangst in einer Leistungssituation erfasst werden sollte. Zum einen ist nicht hinreichend geklärt, welche Auswirkungen die Erfassung von Testangst vor einem Test hat. Konsequenzen könnten sich auf verschiedenste Maße des „Testerlebens“ ergeben. „Testerleben“ bezeichnet hier als Arbeitsbegriff, wie ein Test erlebt und bewertet wird und darüber hinaus die objektive Leistung im Test. Zum anderen wirft die Literatur die Frage auf, wie die vor und nach dem Test berichtete Testangst zu interpretieren ist. Ein Beitrag zur Klärung würde bereits geleistet, indem verglichen würde, wie die vor und nach einem Test berichtete Testangst (state) mit dispositioneller Testängstlichkeit und damit in Verbindung stehenden Dispositionen zusammenhängt. Das Erleben (und der Bericht) von Testangst findet nicht in einem „Vakuum“ statt, sondern ist im Zusammenhang zu anderen psychologischen Prozessen zu sehen. Von besonderer Relevanz sind hierbei (Emotions-)Regulationsprozesse. Testangst kann dabei sowohl ein Bestandteil von Regulationsprozessen sein, aber auch ein Ergebnis dieser. Zweifellos steht der Bericht von Testangst nach einem Test unter dem Eindruck des Testerlebens. Die Relevanz von Regulationsprozessen ist also insbesondere für die *nach* dem Test berichtete Testangst zu klären. Eine schematische Darstellung dieser Konstellation an Fragen findet sich in Abbildung 7.

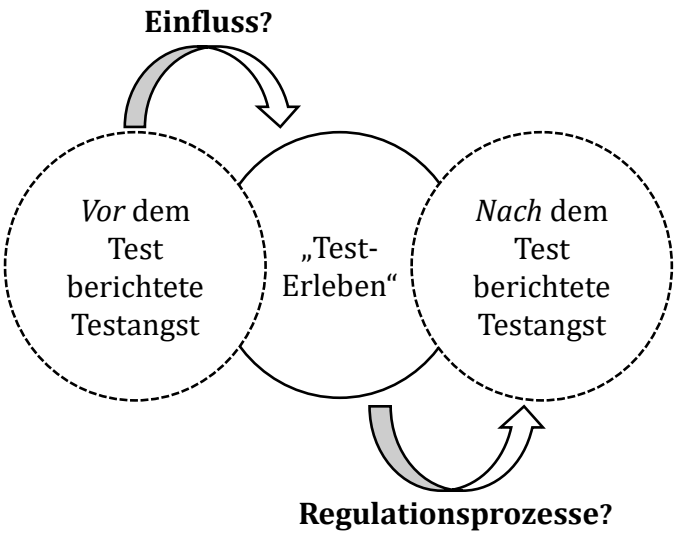

Abbildung 7: Schematische Darstellung der Fragestellungen von Studie 1

Diese Fragestellungen wurden in Studie 1 (Abschnitt 4) untersucht. Die explorativen Analysen und Hypothesen basierten auf der folgenden, impliziten Grundannahme: der Bericht von Testangst kann teilweise als Ausdruck von Regulationsprozessen verstanden werden. Der Bericht von Testangst erlaubt, einen Misserfolg selbstwertdienlich zu attribuieren. Die genaue Bedeutung von Regulationsprozessen für die Interpretation von Testangstwerten ist jedoch klärungsbedürftig.

2.2 Fragestellung 2: Testangst und stereotype threat

Die Fragestellung von Studie 2 basierte auf den in Abschnitt 1.2.2.2.4 dargestellten Fragen und der Heterogenität der Befunde zur Erklärung des STT. Für diese Fragestellung wurde der STT bei Frauen in Bezug auf deren mathematische Begabung herangezogen. Somit waren (da es sich um eine Schülerstichprobe handelte) die targets Mädchen, die nontargets Jungen. Im Fokus stand dabei, ob Testangst einen Beitrag zur Erklärung von STT-Effekten leistet. In Erweiterung der bisherigen Literatur sollten dabei bestimmte Untersuchungsparameter gezielt berücksichtigt werden. Zum einen sollte die Testangst auf Facettenebene erfasst werden, wobei sich die Betrachtung auf den plausibelsten Mediator Besorgtheit konzentrierte. Dadurch sollte eine Konfundierung insbesondere von kognitiven und affektiven Angstprozessen vermieden werden. Außerdem wurden neben targets auch nontargets untersucht. Grund hierfür ist, dass eine theoretische Erklärung des STT auch in der Lage sein müsste, dessen Abwesenheit bei nontargets zu begründen. Falls beispielsweise Testangst den STT mediiert, sollte dies nur für Frauen, nicht aber für Männer zutreffen. Der STT sollte außerdem mit einer subtilen Aktivierung herbeigeführt werden, da explizite Aktivierungen (z. B. „Es ist bekannt, dass Frauen in diesem Test schlechter abschneiden“) in der praktischen Anwendung von Prüfungen und Tests seltener sein dürften (Nguyen & Ryan, 2008). Auch unter der Annahme, dass diese in der Praxis auftreten, würde eine derartige Instruktion bei

gängigen, psychometrischen Testverfahren meist eine Verletzung der im Testmanual vorgesehen Durchführungsanweisungen bedeuten. Subtile Aktivierungen sollten hingegen wesentlich häufiger in der Praxis, also außerhalb kontrollierter Untersuchungen, auftreten. Diese können entweder durch das Priming der eigenen Gruppenzugehörigkeit stattfinden (z. B. indem man in einem vorab ausgegebenen Fragebogen das eigene Geschlecht angibt) oder durch die Mitteilung des jeweils getesteten Bereichs: „Emphasizing test diagnosticity purpose. For example, labeling the test as a diagnostic test or stressing the evaluative nature of the test" (Nguyen & Ryan, 2008, S. 1316). Betrachtet man den STT aus Perspektive der Testängstlichkeitsforschung, so handelt es sich dabei um eine bestimmte Interaktion von Instruktionsvariation und getesteter Domäne. Eine Aufgabe wird als aussagekräftig für eine bestimmte Fähigkeit beschrieben, in welcher targets gemäß eines Stereotyps schlechter sind. Tabelle 11 vergleicht diese Instruktion mit der im Paradigma nach Sarason gängigen Induktion von Bewertungsstress (siehe Abschnitt 1.1.2 sowie 1.2.2.1).

Tabelle 11: Vergleich einer subtilen STT-Aktivierung mit einer evaluativen Instruktion sensu Sarason

Instruktion	Operationalisierung	Quelle
STT subtil*	„participants learned that they would be taking a test diagnostic of their "genuine math abilities" that could indicate their "strengths and weaknesses" in the quantitative domain."	Johns et al. (2008, S. 696)
Evaluativ**	„subjects were also told that how much information a person possesses and how that information is used are important aspects of intelligence."	Sarason et al. (1986, S. 218)

* targets = Frauen; ** stressinduzierende Instruktion eines Wissenstests

Der Vergleich macht deutlich, wie nahe diese Instruktionsformen beieinander liegen. Wesentliche Gemeinsamkeit ist die Beschreibung eines Tests als diagnostisch valide bezüglich der individuellen Fähigkeiten. Eine subtile STT-Induktion ist nun dadurch gekennzeichnet, dass dabei ein gemäß Stereotyp bei targets schwacher Fähigkeitsbereich adressiert wird (z. B. Mathematik bei Frauen). Gewissermaßen wird einer evaluativen Instruktion ein „stereotyper Charakter"[30] verliehen (siehe auch Abschnitt 1.2.2.2.4). Aufgrund dieser Situation ist es untersuchungstechnisch sinnvoll, sowohl eine evaluative, stereotype als auch eine evaluative, nicht-stereotype Instruktion zu implementieren. Fügt man eine nonevaluative Bedingung hinzu, ergibt sich ein experimentelles Design mit drei Bedingungen:

a) Eine evaluative Bedingung, die stereotyp formuliert ist
b) Eine evaluative Bedingung, die stereotypneutral formuliert ist

[30] Im Folgenden wird eine derartige, subtile STT-Aktivierung verkürzt als Instruktion mit „stereotypem Charakter" bezeichnet. Dies drückt aus, dass die Testinstruktion durch die Mitteilung des getesteten Bereichs ein bestimmtes Stereotyp aktivieren soll (und bei targets also den STT auslösen soll).

c) Eine nonevaluative Bedingung, die ebenfalls stereotypneutral formuliert ist

Der Vergleich von Bedingung a) und b) ermöglicht eine Beobachtung der Prozesse bei evaluativen Instruktionen mit und ohne stereotypem Charakter. Diese Unterscheidung ist auch deswegen nötig, weil die typische Induktion von Bewertungsstress – je nachdem welcher Fähigkeitsbereich gemäß Instruktion getestet wird – unbeabsichtigt einen STT auslösen kann. Der Vergleich von Bedingung b) und c) hingegen ist an die Untersuchungen von Sarason angelehnt. Im Idealfall sollte es durch diese drei Bedingungen möglich sein, die Effekte einer stressvollen Instruktion mit (a) und ohne stereoptypem Gehalt (b) zu vergleichen. Die Fragestellungen von Studie 2 sind schematisch in Abbildung 8 dargestellt[31]. Diese Fragestellungen wurden in Studie 2 (Abschnitt 5) untersucht.

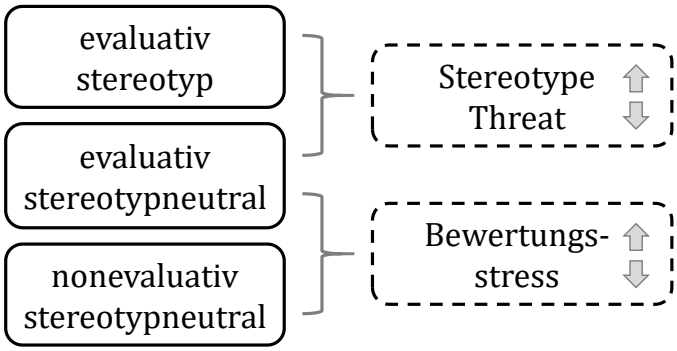

Abbildung 8: Schematische Darstellung der Fragestellungen von Studie 2

2.3 Fragestellung 3: Motivierende Wirkung von Angst

Studie 3 befasste sich mit den in Abschnitt 1.3.2.2 geschilderten Fragen. Die Studie griff dabei die in der Testängstlichkeitsforschung lange Zeit vertretene Idee auf, dass es leistungsförderliche Aspekte von Testängstlichkeit bzw. Testangst gibt. Grundlage sind theoretische Argumente für funktionale Wirkungen von Testängstlichkeit bzw. Testangst sowie Befunde zu den positiven Effekten von reappraisal-Interventionen. Kerngedanke war die Differenzierung des „reinen" Erlebens von Testangst von dessen Bewertung (siehe Abschnitt 1.3.2.1). Hierfür wurde das von Strack et al. (2014) vorgeschlagene Konzept der anxiety motivation (AM) aufgegriffen. Da sich der Fragebogen von Strack et al. (2014) auf das Angsterleben in beruflichen Situationen bezieht, war eine Adaption für den Kontext von Prüfungen und Tests nötig. Wie bereits erwähnt steht dieser Ansatz keineswegs im Gegensatz zu den berichteten negativen Zusammenhängen von Testängstlichkeit, Testangst und Leistung sowie dem maladaptiven Charakter von Testängstlichkeit. Die besagte Unterscheidung zwischen Erleben und Bewertung von Testangst könnte aber zu einem adäquateren

[31] Eine nonevaluative, stereotype Bedingung wurde nicht aufgenommen, da deren Erkenntniswert niedrig wäre. Wenn eine nonevaluative Situation nur wenig Bewertungsstress induziert, dürfte es unerheblich sein, ob die entsprechende Instruktion stereotyp oder stereotypneutral ist.

Verständnis der eigentlichen Wirkung von Testangst beitragen. Der auf Alpert und Haber (1960) zurückgehende Grundgedanke von hemmender und erleichternder Angst soll dadurch einer Revision unterzogen werden. Die Studie sollte dabei zum einen Aufschluss geben über die Bedeutung und den theoretischen Nutzen von AM. Neben den unklaren Zusammenhängen zu anderen Dispositionen sollte auch, ausgehend von den Befunden von Strack und Esteves (2014) sowie Strack et al. (2014) untersucht werden, inwiefern AM den Zusammenhang von Testängstlichkeit bzw. Testangst und unterschiedlichen Leistunsmaßen bzw. -kriterien moderiert. Dieser Teil der Fragestellung von Studie 3 ist in Abbildung 9 veranschaulicht.

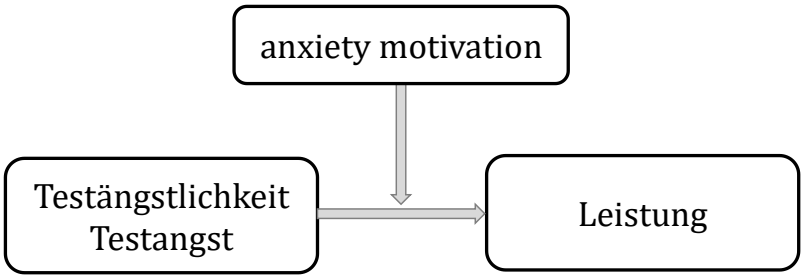

Abbildung 9: Schematische Darstellung einer der Fragestellungen von Studie 3

Zum anderen sollte im Kontext der Forschung zu Kurzinterventionen geprüft werden, inwieweit eine kurze reappraisal-Manipulation zu einer Veränderung der Leistung und der erlebten Testangst führt und ob die Intervention, ähnlich wie andere Kurzinterventionen (Lang & Lang, 2010; Ramirez & Beilock, 2011), den negativen Zusammenhang von Testangst und Leistung aufhebt. Zur Untersuchung eignete sich ein experimentelles Design mit einer reappraisal-Manipulation vor einem Leisungstest. Ferner war die Erhebung von Testängstlichkeit und Testangst sowie der dispositionellen AM erforderlich. Für eine Analyse der Prozesse, die während des Tests stattfanden, musste auch die bezüglich des Tests empfundene leistungsförderliche Wirkung von Angst (state-anxiety motivation) erfasst werden. Diese Fragestellungen wurden in Studie 3 (Abschnitt 5) behandelt.

Im Anschluss wird in den Abschnitten 4, 5 und 6 jede der drei Studien gesondert beschrieben. Dabei werden zunächst die Methoden und die Ergebnisse dargestellt und anschließend die Ergebnisse diskutiert. Zuvor werden in Abschnitt 3 einige methodische Vorgehensweisen beschrieben, die für alle drei Studien gelten.

3. Methodische Vorbemerkungen

Die folgenden Punkte betreffen die Darstellung in allen drei Studien. Die relevanten deskriptiven Statistiken der verwendeten Skalen werden immer zusammen mit den jeweiligen Reliabilitätswerten berichtet. Dabei wird Cronbach's Alpha als Kennwert angegeben. Das vom Commitee of Test Affairs of the Dutch Association of Psychologists vorgeschlagene Bewertungssystem (COTAN; Evers, 2001) wird genutzt, um die Höhe der Reliabilität grob zu bewerten. Im COTAN-System werden dabei unter anderem Grenzwerte beim Einsatz von Verfahren für Forschungszwecke und zur Gewinnung von Informationen auf Gruppenebene genannt (sogenanntes „Niveau 3"). Hierbei wird ein Reliabilitätskoeffizient von < .60 als „unzureichend", zwischen .60 und .70 als „ausreichend" und über .70 als „gut" eingestuft (Evers, 2001). Darüber hinaus werden die einzelnen Subskalen von Verfahren zur Erfassung von Testängstlichkeit und Testangst vereinfachend als „Facetten" bezeichnet, bei Beschreibung statistischer Eigenschaften wird auch auf den Subskalenbegriff zurückgegriffen. In den meisten Analysen wurde dabei ein Fokus auf die im Leistungskontext zentrale Facette Besorgtheit (worry) gelegt. Geschlechtsunterschiede wurden für demographische Variablen (z. B. Noten) sowie für Dispositionen (traits) analysiert, für weitere Variablen nur wenn dies Gegenstand der Fragestellung war. Bei den deskriptiven Statistiken wird zusätzlich der Median berichtet, wenn die Schiefe der jeweiligen Variablen den Bereich von ± .50 überschreitet.

Bei der Berechnung der (Sub-)Skalenwerte wurde jeweils ein konservatives Vorgehen gewählt: Skalenwerte wurden nur gebildet, wenn die vollständige Skala bearbeitet wurde. Die dadurch bedingte schwankende Fallzahl wird jeweils mit berichtet. Vergleiche der in den Untersuchungen vorliegenden Skalenmittelwerte mit Referenzwerten aus den Originalquellen oder sonstigen Untersuchungen werden berichtet, wenn diese für die Interpretation der in den Studien gefundenen Werte hilfreich sind. Alle Werte sind auf zwei Nachkommastellen gerundet, mit Ausnahme der p-Werte (drei Nachkommastellen). Die Konfidenzintervalle der mit PROCESS (Hayes, 2013) berechneten Analysen sind auf drei Nachkommastellen gerundet aufgeführt, wenn dadurch die exakte Grenze des Intervalls erkennbar wird oder durch die Rundung auf zwei Nachkommastellen wichtige Information verloren gehen würde.

Für den Vergleich von Korrelationskoeffizienten wurde zusätzlich eine Prüfung auf signifikante Unterschiede mit dem Online-Tool cocor vorgenommen (http://comparingcorrelations.org/; Diedenhofen & Musch, 2015). Korrelationsvergleiche wurden nur dann durchgeführt, wenn sie statistisch möglich waren: Vergleiche überlappender Korrelationen (Vergleich von r.jk und r.jh) wurden nur für die Teilstichproben berechnet, in denen alle relevanten Variablen erhoben wurden. In den übrigen Fällen wurde ein deskriptiver Vergleich der Korrelationen vorgenommen. Für alle berichteten Mittelwertsunterschiede werden jeweils die Effektstärken mitberichtet. Diese ermög-

lichen es, die Größe eines Mittelwertsunterschieds unabhängig von der Stichprobengröße einzuschätzen (Fritz, Morris & Richler, 2012). Somit ergänzen sie die Information des Signifikanztests, welcher zudem bei zahlreichen statistischen Tests von der Kumulation des alpha-Fehlers betroffen ist. Als Effektmaß für Mittelwertsunterschiede zwischen zwei Gruppen verschiedener Größe wurde Hedges' *g* genutzt, das in der Literatur häufig vereinfachend als *d* bezeichnet wird (Fritz et al., 2012). Zur Berechnung wurde das Online-Tool von Lenhard und Lenhard (2016) (https://www.psychometrica.de/effektstaerke.html) verwendet. Die Interpretation der Effektgröße folgt der Empfehlung von Cohen (1988), wonach $d = .20$ einen kleinen, $d = .50$ einen mittelgroßen und $d = .80$ einen großen Effekt bezeichnet (zitiert nach Eid, Gollwitzer & Schmitt, 2013).

Alle Abkürzungen, sofern nicht allgemein gängig (wie z. B. *M*, *SD*) werden im Abkürzungsverzeichnis (Anhang A) aufgeführt. Um die Maße für Testängstlichkeit (trait) und Testangst (state) einfacher zu unterscheiden, wird im Folgenden ersteres als trait-TÄ, letzteres als state-TA abgekürzt. „Probanden" wurden mit „Pbn" abgekürzt. In den Tabellen zum Bericht der deskriptiven Statistiken und der Reliabilitäten wird jeweils auf die zugehörige Skala verwiesen. Beim tabellarischen Bericht von Interkorrelationen wird vereinfachend auf die zu erfassenden Konstrukte verwiesen, gleichwohl selbstverständlich die jeweiligen Skalen gemeint sind.

Alle betrachteten Variablen wurden auf Normalverteilung geprüft (siehe Anhang F). Da in den allermeisten Fällen keine Normalverteilung vorlag, wurde entweder auf Verfahren zurückgegriffen, die keine Normalverteilung voraussetzen, oder aber gegen Voraussetzungsverletzungen robust sind. Skaleninterkorrelationen sowie weitere Korrelationskoeffizienten sind dabei, wenn nicht anders angegeben, mit der Rangkorrelation nach Spearman berechnet worden. T-tests für unabhängige Stichproben wurden aufgrund ihrer Robustheit auch bei Verletzung der Normalverteilungsannahme eingesetzt (Bortz & Schuster, 2010). Bei allen multiplen Regressionsanalysen wurden ebenfalls die Voraussetzungen geprüft, was die Prüfungen auf Multikollinearität, Homoskedastizität und Normalverteilung der Residuen beinhaltet (Eid et al., 2013). Bezüglich der Multikollinearität wurden für den Toleranz-Wert sowie den Varianz-Inflations-Faktor (VIF) die von Urban und Mayerl (2011) empfohlenen Grenzwerte herangezogen. Da auch die Moderations- und Mediationsanalysen auf multiplen Regressionen beruhen, wurde deren Voraussetzung jeweils analog geprüft. Ergebnisse dieser Prüfungen und etwaige Auffälligkeiten sind in Anhang F vermerkt. Die Mediationen und Moderationen wurden mit der SPSS-Macro PROCESS von Hayes (2013) berechnet. Die Grundlagen dieser Prozeduren sind im Folgenden kurz geschildert.

Mediation und Moderation mit PROCESS

Der „klassische“ Ansatz zur Prüfung von Mediationen geht auf Baron und Kenny (1986) zurück. Dieser beruht darauf, dass die Signifikanz verschiedener Effekte nacheinander geprüft wird, weswegen er auch „causal steps aproach“ (Hayes, 2013) genannt wird. Betrachtet man *X* als unabhängige Variable, *Y* als abhängige Variable und *M* als Mediator, so muss zunächst ein signifikanter Effekt *c* von *X* auf *Y* vorliegen. Ist dies erfüllt, muss der Effekt *a* von *X* auf *M* signifikant sein. Die dritte Bedingung ist, dass der Effekt *b* von *M* auf *Y* signifikant ist, unter Kontrolle des Effekts von *X* (dieser Effekt von *X* wird als *c‘* bezeichnet). Dies wird geprüft durch eine multiple Regression von *Y* auf *X* und *M*. Sofern *c‘* signifikant, aber (im Betrag) kleiner als *c* ist, liegt eine partielle Mediation vor. Ist c‘ kleiner als *c* und nicht signifikant, liegt eine vollständige Mediation vor – der Effekt von *X* auf *Y* geht dann gänzlich auf *M* zurück (Hayes, 2013). Hayes (2013) führt mehrere Probleme dieser Methodik auf. Erstens beinhaltet dieses Vorgehen keine inferenzstatistische Prüfung des indirekten Effektes. Zweitens setzt eine Mediation unter diesem Ansatz die Signifikanz von drei Effekten voraus (*c*, *a* und *b*), wodurch die Power stark absinkt. Drittens lässt sich logisch widerlegen, dass *X* einen Effekt auf *Y* haben *muss*, damit eine Mediation vorliegt. Dies kann beispielsweise der Fall sein, wenn zwei Mediatoren existieren, über die jeweils gleich starke, aber im Vorzeichen entgegengesetzte, indirekte Effekte vorliegen. Hierbei würde ein totaler Effekt von $c = 0$ resultieren.

In den vorliegenden Studien wurde daher der alternative Ansatz, die „conditional process analysis“ (Hayes, 2013), eingesetzt. Dabei werden direkter Effekt (*c‘*) und indirekter Effekt (das Produkt aus *a* und *b*) simultan geschätzt, wobei der totale Effekt *c* von nachrangiger Bedeutung ist. Die Signifikanz des indirekten Effekts wird dabei durch die bootstrapping-Methode getestet. Bootstrapping dient dazu, die Stichprobenkennwerteverteilung des indirekten Effekts *ab* zu ermitteln. Dabei wird *k*-mal aus der Stichprobe *N* eine sog. „bootstrap sample“ *n* zufällig (und mit Zurücklegen) gezogen und der Effekt geschätzt, wobei *k* bis zu 10.000 sein kann. Jede *n* entspricht dabei im Umfang der tatsächlichen Stichprobengröße *N*, durch das „Zurücklegen“ von Fällen unterscheiden sich jedoch die einzelnen $n_{1...k}$ voneinander[32]. Auf Basis der daraus resultierende Verteilung von *ab* kann ein 95%-iges Konfidenzintervall ermittelt werden. Sofern dieses Konfidenzintervall 0 nicht umschließt, gilt die Signifikanz des indirekten Effekts auf dem alpha-Fehlerniveau von .05 als abgesichert. Im Gegensatz zu dem früher gebräuchlichen Sobel-Test setzt die bootsrapping-Methode keine Normalverteilung von *ab* voraus (Hayes, 2013). Auf Basis der Empfehlung von Hayes (2013), dass 5.000 bis 10.000 bootstraps üblicherweise ausreichen, wurde in den nachfolgenden Analysen standardmäßig $k = 5.000$ gesetzt.

[32] So kann im ersten bootstrap sample ein Fall mehrfach enthalten sein, in einem zweiten jedoch fehlen.

PROCESS erlaubt auch eine Berechnung von Moderationseffekten. Im Falle einer einfachen Moderation wird eine multiple Regression der abhängigen Variable *Y* auf die unabhängige Variable *X*, den Moderator *M* und den Interaktionsterm *XM* berechnet. Für *X*, *M* und *XM* werden so Regressionsgewichte b_1, b_2 und b_3 ermittelt. b_1 beschreibt dabei den konditionalen Effekt von *X* auf *Y* wenn $M = 0$, analog ist b_2 der konditionale Effekt von *M* auf *Y* wenn $X = 0$. Ferner gilt, dass der Unterschied im vorhergesagten ŷ-Wert zweier Fälle, die sich in *X* um eine Einheit unterscheiden, sich um b_3 verändert, wenn sich *M* um eine Einheit verändert. Die Interpretation (und graphische Darstellung) der Interaktion erfolgt dabei häufig über die „simple slopes", also den Regressionskoeffizienten von *X* bei einem bestimmten Wert von *M* (Cohen, 2010). Bei einer kontinuierlichen Variable *M* wird dabei für gewöhnlich der konditionale Effekt von *X* betrachtet, wenn *M* den jeweiligen Mittelwert annimmt sowie eine Standardabweichung über und unter dem Mittelwert liegt (bei Mittelwertszentrierung $M = 0$ sowie $M = \pm 1\ SD$) (Hayes, 2013). In den folgenden Analysen werden daher auch simple slopes berichtet. Bei der Interpretation von simples slopes gilt es jedoch bestimmte Aspekte zu beachten. So sind die Ausprägungen von *M* letztlich willkürlich gewählt. Ferner spielt für die Interpretation der simple slopes auch die Verteilung von *M* eine Rolle – so sollte berücksichtigt werden, ob *M* schief verteilt ist, da eine „hohe" Ausprägung im Sinne der simple slopes de facto „nur" einer moderaten Ausprägung der Variable entsprechen könnte (z. B. im Falle eines Bodeneffekts). Eine Lösung bietet die Johnson-Neyman-Technik (JN-Technik). Diese Prozedur generiert „regions of significance", also den Ausprägungsbereich von *M*, in welchem *X* einen signifikanten Effekt auf *Y* aufweist[33]. Bei Ausprägung von *M* außerhalb dieses Bereichs ist kein signifikanter Effekt gegeben. Dementsprechend wird in PROCESS die JN-Technik auch nicht durchgeführt, wenn der Effekt von X entweder auf allen Ausprägungen *oder* auf keiner Ausprägung von *M* signifikant ist. Ist der Moderator dichotom, dann wird in PROCESS der jeweilige Effekt von *X* auf den Stufen von *M* ausgegeben (Hayes, 2013). In erster Linie zur Erleichterung der Interpretation der Koeffizienten, wurden in den folgenden Analysen die kontinuierlichen Prädiktoren in den Modellen mittelwertszentriert (siehe hierzu Hayes, 2013).

Auf Grundlage der einfachen Mediation mit einem Mediator bzw. der einfachen Moderation mit einem Moderator ermöglicht PROCESS durch die automatische Berechnung von Produkttermen bzw. Interaktionen eine Reihe komplexerer Analysen. In Studie 1 und 2 wurde der Fall der moderierten Mediation betrachtet. Hierbei wird die Abhängigkeit eines Mediationseffektes von der Ausprägung eines Moderators geprüft. Somit kann unter Einbezug des Mediators *M* und des Moderators *W* geprüft werden, ob *W* den Effekt von *X* auf *M*, von *M* auf *Y* sowie von *X* auf *Y* moderiert.

[33] So könnte X einen signifikanten Effekt haben, wenn M einen bestimmten Wert M_1 unterschreitet oder überschreitet, oder zwischen zwei Werten M_1 und M_2 liegt, etc.

Abbildung 10 zeigt die Gegenüberstellung vom konzeptuellen und statistischen Modell einer moderierten Mediation mit einem Mediator und einem Moderator (Model 59; Hayes, 2013).

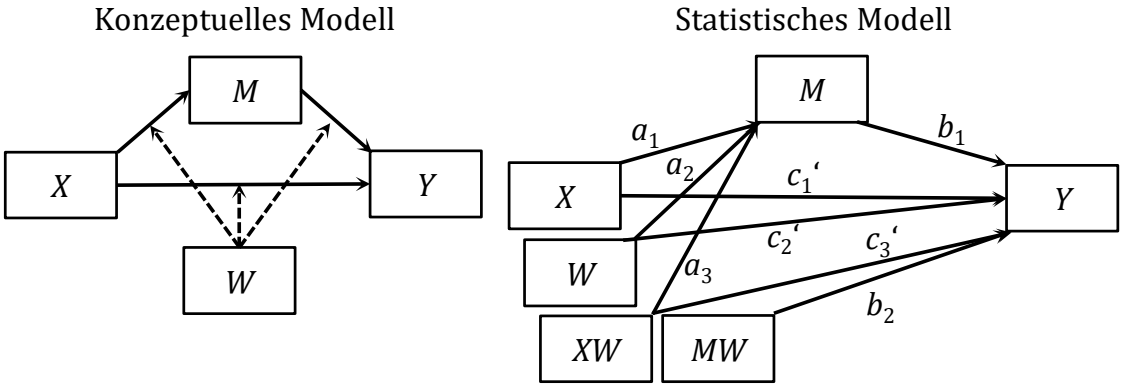

Abbildung 10: Konzeptuelles und statistisches Modell einer moderierten Mediation (Hayes, 2013, S. 455)

Die Betrachtung des konzeptuellen Modells verdeutlicht, dass *W* jeden der Teileffekte moderiert. Aus dem statistischen Modell wird die in PROCESS stattfindende Operation deutlich. Mit einer multiplen Regression wird zum einen *M* vorhergesagt (durch *X*, *W* und *XW*) und anschließend Y (durch *X*, *W*, *XW*, *MW* und *M*). Sowohl direkte als auch indirekte Effekte werden anschließend auf verschiedenen Ausprägungen des Moderators getestet. PROCESS berichtet die unstandardisierten Regressionskoeffizienten, weshalb diese auch in den Ergebnissen in dieser Arbeit berichtet werden.

In Studie 1 sowie in Studie 3 wurden darüber hinaus mit PROCESS moderierte Moderationen gerechnet, also Modelle mit 2 Moderatoren. Diese werden im Kontext von Studie 3 im entsprechenden Abschnitt 6.1.6 beschrieben.

4. Studie 1

4.1 Methode

4.1.1 Kurzüberblick

Ausgangspunkt für Studie 1 war die Moderation des Zusammenhangs von Testangst und Leistung durch den Messzeitpunkt von Testangst. Ziel war zu untersuchen, welchen Effekt die Erfassung von Testangst vor einem Test hat. Darüber hinaus sollte näher betrachtet werden, ob die Testangst vor und nach einem Test in ähnlicher (oder unterschiedlicher) Weise mit bestimmten Eigenschaften und Prozessen in Zusammenhang steht. Drittes Ziel war es, den Zusammenhang der nach einem Test berichteten Testangst und der Leistung näher zu beleuchten.

Studie 1 wurde als Online-Studie durchgeführt. Bei allen Probanden wurden zu Beginn der Selbstwert (SW), die Testängstlichkeit (trait-TÄ), die Selbstwertkontingenz (SWK) und das Selbstkonzept eigener Fähigkeiten (SKEF) erhoben. Kern der Studie bildete die Manipulation: eine Gruppe wurde vor einem kurzen Intelligenztest nach der aktuell erlebten Testangst (state-TA prä) befragt (Bedingung A), die andere Gruppe nicht (Bedingung B). Vor dem Test sollten alle Pbn ihre erwartete Leistung (Anspruchsniveau; AN) angeben. Die Pbn beider Bedingungen wurden im Anschluss an den Test nach ihrer während des Tests erlebten Testangst befragt (state-TA post). Darüber hinaus wurden die Akzeptanz des Verfahrens, subjektiv die Leistung beeinträchtigende Aspekte bzw. self-handicapping (SH-Aspekte) sowie demographische Variablen erhoben.

4.1.2 Herleitung der explorativen Analysen und Hypothesen

Die explorativen Analysen und Hypothesen von Studie 1 ergeben sich unmittelbar aus den in Abschnitt 2.1 geschilderten Fragestellungen. Nach einer kurzen Herleitung werden die explorativen Analysen und Hypothesen konkret formuliert.

Im Rahmen einer explorativen Analyse sollte geprüft werden, welche Effekte die Manipulation, also die Erhebung von Testangst vor einem Test (state-TA prä), auf das Testerleben hat. Hierzu wurden mehrere Variablen zum Erleben vor und während des Tests (retrospektiv) betrachtet, konkret das Anspruchsniveau (vor dem Test) und die Akzeptanz sowie self-handicapping (SH) in Bezug auf den Test (nach dem Test). Auch die retrospektiv während des Tests berichtete Testangst (state-TA post) sowie die objektive und die subjektive Leistung wurden in die Analyse einbezogen.

> **Explorative Analyse:** Analyse der Gruppenunterschiede zwischen Bedingung A und B bezüglich Anspruchsniveau, der Testleistung und der state-Testangst post (alle Facetten), Akzeptanz und self-handicapping-Aspekten.

Die Hypothesen 1a und 1b gingen der Frage nach, durch welche Prozesse die state-TA vor und nach einem Test jeweils beeinflusst wird, wobei die Beziehungen von state-TA prä und post zu stabilen Dispositionen sowie zu SH-Aspekten untersucht wurden. Es ist wahrscheinlich, dass die state-TA prä eher die Erwartungen bezüglich des Tests, die state-TA post hingegen das Erleben während des Tests widerspiegelt. Ausgehend von den Befunden von Ringeisen und Buchwald (2010) wurde angenommen, dass sich für die state-TA prä etwas höhere Zusammenhänge zu stabilen Dispositionen wie Testängstlichkeit ergeben als für state-TA post. Hypothese 1b bezog sich auf den Bericht von SH, also Aspekten, welche die eigene Leistung beeinträchtigt haben. Aus der Perspektive der Selbstregulation kann der Bericht von SH eine Möglichkeit sein, eine subjektiv schlechte Leistung nicht internal oder zumindest nicht stabil zu attribuieren. Betrachtet man darüber hinaus auch den Bericht von state-TA als Ausdruck dieser Regulationsprozesse, so müssten state-TA und SH miteinander kovariieren. State-TA post sollte dabei stärker mit SH-Aspekten in Zusammenhang stehen als state-TA prä, da erstere – als *nach* dem Test erfasste Erlebenskomponente – unter dem Eindruck des Tests steht. Berichtet eine Person SH, um ein subjektiv schlechtes Abschneiden zu rechtfertigen, so besteht nach einem Test mehr subjektiver „Anlass" bzw. „Bedarf" hierfür. Die Analyse der state-TA wurde dabei auf die im Leistungskontext bedeutsamste Facette Besorgtheit konzentriert.

> **Hypothese 1a:** State-Testangst prä (Besorgtheit) hängt stärker mit den stabilen Dispositionen Selbstwert, Selbstwertkontingenz, Selbstkonzept eigener Fähigkeiten und Testängstlichkeit zusammen als state-Testangst post (Besorgtheit).
>
> **Hypothese 1b:** State-Testangst post (Besorgtheit) hängt stärker mit self-handicapping-Aspekten zusammen als state-Testangst prä (Besorgtheit).

Auch die Hypothesen 2a-d nahmen analog die Perspektive ein, dass der Bericht von state-TA, insbesondere von state-TA post, Ausdruck von Regulationsprozessen ist. Aus dieser Perspektive liegt die Vermutung nahe, dass der Zusammenhang zwischen state-TA und Leistung *nicht* (oder nicht nur) auf eine kausale Wirkung der Angst auf die Leistung zurückgeht. Letzteres würde man aus der Interferenzperspektive folgern, beispielsweise weil Testangst kognitive Ressourcen bindet und damit die Leistungsfähigkeit reduziert (siehe Abschnitt 1.2.1.1). Konträr hierzu ist die Vorstellung, dass Testangst eine Konsequenz aus (subjektiv) schlechter Leistung ist. Eine solche Konstellation würde man gemäß der Defizitperspektive erwarten, nach der (lediglich) das Bewusstsein der eigenen unzureichenden Leistung (bzw. Leistungsfähigkeit) die kausale Ursache des Angsterlebens ist. Demzufolge müsste ein negativer Zusammenhang von objektiver Leistung und state-TA durch die subjektive Leistung „aufgeklärt" werden. Mit anderen Worten hieße dies, dass der Zusammenhang von objektiver Leistung und state-TA verschwindet, wenn die subjektive Leistung kontrolliert wird. In einem statistischen Sinne müsste die subjektive Leistung also die Relation von objektiver Leistung und state-TA post vollständig mediieren. Diese Annahme wurde

bislang noch nicht direkt geprüft. In diesem Modell wurde angenommen, dass die subjektive Leistung die state-TA post determiniert. Diese Annahme setzt jedoch voraus, dass ein schlechtes Abschneiden im Test auch als bedrohlich empfunden wird. Es ist wahrscheinlich, dass dies nicht per se der Fall sein muss. In drei weiteren Analysen sollte geprüft werden, ob die Relation von subjektiver Leistung und state-TA post durch drei Variablen moderiert wird.

Bei der Relation von subjektiver Leistung und state-TA post ist erstens davon auszugehen, dass eine (schlechte) subjektive Leistung umso mehr mit state-TA einhergeht, je bedrohlicher ein Scheitern empfunden wird. Personen, die einen „resistenten", also geringer von Erfolg oder Misserfolg abhängigen Selbstwert haben, sollten auch dann keine state-TA berichten, wenn sie subjektiv schlecht abgeschnitten haben. Ein subjektiv schlechtes Ergebnis in einem Test kann den Selbstwert dann tangieren, wenn letzterer „volatil" ist, also die Selbstwertkontingenz hoch ist (siehe Abschnitt 1.2.1.3.1). Subjektive Leistung und state-TA post sollten also umso stärker zusammenhängen, je höher die Selbstwertkontingenz ist.

Es ist zweitens zu erwarten, dass subjektive Leistung und state-TA post umso stärker zusammenhängen, je eher ein Test als aussagekräftig bezüglich der eigenen Fähigkeiten aufgefasst wird. Gemäß der Selbstwerttheorie von Covington (siehe Abschnitt 1.2.1.3.1) hebt die Infragestellung des diagnostischen Werts eines Tests den subjektiven Zusammenhang zwischen einem konkreten Leistungsergebnis und der Einschätzung der subjektiven Fähigkeiten auf. Wird also einem Test die diagnostische Aussagekraft abgesprochen, sollte ein subjektives Versagen im Test keinerlei Grund sein, mit Testangst zu reagieren.

Der Bericht von state-TA post kann (zumindest teilweise) als Versuch interpretiert werden, Erklärungen für die eigene (schlechte) Leistung zu suchen bzw. zu „generieren". Testangst ist dabei nur ein Aspekt, der (subjektiv) für eine schlechte Leistung ursächlich sein kann. Darüber hinaus sind auch zahlreiche nicht mit Testangst assoziierte Zustände oder Prozesse denkbar, die die Leistung objektiv oder subjektiv beeinträchtigt haben können (z. B. Müdigkeit, körperliche Schmerzen). Ebenfalls im Sinne von Covington ermöglicht eine Senkung der eigenen Anstrengung, Rückschlüsse einer schlechten Leistung auf die eigenen Fähigkeiten zu unterbinden. Drittens sollte also die Relation von subjektiver Leistung und state-TA post umso schwächer ausfallen, je niedriger eine Person nach einem Test ihre Anstrengung bei der Bearbeitung einschätzt.

Während in der explorativen Analyse alle Facetten der Testangst inkludiert wurden, beschränkten sich die Hypothesen 1a und 1b auf die Betrachtung der im Leistungskontext theoretisch und empirisch bedeutsamsten Facette Besorgtheit. Da die Hypothesen 2a-d spezifisch auf den Leistungszusammenhang fokussierten, wurde auch hier nur die Besorgtheit betrachtet.

Hypothese 2a: Der Zusammenhang von objektiver Leistung und state-Testangst post (Besorgtheit) wird teilweise oder vollständig durch die subjektive Leistung mediiert.

Hypothese 2b: Der Zusammenhang von subjektiver Leistung und state-Testangst post (Besorgtheit) wird durch die Selbstwertkontingenz moderiert – je höher die Selbstwertkontingenz, desto stärker ist der Zusammenhang.

Hypothese 2c: Der Zusammenhang von subjektiver Leistung und state-Testangst post (Besorgtheit) wird durch die Messqualität moderiert – je höher die Messqualität, desto stärker ist der Zusammenhang.

Hypothese 2d: Der Zusammenhang von subjektiver Leistung und state-Testangst post (Besorgtheit) wird durch die selbst berichtete mangelnde Anstrengung moderiert – je stärker der negative Effekt mangelnder Anstrengung auf das Testergebnis beurteilt wird, desto schwächer ist der Zusammenhang.

4.1.3 Stichprobe

Die Stichprobe wurde im Dezember 2014 durch eine E-Mail (einschließlich einer zeitlich verzögerten weiteren Einladung) über den Verteiler der Justus-Liebig-Universität Gießen rekrutiert. Um eine Selbstselektion hoch testängstlicher Personen zu verhindern, wurde als Thema der Studie „Selbstbild und Motivation“ genannt, wobei aber darauf hingewiesen wurde, dass es um Verhalten in Leistungssituationen gehe (siehe Anhang C). Durch eine Verlosung mehrerer Amazon-Gutscheine sollte ein zusätzlicher Anreiz für die Teilnahme gegeben werden. Die Erhebung wurde mit der kostenfreien Umfragesoftware SoSci Survey (www.soscisurvey.de) durchgeführt. Der Zugang zur Studie wurde Ende Januar 2015 geschlossen.

Der link zum Fragebogen wurde insgesamt von 354 Personen aufgerufen. 50% der Teilnehmer brachen vor Beginn des Intelligenztests ab (N = 177). Weitere 25 Teilnehmer brachen im weiteren Verlauf der Erhebung ab, so dass N = 152 vollständig bearbeitete Datensätze vorlagen (zur Analyse des Dropout siehe Anhang C). Um eine Vergleichbarkeit der Daten zu gewährleisten und in Ermangelung eines adäquaten Kriteriums für die Inklusion wurde konservativ entschieden und nur diese 152 Pbn in die Analysen einbezogen.

Die Stichprobe bestand aus 118 Frauen und 34 Männern (78% vs. 22%). Der Altersdurchschnitt betrug M = 23.24 (Md = 23.00; SD = 4.08). Im Mittel wurde eine Abiturnote von M = 2.15 (SD = .62; N = 149) angegeben. Ein Teil der Pbn gab aktuelle Studiendurchschnittsnoten an. Der bisherige Notendurchschnitt im Bachelor betrug M = 11.18 Notenpunkte (SD = 1.58; N = 62) und der bisherige Notendurchschnitt im Master M = 12.21 Notenpunkte (Md = 12.25; SD = 1.55; N = 24). Die Pbn stammten aus verschiedenen Studiengängen, am stärksten vertreten waren Humanmedizin

(*N* = 19), Veterinärmedizin (*N* = 16), Lehramt (*N* = 16) und Psychologie (*N* = 11)[34]. In Bedingung A waren *N* = 74 Pbn (Bedingung B: *N* = 78). Eine Übersicht über der deskriptiven Statistiken zu besagten Variablen findet sich in Tabelle 12.

Tabelle 12: Deskriptive Statistik zu demographischen Variablen der Stichprobe von Studie 1

	N	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*
Alter	152	18	39	23.24	4.08	16.65	1.12	1.27
Abiturnote	149	1.0	3.5	2.15	.62	.38	.16	-.61
Bachelornote	62	7	14	11.18	1.58	2.50	-.38	-.23
Masternote	24	8	15	12.21	1.55	2.41	-1.06	1.78

Anm.: schwankendes *N* zurückführbar auf freiwillige Angabe der Note, bei Staatsexamensfächern war überdies Frage nach Bachelor- und Masternote nicht beantwortbar. Bachelor- und Masternote in Notenpunkten (0 bis 15)

Lediglich in der Abiturnote lag ein Geschlechtsunterschied vor, der erwartungsgemäß ist: Frauen (*M* = 2.09, *SD* = .60) hatten einen signifikant besseren Notenschnitt als Männer (*M* = 2.35, *SD* = .65), *t* (147) = -2.12, *p* = .036. Insgesamt hatten *N* = 116 Pbn (76 %) Interesse an einer Rückmeldung zu ihrem Ergebnis (*N* = 26 wollten keine Rückmeldung, *N* = 10 beantworteten die entsprechende Frage nicht).

[34] Nicht erhoben wurde das Semester. Aufgrund des Altersdurchschnitts und der Nennung von Bachelor- und Masterstudiengängen ist davon auszugehen, dass die Stichprobe, auch was den Studienfortschritt angeht, heterogen war.

4.1.4 Beschreibung des Untersuchungsablaufs

Nach der Begrüßung wurden die Pbn durch den Online-Fragebogen geführt. Nacheinander wurden auf separaten Seiten der Selbstwert, die Testängstlichkeit, die Selbstwertkontingenz und das Selbstkonzept eigener Fähigkeiten erfasst. Nach der Ankündigung des Intelligenztests (mit Beispielitems) wurden die Pbn zufällig einer Bedingung zugeordnet. In Bedingung A wurden die Pbn nach der Testangst (prä) gefragt, in Bedingung B nicht. Der anschließende Untersuchungsablauf war für beide Bedingungen identisch. Nach der Abfrage des Anspruchsniveaus wurde der Intelligenztest zur Bearbeitung vorgelegt. Nach dem Test wurde die während des Tests erlebte Testangst (post) erhoben, ebenso wie die Akzeptanz des Verfahrens und self-handicapping-Aspekte. Ein achtstelliger Evaluationscode sowie demographische Variablen wurden zum Schluss erhoben, sowie der Wunsch auf Rückmeldung und die Teilnahme am Gewinnspiel. Die Bearbeitung des gesamten Fragebogens dauerte im Schnitt ca. 23 Minuten. Eine schematische Darstellung findet sich in Abbildung 11.

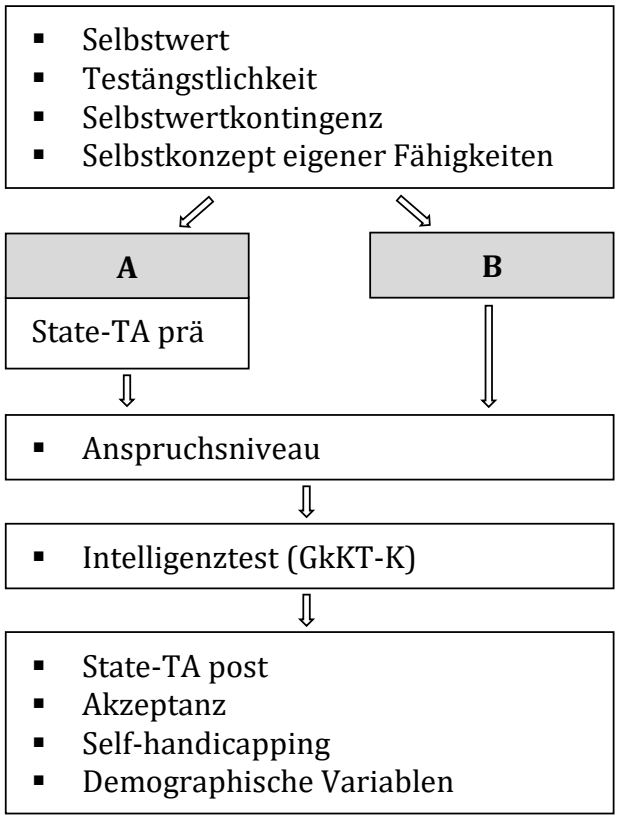

Abbildung 11: Untersuchungsablauf von Studie 1

4.1.5 Darstellung der Erhebungsdetails und Operationalisierung der Konstrukte

Selbstwert

Der Selbstwert wurde mit der modifizierten Rosenberg-Skala von Collani und Herzberg (2003) erhoben. Die 10 Items sind dabei auf einer vierstufigen Skala zu beantworten. Die Itemantworten wurden analog zur Originalquelle von 0 bis 3 kodiert (0 = „trifft gar nicht zu“; 3 = „trifft voll und ganz zu“), woraus ein maximaler Summenscore von 30 resultiert. Die deskriptiven Statistiken sind Tabelle 13 zu entnehmen.

Tabelle 13: Deskriptive Statistiken und Reliabilitätswert für die Skala Selbstwert

	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Selbstwert	0-3	5	30	19.84	6.60	43.58	-.38	-.85	.91

N = 152; Itemzahl: 10; Skalenwert berechnet via Summenscore

Die Reliabilität war gut. Zwischen Frauen und Männern zeigte sich kein signifikanter Unterschied. Eine Untersuchung von Schmitt und Allik (2005) lieferte in einer internationalen Vergleichsstudie mit Übersetzungen der Skala für die deutsche Stichprobe einen Referenzwert von M = 21.73[35] (SD = 4.17; N = 782). Während die Männer der vorliegenden Studie (M = 20.32, SD = 6.64) von diesem Referenzwert nicht abwichen (t (33) = -1.24, p = .225), wiesen die Frauen (M = 19.70, SD = 6.61) einen niedrigeren Wert auf (t (117) = -3.33, p = .001).

Testängstlichkeit

Zur Erfassung der Testängstlichkeit wurde der TAI-G XU von Wacker et al. (2008) eingesetzt. Dieser besteht aus 15 Items, mit denen die Facetten Besorgtheit, Aufgeregtheit, Interferenz und Mangel an Zuversicht erfasst werden. Die Items haben ein vierstufiges Antwortformat (1 = „fast nie"; 4 = „fast immer"). Gemäß der Vorgabe in der Publikation und auch analog zu anderen Testängstlichkeitsskalen wurden Summenscores gebildet, woraus für die Gesamtskala ein maximaler Score von 60 resultiert. Um die Interpretation der Ausprägung relativ zur Antwortskala zu erleichtern, wurden die Skalenmittelwerte zusätzlich aus den Summenscores errechnet. Die deskriptiven Statistiken und Reliabilitäten sind Tabelle 14 zu entnehmen.

Tabelle 14: Deskriptive Statistiken und Reliabilitätswerte für die Skala Testängstlichkeit

	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Besorgtheit	5-20	5	20	12.95	4.01	16.05	.13	-.91	.89
Aufgeregtheit	4-16	4	16	8.28	3.31	10.92	.59	-.47	.88
Interferenz	3-12	3	12	6.13	2.32	5.40	.79	.14	.81
Mangel an Zuversicht	3-12	3	12	7.07	2.29	5.23	.10	-.69	.85
Gesamt	15-60	17	57	34.43	9.65	93.13	.28	-.86	.92

N = 152; Angegeben sind Skalensummenwerte; bei Umrechnung in Skalenmittelwerte ergeben sich folgende Werte: Besorgtheit = 2.59; Aufgeregtheit = 2.07; Interferenz = 2.04; Mangel an Zuversicht = 2.36; Gesamt = 2.30
Md: Aufgeregtheit = 8.00; Interferenz = 6.00

Alle Subskalen und die Gesamtskala wiesen eine gute Reliabilitäten auf. In allen vier Facetten war die volle Skalenbreite ausgeschöpft. Relativ gesehen wiesen Aufgeregtheit und Interferenz Bodeneffekte auf. Für beide Geschlechter getrennt vorgenommene Vergleiche mit den Werten, die bei Wacker et al. (2008) für eine studentische Stichprobe berichtet werden (Frauen: M = 36.2; Männer: M = 32.7) zeigen, dass es keine signifikanten Abweichungen gab (t (117) = -.96, p = .340; t (33) = -.95, p = .352). Erwartungsgemäß zeigte sich bei Frauen (M = 35.35, SD = 9.66) ein höherer Gesamtscore als bei Männern (M = 31.24, SD = 9.04), t (150) = 2.22, p = .028. Auch bei Aufgeregtheit wiesen Frauen (M = 8.61, SD = 3.28) einen höheren Wert auf als Männer (M = 7.15, SD = 3.17),

[35] Der bei Schmitt & Allik (2005) berichtete Mittelwert von 31.73 basiert auf der Kodierung von 1-4 und wurde durch Umrechnung in die in dieser Studie verwendete Kodierung überführt: 31.71 – 10 = 21.73.

t (150) = 2.31, *p* = .022. Frauen (*M* = 6.33, *SD* = 2.39) zeigten überdies bei Interferenz einen höheren Wert als Männer (*M* = 5.44, *SD* = 1.96), *t* (150) = 1.99, *p* = .049. Lediglich marginal signifikant war die höhere Ausprägung der Frauen (M = 13.26, SD = 4.06) gegenüber der der Männer bei Besorgtheit (*M* = 11.85, *SD* = 3.65), *t* (150) = 1.82, *p* = .070. Facetteninterkorrelationen sowie Zusammenhänge mit Noten sind in Tabelle 15 angegeben.

Tabelle 15: Facetteninterkorrelationen bezüglich Testängstlichkeit und Korrelationen mit Noten

	Trait-TÄ			Notendurchschnitt		
	BE	AU	IN	Abitur	Bachelor	Master
Besorgtheit				.06	-.41**	-.11
Aufgeregtheit	.69**			-.01	-.13	-.04
Interferenz	.34**	.36**		.28**	-.24	-.06
Mangel an Zuversicht	.65**	.63**	.36**	.12	-.39**	-.28
Gesamt				.10	-.38**	-.18

N = 152; Korrelationen mit Noten: *N* = 149 (Abitur), 62 (Bachelor), 24 (Master); ** *p* < .01

Zwischen den Facetten zeigten sich erwartungsgemäß moderate bis hohe Zusammenhänge. Während die Zusammenhänge mit der Masternote aufgrund der kleinen Stichprobe wenig belastbar sind, fielen die (inhaltlich) negativen Korrelationen mit der Bachelornote erstaunlich hoch aus. Hingegen zeigten sich keine Zusammenhänge zur Abiturnote, mit Ausnahme der Facette Interferenz, die ebenfalls (inhaltlich) negativ mit dieser korrelierte.

Selbstwertkontingenz

Hierfür wurde die Skala zur Erfassung der Selbstwertkontingenz bei Studierenden (SESKON-ST) von Schöne, Hermann und Stiensmeier-Pelster (in Vorb.) eingesetzt. Da die Abhängigkeit des Selbstwerts von Erfolg bzw. Misserfolg in Leistungssituationen im Fokus stand, wurden nur die beiden Subskalen Leistungskontingenz (5 Items; z. B. „Meine Selbstwertgefühle sind stark davon abhängig, wie ich meine Leistung in der Uni einschätze.“) und Kompetenzkontingenz (8 Items; z. B. „Ich fühle mich minderwertig, wenn andere merken, dass ich etwas nicht gut beherrsche.“) verwendet. Alle Items haben ein siebenstufiges Antwortformat (1 = „stimmt überhaupt nicht“; 7 = „stimmt genau“), wobei einige Items rekodiert werden, so dass eine hohe Ausprägung einer starken Selbstwertkontingenz entspricht. Subskalenwerte werden über den Mittelwert berechnet. Aufgrund der hohen Interkorrelation beider Subskalen von *r* = .85 (*p* < .001) wurden diese unter dem Namen „Selbstwertkontingenz“ (SWK) zusammengefasst und ein Mittelwert aus allen 13 Items gebildet. Die deskriptiven Statistiken sowie Reliabilitäten sind Tabelle 16 zu entnehmen.

Tabelle 16: Deskriptive Statistiken und Reliabilitätswerte für die Skala Selbstwertkontingenz

	Items	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Kompetenz	8	1-7	1.00	6.75	4.28	1.32	1.74	-.37	-.40	.89
Leistung	5	1-7	1.00	6.80	3.95	1.44	2.08	-.17	-.67	.87
Gesamt	13	1-7	1.00	6.54	4.15	1.32	1.74	-.29	-.47	.94

N = 152; Skalenwerte berechnet via Mittelwert

Alle Subskalen und die Gesamtskala wiesen eine gute Reliabilität auf. Sowohl was die Subskalen als auch den Gesamtscore angeht lag die Ausprägung nahe dem Skalenmittelpunkt. Frauen (*M* = 4.25, *SD* = 1.31) berichteten eine etwas höhere Selbstwertkontingenz als Männer (*M* = 3.82, *SD* = 1.33). Dieser Unterschied war nicht signifikant, *t* (150) = 1.65, *p* = .100. Die Werte auf beiden Subskalen wurden überdies verglichen mit den geschlechtsspezifischen Ausprägungen aus einer größeren Stichprobe von Kausch (2013), die sich aus Studierenden und Mitarbeitern der JLU Gießen zusammensetzte (*N* = 592). Sowohl bei Leistungs- als auch Kompetenzkontingenz unterschieden sich Frauen und Männer dieser Studie nicht von besagter Vergleichsstichprobe, lediglich wiesen die Frauen einen marginal signifikant niedrigeren Wert bei der Leistungskontingenz auf als die Vergleichsstichprobe (*M* = 4.06, *SD* = 1.44 vs. *M* = 4.30, *SD* = 1.43), *t* (117) = -1.85, *p* = .068.

Selbstkonzept eigener Fähigkeiten

Zur Erfassung dieses Merkmals wurde die Skala absolutes Selbstkonzept (aS) adaptiert, welche aus den von Dickhäuser et al. (2002) entwickelten Skalen zum akademischen Selbstkonzept stammt. Da diese Items auf den akademischen Kontext abzielen und folglich entsprechend formuliert sind (z. B. „Ich halte meine Begabung für das Studium für … niedrig / hoch“; „Aufgaben im Rahmen meines Studiums fallen mir … schwer / leicht“), wurden sie so angepasst, dass sie unmittelbar das Selbstkonzept in Bezug auf Intelligenz (z. B. „Ich halte meine Begabung im Bereich Intelligenz für … niedrig / hoch“) und logisches Denken (z. B. „Aufgaben die logisches Denken erfordern fallen mir … schwer / leicht“) abbilden. Die Items sind mittels eines siebenstufigen Antwortformats zu bewerten, wobei hohe Werte für ein positiv ausgeprägtes Selbstkonzept stehen. Es wurde ein Skalenmittelwert berechnet. Die deskriptiven Statistiken sind in Tabelle 17 dargestellt.

Tabelle 17: Deskriptive Statistiken und Reliabilitätswert für die Skala Selbstkonzept eigener Fähigkeiten

	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Selbstkonzept e. Fähig.	1-7	2.20	7.00	5.09	1.00	1.00	-.49	.05	.88

N = 152; Itemzahl: 5; Skalenwert berechnet via Mittelwert

Die Reliabilität der Skala war gut. Aufgrund der Modifikation der Skala wurde auf einen Vergleich mit der Stichprobe von Dickhäuser et al. (2002) verzichtet. Zwischen Frauen und Männern lag kein signifikanter Unterschied vor.

Ankündigung und Instruktion des Intelligenztests

Die Instruktion des Intelligenztests erfolgte in Anlehnung an die evaluative Instruktion von Englert et al. (2011). Es wurde darauf hingewiesen, dass der Test Intelligenz erfasse, ein Vergleich mit der Leistung der anderen Versuchsteilnehmer vorgenommen werde und die Pbn auf Wunsch eine individuelle Rückmeldung erhalten können (vollständiger Text siehe Anhang D). Daraufhin wurden drei Beispielitems des Tests präsentiert. Dabei wurde je ein verbales, numerisches und figurales Item vorgelegt. Alle Items stammten aus Aufgabentypen, die im später durchgeführten Test enthalten waren (siehe Ausführungen zum GkKT-K in diesem Abschnitt), sprich Analogien, Zahlenreihen und Puzzle. Hierbei wurden Items verwendet, die in früheren Erhebungen bereits erprobt wurden (entweder zur Entwicklung des GkKT oder im Rahmen studentischer Arbeiten). Die Items wurden so gewählt, dass sie inhaltlich und / oder psychometrisch durchschnittlich anspruchsvoll waren (mittlerer Schwierigkeitsindex).

State-Testangst prä

Zur Erfassung der situativ erlebten Testangst wurde der STAI-SKD (Englert et al., 2011) eingesetzt. Dieser erfasst mit fünf Items die Zustandsangst in den Facetten Besorgtheit (2 Items; z. B. „Ich bin besorgt, dass etwas schiefgehen könnte.“) und Aufgeregtheit (3 Items; z. B. „Ich bin angespannt.“). Der STAI-SKD ist eine Kurzform des State-Trait-Angstinventars STAI von Laux, Glanzmann, Schaffner und Spielberger (1981). Die Items besitzen ein vierstufiges Antwortformat, wobei hohe Werte einen starken Angstzustand repräsentieren. Die Autoren des STAI-SKD sehen die separate Interpretation der beiden Subskalen aufgrund der hohen latenten Interkorrelation in ihrer Untersuchung (r = .74) kritisch. Dennoch ist eine hohe Überlappung eher die Regel als die Ausnahme (siehe Abschnitt 1.1.1.2.1). Da sich darüber hinaus ein empirisches Beispiel für eine separate Interpretation beim Einsatz eben dieses Verfahrens findet (Sommer & Arendasy, 2014) und die Konsistenz mit der theoretisch etablierten Trennung von Besorgtheit und Aufgeregtheit gewahrt bleiben sollte, wurde die Subskalenbildung vorgenommen. Die Facetten Interferenz und Mangel an Zuversicht wurden mit den entsprechenden Subskalen des TAI-G XU (Wacker et al., 2008) erfasst. Da diese auch in Bezug auf das aktuelle Erleben beantwortbar sind, war lediglich eine Änderung der Instruktion sowie der Skalenanker nötig: statt „fast nie“ bis „fast immer“ wurden die Skalenanker des STAI-SKD eingesetzt: „überhaupt nicht“ bis „sehr“. Die state-TA prä Skala

bestand somit aus insgesamt 11 Items. Die deskriptiven Statistiken und Reliabilitäten sind Tabelle 18 zu entnehmen. Um die Interpretation der Ausprägung relativ zur Antwortskala zu erlauben, wurden auch hier die Skalenmittelwerte zusätzlich aus den Summenscores errechnet.

Tabelle 18: Deskriptive Statistiken und Reliabilitätswerte für die Skala state-Testangst prä (nur Bedingung A)

	Items	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Besorgtheit	2	2-8	2	8	3.18	1.37	1.87	1.42	2.13	.72
Aufgeregtheit	3	3-12	3	12	5.43	2.26	5.13	.96	.35	.89
Interferenz	3	3-12	3	12	4.39	2.13	4.52	2.03	3.89	.89
Mangel an Zuversicht	3	3-12	3	12	7.54	2.20	4.83	-.27	-.44	.89
Gesamt	11	11-44	11	39	20.54	5.62	31.57	.96	1.18	.85

N = 74; Angegeben sind Skalensummenwerte; bei Umrechnung in Skalenmittelwerte ergeben sich folgende Werte: Besorgtheit = 1.59; Aufgeregtheit = 1.81; Interferenz = 1.46; Mangel an Zuversicht = 2.51; Gesamt = 1.87
Md: Besorgtheit = 3.00; Aufgeregtheit = 5.00; Interferenz = 3.50; Gesamt = 19.50

Die Reliabilitäten aller Skalen waren gut, insbesondere vor dem Hintergrund der Kürze der Skalen. Erkennbar sind Bodeneffekte bei Besorgtheit und Aufgeregtheit und in besonderer Weise bei Interferenz, nicht aber bei Mangel an Zuversicht. Aus den Items zu Besorgtheit und Aufgeregtheit ergab sich ein gemeinsamer Skalenmittelwert (umgerechnet auf die Kodierung von 1 bis 4) von 1.72 (*SD* = .66). Dieser entspricht fast exakt jenem Wert, den Englert et al. (2011) nach einer bedrohlichen Aufgabenankündigung in einer experimentell induzierten Prüfungssituation ermittelten (*M* = 1.73, *SD* = .47). Vor diesem Hintergrund kann die evaluative Instruktion als gelungen erachtet werden.

Anspruchsniveau

Die erwartete Leistung wurde mit zwei Items erhoben. Dabei wurde die erwartete Punktzahl (0 bis 26) im Test sowie der Leistungsbereich, in dem man vermutlich liegen würde, erhoben („unterdurchschnittlich" bis „überdurchschnittlich", anzugeben auf einer visuellen Analogskala, die im System von 0 bis 100 kodiert war). Die Pbn schätzten im Durchschnitt, dass sie von den 26 Items des GkKT-K *M* = 16.75 (*SD* = 4.76) lösen würden. Überdies schätzten die Pbn ihre Leistung in einem mittleren Bereich ein, *M* = 55.24, *SD* = 14.57. Beide Variablen korrelierten hoch miteinander, $r = .63$ ($p < .001$). Aus diesem Grund wurden beide Variablen z-standardisiert und aus den so transformierten Variablen wiederum ein Mittelwert gebildet, mit dem das Anspruchsniveau operationalisiert wurde.

Gießener kognitiver Kompetenztest – Kurzform

Als Intelligenztest wurde die Kurzversion des Gießener kognitiven Kompetenztest GkKT-K (Ulfert, Ott, Michaelis & Kersting, 2014b) eingesetzt[36]. Sechs der elf Aufgabenblöcke des GkKT (Ulfert, Ott, Michaelis & Kersting, 2014a) bilden in inhaltlich unveränderter, jedoch in ihrer Reihenfolge modifizierter Form den GkKT-K. Jedem der sechs Aufgabenblöcke ist eine separate Instruktion vorgeschaltet, die Aufgabenblöcke selbst unterliegen einer Zeitbegrenzung. Die reine Bearbeitungszeit ohne Instruktionen beträgt insgesamt 8:16 Minuten.

Da es sich beim GkKT bzw. GkKT-K um ein noch wenig verbreitetes Instrument handelt, soll kurz auf dessen Validität eingegangen werden. Skalenanalysen ergaben für den GkKT-K in einer ersten Eichstichproben eine Korrelation mit dem Wonderlic Personnel Test (Wonderlic Inc., 1996) von r_p = .73 (p <. 001) und mit der Abiturnote von r_p = -.22 (p < .01) (N = 170 bzw. 151, Studierende der Physik und Wirtschaftswissenschaften). In einer zweiten Eichstichprobe ergaben sich Korrelationen mit dem BEFKI GC-K (Schipolowski et al., 2013) von r_p = .36 (p < .001; N = 327) und dem MWT-B (Lehrl, 2005) von r_p = .34 (p < .001, N = 167) und mit der Abiturnote von r_p = -.14 (p < .01, N = 458) (Studierende des Fachbereich 09). Cronbach's Alpha betrug jeweils .78 bzw. .71. In der vorliegenden Stichprobe ergaben sich signifikante Zusammenhänge mit der Abiturnote von r = -.22 (p < .01; N = 149) und mit der Masternote von r = .47 (p < .05, N = 24), jedoch kein signifikanter Zusammenhang zur Bachelornote, r = .084 (p = .516, N = 62). Insgesamt sprechen diese Ergebnisse für die Kriteriumsvalidität des GkKT-K. Tabelle 19 sind die deskriptiven Statistiken und die Reliablität des GkKT-K zu entnehmen.

Tabelle 19: Deskriptive Statistiken des und Reliabilitätswert des GkKT-K

	Items	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
GkKT-K	26	4	25	15.80	4.22	17.79	-.30	-.41	.78

N = 152

[36] Eine nähere Beschreibung des GkKT erfolgt im Abschnitt 6.1.5 zur Methode von Studie 3.

Die Reliabilität ist des GkKT-K fiel, geschätzt über die interne Konsistenz, mit einem Wert von α = .78 gut aus. Die Maximalpunktzahl von 26 gelösten Items wurde nicht erreicht. Die Verteilung war linksschief und breitgipflig (siehe Abbildung 12). Insgesamt sprechen diese Befunde dafür, dass der GkKT-K ein für die Stichprobe anspruchsvoller Test ist, der weder Boden- noch Deckeneffekt produziert und somit in der Lage ist, zwischen hoher und niedriger Fähigkeitsausprägung zu differenzieren. Es zeigte sich kein Geschlechtsunterschied zwischen Männern (M = 15.88, SD = 4.84) und Frauen (M = 15.77, SD = 4.04), t (150) = -.14, p = .893.

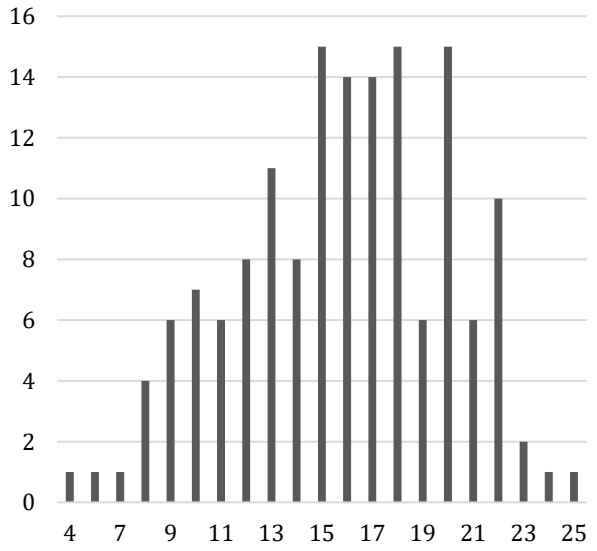

Abbildung 12: Histogramm der Gesamtpunktzahl beim GkKT-K

State-Testangst post

Die im Test erlebte Testangst wurde mit denselben Items erfasst wie die state-TA prä. Dabei wurden die Pbn um eine retrospektive Einschätzung ihres Erlebens während des gerade bearbeiteten Tests gebeten. Der einzige Unterschied bezüglich der Items war, dass die Itemformulierungen aus dem Präsens ins Imperfekt gesetzt wurden (z. B. „Ich war besorgt, dass etwas schiefgehen könnte."; „Ich war angespannt."). Dies wurde sowohl für die Items des STAI-SDK zur Erfassung von Besorgtheit und Aufgeregtheit, als auch für jene des TAI-G XU zur Erfassung von Interferenz und Mangel an Zuversicht durchgeführt. Das Antwortformat wurde nicht geändert. Die deskriptiven Statistiken und Reliabilitäten sind Tabelle 20 zu entnehmen. Um die Interpretation der Ausprägung relativ zur Antwortskala zu ermöglichen, wurden wiederum die Skalenmittelwerte zusätzlich aus den Summenscores gebildet.

Tabelle 20: Deskriptive Statistiken und Reliabilitätswerte für die Skala state-Testangst post (Bedingung A und B)

State-TA post	Items	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	α
Besorgtheit	3	2-8	2	8	3.76	1.73	2.99	.88	-.09	.80
Aufgeregtheit	2	3-12	3	12	6.72	2.62	6.87	.56	-.61	.91
Interferenz	3	3-12	3	12	5.11	2.45	5.98	1.11	.36	.89
Mangel an Zuversicht	3	3-12	3	12	8.46	2.15	4.62	-.33	.06	.89
Gesamt	11	11-44	12	44	24.04	6.74	45.38	.65	.08	.89

N = 152; Angegeben sind Skalensummenwerte; bei Umrechnung in Skalenmittelwerte ergeben sich folgende Werte: Besorgtheit = 1.88; Aufgeregtheit = 2.24; Interferenz = 1.70; Mangel an Zuversicht = 2.82; Gesamt = 2.19
Md: Besorgtheit = 3.00; Aufgeregtheit = 6.00; Interferenz = 4.00; Gesamt = 23.00

Die Reliabilitäten der Subskalen waren gut. In allen vier Subskalen war die komplette Skalenbreite ausgeschöpft, auch im Gesamtwert trat der Maximalwert auf. Rein deskriptiv ergaben sich für state-TA post höhere Werte als für prä. Wie bei der state-TA prä zeigte sich auch bei Mangel an Zuversicht kein Bodeneffekt, während dies bei Interferenz besonders stark der Fall war. Die Facetteninterkorrelationen für state-TA prä und post sind in Tabelle 21 zusammengefasst.

Tabelle 21: Facetteninterkorrelationen bezüglich state-Testangst prä und post

	State-TA prä bzw. post			
	BE	AU	IN	MZ
Besorgtheit	-	.76**	.34**	.38**
Aufgeregtheit	.62**	-	.28**	.36**
Interferenz	.10**	.06	-	.33**
Mangel an Zuversicht	.32**	.50**	.17	-

N = 74 (prä) bzw. 152 (post); Interkorrelationen für state-TA prä sind unter der Diagonalen, für state-TA post über der Diagonalen dargestellt

Bei der state-TA prä fanden sich erwartungsgemäß hohe Korrelationen zwischen Besorgtheit und Aufgeregtheit, etwas überraschend sind die schwachen Korrelationen zwischen Interferenz und den anderen Subskalen. Bei der state-TA post waren die Zusammenhänge von Besorgtheit und Aufgeregtheit höher als bei prä, ebenso von Interferenz mit den anderen Subskalen. Mit Ausnahme der Subskala Interferenz unterscheidet sich das Korrelationsmuster zwischen prä und post nicht erheblich.

Akzeptanz

Zur Erfassung des subjektiven Erlebens des Tests wurde der Akzept!-L von Kersting (2008) eingesetzt. Dieser erfasst mit den Skalen Kontrollierbarkeit (z. B. „Die Testaufgaben waren klar und verständlich.“), Messqualität (z. B. „Der Test misst das, was er misst, zuverlässig.“), Augenscheinvalidität (z. B. „Die Testaufgaben spiegeln Anforderungen wider, die auch im Berufsleben gefordert sind.“) und Belastungsfreiheit (z. B. „Die Testaufgaben waren überwiegend zu schwer für mich.“) die Bewertung eines absolvierten Leistungstests aus Sicht des Testanden. Diese Items sind auf einer sechsstufigen Antwortskala (1 = „trifft nicht zu“; 6 = „trifft genau zu“) zu beantworten, wobei hohe Werte für gute Akzeptanz stehen. Die vier Items jeder Skala werden je über den Mittelwert zusammengefasst. Darüber hinaus wird mit je einem Item auf einer Schulnotenskala die subjektive Leistung in Bezug zu einer gleichaltrigen und schulisch ähnlich gebildeten Vergleichsgruppe sowie eine Gesamtbeurteilung des Verfahrens erfragt. Die deskriptiven Statistiken und Reliabilitäten sind in Tabelle 22 aufgeführt.

Tabelle 22: Deskriptive Statistiken und Reliabilitätswerte für die Skala Akzeptanz

	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Kontrollierbarkeit	1-6	2.25	6.00	5.21	.79	.62	-1.45	2.24	.76
Messqualität	1-6	1.00	5.75	3.55	.90	.80	-.13	-.26	.71
Augenschein-validität	1-6	1.00	5.75	2.50	.91	.82	.49	.21	.74
Belastungsfreiheit	1-6	1.25	6.00	4.13	1.23	1.52	-.29	-.74	.88
Note Verfahren	1-6	1	6	2.78	.89	.78	.92	1.33	-
Subjektive Leistung	1-6	1	6	3.15	1.00	1.00	.57	.62	-

N = 152; Itemzahl: 4 Items, Note & subjektive Leistung jeweils mit einem Item erfasst
Md: Kontrollierbarkeit = 5.50; Note Verfahren = 3.00; Subjektive Leistung = 3.00

Die Reliabilitäten aller Skalen waren gut. Es wurde überwiegend die komplette Skalenbreite ausgeschöpft, mit Ausnahme der Kontrollierbarkeit, bei der ein Minimum von 2.25 vorlag. Dies spricht für ein gegebenes Instruktionsverständnis seitens der Probanden. Die Augenscheinvalidität wurde eher gering eingeschätzt.

Tabelle 23: Skaleninterkorrelationen bezüglich Akzeptanz sowie Korrelationen mit der Note für das Verfahren und subjektiver Leistung

	KB	MQ	AV	BF	Note
Kontrollierbarkeit					
Messqualität	.22**				
Augenscheinvalidität	-.04	.35**			
Belastungsfreiheit	.34**	-.03	.01		
Note Verfahren	-.22**	-.41**	-.21**	-.24**	
Subjektive Leistung	-.19*	-.05	.01	-.51**	.40**

N = 152; Note & subjektive Leistung jeweils mit einem Item erfasst (1 = „sehr gut“, 6 = „ungenügend“)

Zwischen den Skalen zeigten sich keine bis moderate Zusammenhänge (siehe Tabelle 23). Erwartungsgemäß hing die Belastungsfreiheit deutlich mit der subjektiven Leistung zusammen zu r = -.51 (p <.001). Je anstrengender die Aufgaben erlebt wurden, desto schlechter wurde auch die eigene Leistung eingeschätzt. Auffällig ist auch, dass der Test umso besser bewertet wurde, je besser die eigene Leistung eingeschätzt wurde (r = .40, p < .001). Den stärksten Zusammenhang mit der Gesamtbeurteilung des Verfahrens wies die Subskala Messqualität auf, r = -.41 (p < .001) – je niedriger die Messqualität beurteilt wurde, desto schlechter wurde auch der Test bewertet.

Self-handicapping

Die Items zum self-handicapping entstammen einer Untersuchung von Tandler, Schwinger, Kaminski und Stiensmeier-Pelster (2014). Darin wird abgefragt, wie stark insgesamt acht unterschiedliche Prozesse, Zustände oder Bedingungen die eigene Leistung beeinträchtigen bzw. beeinträchtigt haben. Die Bezeichnung „self-handicapping“ unterstellt implizit, dass Pbn intentional

diese Aspekte als Erklärungen für schlechte Leistung aufführen. De facto ist damit aber nicht ausgeschlossen, dass diese Prozesse oder Zustände tatsächlich die Leistung beeinträchtigt haben. Die Bezeichnung soll aus dem Grund beibehalten werden, weil sie die – potenzielle – selbstregulative Komponente im Bericht dieser Beeinträchtigungen verdeutlicht. Jedes einzelne Item (Müdigkeit, Schlechte Laune, etc.) soll getrennt auf einem fünfstufigen Antwortformat (1 = „trifft nicht zu"; 5 = „trifft zu") beurteilt werden. Um abzubilden, inwiefern eine Beeinträchtigung auf Merkmale des Tests selbst zurückgeführt wird, wurde das Item „Mangelhaftigkeit des Tests" hinzugefügt.

Tabelle 24: Deskriptive Statistiken der Items zu self-handicapping-Aspekten

	Min	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*
Müdigkeit	1	5	2.96	1.31	1.71	-.11	-1.05
Schlechte Laune	1	5	1.78	1.05	1.10	1.25	.71
Nervosität	1	5	2.38	1.39	1.93	.59	-.99
Ablenkung durch Umgebung	1	5	2.43	1.43	2.04	.47	-1.17
Krankheit / körperl. Unwohlsein	1	5	1.74	1.18	1.40	1.59	1.48
Stress	1	5	2.36	1.33	1.76	.49	-1.02
Mangelnde Anstrengung	1	5	2.11	1.24	1.54	.90	-.30
Prüfungsangst	1	5	1.82	1.23	1.51	1.36	.67
Mangelhaftigkeit des Tests	1	5	1.53	.82	.67	1.78	3.20

N = 152; gefragt wurde, inwieweit die genannten Aspekte die eigene Leistung beeinträchtigt haben
Md: Schlechte Laune = 1.00; Nervosität = 2.00; Krankheit / körperl. Unwohlsein = 1.00; Mangelnde Anstrengung = 2.00; Prüfungsangst = 1.00; Mangelhaftigkeit des Tests = 1.00

Die meisten Items zum SH waren linkssteil verteilt, wiesen also mehr oder weniger ausgeprägte Bodeneffekte auf (siehe Tabelle 24).

4.1.6 Statistische Verfahren

In Studie 1 wurde zur Prüfung von Hypothese 2a ein moderiertes Mediationsmodell gerechnet (siehe Abschnitt 3). Hierfür wurde die Untersuchungsbedingung als Moderator *W* in das Modell inkludiert. Der Einbezug der Manipulation als Moderator *W* gründet dabei insbesondere auf den unklaren Auswirkungen der vorherigen Erfassung von Testangst (siehe Explorative Analyse), die eine separate Betrachtung der einzelnen Effekte je Bedingung nahelegt. Mittels PROCESS können direkt und indirekte Effekte auf beiden Ausprägungen des Moderators (in diesem Fall die beiden Bedingungen) festgestellt werden. Zur Prüfung der Hypothesen 2b-d wurden moderierte Moderationen berechnet, deren Konzept in Abschnitt 6.1.6 näher beschrieben ist.

4.2 Ergebnisse

4.2.1 Vorbereitende Analysen

Die vorbereitenden Analysen für Studie 2 fokussierten auf den Vergleich der beiden Untersuchungsbedingungen in Bezug auf demographische Variablen und erfasste Dispositionen. Ziel war es zu prüfen, ob die Randomisierung die gewünschte Wirkung erbracht hat. Auf etwaige Geschlechtsunterschiede in betrachteten Variablen wurde bereits eingegangen.

Der Anteil an Männern unterschied sich zwischen den beiden Bedingungen nicht, χ^2 (1, N = 152) = .05, p = .830. Bezüglich des Alters lag kein Unterschied zwischen Bedingung A (M = 23.50, SD = 4.49) und B (M = 23.00, SD = 3.67) vor, t (150) = .75, p = .452. In der Abiturnote zeigte sich kein signifikanter Gruppenunterschied (Bedingung A: M = 2.08, SD = .62; Bedingung B: M = 2.22, SD = .61), t (147) = -1.40, p = .163. Gleiches galt auch für die Bachelornote (Bedingung A: M = 11.06, SD = 1.85; Bedingung B: M = 11.30, SD = 1.26), t (54.90) = -.58, p = .563, sowie für die Masternote (Bedingung A: M = 12.18, SD = 1.72; Bedingung B: M = 12.24, SD = 1.47), t (22), = -.09, p = .929. Darüber hinaus wurden Unterschiede in unterschiedlichen Dispositionen betrachtet. Diese sind in Tabelle 25 aufgeführt.

Tabelle 25: Unterschiede zwischen den Gruppen in den Bedingungen A und B bezüglich dispositioneller Variablen (geprüft via t-Tests für unabhängige Stichproben)

	A (mit prä)		B (ohne prä)				
	M	*SD*	*M*	*SD*	*t*	*p*	*d*
Selbstwert	19.99	6.69	19.71	6.55	.26	.794	-.04
Trait-TÄ							
Besorgtheit	12.53	4.10	13.35	3.90	-1.26	.209	.21
Aufgeregtheit	8.08	3.31	8.47	3.31	-.73	.465	.12
Interferenz	5.97	2.20	6.28	2.44	-.82	.414	.13
Mangel an Zuversicht	7.01	2.40	7.12	2.19	-.27	.785	.05
Gesamt	33.60	9.70	35.22	9.60	-1.04	.301	.17
Selbstwertkontingenz	4.02	1.25	4.28	1.38	-1.24	.218	.20
Selbstkonzept eig. Fähigkeiten	5.14	.97	5.05	1.03	.57	.573	-.10

N: A = 74, B = 78

Zwischen den beiden Bedingungen zeigten sich keinerlei bedeutsame Unterschiede, die Randomisierung war somit erfolgreich.

4.2.2 Ergebnisse der explorativen Analysen und Hypothesenprüfung

4.2.2.1 Explorative Analyse

Zunächst werden die Effekte der Manipulation auf das Erleben vor dem Test (d. h. auf das Anspruchsniveau), die Leistung und die state-TA post sowie die Akzeptanz betrachtet. Die Ergebnisse der Mittelwertsvergleiche sind in Tabelle 26 aufgeführt.

Tabelle 26: Unterschiede zwischen den Gruppen in den Bedingungen A und B bezüglich Anspruchsniveau, Leistung, state-Testangst post und Akzeptanz (geprüft via t-Tests für unabhängige Stichproben).

	A (mit prä)		B (ohne prä)				
	M	*SD*	*M*	*SD*	*t*	*p*	*d*
Anspruchsniveau	.17	.78	-.16	1.02	2.23	.027	-.36
Leistung	16.19	4.17	15.42	4.26	1.12	.264	-.18
State-TA post							
Besorgtheit	3.60	1.65	3.91	1.80	-1.13	.262	.18
Aufgeregtheit	6.41	2.75	7.01	2.47	-1.43	.154	.23
Interferenz	5.08	2.68	5.13	2.22	-.12	.906	.02
Mangel an Zuversicht	8.38	2.27	8.54	2.04	-.46	.648	.07
Gesamt	23.46	7.06	24.59	6.41	-1.03	.303	.17
Akzeptanz							
Kontrollierbarkeit	5.25	.69	5.17	.87	.65	.515	-.10
Messqualität	3.34	.86	3.75	.89	-2.95	.004	.48
Augenscheinvalidität	2.34	.88	2.65	.91	-2.10	.037	.35
Belastungsfreiheit	4.19	1.24	4.06	1.24	.64	.523	-.11
Note Verfahren	2.77	.84	2.78	.94	-.08	.935	.01
Subjektive Leistung	3.08	1.04	3.22	.96	-.84	.402	.14

N: A = 74, B = 78; Anspruchsniveau: z-standardisiert

Es zeigten sich nur sehr wenige Unterschiede zwischen den beiden Bedingungen. Signifikante und kleine Effekte zeigten sich beim Anspruchsniveau sowie zwei Skalen der Akzeptanz. So zeigte sich in A (M = .17, SD = .78) ein signifikant höheres Anspruchsniveau als in B (M = -.16, SD = 1.02), t (150) = 2.23, p = .027, d = -.36. Darüber hinaus wurde in A (M = 3.34, SD = .86) die Messqualität schlechter bewertet als in B (M = 3.75, SD = .89), t (150) = -2.95, p = .004, d = .48. Auch die Augenscheinvalidität wurde in A (M = 2.34, SD = .88) niedriger bewertet als in B (M = 2.65, SD = .91), t (150) = -2.10, p = .037, d = .35.

Im Folgenden werden Gruppenunterschiede bezüglich self-handicapping-Aspekten aufgeführt (Tabelle 27).

Tabelle 27: Unterschiede zwischen den Gruppen in den Bedingungen A und B bezüglich self-handicapping-Aspekten (geprüft via t-Tests für unabhängige Stichproben)

	A (mit prä)		B (ohne prä)				
	M	*SD*	*M*	*SD*	*t*	*p*	*d*
Müdigkeit	2.86	1.26	3.05	1.35	-.88	.381	.15
Schlechte Laune	1.69	1.01	1.87	1.09	-1.07	.284	.17
Nervosität	2.15	1.32	2.60	1.43	-2.03	.044	.33
Ablenkung durch Umgebung	2.32	1.45	2.54	1.40	-.92	.357	.15
Krankheit / körperl. Unwohlsein	1.70	1.27	1.77	1.10	-.35	.730	.06
Stress im Alltag	2.24	1.21	2.47	1.42	-1.08	.284	.17
Mangelnde Anstrengung	2.12	1.33	2.10	1.16	.09	.925	-.02
Prüfungsangst	1.61	1.10	2.03	1.32	-2.12	.036	.35
Mangelhaftigkeit des Tests	1.50	.78	1.56	.86	-.48	.632	.07

N: A = 74, B = 78

Lediglich bei den Aspekten „Nervosität“ und „Prüfungsangst“ zeigten sich Unterschiede, die in beiden Fällen in die gleiche Richtung weisen. In Bedingung A (M = 2.15, SD = 1.32) wurde in geringerem Maße Beeinträchtigung durch Nervosität berichtet als in B (M = 2.60, SD = 1.43), t (150) = -2.03, p = .044, d = .33. In gleicher Weise wurde in Bedingung A (M = 1.61, SD = 1.10) in geringerem Maß Prüfungsangst als beeinträchtigender Prozess berichtet als in B (M = 2.03, SD = 1.32), t (150) = -2.12, p = .036, d = .35. Beide Effekte waren klein.

Bezüglich der explorativen Analyse zeigten sich also relativ wenige Effekte in Form von Mittelwertsunterschieden durch die Messung von state-TA vor dem Test. Keine Effekte zeigten sich auf das Erleben von state-TA post sowie auf das absolute Leistungsniveau. Einige Variablen zeigten Effekte, diese waren jedoch klein. Aufgrund der relativ geringen Effekte der Manipulation wurden in den Hypothesen 1a und 1b sowie 2a-d bei der Betrachtung der state-TA post die Bedingungen A und B zusammengelegt (eine Betrachtung der Ergebnisse zur state-TA post unter Trennung von Bedingung A und B erfolgt in Abschnitt 4.2.3).

4.2.2.2 Hypothese 1a und 1b

Zur Prüfung der Hypothesen 1a und 1b wurden die state-TA prä und post mit den erfassten Dispositionen Selbstwert, Selbstwertkontingenz, Selbstkonzept eigener Fähigkeiten und Testängstlichkeit korreliert. Hierbei wurde (bei der state-TA) nur die im Leistungskontext bedeutsamste Facette Besorgtheit betrachtet. Um eine Vergleichbarkeit des Abstraktionsniveaus der betrachteten Dispositionen zu gewährleisten, wurde bei der Testängstlichkeit nur der Gesamtwert herangezogen. Die Korrelation zwischen state-TA prä und post (nur für Bedingung A) war signifikant, sie lag bei r = .62 (p < .001).

Die Analysen zur Prüfung von Hypothese 1a sind in Tabelle 28 aufgeführt.

Tabelle 28: Bivariate Korrelationen von state-Testangst (Besorgtheit) und dispositionellen Variablen

	State-Testangst	
	prä (nur A)	post (A & B)
Selbstwert	-.26*	-.30**
Selbstwertkontingenz	.21	.43**
Selbstkonzept eigener Fähigkeiten	-.08	-.14
Testängstlichkeit	.30**	.51**

N: A = 74; B = 78; die Prüfung auf signifikante Unterschiede wurde nicht vorgenommen aufgrund lediglich partieller Überlappung der Stichproben

Gemäß Hypothese 1a sollten die erhobenen Dispositionen stärker mit der state-TA prä zusammenhängen als mit der state-TA post. Dieses Befundbild fand sich für keine der vier Dispositionen. State-TA post korrelierte etwas stärker mit dem Selbstwert (r = -.30, p < .001) als state-TA prä (r = -.26, p = .025). Deutliche Unterschiede fanden sich für die Zusammenhänge von state-TA post mit Selbstwertkontingenz (r = .43, p < .001) und Testängstlichkeit (r = .51, p < .001) verglichen mit state-TA prä (r = .21, p = .077 bzw. r = .30, p = .009). Mit dem Selbstkonzept eigener Fähigkeiten lagen für state-TA prä und post keine bedeutsamen Zusammenhänge vor. Insgesamt wurde Hypothese 1a somit nicht bestätigt. Trotz der Unterschiede in der Höhe der Korrelationen von state-TA prä und post mit den vier Dispositionen wiesen die Zusammenhänge jeweils in dieselbe Richtung: so fanden sich positive Zusammenhänge von state-TA prä und post mit Testängstlichkeit und Selbstwertkontingenz, negative Zusammenhänge mit dem Selbstwert und keine signifikanten Zusammenhänge mit dem Selbstkonzept eigener Fähigkeiten.

Dieselbe Betrachtung wurde für Hypothese 1b vorgenommen bezüglich der Zusammenhänge mit den self-handicapping-Aspekten. Die Interkorrelationen sind in Tabelle 29 aufgeführt.

Tabelle 29: Bivariate Korrelationen von state-Testangst (Besorgtheit) mit self-handicapping Aspekten

	Stata-Testangst	
self-handicapping-Aspekte	prä (nur A)	post (A & B)
Müdigkeit	.04	.13
Schlechte Laune	.17	.26**
Nervosität	.50**	.66**
Ablenkung durch Umgebung	.18	.15
Krankheit / körperl. Unwohlsein	.18	.29**
Stress im Alltag	.33**	.25**
Mangelnde Anstrengung	.03	.01
Prüfungsangst	.46**	.52**
Mangelhaftigkeit des Tests	.11	.21*

N für prä = 74; *N* für post = 152; die Prüfung auf signifikante Unterschiede wurde nicht vorgenommen aufgrund lediglich partieller Überlappung der Stichproben

Bei Betrachtung der Korrelationen fallen drei Befunde auf. Mit Ausnahme einiger nicht signifikanter Korrelationen (Müdigkeit, Ablenkung durch Umgebung, Mangelnde Anstrengung) fanden sich

durchweg positive Korrelationen zwischen state-TA und SH-Aspekten. Zweitens fand sich für sechs der neun SH-Aspekte ein höherer Zusammenhang zu state-TA post, verglichen mit state-TA prä (Müdigkeit, schlechte Laune, Nervosität, Krankheit / körperliches Unwohlsein, Prüfungsangst, Mangelhaftigkeit des Tests). Hypothese 1b wurde somit bestätigt. Drittens fällt auf, dass die mit Abstand stärksten Zusammenhänge zu state-TA prä und post mit den SH-Aspekten Nervosität (r = .50 bzw. r = .66, jeweils p < .001) und Prüfungsangst (r = .46 bzw. r = .52, p < .001) vorlagen (diese korrelierten untereinander zu r = .63, p < .001).

4.2.2.3 Hypothese 2a-d

Vor der eigentlichen Analyse zu Hypothese 2a wurde zunächst geprüft, ob state-TA post stärker mit Leistung korreliert als state-TA prä. Entgegen der bisherigen Befundlage korrelierte state-TA post (Bedingung A & B) mit r = -.26, p = .001, nur geringfügig stärker mit der Leistung als state-TA prä (Bedingung A), r = -.25, p = .029. Ein zu erwartendes Bild zeigte sich hingegen bei der subjektiven Leistung: state-TA post (Bedingung A & B) korrelierte stärker mit der subjektiven Leistung (r = .41, p < .001) als state-TA prä (Bedingung A; r = .27, p = .021).

Hypothese 2a befasste sich mit der Frage, ob der Zusammenhang von objektiver Leistung und state-TA post durch die subjektive Leistung mediiert wird. Zur Prüfung dieser Annahme wurde mit PROCESS (Hayes, 2013) ein moderiertes Mediationsmodell (Model 59) berechnet mit dem Prädiktor objektive Leistung (*X*), dem Mediator subjektive Leistung (*M*) und der abhängigen Variable state-TA post (*Y*). Um die Ergebnisse gegen etwaige (unerwartete) Effekte der Manipulation abzusichern wurde die Manipulation als dummy-kodierter Moderator ins Modell inkludiert (siehe Abschnitt 4.1.6). Die beiden unabhängigen Variablen objektive Leistung und subjektive Leistung wurden jeweils mittelwertszentriert.

Aus Tabelle 30 ist ersichtlich, dass die objektive Leistung einen signifikanten Effekt (Koeff. = -.11, p < .001) auf die subjektive Leistung hatte. Eine höhere objektive Leistung ging also mit einer höheren subjektiven Leistung einher. Aus der Betrachtung der Effekte auf die state-TA post geht wiederum hervor, dass die subjektive Leistung einen signifikanten Effekt auf die state-TA post hatte (Koeff. = .62, p = .002): je schlechter die subjektive Leistung, desto höher war die state-TA post. Der Effekt der objektiven Leistung auf die state-TA post war signifikant (Koeff. = -.10, p = .047). Der Effekt der Bedingung und die beiden Interaktionen von objektiver Leistung und subjektiver Leistung mit Bedingung waren nicht signifikant.

Tabelle 30: Zusammenfassung der Regressionsanalysen zur Vorhersage des Mediators subjektive Leistung und der abhängigen Variable state-Testangst post (Besorgtheit)

	Kriterium							
	M (subjektive Leistung)				Y (state-TA post)			
	Koeff.	*SE*	*t*	*p*	Koeff.	*SE*	*t*	*p*
Konstante	-.03	.11	-.24	.809	3.68	.18	19.98	.000
X: objektive Leistung	-.11	.03	-4.04	.000	-.10	.05	-2.00	.047
M: subjektive Leistung	-	-	-	-	.62	.20	3.16	.002
W: Bedingung	.08	.15	.49	.628	.19	.26	.73	.466
Int 1: obj. Leistung x Beding.	.05	.04	1.47	.143	.08	.07	1.22	.224
Int 2: subj. Leistung x Beding.	-	-	-	-	-.03	.27	-.09	.925
	R^2 = .13				R^2 = .20			
	F (3, 148) = 7.18				F (5, 146) = 7.23, p < .001			

N = 152; X und M jeweils mittelwertszentriert; Anzahl der bootstraps: 5.000; Manipulation kodiert mit 0 (Bedingung A) und 1 (Bedingung B)

Die getrennte Betrachtung direkter und indirekter Effekte in den beiden Bedingungen (Tabelle 31) macht deutlich, dass in Bedingung A sowohl ein direkter (Koeff. = -.10, p = .047) als auch ein indirekter Effekt vorlag (Koeff. = -.07), während in Bedingung B weder ein direkter noch ein indirekter Effekt vorlagen. Hypothese 2a wurde somit nur teilweise bestätigt.

Tabelle 31: Direkter und indirekter Effekt des Mediationsmodells zur Vorhersage der state-Testangst post (Besorgtheit) in Bedingung A und B

		Kriterium (Y): state-Testangst post					
	Bedingung	Effekt	*SE*	*t*	*LLCI*	*ULCI*	*p*
Direkter Effekt	A	-.10	.05	-2.00	-.20	-.001	.047
(X: objektive Leistung)	B	-.02	.04	-.42	-.10	.07	.679
		Effekt	*B-SE*		*B-LLCI*	*B-ULCI*	
Indirekter Effekt	A	-.07	.03	-	-.13	-.03	-
(M: subjektive Leistung)	B	-.03	.03	-	-.11	.004	-

N = 152; Anzahl der bootstraps: 5.000; *LLCI* bzw. *ULCI* = untere bzw. obere Grenze des Konfidenzintervalls; *B-LLCI* bzw. *B-ULCI* = untere bzw. obere Grenze des via Bootstrapping ermittelten Konfidenzintervalls

Hypothese 2b bezog sich auf die Annahme, dass der Zusammenhang von subjektiver Leistung und state-TA post umso stärker ist, je höher die Selbstwertkontingenz ist. Hierfür wurde mit PROCESS ein moderiertes Moderationsmodell (Model 3) berechnet (siehe zur Erläuterung Abschnitt 6.1.6). Die Ergebnisse sind in Tabelle 32 dargestellt.

Tabelle 32: Zusammenfassung des Regressionsmodells zur Vorhersage von state-Testangst post (Besorgtheit) durch subjektive Leistung und die Moderatoren Selbstwertkontingenz und Bedingung

	Kriterium (Y): state-Testangst post			
	Koeff.	*SE*	*t*	*p*
Konstante	3.70	.18	21.15	.000
X: subjektive Leistung	.73	.17	4.35	.000
M: Selbstwertkontingenz	.30	.14	2.16	.032
W: Bedingung	.04	.24	.16	.873
Int 1: subj. Leistung x SWK	-.06	.15	-.38	.703
Int 2: subj. Leistung x Beding.	-.31	.25	-1.25	.214
Int 3: SWK x Beding.	.28	.19	1.48	.141
Int 4: subj. Leistung x SWK x Beding.	.27	.19	1.40	.164
		R^2 = .33		
		F (7, 144) = 9.93, p < .001		

N = 152; X und M jeweils mittelwertszentriert; Korrelation subj. Leistung & SWK: *r* = .24 (*p* = .002)

Aus Tabelle 32 geht hervor, dass sowohl subjektive Leistung (Koeff. = .73, $p < .001$) als auch Selbstwertkontingenz (Koeff. = .30, $p = .032$) signifikant positive Effekte auf state-TA post hatten. Keine der Interaktionen wurde signifikant. Eine getrennte Betrachtung der Interaktion von subjektiver Leistung und Selbstwertkontingenz nach Bedingung zeigt, dass in Bedingung A keine signifikante Interaktion vorlagen (Effekt = -.06, $p = .703$), in Bedingung B hingegen eine marginal signifikante Interaktion (Effekt = .21, $p = .085$). Hypothese 2b wurde somit nicht bestätigt.

Hypothese 2c bezog sich auf die Annahme, dass der Zusammenhang von subjektiver Leistung und state-TA post umso stärker ist, je höher die Messqualität des Tests eingeschätzt wird. Analog zu Hypothese 2b wurde mit PROCESS ein moderiertes Moderationsmodell (Model 3) berechnet, dessen Ergebnisse in Tabelle 33 aufgeführt sind.

Tabelle 33: Zusammenfassung des Regressionsmodells zur Vorhersage von state-Testangst post (Besorgtheit) durch subjektive Leistung und die Moderatoren Messqualität und Bedingung

	Kriterium (Y): state-Testangst post			
	Koeff.	*SE*	*t*	*p*
Konstante	3.71	.20	18.92	.000
X: subjektive Leistung	.82	.18	4.44	.000
M: Messqualität	.23	.23	1.02	.310
W: Bedingung	.17	.27	.61	.543
Int 1: subj. Leistung x MQ	.06	.21	.30	.763
Int 2: subj. Leistung x Beding.	-.21	.27	-.77	.441
Int 3: MQ x Beding.	-.25	.31	-.79	.431
Int 4: subj. Leistung x MQ x Beding.	-.05	.32	-.15	.880
		R^2 = .18		
		F (7, 144) = 4.56, p < .001		

N = 152; X und M jeweils mittelwertszentriert; Korrelation subj. Leistung & MQ: *r* = -.05 (*p* = .584)

Lediglich die subjektive Leistung hatte einen signifikanten Effekt auf die state-TA post (Koeff. = .82, $p < .001$). Die Messqualität (Koeff. = .23, $p = .310$) hatte keinen signifikanten Effekt, überdies wurde keine der Interaktionen signifikant. Eine Betrachtung der konditionalen Effekte zeigt, dass in beiden Bedingungen die Interaktion nicht signifikant war. Hypothese 2c wurde somit nicht bestätigt.

Hypothese 2d beinhaltete die Annahme, dass der Zusammenhang von subjektiver Leistung und state-TA post durch die berichtete mangelnde Anstrengung moderiert wird. Inhaltlich bedeutet dies, dass der besagte Zusammenhang umso schwächer ist, je stärker die eigene Testleistung auf mangelnde Anstrengung zurückgeführt wird. Analog zu den Hypothesen 2b und 2c wurde via PROCESS ein moderiertes Moderationsmodell (Model 3) berechnet, dessen Ergebnisse in Tabelle 34 berichtet sind.

Tabelle 34: Zusammenfassung des Regressionsmodells zur Vorhersage von state-Testangst post (Besorgtheit) durch subjektive Leistung und die Moderatoren mangelnde Anstrengung und Bedingung

	Kriterium (Y): state-Testangst post			
	Koeff.	*SE*	*t*	*p*
Konstante	3.64	.19	19.50	.000
X: subjektive Leistung	.78	.18	4.33	.000
M: Mangelnde Anstrengung	.09	.14	.62	.535
W: Bedingung	.23	.26	.87	.386
Int 1: subj. Leistung x MA	-.10	.13	-.78	.437
Int 2: subj. Leistung x Beding.	-.18	.27	-.67	.507
Int 3: MA x Beding.	-.03	.21	-.15	.882
Int 4: subj. Leistung x MA x Beding.	.11	.20	.53	.595
		$R^2 = .18$		
		$F(7, 144) = 4.59, p < .001$		

N = 152; X und M jeweils mittelwertszentriert; Korrelation subj. Leistung & MA: $r = .01$ ($p = .908$)

Wiederum hatte nur die subjektive Leistung einen signifikanten Effekt auf die state-TA post (Koeff. = .78, $p < .001$). Die berichtete mangelnde Anstrengung hatte keinen signifikanten Effekt auf die state-TA post (Koeff. = .09, $p = .535$). Keine der Interaktionen wurde signifikant. In beiden Bedingungen zeigte sich kein signifikanter Interaktionseffekt. Hypothese 2d wurde somit abgelehnt.

4.2.3 Weiterführende Analysen

In den Analysen zu Hypothese 2a-d wurde die Bedingung jeweils als Moderator in die Modelle integriert, um mögliche Effekte der Manipulation registrieren zu können. Demgegenüber wurde bei den Analysen zu Hypothese 1a und 1b bei der Betrachtung von state-TA post und deren Kor-

relaten Bedingung A und B zusammengelegt. Um die Ergebnisse gegen potenzielle Effekte der Bedingungs-Manipulation abzusichern, wurden die Analysen für die Hypothesen 1a und 1b wiederholt, wobei bei Betrachtung der Korrelate von state-TA post die Stichprobe jeweils in die Teilstichproben aus Bedingung A und B ausdifferenziert wurde. In Rekapitulation von Hypothese 1a wurden zunächst die Korrelationen von state-TA prä und post mit stabilen Dispositionen betrachtet (siehe Tabelle 35).

Tabelle 35: Bivariate Korrelationen von state-Testangst (Besorgtheit) und dispositionellen Variablen; state-Testangst post separiert nach Bedingung A und B

	State-Testangst		
	prä (nur A)	post (A)	post (B)
Selbstwert	-.26*	-.23	-.36**
Selbstwertkontingenz[1]	.21	.25*	.54**
Selbstkonzept eigener Fähigkeiten	-.08	-.25*	-.03
Testängstlichkeit	.30**	.49**	.52**

N: A = 74; B = 78; [1] Korrelationsunterschied zwischen post (A) und post (B) $p < .05$

Die Befunde veränderten sich nicht maßgeblich durch die Aufteilung. So waren die Interkorrelationen mit den dispositionellen Variablen teilweise bei B geringfügig höher als in A (für Selbstwert und Testängstlichkeit), teilweise in B niedriger als in A (für Selbstkonzept eigener Fähigkeiten). Der größte und auch einzig signifikante Unterschied in der Interkorrelation von state-TA post zwischen A und B zeigte sich für die Selbstwertkontingenz.

Bezüglich Hypothese 1b fanden sich kaum Unterschiede zwischen den Interkorrelationen von state-TA post in A und B mit den Aspekten von self-handicapping. Ausnahme waren die beiden Aspekte Müdigkeit und Mangelhaftigkeit des Tests, die in Bedingung B etwas stärker mit der state-TA post korrelierten. Dieser Unterschied war jedoch nicht signifikant. Die Korrelationen sind in Tabelle 36 berichtet. Der ursprüngliche Fokus auf die für Leistung bedeutsamste Facette (Besorgtheit) wurde außerdem ergänzt um die Korrelationen der SH-Aspekte zur Facette Interferenz. Dieser explorative Einbezug der Interferenz ermöglichte eine spezifischere Interpretation. Er bot sich insofern an, da sich einige der SH-Aspekte auf störende Einflüsse beziehen, die mental wirksam werden (z. B. Schlechte Laune und Ablenkung durch Umgebung). Dies sind Prozesse, die auch in der Facette Interferenz subsumiert werden (siehe Abschnitt 1.1.1.2.2) und in der Facette Besorgtheit nicht repräsentiert sind. Dabei fällt auf, dass Interferenz am stärksten mit dem Aspekt Ablenkung durch Umgebung korrelierte.

Tabelle 36: Bivariate Korrelationen von state-Testangst (Besorgtheit und Interferenz) mit Aspekten von self-handicapping; state-Testangst post separiert nach Bedingung A und B

	Stata-TA Besorgtheit			State-TA Interferenz		
	prä (nur A)	post (A)	post (B)	prä (nur A)	post (A)	post (B)
Müdigkeit	.04	.04	.20	.19	.17	.27*
Schlechte Laune	.17	.24*	.27*	.25*	.29*	.34**
Nervosität	.50**	.64*	.68*	.29*	.41**	.22
Ablenkung durch Umgebung	.18	.18	.12	.64**	.60**	.54**
Krankheit / körperl. Unwohlsein	.18	.31**	.26**	.28*	.33**	.21
Stress im Alltag	.33**	.24*	.24*	.28*	.20	.31**
Mangelnde Anstrengung	.03	-.01	.03	.34**	.30**	.30**
Prüfungsangst	.46**	.56**	.48**	.29*	.29*	.37**
Mangelhaftigkeit des Tests	.11	.12	.29*	.27*	.20	.16

N für prä = 74; *N* für post = 152; keine sign. Korrelationsunterschiede im Vergleich von post (A) und post (B)

Zusätzlich wurden Korrelationen zwischen den SH-Aspekten und subjektiver Leistung berechnet. Hier waren lediglich die Korrelationen zu Nervosität ($r = .37$, $p < .001$) und Prüfungsangst ($r = .38$, $p < .001$) signifikant. Zu den Aspekten Schlechte Laune ($r = .16$, $p = .057$) sowie Krankheit / körperliches Unwohlsein ($r = .15$, $p = .068$) lagen marginal signifikante Korrelationen vor. Alle weiteren Korrelationen wurden nicht signifikant.

Bezüglich Hypothese 2a wurden die Korrelationen von state-TA post zu subjektiver und objektiver Leistung nach Bedingung A und B differenziert (siehe Tabelle 37).

Tabelle 37: Bivariate Korrelationen von state-Testangst prä und post (Besorgtheit) mit objektiver und subjektiver Leistung, differenziert nach Bedingung A und B

	State-Testangst		
	prä (nur A)	post (A)	post (B)
Leistung			
Objektiv[1]	-.25*	-.43**	-.10
Subjektiv	.27*	.48**	.35**

N: A = 74; B = 78; [1] Korrelationsunterschied zwischen post (A) und post (B) $p < .05$

Die Analyse zeigte einen deutlichen und signifikanten Unterschied zwischen Bedingung A und B in der Interkorrelation der state-TA post mit objektiver Leistung ($r = -.43$, $p < .001$ bzw. $r = -.10$, $p = .372$). Bezüglich der Interkorrelationen mit der subjektiven Leistung fand sich ebenfalls ein Unterschied, der jedoch nicht so ausgeprägt war. Dieses Ergebnis lässt sich vereinbaren mit der Analyse in Tabelle 31, der zufolge ein direkter und indirekter Effekt der objektiven Leistung auf die state-TA post nur in Bedingung A vorlag.

Darüber hinaus wurde die Korrelation zwischen objektiver und subjektiver Leistung in den beiden Bedingungen verglichen. In beiden Bedingungen lag eine signifikante Korrelation vor, wobei

das negative Vorzeichen bedeutet, dass mit einer besseren objektiven Leistung auch eine bessere subjektive Leistungseinschätzung vorgenommen wurde (aufgrund der Kodierung der subjektiven Leistung von 1 bis 6). In Bedingung A (r = -.40, p < .001) fand sich dabei eine etwas höhere Korrelation als in Bedingung B (r = -.26, p = .020), obgleich dieser Unterschied statistisch nicht bedeutsam war (Fishers' z = -.95, p = .341).

In den Analysen zu Hypothese 2b hatte sich Selbstwertkontingenz neben der subjektiven Leistung als bedeutsamer Prädiktor von state-TA post erwiesen. Das heißt, dass mit zunehmender Selbstwertkontingenz auch mehr state-TA post berichtet wurde (unter Kontrolle der subjektiven Leistung). Zwar moderierte die Messqualität nicht die Relation von subjektiver Leistung und state-TA post (Hypothese 2c), jedoch ist es denkbar, dass eine hohe Selbstwertkontingenz nicht *per se* mit mehr state-TA post einhergeht, sondern nur, wenn das Testergebnis auch aus subjektiver Sicht Rückschlüsse auf die eigenen Fähigkeiten erlaubt. Letzteres Urteil spiegelt sich in der erlebten Messqualität des Tests. Analysiert wurde daher, ob die Relation von Selbstwertkontingenz und state-TA post durch die Messqualität moderiert wird. Analog zu den Hypothesen 2b-d wurde wiederum die Bedingung als zweiter Moderator in das Modell inkludiert, so dass eine moderierte Moderation gerechnet wurde. Die subjektive Leistung wurde dabei als Kovariate in das Modell inkludiert, um deren Effekt konstant zu halten (siehe Tabelle 38).

Tabelle 38: Zusammenfassung des Regressionsmodells zur Vorhersage von state-Testangst post (Besorgtheit) durch Selbstwertkontingenz und die Moderatoren Messqualität und Bedingung (Kovariate: subjektive Leistung)

	Kriterium (Y): state-Testangst post			
	Koeff.	*SE*	*t*	*p*
Konstante	3.69	.18	20.83	.000
X: Selbstwertkontingenz	.33	.15	2.28	.024
M: Messqualität	.14	.21	.68	.499
W: Bedingung	.11	.25	.46	.649
Int 1: SWK x MQ	.10	.15	.66	.513
Int 2: SWK x Beding.	.18	.19	.91	.363
Int 3: MQ x Beding.	-.17	.28	-.61	.543
Int 4: SWX x MQ x Beding.	.11	.20	.57	.572
Kovariate: subjektive Leistung	.58	.12	4.68	.000
		R^2 = .32		
		F (8, 143) = 8.36, p < .001		

N = 152; X und M jeweils mittelwertszentriert

Entgegen der Erwartung zeigte sich kein signifikanter Interaktionseffekt. Eine separate Betrachtung zeigte, dass dies für beide Bedingungen zutraf. Somit wurde diese explorative Annahme nicht bestätigt.

4.2.4 Zusammenfassung

Die explorativen Analysen zeigen, dass die Erfassung der state-TA vor dem Test relativ geringe Effekte nach sich zieht. Die Pbn in Bedingung A, bei der die state-TA vor dem Test erfasst wurde, berichteten ein höheres Anspruchsniveau und schätzten Messqualität und Augenscheinvalidität des Tests niedriger ein. Bezüglich der self-handicapping-Aspekte berichteten die Pbn in Bedingung A eine niedrigere Beeinträchtigung durch Nervosität und Prüfungsangst.

Hypothese 1a, wonach die state-TA prä stärker mit Selbstwert, Selbstwertkontingenz, Selbstkonzept eigener Fähigkeiten und Testängstlichkeit korrelieren sollte als state-TA post, wurde nicht bestätigt. Für keine dieser vier Eigenschaften zeigte sich das erwartete Bild. Hypothese 1b wurde bestätigt, da die state-TA post überwiegend stärker mit den self-handicapping-Aspekten korrelierte als die state-TA prä.

Gemäß Hypothese 2a sollte die Relation von objektiver Leistung und state-TA post durch die subjektive Leistung mediiert werden. Es wurden ein direkter und ein indirekter Effekt gefunden, jedoch ausschließlich in Bedingung A. Die Relation von subjektiver Leistung und state-TA post wurde entgegen der Erwartung nicht durch Selbstwertkontingenz, Messqualität und die berichtete Leistungsbeeinträchtigung moderiert (Hypothese 2b, 2c und 2d). Dabei wurde jedoch festgestellt, dass Selbstwertkontingenz, kontrolliert um die subjektive Leistung, ein signifikanter Prädiktor der state-TA post war.

In den weiterführenden Analysen wurden die Ergebnisse der Hypothesen 1a und 1b reanalysiert, wofür die Korrelationen von state-TA post mit dispositionellen Variablen, self-handicapping sowie subjektiver und objektiver Leistung in den Substichproben aus Bedingung A und B verglichen wurden. Die Ergebnisse zu den Hypothesen 1a und 1b veränderten sich dabei nur geringfügig. Vorbereitende Analysen zu Hypothese 2a ergaben, dass der Zusammenhang zwischen state-TA post und objektiver Leistung in Bedingung B nur schwach war (r = -.10), in Bedingung A jedoch substanziell (r = -.43). Eine zusätzliche Analyse konnte die Vermutung, dass die Relation von Selbstwertkontingenz und state-TA post von der Messqualität moderiert wird, nicht bestätigen.

4.3 Diskussion

Im experimentellen Setting von Studie 1 sollten etwaige Effekte der prä-Messung der state-TA analysiert und die state-TA prä und post insbesondere bezüglich der Korrelation zur objektiven Leistung verglichen werden. Zentral war darüber hinaus die systematische Berücksichtigung der subjektiven Leistung als mutmaßlicher Schlüsselvariable in der Determination der state-TA post. Die Prüfung der Defizitperspektive durch das beschriebene Mediationsmodell wurde in dieser Form bislang noch nicht vorgenommen.

4.3.1 Bewertung der explorativen Analysen und Hypothesen

Die explorativen Analysen legen den Schluss nahe, dass die Erfassung von state-TA vor einem Test nur wenig Effekte nach sich zieht. Keine Unterschiede zeigten sich bei der state-TA post sowie bei der Kontrollierbarkeit, der Belastungsfreiheit, der Gesamtbewertung des Tests und der subjektiven Leistung. Insbesondere war auch kein Unterschied in der objektiven Leistung festzustellen. Ein Effekt auf die Leistung und auch auf die state-TA post wäre insofern problematisch, da dies im Allgemeinen zu einer unerwünschten Veränderung von Testergebnissen führen könnte. Wäre dies der Fall, so würde sich eine Erfassung von state-TA vor einem Test geradezu verbieten. Dennoch zeigten sich einige Unterschiede, die Hinweise auf eine veränderte Wahrnehmung der Situation durch die Pbn in Bedingung A geben. Das von den Pbn in Bedingung A berichtete höhere Anspruchsniveau und die niedrigere Beeinträchtigung durch Nervosität und Prüfungsangst scheinen auf den ersten Blick überraschend. Womöglich wurden die Pbn aber durch die Erhebung der state-TA vor dem Test in einem gewissen Sinne geprimt, so dass der Bewertungscharakter der Situation salienter und ihnen die Zielsetzung der Untersuchung (Erleben von Angst während des Tests) bewusster geworden ist.

Das Empfinden, getestet und in einer Studie gezielt einer Stresssituation ausgesetzt zu werden, könnte zu einer Reaktanzreaktion geführt haben. Kray et al. (2001) definierten im Kontext des STT (siehe Abschnitt 1.2.2.2.2) Stereotyp-Reaktanz als „the tendency, to behave in a manner inconsistent with a stereotype“ (S. 942), welche in Reaktion auf die offenkundige Aktivierung von Stereotypen auftritt. Übertragen auf Studie 1 würde Reaktanz bedeuten, sich *nicht* so zu verhalten, wie es dem Bild einer testängstlichen Person entspräche. Das hieße konkret, das Erleben von Testangst als aversiver Emotion im Selbstbericht gewissermaßen zu *negieren* und bestimmte Fragen entsprechend zu beantworten. Möglicherweise wurde der Test dadurch auch stärker als Herausforderung bewertet, was insbesondere das höhere Anspruchsniveau erklären könnte. Auch dies könnte als Reaktanzeffekt gedeutet werden, nämlich als Wunsch, in einer offenkundigen Drucksituation eine gute Leistung zu erbringen. In dieselbe Richtung lässt sich die kritischere Bewer-

tung des Tests deuten, die sich in der niedrigeren Messqualität und Augenscheinvalidität in Bedingung A ausdrückt. Diese beiden Skalen beinhalten ein Urteil über die Aussagekraft und Generalisierbarkeit der Testergebnisse. Überspitzt betrachtet wurden also negative Konsequenzen von Testangst weniger berichtet und gleichzeitig die diagnostische Qualität des Tests strenger bewertet. Dies kann, muss aber keineswegs zu dem Zweck erfolgt sein, einen möglichen Misserfolg external (auf die Unzulänglichkeit des Tests) zu attribuieren.

Ferner widersprechen die Ergebnisse der Annahme in Hypothese 1a, dass state-TA prä stärker mit den erhobenen Dispositionen zusammenhängt als state-TA post. Bei allen vier Eigenschaften (Selbstwert, Selbstwertkontingenz, Selbstkonzept eigener Fähigkeiten, Testängstlichkeit) zeigten sich höhere Korrelationen mit state-TA post als mit state-TA prä. Erwartet wurde, dass das Erleben vor dem Test stärker mit stabilen Eigenschaften korrespondiert. Dass dies nicht auftrat, scheint den Ergebnissen von Ringeisen und Buchwald (2010) zu widersprechen. Jedoch erhoben die beiden Autoren die state-Emotionen rund drei Wochen vor der eigentlichen Prüfung und in derselben Erhebung, in der auch die trait-Testängstlichkeit erhoben wurde. Zwar wurden in Studie 1 Eigenschaften und state-TA auch in der gleichen Erhebung erfasst. Der entscheidende Unterschied ist aber, dass die Messung der state-TA prä in Studie 1 de facto bereits in der Testsituation stattgefunden hat. So wurde die state-TA prä direkt nach der Ankündigung des Tests und den Beispielitems erhoben. Durch die Beispielitems hatten die Pbn einen präzisen Eindruck von der Schwierigkeit und dem Typ der Aufgaben. So unterschied sich das Erleben der state-TA prä und post nicht fundamental, wofür auch die Korrelation zwischen state-TA prä und post (nur Bedingung A) von $r = .62$ ($p < .001$) spricht[37].

Unklar ist aber, warum die state-TA post tendenziell höher mit den erfassten Eigenschaften korrelierte als die state-TA prä. Am deutlichsten ist der Unterschied in den Korrelationen bei der Testängstlichkeit und der Selbstwertkontingenz. Es lässt sich darüber spekulieren, dass die state-TA post sich auf das Erleben während der Bearbeitung des GkKT-K bezog, während Testängstlichkeit sich allgemein auf das Erleben in Prüfungssituationen bezog. Demgegenüber bezog sich state-TA prä zwar auch auf das Erleben in der Testsituation, jedoch nicht spezifisch auf die unmittelbare Aufgabenbearbeitung. Selbstwertkontingenz beschreibt letztlich Reaktionen auf Leistungsresultate (und deren Auswirkungen auf den Selbstwert). Aus dieser Perspektive ist es plausibel, dass Selbstwertkontingenz stärker mit der state-TA post korreliert, die ja auch unter dem Eindruck des Tests (und der subjektiven Leistung) berichtet wird, während die state-TA prä wahrscheinlich eher Leistungserwartungen abbildet.

Die Befunde zu Hypothese 1b sind erwartungsgemäß. So lagen bei den betrachteten Interkorrelationen mit SH-Aspekten in sechs von neun Fällen höhere Zusammenhänge bei state-TA post als

[37] Wobei dies sicherlich auch ein Methodeneffekt ist, da die Items mit Ausnahme des Tempus identisch sind.

bei state-TA prä vor. Eine naheliegende Interpretation ist, dass sich state-TA post und die SH-Aspekte retrospektiv auf die Testbearbeitung bezogen und diese Variablen daher stärker miteinander zusammenhängen. Dass sowohl state-TA prä als auch state-TA post die mit Abstand stärksten Zusammenhänge mit den SH-Aspekten Nervosität und Prüfungsangst aufwiesen, ist aufschlussreich. Dies spricht dafür, dass das Erleben von Testangst auch mit dem Empfinden zusammenhängt, dass diese die eigene Leistung beeinträchtig hat. Auch die in den weiterführenden Analysen betrachteten Korrelationen der Facette Interferenz fügen sich in dieses Bild: im Gegensatz zur Facette Besorgtheit korrelierte Interferenz am deutlichsten mit dem SH-Aspekt Ablenkung durch Umgebung. Insgesamt bedeutet dies, dass das Erleben (sorgenvolle Gedanken oder aber das Abschweifen der eigenen Gedanken) auch mit jenen Prozessen in Verbindung steht, die als leistungsbeeinträchtigend berichtet wurden: Pbn, die während des Tests besorgt waren, berichteten auch, dass Prüfungsangst ihre Leistung beeinträchtigt hat; Pbn, deren Gedanken während des Tests abschweiften, berichteten auch, dass Ablenkung durch die Umgebung sie beeinträchtigt hat. SH-Aspekte sind somit keineswegs Faktoren, die völlig unabhängig vom Testerleben berichtet werden, um als Erklärung für eine suboptimale Leistung zu fungieren. Unterstützung für diesen Schluss liefern die Befunde, dass Testängstliche häufiger aufgabenirrelevante Gedanken haben (z. B. Sarason & Stoops, 1978) und ihre Aufmerksamkeit nicht nur auf die Aufgabenanforderungen, sondern auch auf die eigene Person richten (Wine, 1971). Berichtet also eine testängstliche Person mehr Beeinträchtigung durch schlechte Laune oder Ablenkung durch die Umgebung, kann dies Ausdruck davon sein, dass diese Person einen Selbstfokus („self-preoccupation“) aufweist und objektiv stärker ablenkbar ist als niedrig testängstliche Personen (Eysenck, 1992; Sarason & Sarason, 1990).

Abweichungen der Korrelationsmuster von state-TA prä und post sind erkennbar, aber nicht gravierend. So existieren sowohl zu den betrachteten Dispositionen Selbstwert, Selbstwertkontingenz, Selbstkonzept und Testängstlichkeit als auch zu den SH-Aspekten ähnliche Zusammenhangsmuster zu state-TA prä und post. Dass die Zusammenhänge dieser Variablen zu state-TA prä und post jeweils unterschiedlich hoch ausfielen, lässt sich auf Basis inhaltlicher Überlegungen erklären. Bei der Differenzierung der state-TA in Besorgtheit und Interferenz wird deutlich, dass die berichteten SH-Aspekte durchaus mit unterschiedlichen kognitiven Aspekten von Testangst korrespondierten.

Die Ergebnisse zur Hypothese 2a sind komplex. Bei einer reinen Betrachtung von Bedingung A zeigte sich das erwartete Bild, dass die state-TA prä (r = -.25) schwächer mit der Leistung korrelierte als die state-TA post (r = -.43). Dass die state-TA post demgegenüber in Bedingung B kaum mit der objektiven Leistung korrelierte (r = -.10), ist unerwartet. Eine mögliche Erklärung ist, dass

sich durch die vor dem Test erfolgte Abfrage der state-TA prä der Situationscharakter verändert hat. Es wurde bereits diskutiert, dass die Pbn womöglich eine Reaktanzreaktion gezeigt haben, erkennbar an der strengeren Bewertung des Tests und dem erhöhten Anspruchsniveau. Aus den Daten kann nicht abgeleitet werden, dass die Pbn in A mehr Bewertungsstress erlebt haben als in B. Zu beachten ist, dass es zwischen Bedingung A und B keinen bedeutsamen Mittelwertsunterschied in der state-TA post und auch keinen erheblichen Unterschied in den Standardabweichungen gab, sich aber die Korrelationen mit der objektiven Leistung unterschieden. Eine ähnliche Konstellation fand sich auch in der bereits zitierten Untersuchung von Zeidner (1991). Hier unterschieden sich bei der vor und nach einer Prüfung erhobenen Testängstlichkeit die Mittelwerte nicht und die Standardabweichungen nur geringfügig, jedoch lagen unterschiedlich hohe Korrelationen mit der Leistung vor. Möglicherweise wurde in Studie 1 durch die Erfassung der state-TA prä die Salienz der Bewertungssituation verstärkt, *ohne* dass die Pbn de facto mehr Bewertungsstress verspürt haben. Die Pbn könnten durch die Fragen zur state-TA prä zu einer verstärkten Selbstreflexion angeregt worden sein und womöglich die entsprechenden Prozesse wie die eigene Nervosität oder sorgenvolle Gedanken eher registriert haben, weshalb die berichtete state-TA post stärker mit der objektiven Leistung zusammenhing als in Bedingung B. Die höhere Korrelation von state-TA post mit der subjektiven Leistung in Bedingung A gegenüber Bedingung B (r = .48 vs. .35) weist ebenfalls in diese Richtung. Dafür, dass die Pbn in Bedingung A ihr Erleben (und Verhalten) im Test intensiver beobachtet haben, spricht darüber hinaus auch die stärkere Korrelation von objektiver und subjektiver Leistung in Bedingung A gegenüber Bedingung B (r = -.40 vs. -.26). Die Pbn in A haben also ihre eigene Leistung zutreffender eingeschätzt. Womöglich waren sie emotional stärker in die Aufgaben involviert und konnten dadurch auch besser einschätzen, wie viele Items sie richtig bzw. falsch gelöst hatten.

Die Ergebnisse zu Hypothese 2a unterschieden sich demzufolge – und unerwartet – in Abhängigkeit der Bedingung. Die Analysen zeigen, dass in Bedingung A die objektive Leistung einen direkten sowie einen indirekten, über die subjektive Leistung vermittelten Effekt auf die state-TA post hatte. Somit wurde ein Mediationseffekt gefunden. Demnach sind subjektive *und* objektive Leistung Determinanten der state-TA post. Es besteht folglich ein Zusammenhang zwischen objektiver Leistung und state-TA post, der *nicht* allein durch die subjektive Leistung erklärt werden kann. Einer strengen Auslegung der Defizitperspektive folgend müsste der Zusammenhang zwischen objektiver Leistung und state-TA post verschwinden, wenn die subjektive Leistung kontrolliert wird. Dies war aber nicht der Fall. Im Umkehrschluss bedeutet dies, dass der Zusammenhang von objektiver Leistung und state-TA post tatsächlich kausaler Natur sein *kann*, in der Weise, dass hohe state-TA zu einer objektiven Leistungsbeeinträchtigung führt. In einem statistischen Sinne lässt der direkte Effekt der objektiven Leistung auf die state-TA post auf gemeinsame Varianz schließen, die nicht durch subjektive Leistung erklärt wird. Diese gemeinsame Varianz kann auf

die leistungsbeeinträchtigende Wirkung der Testangst zurückgehen. Dieses Ergebnis ist noch nicht als „Beweis“ für eine kausale Erklärung zu verstehen, lässt aber Raum für die Gültigkeit von Interferenz- *und* Defizitperspektive. Dabei ist der indirekte Effekt über die subjektive Leistung kein Argument gegen die Gültigkeit der Interferenzperspektive. Personen, die aufgrund ihrer Testangst schlecht abschneiden, dürften in den meisten Fällen auch registrieren, dass ihnen die gestellten Aufgaben oder Fragen einer Prüfung Probleme bereiten. Eine wesentliche Einschränkung dieser Interpretation ist, dass sich die Mediation nur in Bedingung A, nicht aber in Bedingung B zeigte. Auf Erklärungen für die Divergenzen zwischen Bedingung A und B wurde oben bereits eingegangen.

Das in Hypothese 2b geprüfte Moderationsmodell zeigt, dass neben der subjektiven Leistung die Selbstwertkontingenz einen signifikanten Effekt auf die state-TA post hatte. Kontrolliert um den Prädiktor subjektive Leistung hatte also auch die „Volatilität“ des Selbstwerts einen Effekt auf die state-TA post. Die erwartete Moderation des Effekts der subjektiven Leistung auf die state-TA post durch die Selbstwertkontingenz trat jedoch nicht auf. Auch Hypothese 2c und 2d wurden nicht bestätigt: der Effekt der subjektiven Leistung auf die state-TA post wurde nicht durch Messqualität und die berichtete mangelnde Anstrengung moderiert. Dass die Abhängigkeit des Selbstwerts von Erfolg oder Misserfolg in Leistungssituationen auch mit der retrospektiv berichteten Testangst einherging, ist plausibel. Dies ist umso beachtlicher, da insbesondere die Items zur Kompetenzkontingenz sich explizit auf sozial sichtbare Misserfolge und deren Auswirkungen auf den Selbstwert beziehen (z. B. „Ich fühle mich minderwertig, wenn andere merken, dass ich etwas nicht gut beherrsche.“): den Pbn wurde mitgeteilt, dass nur sie ihr Ergebnis erfahren würden (und auch nur auf Wunsch), was die Sichtbarkeit eines (Miss-)Erfolgs begrenzt hätte. Unerwartet ist, dass sowohl Selbstwertkontingenz als auch Messqualität den Effekt der subjektiven Leistung auf die state-TA post nicht moderierten. Beide Hypothesen bauten auf dem Gedanken auf, dass (der Bericht von) Testangst in erster Linie eine Konsequenz aus (subjektiv) schlechter Leistung ist, was umso stärker der Fall sein sollte, je empfindlicher der Selbstwert für Misserfolg ist (Selbstwertkontingenz) und je valider der Test eingeschätzt wird (Messqualität). Diese Perspektive ist offenbar zur Beschreibung der stattfindenden Prozesse nicht hinreichend. Jenseits der Situation, dass eine Person Angst erlebt (und berichtet), weil sie viele Aufgaben nicht lösen konnte (Defizitperspektive), liefern die Ergebnisse zu Hypothese 2a Hinweise, dass auch weitere Prozesse stattfinden: eine Person kann viele Aufgaben nicht lösen, weil sie Angst erlebt (Interferenzperspektive). Eine extreme Position, dass (der Bericht von) state-TA post *ausschließlich* ein Produkt von Regulationsprozessen ist, um den Selbstwert vor einem subjektiven Misserfolg zu schützen, lässt sich damit nicht halten.

Die Ergebnisse zu Hypothese 2d passen in dieses Bild. Die selbst berichtete mangelnde Anstrengung moderierte den Effekt der subjektiven Leistung auf die state-TA post nicht. Die subjektive

Beeinträchtigung durch mangelnde Anstrengung war ursprünglich als eine „Alternativattribution“ konzipiert: Pbn, die angeben, sich nicht angestrengt zu haben, haben eine „gute“ Erklärung für eine schlechte Leistung – sie haben schlichtweg keine Mühe in die Aufgabe investiert. Ein schlechtes Ergebnis ermöglicht daher auch keine Rückschlüsse auf ihre Fähigkeiten (während ein gutes Ergebnis die eigenen Fähigkeiten umso mehr belegt). Somit gäbe es für sie keinen Grund, state-TA post zu berichten. In Abschnitt 2.1 wurde die hypothetische Vorannahme explizit formuliert, dass Pbn state-TA als Reaktion auf das eigene Abschneiden im Test berichten. Dass auch Hypothese 2d nicht bestätigt wurde, spricht gegen diese Interpretation. Darüber hinaus scheint diese Vorannahme auch für die berichtete mangelnde Anstrengung nicht angemessen zu sein, da letztere nicht mit der subjektiven Leistung korrelierte. Auch die Korrelationen der state-TA Facetten Besorgtheit und Interferenz mit der Beeinträchtigung durch Nervosität / Prüfungsangst und Ablenkung durch Umgebung sprechen gegen besagte Vorannahme. Die (extreme) Position, dass die state-TA post völlig vom tatsächlichen Testerleben entkoppelt ist, muss auch aufgrund der hohen Korrelation von state-TA prä und post (r = .62) und des bereits berichteten direkten Effekts der objektiven Leistung auf die state-TA post (in Bedingung A) verworfen werden.

Wichtig ist zu betonen, dass die subjektive Leistung keineswegs irrelevant ist in der Prädiktion der state-TA post. Die diskutierten Ergebnisse zeigen aber, dass sich das Erleben von Testangst im Test durchaus im Bericht der state-TA post niederschlägt und letzteres nicht ausschließliches Resultat mechanistischer oder „rationaler“ Regulationsprozesse ist („wenn ich schlecht war, berichte ich Angst“). Wahrscheinlich vereinfacht die Annahme einer Wirkrichtung von subjektiver Leistung auf die state-TA post die tatsächlichen Prozesse übermäßig. Es ist davon auszugehen, dass während des Tests ein fortlaufendes Wechselspiel stattfindet zwischen dem Erleben von Testangst und dem, was später als subjektive Leistung berichtet wird. Beispielsweise kann das Erleben von physiologischen Angstsymptomen als Signal für ein Scheitern an den Anforderungen erlebt werden (siehe auch Bandura, 1977). Umgekehrt werden Schwierigkeiten beim Lösen einer Aufgabe sorgenvolle Kognitionen über das eigene Abschneiden verstärken oder gar erst auslösen. Dies wiederum könnte die Sensibilität für die Richtigkeit der eigenen Antworten steigern.

4.3.2 Limitationen

Im Folgenden sollen einige Limitationen der obigen Interpretationen erörtert werden. Zunächst ist die vergleichsweise kleine Stichprobe zu erwähnen. Da die Studie als Online-Erhebung durchgeführt wurde und die Pbn per E-Mail eingeladen wurden, kann ein Selektionseffekt nicht ausgeschlossen werden. So wurde in der Einladung erwähnt, dass es um Verhalten in Leistungssituationen gehe, was stark ängstliche Personen abgeschreckt haben könnte. Dies ist aber unwahrscheinlich, da sich die Ausprägung der Testängstlichkeit bei Männern und Frauen nicht von den

zitierten Vergleichswerten von Wacker et al. (2008) unterschied. Obwohl es keine theoretische Begründung gibt, warum sich die untersuchten Effekte bei den Geschlechtern unterscheiden sollten, ist die Unterrepräsentation von Männern in der Stichprobe (22%) zu bemängeln.

Eine weitere Limitation stellt die Erhebung in einem Online-Kontext dar. Eine Aufsicht über die Durchführungsbedingungen konnte nicht stattfinden. Eine Maßnahme zur Begrenzung möglicher Störeffekte war die ausschließliche Analyse vollständiger Bearbeitungen. Insgesamt sprechen die Ergebnisse für eine instruktionsgemäße Bearbeitung der eingesetzten Fragebögen sowie des GkKT-K. Dies stützt sich auf die gefundenen Mittelwerte und Reliabilitätskoeffizienten.

Ebenfalls zu den Limitationen gehören die Verteilungen der Messwerte für state-TA prä und post, die tendenziell Bodeneffekte aufwiesen, insbesondere bei der Besorgtheit. Jedoch ist dieses Ergebnis für eine low-stakes Testung erwartungskonform. Der über die Skalen Besorgtheit und Aufgeregtheit berechnete Mittelwert bei der state-TA prä (M = 1.72) entspricht fast genau dem Vergleichswert von Englert et al. (2011) (M = 1.73). Positiv zu werten ist, dass es offensichtlich gelungen ist, eine evaluative Testatmosphäre zu induzieren, obwohl die Online-Erhebung einer realen Prüfungssituation sehr unähnlich war (u. a. kein Untersuchungsraum, kein direkter Kontakt zum Prüfer bzw. Versuchsleiter).

Dass es sich um keine echte Prüfungssituation handelte, hat auch weitergehende Relevanz für die Interpretation. Insbesondere die Bedeutsamkeit der subjektiven Leistung dürfte in einer benoteten Prüfung deutlich höher sein. Eine subjektive Abwertung der Aussagekraft eines Tests sollte in „künstlichen" Testsituationen wie der hier vorliegenden einfacher sein als in einer realen Prüfungssituation mit echten Konsequenzen für den eigenen Notenspiegel. Zwar kann eine akademische Prüfung von einem Prüfling ebenfalls als nicht valide eingeschätzt werden, jedoch wird dies erschwert durch die Note, welche eine formelle Beurteilung auf normativer oder kriterienbezogener Basis darstellt. So kann man die (subjektive) Aussagekraft einer Prüfung relativieren, die Note selbst und deren Konsequenzen sind jedoch nicht veränderbar.

Ohne Zweifel stellen auch die unerwarteten Divergenzen in den Ergebnissen zwischen Bedingung A und B eine Einschränkung für die Generalisierbarkeit der Ergebnisse dar. Auf mögliche Erklärungen wurde im Rahmen der Diskussion bereits eingegangen.

Die wichtigste Limitation ist, dass die subjektive Leistung erfasst wurde, *nachdem* die state-TA post erhoben wurde. Im in Hypothese 2a geprüften Mediationsmodell jedoch war die subjektive Leistung eine unabhängige, state-TA post die abhängige Variable. Hintergrund für diese Reihenfolge in der Erhebung war die Vermutung, dass eine Erfassung der subjektiven Leistung direkt nach dem Test den späteren Bericht der state-TA post in unbestimmter Weise beeinflusst hätte.

Die Untersuchung der vermuteten Regulationsprozesse bezüglich des Zusammenhangs von subjektiver Leistung und state-TA war zwar expliziter Untersuchungsgegenstand. Jedoch sollte eine bewusste „Provokation“ derartiger Prozesse durch die vorherige Erfassung der subjektiven Leistung verhindert werden. Die letztlich gewählte Reihenfolge stellte somit in der Untersuchungsplanung die Variante mit den wenigsten Nachteilen dar. Für eine spezifischere Analyse möglicher Reihenfolgeeffekte ist eine Variation der Erhebungsabfolge in einem modifizierten Untersuchungsdesign nötig. Darüber hinaus wurde bereits darauf eingegangen, dass ein ausschließlich unidirektionaler Kausalzusammenhang zwischen subjektiver Leistung und dem Bericht von state-TA post ohnehin unwahrscheinlich ist, da vermutlich komplexe Wechselwirkungen vorliegen.

Zwei weitere Limitationen beziehen sich auf die SH-Aspekte. Es wurde zwar gefragt, inwieweit verschiedene Einflüsse die eigene Leistung beeinträchtigt haben; nicht gefragt wurde jedoch, ob sich die Pbn – trotz dieser Einflüsse – in der Lage gefühlt haben, ihre bestmögliche Leistung abzurufen. Hier wäre eine Frage hilfreich gewesen, welches Ergebnis man unter optimalen Bedingungen (keine äußeren Störungen, gute Konzentrationsfähigkeit, normales Wohlbefinden) hätte erreichen können. Eine derartige Frage wäre darüber hinaus ein direkterer Indikator für die Attribution eines subjektiv schlechten Ergebnisses auf externale Ursachen (z. B. Ablenkung durch Umgebung) oder internale, aber variable Ursachen (z. B. Anstrengung). Zweitens kann beim SH-Aspekt Prüfungsangst nicht eindeutig behauptet werden, dass sich die Beeinträchtigung der Leistung auf das Erleben von Prüfungsangst *während des Tests* bezog und nicht auf allgemeine Prüfungsangst aufgrund aktuell im Studium bevorstehender Prüfungen. Allerdings korrelierte der SH-Aspekt Nervosität, der sich eindeutiger dem eigentlichen Testerleben zuordnen lässt, hoch mit dem SH-Aspekt Prüfungsangst, $r = .63$ ($p < .001$). Es ist somit wahrscheinlich, dass sich auch letzterer Aspekt vorwiegend auf das Erleben im Test bezog.

Schließlich stellt die tatsächliche emotionale Involvierung und subjektive Relevanz des eigenen Testergebnisses einen „missing link“ in der Interpretation der Ergebnisse dar. Die berichtete Augenscheinvalidität und Messqualität des GkKT-K liefern hierfür nur indirekte Informationen. Dafür, dass die Pbn den Test und dessen Ergebnis ernst genommen haben, spricht aber der positive Zusammenhang von Selbstwertkontingenz und state-TA prä ($r = .21$, $p = .077$) sowie der Wunsch nach einer Ergebnisrückmeldung bei der Mehrzahl der Pbn (76 %).

4.3.3 Implikationen

4.3.3.1 Theoretische und praktische Implikationen

Zunächst sollen die theoretischen Beiträge in Bezug auf Interferenz- und Defizitperspektive diskutiert werden. Die Ergebnisse (in Bedingung A) lassen Raum sowohl für die Interferenz- als auch

die Defizitperspektive. Subjektive Leistung war eine bedeutsame Determinante der state-TA post, allerdings nicht die einzige. Auch die objektive Leistung wies einen direkten Effekt auf die state-TA post auf, was möglicherweise auf einen kausalen Effekt der state-TA post auf die objektive Leistung zurückzuführen ist. Somit stützen die Ergebnisse das in Abschnitt 1.2.1.4 gezogene Fazit, dass beide Theorieansätze nicht als konkurrierend, sondern als sich ergänzend betrachtet werden sollten. Dabei gilt zu beachten, dass das geprüfte Mediationsmodell keine unidirektionale Kausalität unterstellt. State-TA post wurde in dem Modell als abhängige Variable modelliert, da deren Determination im Fokus der Fragestellung stand. Diese Variable besitzt aber im Modell eine doppelte theoretische Rolle: state-TA ist einerseits kausale Ursache schlechter Leistung, aber auch beeinflusst durch die subjektive Leistung, die wiederum mit der objektiven Leistung in Zusammenhang steht. Ursache dieser scheinbar paradoxen Situation ist, dass die state-TA post als Selbstberichtsmaß die während des Tests erlebte Testangst abbildet, aber als retrospektiv erhobene Variable unvermeidbar unter dem Eindruck des Testerlebens steht. Greenwald et al. (2002) zufolge sind Selbstberichte zwei wesentlichen Einschränkungen unterworfen, nämlich „introspective limits“ und „response factors“. So wird angenommen, dass bestimmte Aspekte des Erlebens dem bewussten Bericht unzugänglich sind, während andere Aspekte, die zugänglich sind, im Bericht durch dritte Variablen verzerrt werden. Dies sind beispielsweise „demand characteristics“, also Hinweisreize auf ein vermutetes Untersuchungsziel, das sich auf das Erleben und Verhalten eines Pbn auswirkt (Orne, 1962). Ein weiterer Faktor ist die soziale Erwünschtheit (Crowne & Marlowe, 1960). Darüber hinaus erleben Teilnehmer psychologischer Untersuchungen nach Rosenberg (1965) häufig eine „evaluation apprehension“: an „active, anxiety-toned concern that he [gemeint: der Pbn] win a positive evaluation from the experimenter, or at least that he provide no grounds for a negative one.” (S. 29). Dieses Gefühl, begutachtet und „diagnostiziert” zu werden, wird durch evaluative Instruktionen noch verstärkt. Es könnte, verbunden mit den bereits beschriebenen Reaktanzeffekten, zu einer Verzerrung der selbst berichteten Testangst führen. Sowohl die Erfassung von state-TA vor dem Test (und ihre möglichen Effekte) als auch die Determination der state-TA post sind hier potenziell betroffen.

In beiden Fragestellungen könnten alternative Zugänge zur Erfassung von Testängstlichkeit bzw. Testangst, kombiniert mit den etablierten Paradigmen aus der Testängstlichkeitsforschung (z. B. evaluative vs. nonevaluative Testinstruktionen), theoretisch sehr ergiebig sein. Beispiele hierfür sind der Einbezug von Verhaltensbeobachtungen zur Erfassung des Ausdrucks von Angst (siehe z. B. Beltzer et al., 2014) und der Einsatz von Impliziten Assoziationstests (IAT). Greenwald, McGhee und Schwartz (1998) definieren das Prinzip von IATs wie folgt: „The IAT procedure seeks to measure implicit attitudes by measuring their underlying automatic evaluation.” (S. 1464). Beim IAT werden bestimmte Einstellungen, aber auch Eigenschaften gemessen auf Basis der Schnelligkeit, mit der eine Person auf bestimmte Konzepte reagiert. Bei IATs wird die Assoziation

zwischen einem Zielkonzept (target concept, z. B. „BLACK“ und „WHITE“) und bestimmten Attributen erfasst (z. B. „pleasant“ und „unpleasant“). Dabei werden verschiedene Kombinationen von target und Attribut vorgegeben und die Reaktionszeiten verglichen (z. B. Reaktion auf „BLACK“ & „pleasant“ sowie „WHITE“ & „unpleasant“). Im Kern wird angenommen, dass eine Person auf stärkere Assoziationen schneller reagiert. Beispielsweise würde die Existenz eines impliziten Vorurteils gegenüber Afroamerikanern abgeleitet werden, wenn die Reaktion auf die Assoziation des Konzepts „WHITE“ und des Attributs „pleasant“ schneller erfolgt als jene auf die Assoziation von „BLACK“ und „pleasant“ (Greenwald et al., 1998). IATs wurden auch entwickelt, um Vorstellungen über die eigene Person zu erfassen, wie z. B. den Selbstwert (Greenwald & Farnham, 2000). Egloff und Schmukle (2002) entwickelten einen IAT zur Erfassung von Ängstlichkeit (bzw. des bipolaren Konstrukts Ängstlichkeit vs. Gelassenheit). In diesem werden als Zielkonzepte „Me“ und „Others“ benutzt und kombiniert mit den Attributen „Anxiety“ (Zielbegriffe sind „afraid“, „nervous“) sowie „Calmness“ (Zielbegriffe sind „relaxed“, „balanced“). Dieser sogenannte IAT-Anxiety zeigte prädiktive Validität für die Leistungsverschlechterung bei einem Aufmerksamkeitstest nach einer Misserfolgsrückmeldung (Studie 3, N = 62 Studierende). Darüber hinaus zeigte sich ein bedeutsamer Zusammenhang zwischen der mit dem IAT erfassten Ängstlichkeit und dem Angstausdruck bei einer Präsentationsaufgabe – demgegenüber sagte die mit dem STAI erfasste Ängstlichkeit die retrospektiv berichtete state-Angst während der Präsentation vorher (Studie 4, N = 33 Studierende). IATs könnten also ein Weg sein, die Frage nach der Kausalität in der Relation von Testangst und Leistung ohne die genannten Einschränkungen von Selbstberichtsdaten zu untersuchen.

Eine praktische Implikation für Forscher und Praktiker, die in verschiedenen Settings (z. B. Coaching und Therapie) Testangst erfassen, betrifft die Frage, wann Testangst erfasst werden sollte. Bei der Wahl des Messzeitpunkts sollten stets inhaltliche Argumente im Vordergrund stehen und berücksichtigt werden, dass state-TA prä und post – in Teilen – unterschiedliche Prozesse abbilden. Emotionen unterscheiden sich vor, während und nach einem Test (siehe Folkman & Lazarus, 1985, die bei einer Abschlussprüfung anticipatory, waiting und outcome stage unterschieden). Je nach Leistungssituation ist genau zu prüfen, welche Reihenfolge der Erfassung gewählt werden sollte und welche Effekte dies nach sich ziehen könnte. Beispielsweise dürfte die state-TA vor und nach einer Instruktion von teilweise unterschiedlichen Faktoren abhängen. Wird state-TA nach einer Instruktion erhoben, dürfte sie stark durch Komplexität und Schwierigkeit etwaiger Beispielaufgaben beeinflusst sein. Unterschiede zwischen diesen Phasen entstehen auch durch einen unterschiedlichen Grad an Ambiguität bezüglich des Tests und des erwarteten Ergebnisses (Folkman & Lazarus, 1985). Im Falle von Studie 1 war die Situation zum Zeitpunkt der prä-Messung von eher geringer Ambiguität, da Typ und Schwierigkeit der Aufgaben vorab vermittelt wurden. Auch bei der Messung von Testangst nach einem Test ist genau zu unterschieden, ob sich die

Fragen retrospektiv auf das Erleben während des Tests beziehen oder auf den gegenwärtigen Zustand. Auf die komplexe Natur der retrospektiven Einschätzung wurde bereits eingegangen. Die Erfassung des aktuellen Zustands nach dem Test bezöge sich demgegenüber auf die waiting stage sensu Folkman und Lazarus (1985), in der der Test absolviert, aber noch keine Leistungsrückmeldung erfolgt ist. In jedem Fall sollten Forscher genau berichten, wann Testangst innerhalb einer Untersuchung erfasst wurde (Zeidner, 1991) und wie die Items formuliert waren (in Bezug auf den gegenwärtigen Zustand oder retrospektiv). Dies findet zwar meist statt. Jedoch wurden die Implikationen des Messzeitpunkts bislang kaum untersucht, obwohl bereits Zeidner (1991) darauf hinwies, dass dieser Aspekt eine Ursache sein könnte für die zuweilen heterogene Befundlage bezüglich der Relation von Angst und Leistung. Die eher geringen Auswirkungen der Erfassung der state-TA prä (explorative Analyse) und die hohe Konvergenz von state-TA prä und post (Hypothese 1a und 1b) sprechen dafür, dass die Erfassung von state-TA vor und nach einem Test eine pragmatische Lösung sein könnte. Eine Einschränkung für diese Option ist sicherlich der nicht vollständig erklärbare Korrelationsunterschied zwischen der state-TA post und Leistung in Bedingung A und B. Ob die Erfassung der state-TA prä also subtilere Effekte (z. B. die oben angesprochene Reaktanz oder demand effects) auslöst, muss zukünftige Forschung beantworten.

Die oben diskutierten Unterschiede zwischen den Bedingungen (erhöhtes Anspruchsniveau und geringere Beeinträchtigung durch Nervosität und Prüfungsangst) sind als Reaktanzeffekte interpretierbar, lassen aber Raum für eine alternative Interpretation. Möglicherweise wird es von Personen *positiv* erlebt, wenn sie vor einem Test nach ihrem Befinden, ihrer Aufregung und ihren Selbstzweifeln befragt werden. Parallelen sind aus der Personalauswahl bekannt. Gilliland (1993) formulierte ein Modell, welches die Determinanten von Bewerberreaktionen auf ein Auswahlverfahren beschreibt. Demnach wird die subjektive Fairness eines Auswahlverfahrens unter anderem von der Rücksichtnahme und wertschätzenden Behandlung durch die auswählende Person beeinflusst („Interpersonal effectiveness of administrator"). Übertragen auf Tests und Prüfungen könnte das bedeuten, dass Testanden Fragen nach ihrer gegenwärtigen Testangst *nicht* als belastend, sondern – im Gegenteil – positiv wahrnehmen, da diese ausdrücken, dass der Testleiter sich über den Stress in der Situation bewusst ist. Analog dürfte die Frage „Sind Sie aufgeregt?" vor einer mündlichen Prüfung für manchen Studierenden entlastend sein, da sie sowohl Empathie als auch das Zugeständnis von Aufgeregtheit als normaler Reaktion signalisiert. Die praktische Implikation wäre, dass Personen *per se* vor einer Prüfung nach ihrer Angst befragt werden sollten. Allerdings stehen derartige „wünschenswerte" Effekte im Kontrast zur stärkeren Korrelation von state-TA post mit objektiver Leistung in Bedingung A, welche womöglich mit der erhöhten Salienz von Testangst erklärt werden kann. Diese Gedanken zu elaborieren ist Aufgabe zukünftiger Forschung.

Kern der Fragestellung war neben den Effekten der Erfassung von Testangst vor dem Test die Bedeutung von Regulationsprozessen für die nach dem Test berichtete Testangst. Grundlage war folgende Vermutung: „examinees who feel they may have performed poorly may rationalize their low performance by reporting higher levels of anxiety“ (Zeidner, 1991, S. 107). Die Ergebnisse von Studie 1 erfordern eine Relativierung dieser Vermutung. Wahrscheinlich sind die stattfindenden Prozesse komplexer. Die nach dem Test berichtete Testangst ist kein ausschließliches Resultat der subjektiven Leistung und des Bemühens, eine subjektive Erklärung für ein schlechtes Ergebnis zum Schutz des Selbstwerts zu generieren. Diese „Rationalisierung“ wurde prinzipiell auch angenommen für die Variablen zum self-handicapping. Deren Bezeichnung unterstellt bereits, dass es sich um Aspekte handelt, welche zum Zwecke berichtet werden, ein subjektiv schlechtes Ergebnis zu rechtfertigen (und umgekehrt nicht berichtet werden, wenn das Ergebnis subjektiv gut war). Bezüglich der Beziehung von state-TA post und SH-Aspekten ist jedoch auch eine einfachere Erklärung möglich: das Erleben von Testangst wird unmittelbar als beeinträchtigend erlebt, d. h. sobald eine Person sorgenvolle Gedanken oder Nervosität erlebt, fühlt sie sich davon in ihrer Leistungsfähigkeit eingeschränkt. Die Vorannahme, dass Testangst zum Schutze des Selbstwerts berichtet wird, setzt nicht nur voraus, dass einer Person ein Testergebnis wichtig ist. Darüber hinaus muss der Bericht von Testangst – dieser Logik folgend – überhaupt funktional zum Schutz des Selbstwerts vor den Implikationen eines Misserfolgs sein. Es gibt empirische Befunde, die für diese „Schutzfunktion“ sprechen (siehe Abschnitt 1.2.1.3.1). Das eigene Versagen auf das singuläre Erleben von Testangst zu attribuieren erfordert aber die zusätzliche Überzeugung, in zukünftigen Testsituationen weniger Angst zu haben und besser abzuschneiden. Ob dies bei dispositionell testängstlichen Personen grundsätzlich angenommen werden kann, ist fraglich.

Darüber hinaus ist es denkbar, dass Personen – sofern sie nach selbstwertdienlichen Erklärungen für eine schlechte Leistung „suchen“ – andere Faktoren benennen, welche einer optimalen Leistung im Wege standen. Dies könnten beispielsweise Frustration über den Test oder Enttäuschung über die eigene Leistung sein. Die verfügbaren Daten liefern Hinweise darauf, ob Zweifel an der Qualität des Tests einer dieser Faktoren gewesen sein könnte: die subjektive Leistung korrelierte immerhin zu $r = .40$ ($p < .001$) mit der Note für das Testverfahren. Zweifel an der Aussagekraft des Tests scheinen jedoch angesichts der nicht signifikanten Korrelationen von Messqualität (und auch Augenscheinvalidität) mit der subjektiven Leistung in diesem Kontext nicht von Bedeutung gewesen zu sein. Überdies korrelierte auch keiner der SH-Aspekte (mit Ausnahme von Nervosität und Prüfungsangst) signifikant mit der subjektiven Leistung, was zu erwarten gewesen wäre, wenn die SH-Aspekte (maßgeblich) in Reaktion auf ein schlechtes Ergebnis berichtet worden wären. Die Korrelationen von subjektiver Leistung mit Nervosität und Prüfungsangst sprechen dafür, dass sich die Pbn tatsächlich von Testangst beeinträchtigt fühlten (je stärker die Beeinträchtigung,

umso schlechter die subjektive Leistung)[38]. Selbstverständlich erlaubt die Erhebung via Selbstberichtsmaßen nicht die tiefer gehende Differenzierung, ob die Pbn sich subjektiv *tatsächlich* von einem Einfluss (z. B. Ablenkung durch Umgebung) gestört fühlten, oder diesen spezifischen Einfluss im Bericht übertrieben haben oder auch nur angegeben haben, weil er zur Auswahl stand.

Unzweifelhaft ist die Entstehung und Wirkung von Testangst ein komplexer Prozess. Bei der Frage, wie retrospektiv berichtete Testangst zu interpretieren ist, liegt die Wahrheit vermutlich „in der Mitte“: weder ist das Maß vollständiger Ausdruck von self-handicapping, noch ist es eine von der subjektiven Leistung unabhängige Wiedergabe des eigenen Erlebens. Insbesondere letztere Annahme ist äußerst unrealistisch, mit Blick auf die komplexen Prozesse der Entstehung und Wirkung von Testangst (siehe z. B. die S-REF Theorie, Abschnitt 1.1.3) und die diskutierten Grenzen des Selbstberichts. Die nonverbalen und verbalen Reaktionen eines Prüfers in einer mündlichen Prüfung, aber auch die subjektive Schwierigkeit von Testaufgaben bilden einen permanenten Informationsfluss bezüglich der subjektiven Leistung. Die von Schwarzer (2000) beschriebene „Selbstvoreingenommenheit“ (S. 105) Testängstlicher, der Aufmerksamkeitsbias (Bar-Haim et al., 2007) und auch die Hypervigilianz (Eysenck, 1992) dürften dazu beitragen, dass Testängstliche Hinweisreizen auf ihr Abschneiden besondere Aufmerksamkeit zuwenden. Dass also Testangst und subjektive Leistung miteinander kovariieren *müssen* liegt auf der Hand.

Wie bedeutsam diese Wechselwirkungen sind, zeigt eine Studie von Ortner und Caspers (2011). Eine Stichprobe von $N = 110$ Schülern (Alter zwischen 16 und 20) wurde anhand des TAI-G XU am Median in niedrig und hoch Testängstliche aufgeteilt und bearbeitete den adaptiven Matrizentest AMT (Hornke, Küppers & Etzel, 2000). Der Test wurde entweder in der adaptiven Form (computerbasiertes adaptives Testen, CAT) oder mit fixierter Itemreihenfolge (fixed item testing, FIT, d. h. ansteigende Itemschwierigkeit) durchgeführt. Die adaptive Variante wurde ohne (Standardinstruktion) oder mit vorheriger Information über das Prinzip des adaptiven Testens instruiert. Beim adaptiven Testen wird die Itemschwierigkeit eines vorgelegten Items durch die Schwierigkeit der bislang gelösten Items bestimmt (Schmidt-Atzert & Amelang, 2012). Dabei steigt die Schwierigkeit der Items – im Vergleich zur traditionellen Itemabfolge von leichten zu schweren Items – schneller an. Dies könnte zu vermehrter Frustration und bei Testängstlichen zu erhöhter Testangst führen (Ortner & Caspers, 2011). Tatsächlich zeigte sich ein deutlicher Leistungsunterschied zwischen hoch und niedrig Testängstlichen beim CAT, aber nicht beim FIT. Dieser Leistungsunterschied war kleiner in der Teilgruppe, die vorab über das Prinzip des CAT aufgeklärt wurde (Ortner & Caspers, 2011). Generell könnte die retrospektive Erfassung von Testangst spezifischere Informationen liefern, wenn gezielt auf einzelne Aspekte des Erlebens eingegangen

[38] Natürlich wäre auch die Lesart denkbar, dass die Pbn *ausschließlich* Testangst als beeinträchtigenden Faktor berichtet haben, um ihr Ergebnis zu rechtfertigen. Die übrigen Ergebnisse sprechen allerdings nicht dafür.

wird und diese differenziert werden, wie z. B. die Reaktion auf schwierige Items und die Testangst in verschiedenen Phasen des Tests.

4.3.3.2 Ausblick – weitere Forschung

Aus den aufgeführten Limitationen sowie den theoretischen und praktischen Implikationen leiten sich einige Ansätze für die weitere Forschung ab. So sollte in zukünftigen Studien eine präzise Gegenüberstellung verschiedener Erhebungsvarianten bei der Erfassung von Testangst (state) erfolgen. Insbesondere bei der nach dem Test erfassten state-TA sollten Untersuchungen klären, wie sich ein retrospektiver Bericht nach dem Test von der gegenwärtig erlebten Testangst nach dem Test unterscheidet. Die von Zeidner (1991) vermuteten Einflüsse des Messzeitpunkts auf den Zusammenhang von Angst und Leistung könnten so besser verstanden werden.

Auch zukünftig werden experimentelle Designs in der Forschung zu Testangst unverzichtbar sein. Das bedeutet auch, dass die Generalisierbarkeit von Experimenten unter (mehr oder weniger) kontrollierten Bedingungen auf echte Prüfungssituationen weiterhin erforscht werden sollte. Insbesondere in der Frage, inwieweit Testangst (auch) als Ursache für schlechte Leistung berichtet wird, spielt die subjektive Bedeutsamkeit eines Testergebnisses eine wichtige Rolle.

Darüber hinaus muss weiter untersucht werden, in welchem Maße und unter welchen Bedingungen der Bericht von Testangst oder anderen Störeinflüssen Ausdruck von self-handicapping ist. Hier sind zukünftig Untersuchungen mit innovativen Designs notwendig, wobei ein Rückgriff auf bereits genutzte Paradigmen sinnvoll ist. Insbesondere die Induktion von Erfolg und Misserfolg in Testsituationen (z. B. Eckert, Schilling & Stiensmeier-Pelster, 2006) sei hier genannt. Auch die Ergänzung von Selbstberichtsdaten durch Informationen aus alternativen Erhebungsmethoden (z. B. IATs) könnte einen theoretischen Mehrwert liefern. Dies könnte Rückschlüsse darauf erlauben, ob bestimmte Störeinflüsse oder Testangst subjektiv tatsächlich als leistungshemmend erlebt wurden oder „lediglich“ berichtet werden, um den Selbstwert zu schützen.

Die zukünftige Forschung sollte stets berücksichtigen, welche Effekte bestimmte „Eingriffe“ vor einem Test auf die Ergebnisse haben können. Dies betrifft nicht nur die Erfassung von Testangst vor einem Test, sondern auch die Informationen, die einem Testanden in einer Testinstruktion gegeben werden (z. B. Art und Schwierigkeit der Beispielitems oder die Beschreibung des getesteten Bereichs). Diese Hinweisreize können den Bewertungscharakter einer Testsituation – gewollt oder ungewollt – verstärken oder reduzieren.

5. Studie 2

5.1 Methode

5.1.1 Kurzüberblick

Ausgangspunkt für Studie 2 war die fragliche Relevanz von Testangst bei der Erklärung des Stereotype Threat (STT). Ziel war dabei, den evaluativen Charakter einer Testinstruktion sowie den stereotypen Gehalt derselbigen gezielt zu manipulieren und deren Effekt auf das Erleben im Test und die Testleistung von targets und nontargets (hier Mädchen und Jungen) zu prüfen. Ziel war auch zu klären, inwiefern Testangst im Kontext des STT erwartete Leistungsunterschiede erklären kann.

Studie 2 wurde in Papier-Bleistuft-Erhebungen durchgeführt, wobei die Studie auf zwei Erhebungszeitpunkte aufgeteilt wurde. In einer Vorerhebung wurde lediglich die Testängstlichkeit (trait-TÄ) erfasst. In der Haupterhebung, die ca. zwei Wochen später erfolgte, wurden demographische Variablen erhoben (Alter, Geschlecht, letzte Zeugnisnote in sowie Interesse an den Fächern Mathematik und Deutsch) und anschließend einige Fragen zu Interessen (Michaelis, Ott, Palmer, Ulfert & Kersting, 2013)[39]. Danach erfolgte die Ankündigung des Intelligenztests, welche gleichzeitig die Manipulation darstellte. Die Pbn erhielten entweder eine evaluativ-stereotype (Bedingung A), eine evaluativ-nonstereotype (Bedingung B) oder eine nicht-evaluative Testinstruktion (Bedingung C). Nach der Vorgabe von Beispielitems wurde die state-Testangst (state-TA) erfasst sowie Annäherungs- und Vermeidungsziele. Anschließend wurde der Test durchgeführt. Nach dem Test wurde die Akzeptanz des Verfahrens, das Flow-Erleben während des Tests sowie die subjektive Bedrohung durch das Stereotyp erhoben. Die Daten von Vor- und Hauptuntersuchung wurden mittels eines sechsstelligen Pbn-Codes verknüpft.

Die Rekrutierung der Stichprobe, die organisatorische Planung und Leitung der Erhebungen und die Eingabe der Daten wurden von Eva Tuschen und Christin Köhler im Rahmen ihrer Masterarbeiten durchgeführt (Köhler, 2015; Tuschen, 2014). Beide Masterarbeiten wurden vom Verfasser dieser Arbeit betreut, welcher auch für die inhaltliche Planung maßgeblich verantwortlich war. Als Stichprobe wurden von Frau Tuschen und Frau Köhler insgesamt neun Schulklassen eines Gymnasiums in Gießen rekrutiert.

[39] Dabei handelte es sich um sechs Items zu je einem Interessensbereich, die als Bestandteil des dargestellten Untersuchungsziels (FITAS) vorgelegt wurden. Diese Items wurden nicht weiter ausgewertet.

5.1.2 Herleitung der Hypothesen

Die Hypothesen von Studie 2 leiten sich ab aus den in Abschnitt 2.2 dargestellten Fragestellungen. Alle Hypothesen befassten sich mit dem STT (Bedingung A vs. B) bzw. der Manipulation des evaluativen Charakters des Tests (Bedingung B vs. C). Die nach Geschlechtern getrennte Betrachtung der Effekte diente dazu, Prozesse des STT von anderen Prozessen, die beispielsweise „nur" auf die Variation des Bewertungsstresses zurückgehen, zu differenzieren. Analysiert wurden sowohl Unterschiede im absoluten Erleben vor und nach dem Test als auch Effekte auf die Leistung. Darüber hinaus sollte jedoch auch die Relation von state-Testangst und Leistung betrachtet werden. In allen Analysen sollte eine Differenzierung zwischen targets und nontargets stattfinden um Prozesse identifizieren zu können, die sich auf die targets beschränken und somit auch einen etwaigen STT erklären können.

Hypothese 1 befasste sich mit den Effekten der Manipulation auf das Erleben vor dem Test. Dabei handelte es sich um Maße, die nach der Instruktion des Tests, jedoch vor Beginn der eigentlichen Bearbeitung erhoben wurden. Ausgehend von den in Abschnit 1.2.2.2.3 beschriebenen Befunden wurde erwartet, dass sich in Bedingung A ein Geschlechtsunterschied in der vor dem Test berichteten Testangst ergibt (state-TA). Zudem wurde auf Basis der Ergebnisse von Brodish und Devine (2009) angenommen, dass Mädchen in Bedingung A in höherem Maß Leistungsvermeidungsziele berichten als Jungen. Geschlechtseffekte in beiden Variablen wurden in B und C nicht erwartet, da hier kein stereotyper Gehalt der Instruktion vorlag. Insgesamt sollte das Niveau an state-TA und Leistungsvermeidungszielen bei A und B höher sein als in C, da in letzterer Bedingung kein Bewertungsstress induziert wurde. Ein im Vergleich zu den Vermeidungszielen umgekehrtes Bild wurde bei den Annäherungszielen erwartet, weshalb diese in die Prüfung aufgenommen wurden.

> **Hypothese 1:** In Bedingung A berichten Mädchen mehr state-Testangst (Besorgtheit) und in stärkerem Maße Vermeidungsziele als Jungen. Keine derartigen Geschlechtsunterschiede zeigen sich in Bedingung B und C. Darüber hinaus liegt in A und B ein höheres Niveau an state-Testangst (Besorgtheit) und Vermeidungszielorientierung vor als in C.

Die Hypothesen 2a-c befassten sich mit den Auswirkungen der Manipulation auf die Leistung. Da Mädchen die targets waren, wurde ein Test zum rechnerischen Denken als Leistungsmaß gewählt. Der erste Teil der Hypothese beinhaltete den eigentlichen STT-Effekt: es wurde angenommen, dass Mädchen in A – und nur hier – schlechter abschneiden als Jungen. Ein entsprechender Geschlechtsunterschied sollte in B nicht auftreten, da die Instruktion stereotypneutral war. Ferner gab es keine Begründung, warum in C ein Geschlechtsunterschied auftreten sollte, sofern ein Test eingesetzt wird, bei dem üblicherweise Männer und Frauen gleich gut abschneiden. Der zweite Teil der Hypothese befasste sich mit einer möglichen Erklärung des STT. Fokus lag hierbei auf

dem Zusammenhang von state-Testangst und Testleistung in den drei Bedingungen, wobei zusätzlich zwischen targets und nontargets unterschieden wurde. Folgende Annahmen wurden getroffen: Wenn Testangst den STT erklärt, sollte sich ein negativer Zusammenhang von Testangst und Leistung bei targets in A ergeben. Dabei sollte die Annahme geprüft werden, dass Testangst den Effekt der Manipulation (A vs. B) auf die Leistung mediiert. Eine derartige Mediation sollte bei nontargets *nicht* vorliegen, da bei diesen die Manipulation (A vs. B) *keinen* Effekt auf die Testangst haben sollte. Dies bedeutet aber *nicht*, dass es in A bei nontargets *keinen* negativen Zusammenhang von Testangst und Leistung geben kann: bei beiden Geschlechtern wäre ein negativer Zusammenhang von Testangst und Leistung plausibel. Offen war dabei, ob der Zusammenhang bei targets und nontargets gleich hoch sein würde. Zur Prüfung dieser Annahme eignete sich eine moderierte Mediation, wobei geprüft werden kann, ob die einzelnen Effekte des Mediationsmodells für die jeweiligen Geschlechter verschieden sind. Es wurde erwartet, dass in B *keine* negativen Stereotype über targets aktiviert werden. In B sollte sich aber bei *beiden* Geschlechtern ein negativer Zusammenhang von Testangst und Leistung ergeben. Weniger eindeutig war die Erwartung für Bedingung C. Am wahrscheinlichsten war ein nicht signifikanter Zusammenhang von Testangst und Leistung, da in der nonevaluativen Instruktion keine Angst induziert wurde. Analog zu B wurden auch in C keine Geschlechtsunterschiede vermutet. Diese Analyse konzentrierte sich dabei auf Besorgtheit als der Facette, die am engsten mit Leistung in Zusammenhang steht.

> **Hypothese 2a:** In A zeigen Mädchen eine schlechtere Leistung als Jungen. In B und C hingegen schneiden Mädchen und Jungen gleich gut ab.
> **Hypothese 2b:** State-Testangst (Besorgtheit) mediiert bei Mädchen, aber nicht bei Jungen den Effekt der STT-Manipulation (A vs. B) auf die Leistung. Die STT-Manipulation (A vs. B) hat nur bei Mädchen einen Effekt auf die state-Testangst (Besorgtheit) sowie die Leistung.
> **Hypothese 2c:** In A und B wird eine negative Relation von state-Testangst (Besorgtheit) und Leistung bei beiden Geschlechtern erwartet. In Bedingung C wird bei beiden Geschlechtern kein signifikanter Zusammenhang von state-Testangst (Besorgtheit) und Leistung erwartet.

Die Hypothesen 3a-c betrachteten das Erleben nach dem Test. Hierbei wurde eine Reihe von Variablen betrachtet, die alternative oder zusätzliche Erklärungen für STT-Effekte liefern könnten. Dabei wurde erstens untersucht, inwiefern vermutete Leistungseffekte im Sinne des STT mit Geschlechtsunterschieden im Flow-Erleben einhergehen: „In flow, a person is fully concentrated on the task at hand. There is a feeling that action and awareness merge in a single beam of focused consciousness." (Csikszentmihalyi, 2014, S. 24). Flow tritt bei Aktivitäten auf, in denen sowohl die Anforderung (challenge) als auch die zur Verfügung stehenden Fähigkeiten bzw. Fertigkeiten (skill) hoch sind (Angst läge demnach vor, wenn challenge hoch und skill niedrig ist, Langeweile

läge im umgekehrten Fall vor) (Csikszentmihalyi, 2014). Dieses Konstrukt wurde mit aufgenommen, da in der Untersuchung von Lang und Lang (2010) das Flow-Erleben die Reduktion des Zusammenhangs von Testängstlichkeit und Leistung durch das Kompetenzpriming (genauer: deren Interaktion) erklärte. So ist es denkbar, dass der STT mit einem reduzierten Flow-Erleben einhergeht. Während Flow für eine hohe Involvierung in die Aufgabenbearbeitung steht, ist beim STT eine Ablösung (disengagement) von den Aufgabenanforderungen denkbar (siehe hierzu Pennington et al., 2016). Ein derartiger Effekt ließe sich gut mit fehlenden Effekten auf Testangst vereinbaren. Zweitens sollten Effekte auf das Erleben und die Bewertung des Tests selbst geprüft werden, wobei das Belastungserleben im Fokus stand (Kersting, 2008). Drittens sollte mit direkten Fragen geprüft werden, ob die Probanden die Befürchtung hatten, dass ihr Geschlecht bei der Beurteilung ihrer Leistung eine Rolle spielt. Diese Variablen zur subjektiven Bedrohung durch das Stereotyp stellten einen Versuch dar, die Bewertungsangst (Testangst) von der konkreteren Angst, ein Stereotyp zu bestätigen, zu trennen. Um die Manipulation nicht von vornherein offenzulegen, erfolgten diese Fragen erst am Ende der Erhebung. Anhand des oben genannten Designs sollte somit differenziert werden, welche Manipulation (Bewertungsstress mit und ohne stereotypem Gehalt) sich auf welche Form der Besorgtheit auswirkt.

> **Hypothese 3a:** In A berichten Mädchen ein geringeres Flow-Erleben als Jungen, während sich in B und C kein Geschlechtsunterschied zeigt.
> **Hypothese 3b:** Mädchen berichten in A ein höheres Belastungserleben als Jungen, während sich in Bedingung B und C hier kein Geschlechtsunterschied ergibt.
> **Hypothese 3c:** Mädchen berichten in Bedingung A in höherem Maße als Jungen Sorgen, dass ihr Geschlecht für die Bewertung ihrer Leistung eine Rolle spielt, nicht jedoch in B und C.

5.1.3 Stichprobe

Die Stichprobe wurde im Mai und Juni 2014 rekrutiert. Den Lehrern der Klassen wurde mitgeteilt, dass den Schülern das genaue Untersuchungsziel vorab nicht offengelegt werden kann, um deren unvoreingenommene Teilnahme zu gewährleisten. Den Schülern wurde die gesamte Studie unter dem Titel „Fähigkeiten- und Interessentest für Ausbildung und Studium (FITAS)“ vorgestellt, wobei der Kontext von Ausbildungs- und Studienwahl Interesse und Kooperationsbereitschaft auf Seiten der Schüler erhöhen sollte. Insgesamt nahmen an der Vorerhebung N = 188 Schüler teil. Hiervon nahmen N = 168 Schüler an der späteren Haupterhebung teil (Dropout N = 20 ≙ 10.64 %). Diese 168 Pbn wurden in die Analysen inkludiert. Die Stichprobe setzte sich aus acht verschiedenen Klassen der 10. (N = 151) und einem Kurs der 11. Jahrgangsstufe (N = 17) des besagten Gymnasiums zusammen. Von diesen Pbn waren N = 94 weiblich und N = 73 männlich (56% bzw.

44%; 1 Person ohne Angabe). Das Alter lag im Schnitt bei *M* = 16.11 (*SD* = .68) Jahren. Die Pbn wurden zudem nach ihrer letzten Zeugnisnote in Mathematik und Deutsch gefragt sowie (mit je einem Item) nach ihrem Interesse im Fach Mathematik und Deutsch. Die deskriptiven Statistiken sind in Tabelle 39 dargestellt (die Kennwerte pro Bedingung sind in Abschnitt 5.2.1 aufgeführt).

Tabelle 39: Deskriptive Statistiken der letzten Zeugnisnote und Interesse in Mathematik und Deutsch

	N	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*
Zeugnisnote Mathe	166	1	15	9.14	3.18	10.11	-.33	-.62
Zeugnisnote Deutsch	167	1	14	8.94	2.76	7.61	-.63	.22
Interesse Mathe	168	1	5	3.40	1.14	1.31	-.33	-.60
Interesse Deutsch	168	1	5	3.15	1.06	1.11	-.28	-.45

Noten in Punkten von 1-15; Interesse am Fach von 1 = „sehr gering“ bis 5 = „sehr hoch“
Md: Zeugnisnote Deutsch = 9.00

Um eine möglichst hohe Kontrolle der Manipulation zu sichern, wurden die Schulklassen jeweils geschlossen und zufällig einer Bedingung zugeordnet, wobei drei Klassen je Bedingung erhoben wurden. Bedingung A enthielt *N* = 53 Pbn (Klasse 1-3), Bedingung B *N* = 57 Pbn (Klasse 4-6) und Bedingung C *N* = 58 Pbn (Klasse 7-9). Darüber hinaus wurde noch die Anzahl an Schulstunden (pro Woche) in Mathematik und Deutsch erfasst. Pbn, die mehr als vier Stunden Mathematik bzw. Deutsch angaben (sprich fünf oder sechs Stunden), besuchten Förderunterricht. Dies waren in Mathematik *N* = 9 und in Deutsch *N* = 3.

5.1.4 Beschreibung des Untersuchungsablaufs

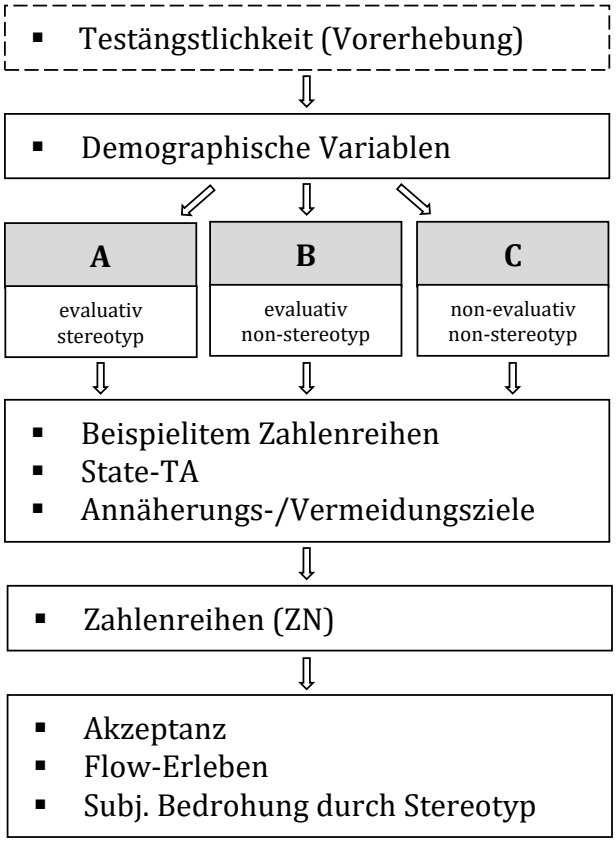

Abbildung 13: Untersuchungsablauf von Studie 2

Die Studie teilte sich auf in eine Vor- und Haupterhebung, welche einen Zeitabstand von ein bis zwei Wochen zueinander hatten. Alle Klassen wurden jeweils geschlossen und zufällig einer der drei Bedingungen zugeordnet, wodurch ein quasi-experimentelles Design vorlag. Die Erhebungen wurden innerhalb von regulären Schulstunden in Form einer Gruppentestung durchgeführt. In der Vorerhebung wurden die Testängstlichkeit sowie ein Pbn-Code erhoben. In der Haupterhebung wurden zunächst demographische Variablen erhoben und allgemeine Fragen zu Interessen (vgl. den kommunizierten Untersuchungszweck „FITAS"). Nach der Ankündigung des Tests, die der Manipulation diente, wurden Beispielitems vorgelegt. Anschließend wurde die state-Testangst sowie Annäherungs- und Vermeidungsziele erfasst. Danach wurde der Test durchgeführt, wobei die Pbn während der Bearbeitung mehrmals die verbleibende Zeit mitgeteilt bekamen. Nach dem Test wurden Fragen zum Testerleben gestellt, konkret zur Akzeptanz des Tests und zum Flow-Erleben. Dann folgten zwei Fragen zur subjektiven Bedrohung durch das Stereotyp (siehe Abbildung 13). Die Pbn konnten zum Schluss ihre E-Mail-Adresse optional angeben, um auf Wunsch eine Rückmeldung zu ihrem Ergebnis zu erhalten. Die Vorerhebung dauerte wenige Minuten, die Haupterhebung ca. 30 Minuten. Entgegen der Absprache erwähnte die Lehrkraft einer Klasse in Bedingung A bei der Vorerhebung (nicht aber in der Haupterhebung), dass es in der Studie um Prüfungsangst gehe. Darauf wird an späterer Stelle nochmals eingegangen.

5.1.5 Darstellung der Erhebungsdetails und Operationalisierung der Konstrukte

Nun folgt eine Beschreibung der gewählten Instrumente. Der Bericht von Geschlechtsunterschieden erfolgt jeweils nur, wenn diese nicht hypothesenrelevant sind. Andernfalls wird darauf in den Ergebnissen eingegangen. Sofern nicht anders dargestellt, beziehen sich die Befunde auf die gesamte Stichprobe.

Testängstlichkeit

Wie in Studie 1 wurde auch in Studie 2 der TAI-G XU (Wacker et al., 2008) zur Erfassung der dispositionellen Testängstlichkeit eingesetzt. Die deskriptiven Statistiken und Reliabilitäten sind Tabelle 40 zu entnehmen.

Tabelle 40: Deskriptive Statistiken und Reliabilitätswerte für die Skala Testängstlichkeit

	N	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Besorgtheit	164	5-20	5	20	13.23	3.87	15.00	-.08	-.97	.86
Aufgeregtheit	158	4-16	4	15	7.14	2.63	6.92	.86	.08	.83
Interferenz	166	3-12	3	12	6.08	2.28	5.20	.42	-.58	.80
Mangel an Zuversicht	163	3-12	3	12	7.73	2.25	5.04	-.14	-.83	.85
Gesamt	150	15-60	18	52	34.08	8.14	66.26	.15	-1.13	.88

Angegeben sind Skalensummenwerte; bei Umrechnung in Skalenmittelwerte ergeben sich folgende Werte: Besorgtheit = 2.65; Aufgeregtheit = 1.78; Interferenz = 2.03; Mangel an Zuversicht = 2.58; Gesamt = 2.27
Md: Aufgeregtheit = 6.00

Die Subskalen sowie die Gesamtskala wiesen eine gute Reliabilität auf, die volle Skalenbreite war fast in allen Subskalen ausgeschöpft. Verglichen mit den anderen Subskalen war bei Aufgeregtheit und Interferenz ein Bodeneffekt zu beobachten. Bezogen auf den Gesamtwert wiesen die weiblichen Pbn (M = 35.75, SD = 8.46) einen sign. höheren Wert auf als die männlichen (M = 32.09, SD = 7.30), t (147) = -2.78, p = .006. Die Werte lagen jeweils sehr nahe an jenen aus der Studierenden-Stichprobe von Wacker et al. (2008) für Frauen (M = 36.2) und Männer (M = 32.7). Bei Besorgtheit berichteten Mädchen (M = 14.06, SD = 3.86) höhere Werte als Jungen (M = 12.29, SD = 3.66), t (161) = -2.98, p = .003. Gleiches gilt für Aufgeregtheit (Mädchen: M = 7.84, SD = 2.78, Jungen: M = 6.24, SD = 2.09), t (155) = -3.96, p < .001, sowie für Mangel an Zuversicht (Mädchen M = 8.15, SD = 2.32; Jungen: M = 7.21, SD = 2.06), t (160) = -2.70, p = .008.

Die Interkorrelationen der Subskalen sowie Korrelationen mit Noten sowie Fachinteresse sind in Tabelle 41 angegeben.

Tabelle 41: Facetteninterkorrelationen bezüglich Testängstlichkeit und Korrelationen mit Zeugnisnoten und Fachinteresse

	Trait-TÄ			Zeugnisnote		Interesse	
	BE	AU	IN	Mathe	Deutsch	Mathe	Deutsch
Besorgtheit				-.24**	-.11	-.35**	.11
Aufgeregtheit	.55**			-.06	-.01	-.15	.06
Interferenz	.27**	.24**		-.27**	-.24*	-.28**	-.16*
Mangel an Zuversicht	.46**	.37**	.32**	-.32**	-.23*	-.50**	-.01
Gesamt				-.29**	-.18*	-.43**	.03

N = 148-166; * p < .05; ** p < .01

Es zeigten sich zwischen den Facetten erwartungsgemäß moderate bis hohe Zusammenhänge. Es lagen für alle Facetten (abgesehen von Aufgeregtheit) signifikante negative Zusammenhänge mit der Mathematiknote vor, teilweise auch mit der Deutschnote. Deutliche negative Zusammenhänge waren zwischen Testängstlichkeit und dem Interesse an Mathematik vorhanden. Sowohl bezüglich der Note als auch dem Interesse lagen somit für Mathematik stärkere und konsistentere Zusammenhänge zu Testängstlichkeit vor als für Deutsch.

Ankündigung des Tests (Manipulation)

Anschließend wurde der Test angekündigt. Bedingung A und B erhielten evaluative Instruktionen, Bedingung C eine nonevaluative Instruktion. Vorbild für diese Kontrastierung waren Elemente evaluativer bzw. nicht evaluativer Instruktionen (siehe Abschnitt 1.1.2 sowie 1.2.2) aus Untersuchungen mit entsprechenden Manipulationen (insbesondere Coy et al., 2011; Englert et al., 2011; Meijer & Oostdam, 2007, 2011). An besagten Arbeiten orientierten sich auch weitere Textelemente der Instruktionen. So enthielten alle drei Instruktionen eine globale Aufforderung, sich anzustrengen sowie den Hinweis, sich nicht zu lange an einer Aufgabe aufzuhalten. Dies sollte eine ernsthafte Bearbeitung in allen drei Bedingungen sichern. Eine explizite Induktion von Misserfolg sollte durch den Hinweis vermieden werden, dass man wahrscheinlich nicht alle Aufgaben in der gegebenen Zeit bearbeiten kann. Wichtigster Unterschied zwischen Bedingung A bzw. B einerseits und C andererseits war, dass in letzterer die diagnostische Aussagekraft des Tests eingeschränkt wurde, indem dieser als sich in Entwicklung befindlich tituliert wurde (vgl. Deffenbacher & Hazaleus, 1985). Auch wurde das Angebot einer individuellen Leistungsrückmeldung (vgl. Studie 1) nicht gegeben (erst am Ende der Befragung). In Bedingung A wurde (gegenüber Bedingung B) eine subtile Aktivierung des STT nach Nguyen und Ryan (2008) angestrebt. Der einzige Unterschied zwischen Bedingung A und B war somit der in der Formulierung adressierte Leistungsbereich (mathematisches vs. schlussfolgerndes Denken). Daraus ergaben sich drei Instruktionsvarianten, die in Tabelle 42 vollständig wiedergegeben sind.

Tabelle 42: Variationen der Testankündigung in Studie 2

	Bedingung	Instruktionstext
A	evaluativ stereotyp	Der folgende Test misst die Fähigkeit im mathematischen Denken. Der Test lässt Aussagen über deine individuellen Stärken und Schwächen im Bereich des mathematischen Denkens zu. Gib also dein Bestes. Du wirst es möglicherweise nicht schaffen alle Aufgaben in der vorgegebenen Zeit zu bearbeiten. Halte dich nicht zu lange an einer Aufgabe auf und nutze die Zeit gut. Du erhältst später wenn du möchtest eine Rückmeldung über dein Ergebnis, anhand der du deine Leistung mit der von den anderen Teilnehmern vergleichen kannst.
B	evaluativ stereotypneutral	Der folgende Test misst die Fähigkeit im schlussfolgernden Denken. Der Test lässt Aussagen über deine individuellen Stärken und Schwächen im Bereich des schlussfolgernden Denkens zu. Gib also dein Bestes. Du wirst es möglicherweise nicht schaffen alle Aufgaben in der vorgegebenen Zeit zu bearbeiten. Halte dich nicht zu lange an einer Aufgabe auf und nutze die Zeit gut. Du erhältst später wenn du möchtest eine Rückmeldung über dein Ergebnis, anhand der du deine Leistung mit der von den anderen Teilnehmern vergleichen kannst.
C	nonevaluativ stereotypneutral	Der folgende Test misst die Fähigkeit im schlussfolgernden Denken. Der Test befindet sich momentan in der Entwicklung. Ziel ist es herauszufinden, wie schwer bzw. wie gut dieser Test ist und ob er für Forschungsprojekte und die Berufsberatung geeignet ist. Gib also dein Bestes. Du wirst es möglicherweise nicht schaffen alle Aufgaben in der vorgegebenen Zeit zu bearbeiten. Halte dich nicht zu lange an einer Aufgabe auf und nutze die Zeit gut.

Anm.: alle Testankündigungen wurden auf einem separaten Blatt vor den Beispielitems vorgelegt.

Allgemeine Instruktion des Tests

Nach der manipulierten Einführung folgte die Testinstruktion. Hier wurde die Originalinstruktion des Subtests „Zahlenreihen“ (ZN) aus dem Wilde-Intelligenz-Test 2 (WIT-2) von Kersting, Althoff und Jäger (2008) eingesetzt. Besagter Subtest wurde auch als Leistungsmaß eingesetzt (siehe separate Beschreibung in diesem Abschnitt). An der Originalinstruktion wurde eine Modifikation vorgenommen. Die ursprüngliche ZN-Instruktion beinhaltet zwei sehr einfache Items. Um den Bewertungsstress zu erhöhen wurde das zweite Beispiel durch ein schwierigeres ersetzt. Dieses wurde aus dem Intelligenz-Struktur-Test Screening (IST-Screening) von Liepmann, Beauducel, Brocke und Nettelnstroth (2012) entnommen und besitzt laut Manual eine mittlere Schwierigkeit (Item 14, p = .64). Vorteil dieses Items war, dass es denselben Itemtypus, nämlich eine Zahlenreihe, darstellt. Das Item aus dem IST-Screening besteht aus sieben vorgegebene Zahlen mit der Anweisung, die fehlende achte Zahl zu ergänzen. Um die Länge der beiden Beispielitems konstant zu halten, wurde beim ersten, originalen Beispielitem die Zahlenreihe um eine Zahl verlängert (von sechs auf sieben). Dadurch sollte trotz Modifikation der Zweck der allgemeinen Instruktion (das Verständnis des Aufgabenprinzips) erfüllt werden.

State-Testangst

Zur Erfassung der Testangst vor dem Test wurden die Items des TAI-G XU von Wacker et al. (2008) eingesetzt. Um anstelle der Testängstlichkeit (trait) das akute Erleben von Testangst (state) abbilden zu können, wurde die Instruktion modifiziert. So wurden zwei Formulierungen verändert:

„Gefühle und Gedanken in Prüfungssituationen“ wurde geändert in „Gefühle und Gedanken jetzt vor dem Test“; „…wie Sie sich im allgemeinen in Prüfungssituationen (Tests, Klausuren, mündlichen Prüfungen) fühlen und was Sie dabei denken“ wurde geändert in „wie du dich jetzt vor dem Test fühlst und was du dabei denkst“. Das vierstufige Antwortformat von „fast nie“ bis „fast immer“ wurde entsprechend geändert in „trifft gar nicht zu“ bis „trifft völlig zu“. Eine Veränderung der Itemformulierungen selbst war nicht notwendig, da diese sich auch im Hinblick auf das aktuelle Erleben beantworten lassen (vgl. Studie 1). Die deskriptiven Statistiken und Reliabilitäten sind Tabelle 43 zu entnehmen. Für eine Interpretation der Ausprägung relativ zur Antwortskala wurden aus den Summenscores zusätzlich Skalenmittelwerte errechnet.

Tabelle 43: Deskriptive Statistiken und Reliabilitätswerte für die Skala state-Testangst

	N	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Besorgtheit	164	5-20	5	20	10.08	3.99	15.89	.75	-.23	.90
Aufgeregtheit	162	4-16	4	14	5.97	2.32	5.37	1.25	.85	.81
Interferenz	166	3-12	3	12	6.15	2.25	5.05	.61	.03	.83
Mangel an Zuversicht	166	3-12	3	12	6.80	1.90	3.62	.31	-.37	.82
Gesamt	158	15-60	15	57	28.88	8.20	67.23	.87	.33	.90

Angegeben sind Skalensummenwerte; bei Umrechnung in Skalenmittelwerte ergeben sich folgende Werte:
Besorgtheit = 2.02; Aufgeregtheit = 1.49; Interferenz = 2.05; Mangel an Zuversicht = 2.27, Gesamt = 1.93
Md: Besorgtheit = 9.00; Aufgeregtheit = 5.00; Interferenz = 6.00; Gesamt = 27.00

Alle Subskalen sowie die Gesamtskala wiesen eine gute Reliabilität auf. Mit Ausnahme von Aufgeregtheit, wo sich auch am deutlichsten ein Bodeneffekt zeigte, war die komplette Skalenbreite ausgeschöpft. Somit kann davon ausgegangen werden, dass durchaus ein gewisses Stressniveau induziert wurde[40].

Tabelle 44. Facetteninterkorrelationen bezüglich state-Testangst

	State-Testangst		
	Besorgtheit	Aufgeregtheit	Interferenz
Aufgeregtheit	.71**		
Interferenz	.11	.30**	
Mangel an Zuversicht	.49**	.57**	.37**

N = 159-165, * $p < .05$; ** $p < .01$

Erwartungsgemäß zeigte sich eine hohe Korrelation zwischen Besorgtheit und Aufgeregtheit. Es lagen auch zwischen den übrigen Facetten positive Zusammenhänge vor (siehe Tabelle 44).

[40] Es sei hier eingeräumt, dass die Induktion von Bewertungsstress in Bedingung C nicht angestrebt wurde. Auf die Ausprägung der state-Testangst je Bedingung wird in Abschnitt 5.2.2 eingegangen.

Annäherungs- und Vermeidungsziele

Die Leistungszielorientierung wurde mit einem Fragebogen von Bachmann (2009) erfasst. Dabei handelt es sich um eine deutsche Übersetzung des Achievement Goal Questionnaire (Elliot & McGregor, 2001; siehe auch Abschnitt 1.1.1.3.1), welcher jedes Feld der 2x2-Taxonomie der Zielorientierung nach Elliot und McGregor (2001) mit je drei Items erfasst. „Mastery"-Ziele sind dabei als „Lernziele", „performance"-Ziele hingegen als „Leistungsziele" bezeichnet. Die Items zu den Lernzielen passten inhaltlich nicht für das Erleben in der Testsituation, da sie das Beherrschen von Lernmaterial, also das Erreichen eines absoluten bzw. intrapersonalen Standards adressieren (z. B. „Beim Lernen … ist es wichtig für mich, die Inhalte so gut wie möglich zu verstehen."). In der Auseinandersetzung mit einem gänzlich unbekannten Test ist es wahrscheinlicher, dass normative Standards an die eigene Leistung gelegt werden. Daher wurden ausschließlich die beiden Skalen zu Leistungszielen eingesetzt (Annäherung und Vermeidung), wodurch zwei der vier in der Taxonomie unterschiedenen Ziele erfasst wurden. Die ursprüngliche Skala bezieht sich in ihrer Formulierung stark auf den Lernalltag von Schülern (z. B. „Beim Lernen … ist es für mich wichtig, besser zu sein als andere Lernende."). Da sich Annäherungs- und Vermeidungsorientierung in dieser Studie konkret auf den Test beziehen sollten, wurden sowohl der Itemstamm als auch die Itemformulierungen modifiziert (z. B. „Beim Bearbeiten des Tests … ist es für mich wichtig, besser zu sein als andere."). So wurde statt des globalen Vergleichs mit anderen Lernenden der direkte Bezug zu den anderen Testanden bzw. deren Leistung hergestellt. Aus dem AGQ-D von Bachmann (2009) wurden also die beiden modifizierten Skalen zu Leistungsannäherungs- und Leistungsvermeidungszielen, die je drei Items umfassen, eingesetzt. Jedes Item muss dabei auf einer siebenstufigen Antwortskala (1 = „stimmt gar nicht"; 7 = „stimmt ganz genau") beantwortet werden. Die deskriptiven Statistiken sowie Reliabilitäten sind in Tabelle 45 aufgeführt.

Tabelle 45: Deskriptive Statistiken und Reliabilitätswerte für die Skala Leistungszielorientierung

	N	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Annäherung	167	1-7	1.00	7.00	3.42	1.58	2.48	.16	-.91	.88
Vermeidung	166	1-7	1.00	6.33	3.33	1.24	1.53	.11	-.37	.63

Itemzahl = 3 (beide Skalen); Interkorrelation: r = .69**

Die Reliabilität der Skala Annäherung war gut, die der Skala Vermeidung ausreichend. Verglichen mit der von Bachmann (2009) berichteten internen Konsistenz von α = .65 ist dieser eher niedrige Wert jedoch nicht überraschend. Die Mittelwerte beider Skalen befanden sich knapp unter der Skalenmitte von 4. Obwohl die Skaleninterkorrelation hoch war, wurde die Trennung beider Skalen beibehalten, auch um die Konsistenz zur Theorie zu wahren – so berichtet auch Bachmann (2009) eine ähnlich hohe Korrelation von r = .68.

Zahlenreihen

Eingesetzt wurde der Subtest „Zahlenreihen“ (ZN) aus dem WIT-2 (Kersting et al., 2008). Dieser Subtest erfüllte die Anforderung der Fragestellung. Demnach sollte sich ein STT in Bezug auf mathematische Fähigkeiten bei Aufgaben mit zahlengebundenem Inhalt zeigen. Zudem unterscheiden sich laut Manual Frauen und Männer in der Leistung in diesem Subtest nicht signifikant voneinander (basierend auf *N* = 191 Frauen und 46 Männern im Alter von 18 Jahren, mit Abitur). Das bedeutet, dass unter „neutralen“ (sprich nicht den STT auslösenden) Testbedingungen keine Geschlechtsunterschiede auftreten sollten. Drei kleinere Änderungen am Testmaterial wurden vorgenommen. Erstens wurde (wie bereits erwähnt) das zweite Beispielitem ausgetauscht, zweitens wurde (wie auch beim übrigen Testmaterial) die Anrede von „Sie“ in das für Schüler geeignetere „du“ geändert. Drittens wurde eine enthaltene Erläuterung zum Antwortbogen entfernt, da die Pbn die Lösungen direkt auf dem Aufgabenblatt in ein freies Feld eintragen sollten, und nicht wie vorgesehen auf ein separates Antwortblatt. Der Subtest ZN besteht aus 20 Zahlenreihen, die innerhalb von 10 Minuten zu bearbeiten sind. Pro richtig gelöstem Item wird ein Punkt vergeben, wodurch maximal 20 Punkte erreichbar sind. Die deskriptiven Statistiken, Reliabilität sowie Ergebnisse zur Kriteriumsvalidität sind in Tabelle 46 dargestellt.

Tabelle 46: Deskriptive Statistiken, Reliabilitätswert und Kriteriumsvalidität des Subtests Zahlenreihen

	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Zahlenreihen	1-20	0	20	11.67	4.27	18.20	-.27	-.56	.85
Korrelationen									
	Zeugnisnote				Interesse				
	Mathematik		Deutsch		Mathematik		Deutsch		
Zahlenreihen	.28**		.06		.41**		-.22**		

N = 166-168 (bei Korrelationen); Noten in Punkten von 1-15; Interesse am Fach von 1 = „sehr gering“ bis 5 = „sehr hoch“

Die deskriptiven Statistiken zeigen, dass die volle Spanne möglicher Leistungen erbracht wurde (Abbildung 14 zeigt die Verteilung der erreichten Punktzahl). Die Fälle mit drei oder weniger korrekt gelösten Items (insgesamt fünf) wurden separat analysiert. In allen fünf Fällen wurde auf Basis der vollständigen Bearbeitung der Fragebogenbatterie von einer ernsthaften Bearbeitung ausgegangen, weshalb diese Fälle inkludiert wurden. Auch die Korrelationen mit Noten bzw. Interesse waren erwartungsgemäß, insofern sich ein höherer Zusammenhang mit der Note und dem Interesse in Mathematik gegenüber der Note und dem Interesse in Deutsch fand. In der Stichprobe fand sich ein Mittelwert von 11.67. Bezogen auf die entsprechende Normgruppe (Personen, welche die (Fach-)Hochschulreife haben oder diese anstreben im Alter von 14-17 Jahren) findet sich für den Rohwert 11 ein Standartwert von 102, für den Rohwert 12 ein Standardwert von 105. Die Stichprobe lag also marginal über dem Durchschnitt (Standardwert = 100). Insgesamt sprechen diese Befunde dafür, dass die ZN instruktionsgemäß und ernsthaft bearbeitet wurden.

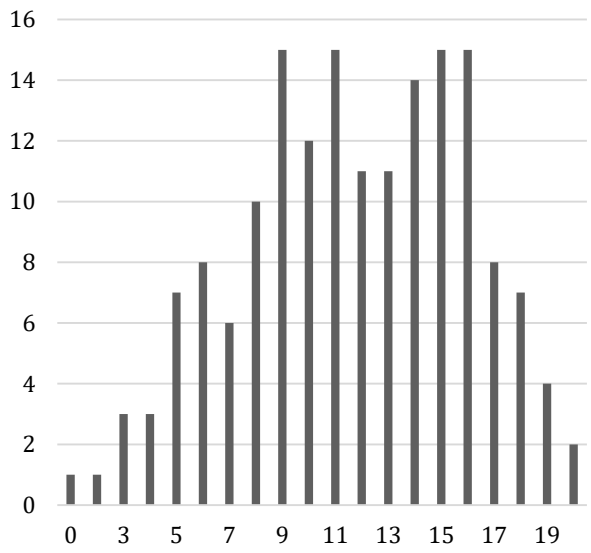

Abbildung 14: Histogramm der Gesamtpunktzahl im Subtest Zahlenreihen; N = 168

Akeptanz

Das subjektive Erleben des Tests wurde wie in Studie 1 mit dem Akzept!-L von Kersting (2008) erfasst (siehe Abschnitt 4.1.5). Die deskriptiven Statistiken und Reliabilitäten sind in Tabelle 47 aufgeführt.

Tabelle 47: Deskriptive Statistiken und Reliabilitätswerte für die Skala Akzeptanz

	N	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Kontrollierbarkeit	166	1-6	1.00	6.00	5.30	.80	.63	-2.30	8.32	.70
Messqualität	153	1-6	1.00	6.00	3.34	.97	.93	-.16	-.35	.69
Augenschein-validität	163	1-6	1.00	6.00	2.40	.99	.98	.42	-.17	.74
Belastungsfreiheit	164	1-6	1.25	6.00	4.22	1.17	1.38	-.48	-.46	.82
Note Verfahren	167	1-6	1	6	3.16	1.17	1.38	.75	.06	-
Subjektive Leistung	162	1-6	1	6	3.07	1.14	1.29	.56	.04	-

Itemzahl: 4 Items, Note & subjektive Leistung jeweils mit einem Item erfasst
Md: Kontrollierbarkeit = 5.50; Note Verfahren = 3.00; Subjektive Leistung = 3.00

Die Reliabilitäten waren ausreichend bis gut. Die Kontrollierbarkeit wurde sehr hoch eingeschätzt, was dafür spricht, dass bei den Pbn ein hinreichendes Instruktionsverständnis vorlag.

Tabelle 48: Skaleninterkorrelationen bezüglich Akzeptanz sowie Korrelationen mit der Note für das Verfahren und subjektiver Leistung

	KB	MQ	AV	BF	Note
Kontrollierbarkeit					
Messqualität	.10				
Augenscheinvalidität	.03	.30**			
Belastungsfreiheit	.42**	.10	.13		
Note Verfahren	-.20*	-.31**	-.49**	-.39**	
Subjektive Leistung	-.24**	-.31**	-.23**	-.51**	.54**

N = 148-165; Note & subjektive Leistung jeweils mit einem Item erfasst (1 = „sehr gut“, 6 = „ungenügend“)

Zwischen den Skalen lagen teilweise keine, teilweise moderate Zusammenhänge vor (siehe Tabelle 48). Von den Skalen hing die Belastungsfreiheit am stärksten mit der subjektiven Leistung zusammen zu r = -.51 (p < .001). Je belastender bzw. anstrengender der Test empfunden wurde, desto schlechter war auch die subjektive Leistung. Überdies korrelierte die Note für das Verfahren deutlich mit der subjektiven Leistung zu r = .54 (p < .001). Die Note für das Verfahren hing bezüglich der Skalen am stärksten mit der Augenscheinvalidität zusammen zu r = -.49 (p < .001): je höher die Augenscheinvalidität beurteilt wurde, desto besser wurde auch der Test bewertet.

Flow-Erleben

Das Flow-Erleben bezüglich der Testbearbeitung wurde mit der Flow-Kurzskala (FKS) von Rheinberg und Vollmeyer (2003) erfasst. Die FKS besteht aus insgesamt 13 Items. Die Erfassung des Flow-Erlebens unterteilt sich dabei in die Unterskalen „Absorbiertheit“ (4 Items; z. B. „Ich fühle mich optimal beansprucht.“) und „Glatter automatisierter Verlauf“ (6 Items; z. B. „Ich habe keine Mühe, mich zu konzentrieren.“) (Rheinberg, Vollmeyer & Engeser, 2003). Diese können zu einem Flow-Gesamtwert zusammengefasst werden. Darüber hinaus bilden drei weitere Items eine Besorgnisskala (z. B. „Ich darf jetzt keine Fehler machen.“) (Rheinberg & Vollmeyer, 2003). Alle Items besitzen ein siebenstufiges Antwortformat (1 = „trifft nicht zu“; 7 = „trifft zu“). Da die ursprünglichen Items sich auf das Erleben in der Gegenwart beziehen, wurden sie ins Imperfekt gesetzt, um das Erleben während des Tests in der Retrospektive abbilden zu können (z. B. „Ich fühlte mich optimal beansprucht.“). Da sich die Flow-Items auch auf körperliche Aktivitäten beziehen können, wurden aus zwei Items Referenzen zu körperlichen Aktivitäten bzw. Bewegungen entfernt (Item 2 und 7). Dadurch wurde der Skaleninhalt auf die kognitiven Prozesse bei der Aufgabenbearbeitung beschränkt. Die Unterskalen „Absorbiertheit“ und „Glatter automatischer Verlauf“ wurden

zu einer Skala „Flow Gesamt“ zusammengefasst. Die deskriptiven Statistiken und Reliabilitäten sind Tabelle 49 zu entnehmen.

Tabelle 49: Deskriptive Statistiken und Reliabilitätswerte für die Skala Flow-Erleben

	N	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Flow Gesamt	158	1-7	1.60	7.00	4.47	1.22	1.49	-.39	-.60	.87
Besorgnis	168	1-7	1.00	7.00	1.97	1.33	1.76	1.82	3.24	.81

Itemzahl: Flow Gesamt (10), Besorgnis (3)
Md: Besorgnis = 1.33

In beiden Skalen war die Reliabilität gut. Bezüglich Besorgnis war ein deutlicher Bodeneffekt zu beobachten, der sich durch die extreme Formulierung des Items erklären lässt (z. B. „Es stand für mich Wichtiges auf dem Spiel.“). Flow Gesamt und Besorgnis korrelierten nicht signifikant miteinander, $r = .06$ ($p = .431$).

Variablen zur subjektiven Bedrohung durch das Stereotyp

Mit zwei Items sollte erhoben werden, inwiefern sich die Pbn selbst als „Opfer“ stereotyper Erwartungen gesehen haben bzw. entsprechende Sorgen hatten. Konkret sollten die Items abbilden, inwiefern die Pbn glaubten, dass das eigene Geschlecht für die Bewertung der eigenen Leistung eine Rolle spielt („Ich habe Bedenken, dass ich aufgrund meines Geschlechts schlechter bewertet werde.“; abgekürzt als „Bewertung“) und ob die eigene Leistung mit dem eigenen Geschlecht in Verbindung gebracht werden würde („Ich habe Bedenken, dass der Versuchsleiter meine Ergebnisse auf mein Geschlecht zurückführt, wenn ich schlecht abgeschnitten habe.“; abgekürzt als „Attribution“). Beide Items sollten mittels eines fünfstufigen Antwortformats beantwortet werden (1 = „trifft nicht zu“; 5 = „trifft genau zu“). Gezielt wurde dabei so formuliert, dass die Beantwortung in Bezug auf das eigene Geschlecht erfolgen konnte. Alternativ wäre eine Frage danach möglich gewesen, ob man glaube, dass Mädchen schlechter bewertet werden würden als Jungen. Dies war jedoch nicht das Ziel, da dann die Frage für weibliche und männliche Pbn sehr unterschiedliche Bedeutung gehabt hätte. Die Items wurden am Ende der Fragebogenbatterie positioniert, um Reaktanz oder andere, unerwünschte Effekte auf die Beantwortung der Fragen zu vermeiden. Die beiden Items wurden separat analysiert und werden im Kontext der Hypothesenprüfung berichtet.

5.1.6 Statistische Verfahren

Wie in Abschnitt 3 dargelegt, wurden die Voraussetzungen auch für die berichteten Varianzanalysen geprüft. Bei bedeutsamen Verletzungen wird zusätzlich das Ergebnis einer nonparametrischen Prüfung berichtet. Für die Hypothese, dass der STT-Effekt bei targets, aber nicht bei nontargets durch die state-TA mediiert wird, war eine moderierte Mediation (mit dem Moderator Geschlecht) vorgesehen, berechnet mit der SPSS-Macro PROCESS (Hayes, 2013). Die Voraussetzungsanalysen sowie deren Interpretation und Implikationen erfolgten nach demselben Schema wie in Studie 1 und 3 (siehe Abschnitt 3). Eine Erläuterung der moderierten Mediation findet sich ebenfalls in Abschnitt 3.

5.2 Ergebnisse

5.2.1 Vorbereitende Analysen

Die vorbereitenden Analysen in Studie 2 gliedern sich in zwei Teile. Zunächst wurden die drei Bedingungen – analog zum Vorgehen in Studie 1 und 3 – auf Unterschiede in den vorab erhobenen Variablen (Note und Interesse bezüglich Mathematik und Deutsch sowie Testängstlichkeit) verglichen. Da das Geschlecht in Studie 2 eine wichtige unabhängige Variable war, wurden in einem zweiten Schritt für jede Bedingung Geschlechtsvergleiche vorgenommen. Die Ergebnisse mehrerer einfaktorieller Varianzanalysen für den Vergleich der drei Bedingungen auf besagten Variablen sind in Tabelle 50 aufgeführt.

Tabelle 50: Unterschiede zwischen den Gruppen in den Bedingungen A, B und C bezüglich Testängstlichkeit und weiterer Variablen (geprüft via einfaktoriellen Varianzanalysen)

		Bedingung A		Bedingung B		Bedingung C				
	N	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*	η_p^2
Note										
Mathematik	166	8.33	3.00	9.04	3.34	9.97	3.02	3.75	.026	.044
Deutsch[1]	167	8.79	2.30	8.58	3.50	9.43	2.24	1.49	.227	.018
Interesse										
Mathematik[2]	168	3.13	1.33	3.54	.91	3.52	1.14	2.24	.110	.026
Deutsch	168	2.96	.98	3.14	1.13	3.34	1.04	1.85	.161	.022
Trait-TÄ										
Besorgtheit	164	13.06	3.89	13.63	3.90	12.98	3.87	.47	.627	.006
Aufgeregtheit	158	6.90	2.50	7.46	2.59	7.02	2.83	.70	.514	.009
Interferenz	166	6.31	2.24	6.16	2.37	5.79	2.24	.76	.470	.009
Mangel an Zuversicht	163	7.90	2.13	7.47	2.35	7.83	2.26	.56	.573	.007
Gesamt	150	33.94	8.03	34.69	8.11	33.52	8.40	.27	.762	.004

mit Kruskal-Wallis-Test aufgrund Verletzung der Varianzhomogenität wiederholt: [1] $p = .372$; [2] $p = .197$

Insgesamt lagen zwischen den drei Bedingungen keine signifikanten Unterschiede vor, mit Ausnahme der Mathematiknote. Post hoc Vergleiche via Tukey-HSD Test zeigten, dass die Pbn in Bedingung A eine signifikant schlechtere Note aufwiesen als die Pbn in Bedingung C. Erklärung hierfür ist, dass die beiden Klassen mit den schlechtesten Noten in Mathematik in Bedingung A waren. Dazu passt, dass die Pbn in A das niedrigste Interesse in Mathematik aufwiesen, wenngleich dieser Unterschied nicht signifikant war. Auch hier zeigte ein Klassenvergleich, dass die einzige Klasse, die im Schnitt ein Mathematikinteresse unter dem Skalenmittelpunkt angegeben hat, in Bedingung A war.

Da Effekte des Geschlechts als unabhängiger Variable gemeinsam mit den Effekten der Manipulationen (A vs. B bzw. B vs. C) betrachtet wurden, wurde zusätzlich eine Betrachtung der Geschlechtsunterschiede in besagten Variablen *pro Bedingung* vorgenommen. Dies ist aussagekräftiger als eine globale Betrachtung von Geschlechtsunterschieden. Diese Vergleiche sind in Tabelle 51 dargestellt. Aus Gründen der Übersichtlichkeit sind die Facetten der Testängstlichkeit nicht mit aufgeführt.

Tabelle 51: Geschlechtsunterschiede in den Bedingungen A, B und C bezüglich Testängstlichkeit und weiteren Variablen (geprüft via t-Tests für unabhängige Stichproben)

		männlich		weiblich				
	N	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Note Mathematik								
A	50	9.48	2.59	7.44	3.07	2.51	.016	-.71
B	57	8.77	3.05	9.20	3.55	-.47	.643	.13
C	58	9.89	3.20	10.03	2.90	-.18	.859	.05
Interesse Mathematik								
A	52	3.71	1.23	2.71	1.21	2.92	.005	-.82
B	57	3.73	.88	3.43	0.92	1.22	.230	-.33
C	58	3.78	1.12	3.29	1.13	1.64	.106	-.44
Note Deutsch								
A	51	8.46	2.70	9.00	1.90	-.82	.418	.23
B	57	7.27	2.90	9.40	3.62	-2.32	.024	.63
C	58	8.96	2.14	9.84	2.28	-1.50	.139	.40
Interesse Deutsch								
A	52	2.50	.98	3.32	0.82	-3.30	.002	.92
B	57	2.32	1.00	3.66	0.87	-5.34	.000	1.45
C	58	2.93	1.00	3.71	0.94	-3.08	.003	.81
Testängstlichkeit								
A	46	31.65	7.22	36.52	8.25	-2.13	.039	.63
B	55	32.91	7.83	35.88	8.20	-1.34	.186	.37
C	48	31.71	7.11	34.93	9.17	-1.33	.192	.39

Noten in Punkten von 1-15; Interesse am Fach von 1 = „sehr gering“ bis 5 = „sehr hoch“; Skalenrange Testängstlichkeit von 15 bis 60

Aus diesem Vergleich geht hervor, dass sich nur in Bedingung A ein signifikanter Geschlechtsunterschied in der Mathematiknote sowie dem Interesse an Mathematik zeigte: demnach waren Mädchen hier schlechter in Mathematik und wiesen ein niedrigeres Interesse am Fach auf, wobei die Effekte substanziell sind. Über die Ursache der schlechteren Noten der Mädchen gerade in dieser Bedingung kann zwar nur spekuliert werden (z. B. strengere Lehrer). Jedoch legt das gleichzeitig schwächer ausgeprägte Interesse an Mathematik die Vermutung nahe, dass es in dieser Bedingung tatsächlich einen Geschlechtsunterschied in den mathematischen Fähigkeiten zuungunsten der Mädchen gab. Diese Tatsache erschwert die Rückführung etwaiger Geschlechtsunterschiede in Bedingung A auf den STT erheblich. In die Hypothesenprüfungen wurden dementspre-

chend nur die Bedingungen B und C einbezogen, wohingegen Bedingung A lediglich in den weiterführenden Analysen berücksichtigt wurde. In den Analysen, die Testangst beinhalten, wurde auf die im Leistungskontext zentrale Facette Besorgtheit zurückgegriffen.

5.2.2 Hypothesenprüfung

5.2.2.1 Hypothese 1

Zur Prüfung der Hypothese 1 wurde für jede der betrachteten abhängigen Variablen (state-Testangst sowie Annäherungs- und Vermeidungszielorientierung) eine 2 (Bedingung B vs. C) x 2 (Geschlecht) ANOVA gerechnet. Sofern sich ein signifikanter Haupteffekt oder eine Interaktion ergab, wurde im Anschluss eine einfaktorielle ANOVA über die vier Zellen (Jungen in B, Mädchen in B, Jungen in C, Mädchen in C) durchgeführt, um insbesondere auf die erwarteten Geschlechtseffekte innerhalb der Bedingungen hin prüfen zu können. Hierbei wurde auf Basis der Empfehlung von Field (2013) im Falle homogener Varianzen auf den post-hoc Test nach Gabriel zurückgegriffen. Die deskriptiven Statistiken für die Variablen zum Erleben vor dem Test sind in Tabelle 52 aufgeführt.

Tabelle 52: Mittelwerte und Standardabweichungen bezüglich state-Testangst (Besorgtheit), Annäherungszielen und Vermeidungszielen bei den Pbn in Bedingung B und C getrennt nach Geschlecht

		Jungen		Mädchen		
	N	*M*	*SD*	*M*	*SD*	*d*
State-Testangst						
B	57	9.64	3.55	9.74	3.67	.03
C	55	9.42	3.98	8.83	4.15	-.15
Annäherungsziele						
B	57	3.46	1.84	4.04	1.60	.34
C	57	3.31	1.54	2.86	1.36	-.31
Vermeidungsziele						
B	57	3.12	1.29	3.75	1.04	.55
C	56	3.18	1.32	2.84	1.20	-.27

Skalenrange der state-TA von 5 bis 20; Skalenrange der Leistungsziele von 1 bis 7

Die Ergebnisse der zweifaktoriellen ANOVA für die abhängige Variable state-TA sind in Tabelle 53 zusammengefasst.

Tabelle 53: Zweifaktorielle Varianzanalyse (Bedingung B vs. C x Geschlecht) für die abhängige Variable state-Testangst (Besorgtheit)

	SS	*df*	*MS*	*F*	*p*	η_p^2
Bedingung: B vs. C	8.67	1	8.67	.58	.446	.005
Geschlecht	1.63	1	1.63	.11	.741	.001
Interaktion	3.35	1	3.35	.23	.635	.002
Fehler	1602.26	108	14.84			

N = 112; R^2 = .009

Keiner der Effekte wurde signifikant. Es zeigten sich somit weder in Bedingung B noch Bedingung C signifikante Geschlechtseffekte, was erwartungskonform war. Entgegen der Erwartung war der Haupteffekt der Bedingung (höhere state-TA in B) ebenfalls nicht signifikant.

Die Ergebnisse der entsprechenden Analyse für Annäherungs- und Vermeidungszielorientierung sind in Tabelle 54 zusammengefasst.

Tabelle 54: Zweifaktorielle Varianzanalysen (Bedingung B vs. C x Geschlecht) für die abhängigen Variablen Annäherungs- und Vermeidungszielorientierung

	SS	*df*	*MS*	*F*	*p*	η_p^2
Annäherung						
Bedingung: B vs. C	12.12	1	12.12	4.89	.029	.043
Geschlecht	.13	1	.13	.05	.821	.000
Interaktion	7.34	1	7.34	2.96	.088	.026
Fehler	272.89	110	2.48			
Vermeidung						
Bedingung: B vs. C	4.95	1	4.95	3.45	.066	.031
Geschlecht	.60	1	.60	.42	.519	.004
Interaktion	6.40	1	6.40	4.46	.037	.039
Fehler	156.63	109	1.44			

N = 114 bzw. 113; R^2 = .079 bzw. .083

Bezüglich der Annäherungsziele zeigte sich ein signifikanter Haupteffekt der Bedingung, F (1, 110) = 4.89, p = .029. In Bedingung B zeigte sich ein höherer Wert als in Bedingung C (M_B = 3.81, SD_B = 1.71; M_C = 3.06, SD_C = 1.45). Die Interaktion war marginal signifikant, F (1, 110) = 2.96, p = .088. Die im Anschluss durchgeführte einfaktorielle ANOVA war dementsprechend ebenfalls signifikant, F (3, 110) = 3.14, p = .028. Einzelvergleiche zeigten, dass Mädchen in B eine höhere Ausprägung aufwiesen als Mädchen in C (p = .018). Bezüglich der Vermeidungsorientierung zeigte sich ein marginal signifikanter Haupteffekt der Bedingung (M_B = 3.51, SD_B = 1.17; M_C = 3.00, SD_C = 1.26), F (1, 109) = 3.45, p = .066. In Bedingung B war die Vermeidungsorientierung höher als in Bedingung C. Überdies zeigte sich eine signifikante Interaktion, F (1, 109) = 4.46, p = .037. Die anschließend durchgeführte ANOVA war signifikant, F (3, 109) = 3.31, p = .023. Einzelvergleiche

zeigten, dass Mädchen in B wiederum eine höhere Vermeidungszielorientierung berichteten als Mädchen in C (p = .017). Die beiden Interaktionen sind optisch in Abbildung 15 gut zu erkennen.

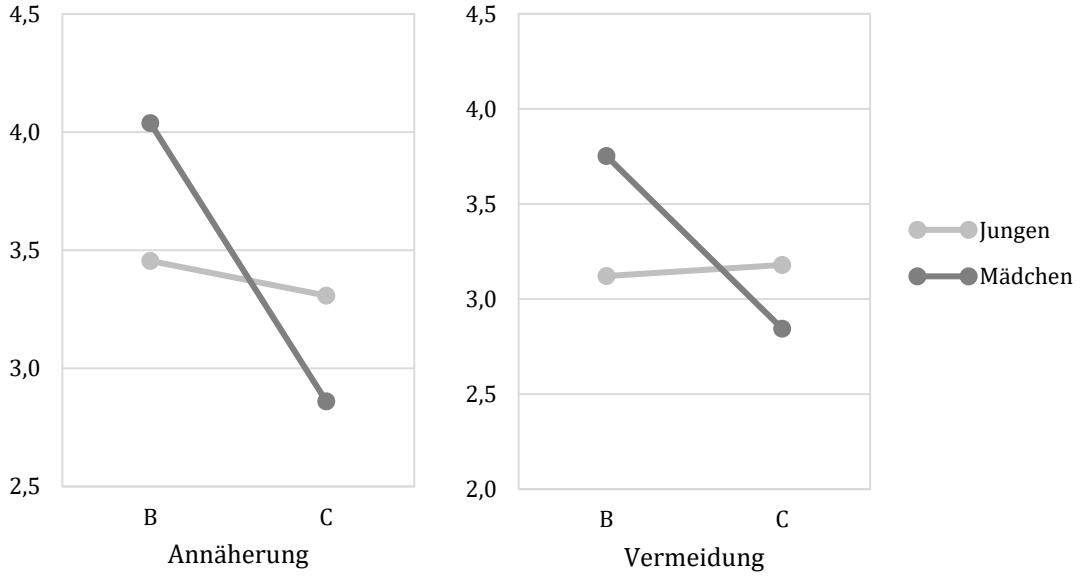

Abbildung 15: Vergleich der Annäherungs- und Vermeidungszielorientierung bei Mädchen und Jungen in B und C

Während der Haupteffekt der Bedingung bei der Vermeidungsorientierung erwartungskonform war, war der entsprechende Effekt bei der Annäherungsorientierung unerwartet, was auch für die Interaktionseffekte gilt. Insgesamt ist dieses Befundbild unerwartet. Ein erwarteter Haupteffekt bei der state-Testangst war nicht zu beobachten. Hypothese 1 war somit überwiegend abzulehnen.

5.2.2.2 Hypothesen 2a-c

Gemäß Hypothese 2a sollte sich in B und C eine vergleichbare Leistung von Jungen und Mädchen zeigen. Die deskriptiven Statistiken sind in Tabelle 55, die Ergebnisse der zweifaktoriellen ANOVA in Tabelle 56 zusammengefasst.

Tabelle 55: Mittelwerte und Standardabweichungen bezüglich der Leistung bei den Pbn in Bedingung B und C getrennt nach Geschlecht

		Jungen		Mädchen		
	N	*M*	*SD*	*M*	*SD*	*d*
Leistung Zahlenreihen						
B	57	14.00	4.19	10.74	4.39	-.76
C	58	12.67	3.92	11.19	4.02	-.37

Maximal erreichbare Punktzahl = 20

Tabelle 56: Zweifaktorielle Varianzanalyse (Bedingung B vs. C x Geschlecht) für die abhängige Variable Leistung bei Zahlenreihen

	SS	*df*	*MS*	*F*	*p*	η_p^2
Bedingung: B vs. C	5.44	1	5.44	.32	.575	.003
Geschlecht	156.12	1	156.12	9.09	.003	.076
Interaktion	22.21	1	22.21	1.29	.258	.012
Fehler	1907.52	111	17.19			

N = 115; R^2 = .084

Insgesamt zeigte sich lediglich ein signifikanter Geschlechtseffekt (Jungen: *M* = 13.27, *SD* = 4.06; Mädchen: *M* = 10.96, *SD* = 4.19). Eine anschließende einfaktorielle ANOVA war signifikant, *F* (3, 111) = 3.40, *p* = .020. Anschließende Einzelvergleiche zeigten, dass die Jungen in B bessere Leistungen erzielten als Mädchen in B (*p* = .026). Der Leistungsunterschied zwischen Jungen in B und Mädchen in C war marginal signifikant (*p* = .094). Der Haupteffekt des Geschlechts ist in Abbildung 16 gut zu erkennen. Dieses Ergebnis widerspricht der Erwartung, weshalb Hypothese 2a abgelehnt wurde.

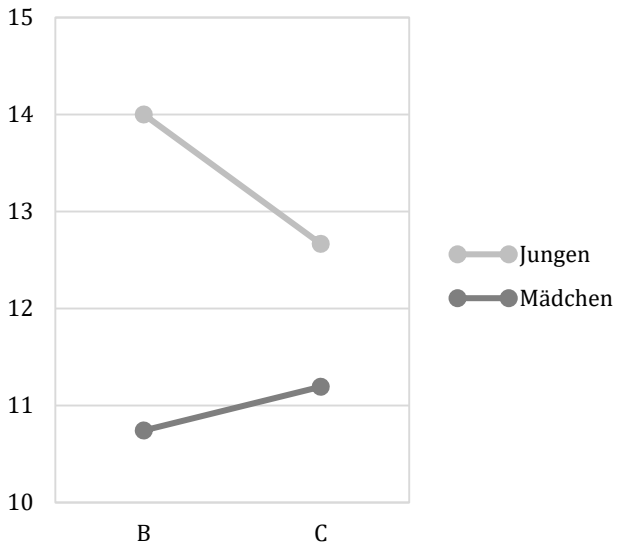

Abbildung 16: Vergleich der Leistung in Zahlenreihen bei Mädchen und Jungen in B und C

Aufgrund der besonderen Datensituation in Bedingung A (Geschlechtsunterschied in der Mathematiknote), wurde auf die geplante Mediationsanalyse zur direkten Prüfung von Hypothese 2b verzichtet (siehe hierzu die weiterführenden Analysen). Hypothese 2c hingegen wurde (mit Ausnahme von Bedingung A) geprüft durch die Berechnung bivariater Korrelationen zwischen state-TA und Leistung, separiert nach Bedingung und Geschlecht (siehe Tabelle 57).

Tabelle 57: Bivariate Korrelationen von state-Testangst (Besorgtheit) und Leistung in Zahlenreihen in Bedingung B und C, getrennt nach Geschlecht

	State-Testangst	
Zahlenreihen	Bedingung B	Bedingung C
Männlich	.26	-.36#
Weiblich	.26	-.06

N = 22-35; #*p* < .10 ; Korrelationsunterscheide zwischen den Geschlechtern je Bedingung n. s.;

Entgegen der Erwartung zeigte sich in B bei beiden Geschlechtern eine positive Korrelation zwischen state-TA und Leistung. Obgleich nicht signifikant, widerspricht der positive Zusammenhang der Vermutung, dass sich in B eine negative Korrelation zwischen state-TA und Leistung zeigt. Die Befunde in Bedingung C entsprachen nur für die Mädchen der Erwartung, für die sich eine nicht signifikante Korrelation mit Leistung fand. Der marginal signifikante, negative Zusammenhang von r = -.36 (p = .074), widerspricht der Erwartung. Hypothese 2c war somit abzulehnen.

5.2.2.3 Hypothese 3a-c

Zur Prüfung der Hypothesen 3a und 3b wurden abermals zweifaktorielle ANOVAs durchgeführt. Die deskriptiven Statistiken für die abhängigen Variablen Flow Gesamt und (Flow-)Besorgnis sowie Belastungsfreiheit sind in Tabelle 58 zusammengefasst.

Tabelle 58: Mittelwerte und Standardabweichungen bezüglich Flow Gesamt, Flow Besorgnis und Belastungsfreiheit bei den Pbn in Bedingung B und C getrennt nach Geschlecht

		Jungen		Mädchen		
	N	*M*	*SD*	*M*	*SD*	*d*
Flow Gesamt						
B	55	4.93	1.19	4.39	1.06	-.49
C	53	4.65	1.12	4.22	1.20	-.37
Flow Besorgnis						
B	57	1.88	1.50	2.34	1.32	.33
C	58	1.93	1.21	1.70	1.18	-.19
Belastungsfreiheit						
B	57	4.85	0.86	3.77	1.37	-.90
C	56	3.97	1.10	4.21	1.15	.21

Skalenrange der Flow-Skalen von 1 bis 7; Skalenrange Belastungsfreiheit von 1 bis 6

Die Ergebnisse der drei zweifaktoriellen ANOVAs sind in Tabelle 59 aufgeführt.

Tabelle 59: Zweifaktorielle Varianzanalyse (Bedingung B vs. C x Geschlecht) für die abhängigen Variablen Flow Gesamt, Flow Besorgnis und Belastungsfreiheit

	SS	*df*	*MS*	*F*	*p*	η_p^2
Flow Gesamt						
Bedingung: B vs. C	1.28	1	1.28	.99	.322	.009
Geschlecht	6.25	1	6.25	4.82	.030	.044
Interaktion	.09	1	.09	.07	.798	.001
Fehler	134.84	104	1.30			
Flow Besorgnis						
Bedingung: B vs. C	2.49	1	2.49	1.48	.226	.013
Geschlecht	.39	1	.39	.23	.630	.002
Interaktion	3.33	1	3.33	1.99	.161	.018
Fehler	186.16	111	1.68			
Belastungsfreiheit						
Bedingung: B vs. C	1.35	1	1.35	1.00	.320	.009
Geschlecht	4.88	1	4.88	3.60	.061	.032
Interaktion	11.91	1	11.91	8.78	.004	.075
Fehler	147.93	109	1.36			

N = 108-115; R^2 = .051; .038; .102

Bezüglich der Gesamtskala des Flowerlebens zeigte sich ein signifikanter Haupteffekt des Geschlechts, *F* (1, 104) = 4.82, *p* = .030. Jungen (*M* = 4.79, *SD* = 1.15) berichteten höhere Werte als Mädchen (*M* = 4.31, *SD* = 1.12). Die anschließende einfaktorielle ANOVA war nicht signifikant, *F* (3, 104) = 1.87, *p* = .139. Bezüglich der Skala Besorgnis zeigten sich keine signifikanten Effekte der zweifaktoriellen ANOVA. Hypothese 3a wurde somit nicht bestätigt.

Bei der Belastungsfreiheit zeigte sich ein marginal signifikanter Haupteffekt des Geschlechts, *F* (1, 109) = 3.60, *p* = .061, sowie eine signifikante Interaktion, *F* (1, 109) = 8.78, *p* = .004. Die einfaktorielle ANOVA zum Vergleich der vier Zellen war signifikant, *F* (3, 109) = 4.13, *p* = .008. Einzelvergleichen war zu entnehmen, dass Jungen in B mehr Belastungsfreiheit angaben als Mädchen in B (*p* = .005). Der Mittelwertsunterschied zwischen Jungen in B und Jungen in C war marginal signifikant (*p* = .059). Die Interaktion lässt sich in Abbildung 17 gut erkennen. Hypothese 3b war somit ebenfalls abzulehnen.

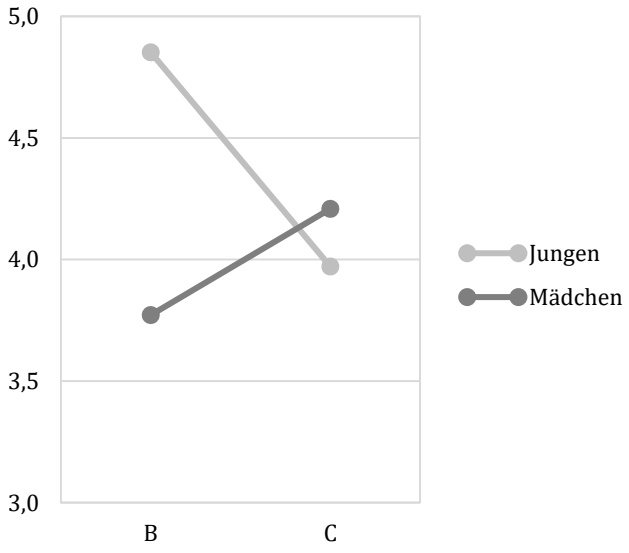

Abbildung 17: Vergleich der Belastungsfreiheit bei Mädchen und Jungen in B und C

Hypothese 3c rekurrierte auf die zwei Items zum Bedenken, dass man aufgrund des eigenen Geschlechts schlechter bewertet werde („Bewertung“) sowie zum Bedenken, dass ein schwaches Abschneiden auf das eigene Geschlecht zurückgeführt werde („Attribution“). Die entsprechenden Mittelwerte sind Tabelle 60 zu entnehmen.

Tabelle 60: Mittelwerte und Standardabweichungen für die Items zur subjektiven Bedrohung durch das Stereotyp bei den Pbn in Bedingung B und C getrennt nach Geschlecht

		männlich		weiblich		
	N	*M*	*SD*	*M*	*SD*	*d*
Bewertung						
B	57	2.14	1.61	1.23	.60	-.83
C	57	1.96	1.31	1.39	.67	-.56
Attribution						
B	56	1.59	.85	1.29	.72	-.39
C	57	1.88	1.34	1.61	1.12	-.22

Bewertung: Ich habe Bedenken, dass ich aufgrund meines Geschlechts schlechter bewertet werde.
Attribution: Ich habe Bedenken, dass der Versuchsleiter meine Ergebnisse auf mein Geschlecht zurückführt, wenn ich schlecht abgeschnitten habe. (Skala von 1-5)

Aufgrund der Verletzung der Varianzhomogenitätsannahme wurde die zweifaktorielle ANOVA nicht durchgeführt. Statt der vorgesehenen einfaktoriellen ANOVA über die vier Zellen wurde der Welch-Test gerechnet (Field, 2013), der für die abhängige Variable Bewertung ein signifikantes Ergebnis lieferte, F (3, 51.09) = 3.96, p = .013. Die anschließenden Einzelvergleiche via der Prozedur nach Games-Howell (Field, 2013) erbrachten einen marginal signifikanten Geschlechtsunterschied in B (p = .079) sowie zwischen den Jungen in C und den Mädchen in B (p = .056). Bei der Attribution war ebenfalls die Varianzhomogenitätsannahme verletzt. Der Welch-Test wurde nicht signifikant, F (3, 55.04) = 1.78, p = .162. Beide Befunde entsprachen nicht der Erwartung, weshalb Hypothese 3c abzulehnen war.

5.2.3 Weiterführende Analysen

Aufgrund des Geschlechtsunterschieds in der Mathematiknote sowie im Interesse an Mathematik in Bedingung A wurde die ursprünglich geplante Kontrastierung von Bedingung A und B in der eigentlichen Hypothesenprüfung nicht vorgenommen. Stattdessen wird nun in weiterführenden Analysen ein modifiziertes Analyseschema der Hypothesenprüfung berichtet. Effekte der Manipulation (A vs. B) auf die abhängigen Variablen wurden geprüft unter Kontrolle der Mathematiknote. Die dabei ermittelten Ergebnisse wurden in Bezug auf die Hypothesenprüfung bewertet. Die deskriptiven Statistiken der abhängigen Variablen für Bedingung A sind in Tabelle 61 getrennt nach Geschlecht aufgeführt.

Tabelle 61: Deskriptive Statistiken der abhängigen Variablen bei den Pbn in Bedingung A getrennt nach Geschlecht

		männlich		weiblich		
	N	*M*	*SD*	*M*	*SD*	*d*
State-TA	51	10.08	3.88	12.82	3.79	.71
Annäherung	52	3.67	1.55	3.14	1.44	-.36
Vermeidung	52	3.60	1.40	3.45	1.11	-.12
Leistung	52	12.83	3.52	9.57	4.33	-.82
Flow Gesamt	49	4.56	1.40	4.28	1.30	-.21
Flow Besorgnis	52	1.93	1.59	1.94	1.25	.01
Belastungsfreiheit	50	4.46	1.07	4.34	1.11	-.11
Bewertung	52	2.38	1.61	1.25	.80	-.91
Attribution	52	1.83	1.20	1.36	.83	-.46

Im Folgenden werden die Ergebnisse der 2 (Bedingung A vs. B) x 2 (Geschlecht) ANCOVAs berichtet. Im Falle signifikanter Effekte, die nicht auf die Mathematiknote zurückgehen, wurde der Geschlechtsunterschied in Bedingung A mit einer einfaktoriellen ANCOVA unter Kontrolle der Mathematiknote geprüft (Bedingung B wurde hier nicht inkludiert, da Geschlechtseffekte darin schon Gegenstand der Hypothesenprüfung waren). Auf einfaktorielle ANOVAs mit den vier Zellen (Jungen und Mädchen in A und B) wurde dementsprechend verzichtet.

Bezüglich der state-TA war ein signifikanter Effekt der Mathematiknote festzustellen, F (1, 101) = 4.09, p = .046, sowie ein signifikanter Effekt der Bedingung, F (1, 101) = 5.92, p = .017. Dieser Effekt ging auf die relativ hohe Ausprägung der state-TA bei den Mädchen in A zurück. In der anschließenden ANCOVA für Bedingung A war der Effekt des Geschlechts nicht signifikant, F (1, 46) = 2.50, p = .121. Bezüglich der Annäherungsziele zeigte sich lediglich ein signifikanter Effekt der Mathematiknote, F (1, 102) = 17.94, p < .001. Gleiches galt für die Vermeidungsziele, wobei der Effekt der Mathematiknote hier marginal signifikant wurde, F (1, 102) = 2.94, p = .089. Mit Ausnahme der höheren state-TA bei Mädchen in A sprechen diese Befunde gegen die Hypothese 1.

Bei den Zahlenreihen zeigte sich ein signifikanter Effekt der Mathematiknote, F (1, 102) = 5.99, p = .016, sowie des Geschlechts, F (1, 102) = 13.71, p < .001. Eine ANCOVA für Bedingung A lieferte einen knapp nicht signifikanten Geschlechtseffekt, F (1, 47) = 4.04, p = .05. Die Tatsache, dass sich der Geschlechtsunterschied in der Leistung auch bei Kontrolle der Mathematiknote noch abzeichnete, spricht für dessen Robustheit, was den auf Bedingung A bezogenen Teil von Hypothese 2a bestätigt. Damit lag sowohl in Bedingung A (erwartet) als auch B (unerwartet) ein Geschlechtsunterschied in der Leistung zugunsten der Jungen vor. Dadurch war von diesem Standpunkt auch Hypothese 2b abzulehnen, die ja vorhandene (in A) bzw. abwesende (in B) Geschlechtsunterschiede erwartete. In Bedingung A fand sich sowohl bei den Jungen (r = -.35, p = .093) als auch bei den Mädchen (r = -.17, p = .401) eine negative Korrelation mit der Leistung, die jedoch in beiden

Fällen – möglicherweise aufgrund der relativ geringen Fallzahl – nicht signifikant war (der Korrelationsunterschied war n. s., Fisher's z = -.65, p = .517). Dieses tendenziell zu erwartende Korrelationsmuster widerspricht den positiven Zusammenhängen von state-TA und Leistung in Bedingung B. Bezüglich Bedingung A wurde Hypothese 2c somit bestätigt, nicht jedoch in Bedingung B.

Um ein spezifischeres Bild der Zusammenhänge von state-TA und Leistung zu erhalten, wurden zusätzlich die Korrelationskoeffizienten pro Klasse berechnet. Diese sowie die Mittelwerte und Reliabilitäten der Zahlenreihen und der state-TA pro Klasse sind in Tabelle 62 aufgeführt.

Tabelle 62: Korrelation von state-Testangst (Besorgtheit) und Zahlenreihen sowie Mittelwerte und Reliabilitäten von Zahlenreihen und state-Testangst (Besorgtheit) separiert nach den Klassen

				Zahlenreihen			State-TA		
Bedingung	Klasse	*N*	*r*	*M*	*SD*	*α*	*M*	*SD*	*α*
	1	17	-.25	11.59	3.14	.79	9.82	3.99	.86
A	2	19	-.46#	11.35	5.40	.90	11.79	4.05	.89
	3	16	-.38	10.25	3.66	.83	13.00	3.46	.84
	4	16	.39	11.50	4.12	.82	8.88	3.56	.79
B	5	21	.06	10.91	4.95	.91	8.24	1.61	.40
	6	20	.25	13.55	4.26	.86	11.90	4.17	.92
	7	15	-.19	14.53	2.70	.63	8.73	3.15	.91
C	8	21	.11	9.55	4.26	.84	10.52	4.81	.94
	9	19	-.36	12.43	3.17	.73	7.84	3.39	.92

r: Korrelation von state-TA und Zahlenreihen; *N* basiert auf Korrelation von state-TA und Zahlenreihen; #p < .10;
Zahlenreihen: M = 11.09, SD = 4.23 (A); M = 12.00, SD = 4.56 (B); M = 11.88, SD = 4.01 (C)
State-TA: M = 11.52, SD = 4.00 (A); M = 9.70, SD = 3.60 (B); M = 9.11, SD = 4.05 (C)

Bezüglich des Flowerlebens lag bei der Gesamtskala nur ein Haupteffekt der Mathematiknote vor, F (1, 97) = 7.08, p = .009. Keine signifikanten Effekte zeigten sich bei der (Flow-)Besorgnis. Hypothese 3a ist daher abzulehnen. Bei der Belastungsfreiheit zeigte sich ein signifikanter Effekt der Mathematiknote, F (1, 100) = 4.00, p = .048. Darüber hinaus lag ein signifikanter Effekt des Geschlechts vor, F (1, 100) = 6.10, p = .015, sowie eine signifikante Interaktion zwischen der Bedingung und Geschlecht, F (1, 100) = 5.64, p = .019. Der geringe Unterschied in der Belastungsfreiheit in Bedingung A zeigt, dass letztgenannte Effekte auf den Geschlechtsunterschied in Bedingung B zurückzuführen waren. Dementsprechend war auch der Geschlechtseffekt in Bedingung A in der anschließenden ANCOVA nicht signifikant. Hypothese 3b war daher abzulehnen.

Bei beiden Items zur subjektiven Bedrohung durch das Stereotyp konnten keine homogenen Varianzen angenommen werden, weshalb die Ergebnisse dieser Analyse unter Vorbehalt zu betrachten sind. Bezüglich des Aspekts Bewertung zeigte sich in der zweifaktoriellen ANCOVA ein signifikanter Effekt des Geschlechts, F (1, 102) = 19.81, p < .001. Die anschließende ANCOVA in Bedingung A bestätigte einen signifikanten Geschlechtsunterschied, F (1, 47) = 8.81, p = .005. Auch beim

Aspekt Attribution lag ein signifikanter Geschlechtseffekt vor, F (1, 101) = 6.19, p = .015. Die anschließende ANCOVA in Bedingung A bestätigte einen signifikanten Geschlechtseffekt, F (1, 47) = 4.38, p = .042. Bei beiden Aspekten berichteten Jungen höhere Werte als Mädchen, weshalb Hypothese 3c abgelehnt wurde.

5.2.4 Zusammenfassung

Um ein vollständiges Bild der Ergebnisse zu erhalten, wurden neben der expliziten Hypothesenprüfung auch die weiterführenden Analysen zu Bedingung A berücksichtigt.

Die Analysen zu Hypothese 1 ergaben ein heterogenes Bild. Bezüglich der state-TA traten erwartungsgemäß keine Geschlechtseffekte in Bedingung B und C auf. Mädchen berichteten in A mehr state-TA als Jungen, jedoch wurde dieser Geschlechtsunterschied bei Kontrolle der Mathematiknote nicht signifikant. Diese Ergebnisse sind tendenziell hypothesenkonform, wobei sich jedoch in B konträr zur Erwartung keine höhere state-TA zeigte als in C. Bei den Annäherungszielen trat (bei Analyse von B und C) ein Haupteffekt der Bedingung (höhere Werte in B) auf sowie eine marginal signifikante Interaktion. Dies geht darauf zurück, dass Mädchen in B in stärkerem Maße Annäherungsziele berichteten als in C. Bei den Vermeidungszielen trat in ähnlicher Weise ein marginal signifikanter Haupteffekt der Bedingung (B vs. C; höhere Werte in B) auf sowie eine signifikante Interaktion. Auch hier berichteten die Mädchen in B eine höhere Vermeidungszielorientierung als in C. Diese Befunde sind unerwartet. Der erwartete Geschlechtseffekt in A bei den Vermeidungszielen zeigte sich in den weiterführenden Analysen nicht, auch bei den Annäherungszielen wurde nur der Effekt der Mathematiknote signifikant. Die Ergebnisse zu den Leistungszielen widersprechen der Vorannahme. Hypothese 1 ist somit insgesamt abzulehnen.

Auch die Ergebnisse zu den Hypothesen 2a-c sind überwiegend unerwartet. Die Analyse von Bedingung B und C zeigte einen signifikanten Geschlechtseffekt in der Leistung auf. Zwar lag keine signifikante Interaktion vor, doch zeigte ein direkter Vergleich, dass der Geschlechtsunterschied in der Leistung in Bedingung B mehr als doppelt so groß war wie jener in Bedingung C (d = -.76 vs. -.37). Dies entspricht einem Leistungsvorsprung der Jungen von 3.26 gelösten Items in B und 1.48 Items in C. Diese Effekte sind unerwartet. Erwartet wurde hingegen ein Leistungsunterschied in A, der sich auch zeigte. In der Größe ist dieser ähnlich ausgeprägt ist wie in Bedingung B (d = -.82; 3.26 Punkte Unterschied; siehe Abbildung 18). Unter Kontrolle der konfundierenden Note in Mathematik wurde dieser Unterschied in A nur knapp nicht signifikant (p = .05). Aufgrund des Befunds in A (erwartet) sowie B und C (unerwartet) ist Hypothese 2a abzulehnen. Da eine STT-Manipulation (A vs. B) mit Leistungsunterschieden in A bei Abwesenheit von ebensolchen in B einhergehen sollte, resultiert daraus automatisch die Ablehnung von Hypothese 2b. Vor diesem Hintergrund kann auch der tendenzielle Geschlechtsunterschied bei der state-TA in A *nicht* als

Stütze für Hypothese 2b herangezogen werden, da in B kein Effekt bei der state-TA (jedoch ein Leistungsunterschied) vorlag. Die Befunde zu den bivariaten Korrelationen von state-TA und Leistung sind ebenfalls erwartungsinkongruent. Die Richtung der Zusammenhänge war relativ unsystematisch in den einzelnen Zellen (Geschlecht x Bedingung). Lediglich in Bedingung A zeigte sich erwartungsgemäß bei beiden Geschlechtern ein negativer Zusammenhang von state-TA und Leistung, der jedoch bei den Jungen höher war (jedoch nicht signifikant höher). In Bedingung B lag bei Jungen und Mädchen ein positiver Zusammenhang von state-TA und Leistung vor, der der Erwartung diametral widerspricht. Während bei den Mädchen in C erwartungsgemäß eine Korrelation nahe 0 zwischen state-TA und Leistung vorlag, lag bei den Jungen ein negativer Zusammenhang vor. Insgesamt ist Hypothese 2c abzulehnen.

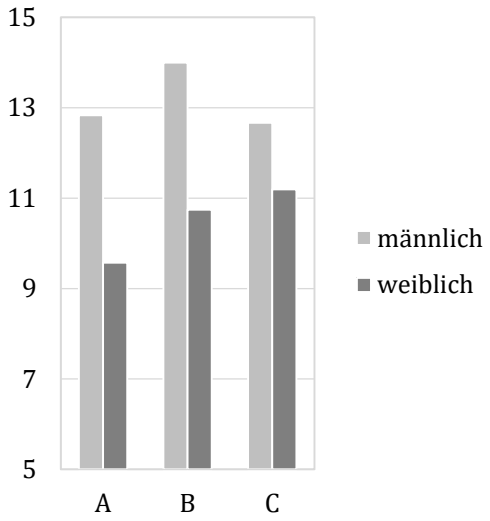

Abbildung 18: Geschlechtsvergleich der Leistung in Zahlenreihen in Bedingung A, B und C

Schließlich sind auch die Ergebnisse zu Hypothese 3a-c überwiegend unerwartet. Bei Betrachtung von Bedingung B und C lag bei der Gesamtskala des Flowerlebens ein Haupteffekt des Geschlechts vor: Jungen wiesen insgesamt höhere Werte auf als Mädchen, wohingegen sich bei der (Flow-)Besorgnis keinerlei Effekte zeigten. Durch die Berücksichtigung der weiterführenden Analysen (sprich Bedingung A) veränderte sich das Bild nicht wesentlich: hier wies lediglich die Mathematiknote signifikante Effekte auf. Da ursprünglich keine Geschlechtseffekte in B und C, jedoch in A erwartet wurden, war Hypothese 3a zu verwerfen. Bei der Belastungsfreiheit lag hingegen ein marginal signifikanter Effekt des Geschlechts sowie eine signifikante Interaktion (Bedingung B vs. C x Geschlecht) vor. Der Interaktionseffekt lässt sich auf einen deutlichen Geschlechtsunterschied in der Belastungsfreiheit in B zurückführen, der in C nicht auftrat: Jungen berichteten weniger Belastungserleben (d. h. mehr Belastungsfreiheit) als Mädchen. In Bedingung A zeigte sich kein Geschlechtsunterschied. Hypothese 3b ist somit abzulehnen. Die Analyse der Items zur subjektiven Bedrohung durch das Stereotyp erbrachte weitere unerwartete Resultate. Beim Aspekt Bewertung zeigte sich in Bedingung B ein Geschlechtsunterschied in der Form, dass Jungen eher glaubten, aufgrund ihres Geschlechts schlechter bewertet zu werden. Auch in Bedingung A lag bei Kontrolle der Mathematiknote ein Geschlechtsunterschied in gleicher Richtung vor. Während sich bei Betrachtung von Bedingung B und C keinerlei Effekte bezüglich der Attribution zeigten, lag in A ein Geschlechtsunterschied vor in der Weise, dass Jungen eher dachten, dass ein schlechtes Ergebnis ihrerseits auf ihr Geschlecht zurückgeführt würde. Da diese Effekte jeweils genau *entgegen* der Erwartung ausfielen, ist Hypothese 3c zu verwerfen.

5.3 Diskussion

Die Besonderheit von Studie 2 war die Kombination eines Designs zur Untersuchung des STT (stereotype vs. stereotypneutrale Instruktion) mit dem „traditionellen" Paradigma aus der Testängstlichkeitsforschung (evaluative vs. nonevaluative Instruktion). Ziel war, die Relevanz von Testangst für die Erklärung des STT zu untersuchen. Hierbei sollten durch die Berücksichtigung von targets und nontargets konkretere Rückschlüsse auf Erklärungsmechanismen des STT ermöglicht werden.

5.3.1 Bewertung der Hypothesen

Die Ergebnisse zu Hypothese 1 sind heterogen und erfordern eine separate Betrachtung der abhängigen Variablen. Bei der state-TA traten in Bedingung B und C keine Geschlechtsunterschiede auf, wohingegen Mädchen in A eine höhere Ausprägung der state-TA berichteten. Dieses auf den ersten Blick erwartungskonforme Ergebnis wird relativiert durch den nicht aufgetretenen Niveauunterschied in der state-TA zwischen Bedingung B und Bedingung C. Lediglich die Mädchen in Bedingung A berichteten einen erhöhten Wert bei der state-TA, wobei der Geschlechtsunterschied in A bei Kontrolle der Mathematiknote nicht mehr signifikant war. Dass die Subgruppe, welche die mit Abstand schlechteste Note in Mathematik aufwies (Mädchen in Bedingung A), auch den höchsten Wert in der state-TA berichtete, ist vor dem Hintergrund des mathematischen Charakters des Tests nicht überraschend. Aus der Berücksichtigung der Mathematiknote resultiert, dass der Geschlechtsunterschied in A (bei nicht vorhandenem Unterschied in B und C) *nicht* auf einen erwarteten STT zurückgeführt werden sollte.

Auch bezüglich der Leistungsziele sind die Ergebnisse unerwartet. Bei Kontrastierung von Bedingung B und C lagen Hinweise für Interaktionseffekte bei den Annäherungs- *und* den Vermeidungszielen vor. Während sich die Ausprägung beider Ziele bei den Jungen zwischen B und C kaum unterschied, berichteten Mädchen in B eine höhere Annährungszielorientierung und eine höhere Vermeidungszielorientierung als in C. Keine Geschlechtseffekte für die Leistungszielorientierung fanden sich in Bedingung A. Dabei ist insgesamt zu beachten, dass beide Geschlechter bei der Leistungszielorientierung insgesamt eher niedrige Werte angaben. Den höchsten Wert berichteten Mädchen in B bezüglich der Annäherungszielorientierung (M = 4.04), sie lagen damit knapp über dem Skalenmittelpunkt von 4. Das wirft die Vermutung auf, dass die Pbn emotional wenig in die Testung involviert waren. Die bereits berichteten durchschnittlich 11.67 gelösten Items bei den Zahlenreihen weisen zumindest darauf hin, dass die Pbn den Test instruktionsgemäß und ausreichend motiviert bearbeitet haben, auch wenn sie dem Ergebnis möglicherweise keine hohe subjektive Bedeutung beigemessen haben.

Der Haupteffekt der Bedingung (B vs. C) mit höheren Werten in der Vermeidungsorientierung in B ist zunächst erwartungsgemäß. Dieser ging allerdings darauf zurück, dass Mädchen in B mehr Vermeidungszielorientierung berichteten als in C. Ebenfalls unerwartet ist, dass Mädchen in B mehr Annäherungszielorientierung berichteten als in C, was auch den Haupteffekt der Bedingung (B vs. C) bei den Annäherungszielen erklärt. Bei der Vermeidungsorientierung konnte der erwartete Bedingungsunterschied über die Geschlechter hinweg nicht bestätigt werden. Das scheinbar paradoxe Ergebnis, dass Mädchen bei den Annäherungszielen *und* bei den Vermeidungszielen höhere Werte berichteten, ist durch die hohe Interkorrelation beider Variablen in der Gesamtstichprobe erklärbar (r = .69). Beide Ziele wurden keineswegs konträr zueinander gesetzt, sondern gingen stark miteinander einher.

Die Ergebnisse zu den Leistungszielen in Bedingung B und C geben Hinweise darauf, dass Mädchen auf die beiden Instruktionen unterschiedlich reagiert haben, während dies bei Jungen (diesbezüglich) weniger der Fall war. Möglicherweise wurde der Leistungscharakter von den Mädchen in B deutlicher empfunden als in C[41].

Die zentrale abhängige Variable waren die Zahlenreihen. In allen drei Bedingungen zeigten sich Geschlechtsunterscheide zugunsten der Jungen. Obgleich die Interaktion zwischen Bedingung (B vs. C) und Geschlecht nicht signifikant wurde, war der Effekt in Bedingung B doppelt so groß wie in Bedingung C. Auch in Bedingung A zeigte sich unter Kontrolle der Mathematiknote noch ein knapp nicht signifikanter Unterschied. Lediglich in Bedingung A wurde ein Geschlechtseffekt erwartet, weshalb auch Hypothese 2a abzulehnen war. Bemerkenswert ist dabei, dass selbst in der nonevaluativen Bedingung C ein (wenn auch nicht signifikanter) Geschlechtsunterschied auftrat, während die Zahlenreihen laut Manual des WIT-2 keinen Geschlechtseffekt aufweisen (Kersting et al., 2008). Insgesamt zeigte sich somit in A *und* B ein Bild, das *nur* für A erwartet wurde. Zu beachten ist, dass Mädchen in Bedingung A schlechtere Noten in Mathematik und weniger Interesse an Mathematik hatten als Jungen, wohingegen diesbezüglich in Bedingung B keine Geschlechtsunterschiede bestanden. Der Vergleich der Bedingungen B und C ähnelt formell (mit Ausnahme der nicht signifikanten Interaktion) einem STT-Effekt, der sich in diesem Fall durch einen Leistungsunterschied zwischen targets und nontargets aufgrund einer experimentellen Manipulation ausdrückt. Dies wirft die wichtige Frage auf, ob die – *eigentlich* als stereotypneutral formulierte – Instruktion in Bedingung B entgegen der Erwartung Geschlechtsstereotype aktiviert bzw. den STT ausgelöst hat. Möglicherweise war die Konnotation des Begriffs „schlussfolgerndes Denken“ näher an „mathematisches Denken“ als erwartet. Dies ist nicht unwahrscheinlich, wenn

[41] Die Mädchen in Bedingung B berichteten überdies auch ein höheres Niveau an state-TA als in C, wobei dieser Unterschied jedoch nicht signifikant ist.

man sich die Verwandtschaft zu Begriffen wie „logisches Schlussfolgern“ und „logisches Denken“ vor Augen führt – Formulierungen, die in der Untersuchungsplanung als zu stark stereotyp „männlich“ verworfen wurden.

Somit könnte eine vorläufige (wohlgemerkt spekulative) Schlussfolgerung lauten, dass in Bedingung B ein STT aufgetreten ist. Eine Besonderheit fällt dabei jedoch auf: während sich die Leistung der Mädchen in B und C nur geringfügig unterschied, lösten die Jungen in B im Schnitt 1.33 (d = -.33) Items mehr als die Jungen in C. Obwohl der Unterschied statistisch nicht bedeutsam war, führt er zur Frage, ob es sich tatsächlich um einen STT-Effekt handelt. Dabei sollten die targets (in diesem Fall Mädchen) in der Experimentalgruppe schlechter abschneiden als in der Kontrollgruppe. Wie jedoch gerade beschrieben unterschied sich die Leistung der Mädchen in Bedingung B nur wenig von der in Bedingung C.

Eine alternative Erklärung ist der sog. „stereotype boost“ (STB). Dieses Phänomen bezeichnet eine Leistungssteigerung, die infolge der Aktivierung eines positiven Stereotyps auftritt (Shih et al., 2011). Ein Beispiel für diesen neueren Forschungsstrang ist die Arbeit von Shih, Ambady, Richeson, Fujita und Gray (2002). Die Autoren legten N = 73 asiatischen Amerikanern in drei verschiedenen Bedingungen einen Mathematiktest vor (Studie 1). Die Studie hatte das Stereotyp zum Gegenstand, dass Asiaten besondere Begabung in Mathematik besitzen. Bei der ersten Gruppe wurde auf subtile Weise die Zugehörigkeit zur Gruppe der asiatischen Amerikaner aktiviert (durch die Frage, seit wie vielen Generationen die eigene Familie bereits in Amerika lebt), bei der zweiten Gruppe wurde diese offenkundig aktiviert (u. a. durch die Information, dass besagtes Stereotyp Untersuchungsgegenstand sei). Die dritte Gruppe erhielt keine derartige Manipulation. Es zeigte sich im Vergleich zu den letzteren beiden Gruppen eine bessere Leistung bei den Pbn, die eine subtile Aktivierung des Stereotyps erfuhren[42]. Shih et al. (2011) argumentieren, dass auch beim STB die Form der Aktivierung eine wichtige Rolle spielt. Effekte wie der STB machen die duale Natur von Stereotypen deutlich: negative Überzeugungen über Fähigkeiten oder Eigenschaften einer Gruppe sind verknüpft mit entsprechend positiven Überzeugungen über Personen, die *nicht* in dieser Gruppe sind – wenn also das Stereotyp besagt, dass Frauen schlecht in Mathematik sind, bedeutet es gleichzeitig, dass Männer gut in Mathematik sind (Smith & Johnson, 2006; Walton & Cohen, 2003).

Walton und Cohen (2003) führten eine Metaanalyse zum sog. „stereotype lift“ (STL) durch, ein dem STB verwandtes Phänomen. Der STL bezeichnet eine Verbesserung der Leistung infolge der Aktivierung von negativen Stereotypen über eine andere Gruppe. Die Autoren bezogen überwie-

[42] Die Autoren führten die mit der Kontrollgruppe vergleichbare Leistung in der offenkundigen Aktivierung auf den Druck zurück, einem bestimmten Stereotyp gerecht zu werden.

gend Studien in die Analyse ein, die aus dem STT-Kontext stammen, in denen jedoch auch nontargets untersucht wurden. Sie berechneten unter anderem den Leistungsunterschied zwischen nontargets in einer stereotypen gegenüber einer stereotypneutralen Bedingung (konkret: Bedingungen, in denen die Verknüpfung zwischen einem Test und einem Stereotyp aufgehoben wurde[43]). Die Autoren ermittelten einen Gesamteffekt von d = .24 (k = 28) für den STL. Die Studien wurden darüber hinaus danach kodiert, wie die Aktivierung des Stereotyps erfolgte. Dabei wurde differenziert, ob ein Test lediglich als diagnostisch für eine bestimmte Fähigkeit präsentiert wurde (vgl. die indirekte und subtile Aktivierung gemäß Nguyen & Ryan, 2008) oder aber zusätzlich auf Gruppenunterschiede (zw. targets und nontargets) hingewiesen wurde. Die Effektstärken unterschieden sich kaum (d = .11 bzw. .07). Die Autoren schlussfolgerten daraus, dass die evaluative Darstellung der besagten Tests bereits das entsprechende Stereotyp aktiviert, da die explizite Betonung von Geschlechtsunterschieden den Effekt nicht vergrößerte. Allerdings ist die Trennung von STL und STB sehr unscharf: „it is conceivable, that as a negative outgroup stereotype is induced, a positive ingroup stereotype is concurrently induced." (Shih et al., 2011, S. 143). Bei der Interpretation der vorliegenden Befunde muss auch in Erwägung gezogen werden, dass in Bedingung C ein STT aufgetreten ist. So wurde zwar in der Testinstruktion der Bewertungscharakter des Tests reduziert, jedoch wurde auch hier die Formulierung „schlussfolgerndes Denken" genutzt, wenn auch nur einmal (und nicht wie in Bedingung B zweimal). Diese Situation verdeutlicht die Schwierigkeit, die genauen Auslöser von Effekten im Kontext des STT zu identifizieren.

Wie stark STT-Effekte (und ggf. STB- bzw. STL-Effekte) sind, die allein durch die Mitteilung des diagnostischen Charakters eines Tests ausgelöst werden, ist schwer zu beziffern. Nguyen und Ryan (2008) unterschieden in ihrer Metaanalyse nicht zwischen verschiedenen Varianten der subtilen bzw. indirekten Aktivierung des STT. Auch neuere Metaanalysen (Flore & Wicherts, 2015; Picho, Rodriguez & Finnie, 2013; Stoet & Geary, 2012; Walton & Spencer, 2009) differenzierten nicht nach der genauen Form der impliziten Aktivierung. Trotz der zunehmenden metaanalytischen Aufarbeitung des Forschungsgebietes gibt es keine systematische Betrachtung dazu, ob eine implizite Aktivierung durch die Nennung des getesteten Bereichs denselben Effekt auf die Leistung hat wie das Priming der eigenen Gruppenzugehörigkeit.

Da der erwartete STT-Effekt (Leistungsunterschied in A, aber nicht in B) nicht gefunden wurde, wurde das formulierte Mediationsmodell nicht berechnet (Hypothese 2b). Die bivariaten Korrelationen der state-TA mit Leistung über die 6 Zellen (je Geschlecht und Bedingung) lieferten äußerst unerwartete Ergebnisse (Hypothese 2*c*). Der negative Zusammenhang zwischen state-TA

[43] Aufhebung bedeutete, dass in der Kontrollgruppe der diagnostische Charakter des Tests negiert wurde oder aber Gruppenunterschiede in den Testergebnissen negiert wurden.

und Leistung bei Jungen und Mädchen in A wurde erwartet. Jedoch sind sowohl die positiven Korrelationen von state-TA mit Leistung bei beiden Geschlechtern in Bedingung B als auch die negative Korrelation bei den Jungen in Bedingung C aus mehreren Gründen überraschend. Erstens gibt es keine theoretische Begründung, warum sich die Korrelation von state-TA und Leistung in den drei Bedingungen in *dieser* Weise zwischen den Geschlechtern unterscheiden sollte. Zweitens widersprechen insbesondere die positiven Korrelationen von state-TA und Leistung in Bedingung B bei beiden Geschlechtern den in Abschnitt 1.2 berichteten metaanalytischen Befunden.

Diese Befunde sind insbesondere dahingehend überraschend, da sich in der Testängstlichkeit keinerlei Unterschiede zwischen den Bedingungen ergaben. Auch eine Aufspaltung des Befunds auf die einzelnen Klassen liefert keine systematischen Erklärungen für den Effekt (siehe Tabelle 62). So zeigten sich in allen drei Klassen in Bedingung A negative Korrelationen zwischen state-TA und Leistung. Auch in Bedingung B und C waren die Befunde relativ gleichsinnig. In zwei von drei Klassen zeigte sich ein positiver (in B) bzw. ein negativer Zusammenhang (in C). Insgesamt scheint das Bild, das sich für die drei Bedingungen zeigt, *innerhalb* der Bedingungen relativ konsistent zu sein. Die große Bandbreite an Korrelationen (-.46 bis .11) spricht für die Heterogenität *dieses* Effekts über die gesamte Stichprobe hinweg. Die Mittelwerte unterschieden sich zwar bei den Zahlenreihen und der state-TA zwischen den Klassen erkennbar, aber zwischen den Bedingungen (über die Geschlechter hinweg) nicht gravierend. Darüber hinaus zeigten sich bei den Reliabilitäten keine Auffälligkeiten, mit Ausnahme der sehr niedrigen Reliabilität für die state-TA in Klasse 5. Zu beachten ist, dass sich die höchste Korrelation in Klasse 2 fand, also jener Klasse in der ein Lehrer bei der Vorerhebung erwähnte, dass es bei der Studie um Prüfungsangst gehe. Jedoch war die Leistung von Klasse 2 nicht auffällig schlecht oder gut (M = 11.35). Ursache für die Divergenzen waren möglicherweise weitere Variablen. Dies könnte beispielswiese das schwankende Verhältnis von Jungen und Mädchen in den Klassen gewesen sein (siehe Anhang G) oder aber Unterschiede im Klassenklima, unterschiedliche Lehrer oder auch Unterschiede in der soziodemographischen bzw. sozioökonomischen Zusammensetzung der Klassen. Da alle Klassen aus derselben Schule erhoben wurden, sind auch spezifische Merkmale der Schule eine mögliche Ursachen für die Ergebnisse in diesem Bereich. Auf eine weitere Subgruppenbildung (Geschlechtsvergleiche pro Klasse) musste verzichtet werden, da die Zellen nicht ausreichend besetzt waren. Ob also Geschlechtsunterschiede innerhalb der Klassen die Heterogenität in den Korrelationen erklären könnten, bleibt offen.

Hypothese 3a bezog sich auf das Flow-Erleben. Auffällig ist, dass beim Vergleich von Bedingung B und C ein Haupteffekt des Geschlechts auftrat: Jungen berichteten insgesamt ein höheres Flow-Erleben als Mädchen. Dieser Befund passt zum signifikanten Haupteffekt des Geschlechts bei der

abhängigen Variable Zahlenreihen (bei Betrachtung von Bedingung B und C). Dies lässt darauf schließen, dass sich Jungen bei der Testbearbeitung „wohler" gefühlt haben und ihnen die Aufgaben leichter gefallen sind, was sich auch an ihrer objektiven Leistung feststellen lässt. Wie auch der Leistungsbefund ist dieses Ergebnis unerwartet. In Bedingung A lag der erwartete Geschlechtseffekt im Flow-Erleben bei Kontrolle der Mathematiknote nicht vor. Keinerlei Effekte zeigten sich bei der Skala Besorgnis.

Ebenfalls unerwartet ist der Interaktionseffekt beim Belastungserleben (Hypothese 3b). Der deutliche Geschlechtsunterschied in Bedingung B ging auf einen sehr hohen Wert in der Belastungsfreiheit bei den Jungen in B zurück (M = 4.85). Dieses Ergebnis deckt sich mit der oben aufgestellten Vermutung, dass die Jungen in B einen STB bzw. STL erlebten. Eine theoretische Erklärung für STL und STB sind soziale Abwärtsvergleiche mit der Gruppe, über die ein negatives Stereotyp existiert, in diesem Fall Mädchen (Shih et al., 2011; Walton & Cohen, 2003). Auch Stereotype können ein Maßstab sein, auf denen ein sozialer Vergleich beruht (Wood, 1996). Eine Ursache für Leistungsverbesserungen und das geringe Belastungserleben könnte eine erhöhte Selbstwirksamkeit sein (Walton & Cohen, 2003). So ist im Kontext der Entwicklungsregulation bekannt, dass soziale Abwärtsvergleiche im Alter – also einer Lebensphase mit zunehmenden Entwicklungsverlusten – zur Aufrechterhaltung der Selbstwirksamkeit beitragen (Heckhausen, 1991). Bei der gemeinsamen Betrachtung der Leistung und der Belastungsfreiheit sind überdies die Mittelwerte bei den Mädchen in den beiden Bedingungen aufschlussreich: dass sich die Mädchen in beiden Bedingungen (B und C) nicht signifikant unterschieden, spricht eher für eine Leistungs*verbesserung* (und damit einen STL bzw. STB) bei den Jungen als eine Leistungs*verschlechterung* (und damit einen STT) bei den Mädchen. Ein möglicher STL- bzw. STB-Effekt erklärt auch den überraschenden Befund, dass die Jungen in C – also der nonevaluativen Bedingung – eine *geringere* Belastungsfreiheit (also eine höhere Belastung) angaben als in B. Eine Betrachtung von Bedingung A verdeutlicht wiederum die schwierige Interpretation des dortigen Befunds. Dort unterschied sich die Belastungsfreiheit zwischen den Geschlechtern nicht, weshalb erstere als Erklärung des Geschlechtsunterschieds in der Leistung (zumindest in Bedingung A) nicht herangezogen werden kann.

Will man ein Zwischenfazit ziehen, so zeigten sich bezogen auf die Leistung in beiden evaluativen Bedingungen (A und B) deutliche Geschlechtsunterschiede und in der nonevaluativen Bedingung (C) ein abgeschwächter Geschlechtsunterschied zugunsten der Jungen. Beim Vergleich von Bedingung B und C, in denen Mädchen und Jungen sowohl eine vergleichbare Note in Mathematik als auch ein vergleichbares Interesse an Mathematik hatten, zeigen sich deutliche Hinweise, dass Jungen und Mädchen unterschiedlich auf die Instruktionen reagiert haben. Die marginal signifikante Interaktion bei der Annäherungszielorientierung und die signifikanten Interaktionen bei der Ver-

meidungszielorientierung und der Belastungsfreiheit lassen darauf schließen, dass Mädchen Bedingung B als stressintensiver wahrgenommen haben als Bedingung C. Die Interaktionen bei den Leistungszielen lassen darüber hinaus vermuten, dass Mädchen in B einen STT in dem Sinne erlebt haben, dass sie sich stärker unter Druck gesetzt fühlten. Dass dabei bei *beiden* Leistungszielen eine Interaktion vorlag, ist aufgrund der hohen Interkorrelation beider Ziele keineswegs widersprüchlich. Demgegenüber weist die Interaktion bei der Belastungsfreiheit darauf hin, dass der deutliche Leistungsunterschied in B ein Ergebnis einer Leistungssteigerung bei den Jungen ist. Der Leistungsunterschied in Bedingung A kann nicht zwingend auf einen STT zurückgeführt werden. Zwar zeigte sich auch bei Kontrolle der Mathematiknote noch ein knapp nicht signifikanter Leistungsunterschied, aber sowohl bei der Leistungszielorientierung (Annäherung und Vermeidung), beim Flowerleben, der (Flow-)Besorgnis und auch bei der Belastungsfreiheit zeigten sich keine Geschlechtsunterschiede in Bedingung A. Diese hätten auf einen geschlechtsspezifischen Effekt der Manipulation zurückgeführt werden können. Zwar berichteten die Mädchen in A mehr state-TA als die Jungen, jedoch ist dieser Unterschied bei Kontrolle der Mathematiknote nicht signifikant.

Insgesamt ist der Beitrag der state-TA zur Erklärung der Effekte fragwürdig. So zeigten sich zwar erwartungsgemäß keine Geschlechtsunterschiede in der state-TA in den Bedingungen B und C, jedoch auch kein Unterschied in Bedingung A (bei Kontrolle der Mathematiknote). Da der erwartete Unterschied in der state-TA in Bedingung A nicht auftrat, wurde keine Mediation berechnet. Auch die Zusammenhänge von state-TA und Leistung werfen eher Fragen auf, als dass sie zu der Erklärung der Ergebnisse beitragen. Auch bei konservativer Betrachtung der Ergebnisse (d. h. nur Vergleich von Bedingung B und C) und der Annahme, dass in B ein STT (oder STB/STL) aufgetreten ist, kann die state-TA nicht zur Erklärung herangezogen werden. Auf die Implikationen dieser Ergebnisse wird an späterer Stelle nochmals eingegangen.

Schließlich sollen noch die – ebenfalls unerwarteten – Ergebnisse bezüglich der subjektiven Bedrohung durch das Stereotyp reflektiert werden (Hypothese 3c). Interessant sind die Ergebnisse insbesondere bezüglich des Aspekts Bewertung („Ich habe Bedenken, dass ich aufgrund meines Geschlechts schlechter bewertet werde“). In B zeigte sich ein signifikant höherer Wert bei den Jungen (d = -.83), und auch in C berichteten Jungen einen höheren Wert als Mädchen (wenn auch nicht signifikant; d = -.56). Auch in Bedingung A zeigte sich ein deutlicher Unterschied (d = -.91). Vorab vermutet wurde, dass die Mädchen – vom bekannten negativen Stereotyp „bedroht“ – erwarten, schlechter bewertet zu werden aufgrund ihres Geschlechts. Stattdessen schienen sich die Jungen mehr Gedanken über eine schlechtere Bewertung aufgrund ihres Geschlechts zu machen. Einerseits könnte das bedeuten, dass Jungen einen „umgekehrten“ bias befürchten, weil von ihnen gute Leistung erwartet wird und sie daher besonders streng beurteilt werden. Umgekehrt ist es

aber auch denkbar, dass Mädchen zwar Befürchtungen haben, dass sie schlechter bewertet werden. Unter Umständen berichten sie diese aber nicht, da sie nicht eingestehen wollen, dass ein Stereotyp sie beeinflusst, dessen Inhalt sie womöglich ohnehin ablehnen (siehe Abschnitt 1.2.2.2). Daher muss in Erwägung gezogen werden, dass die beiden Aussagen zu Bewertung und Attribution bei Jungen und Mädchen eine unterschiedliche Bedeutung hatten. Möglicherweise ist das Stereotyp „Mädchen sind schlecht in Mathe" verknüpft mit weiteren impliziten Stereotypinhalten, wie z. B. „Mädchen werden in Mathe milder benotet, weil sie schlechter sind" oder „Jungen werden strenger benotet".

Mädchen sind dabei keineswegs die einzige Gruppe, über die negative Stereotype existieren. Eine Studie von Latsch und Hannover (2014) nahm die gesellschaftliche Debatte zum „Schulversagen" von Jungen zum Ausgangspunkt (siehe z. B. Kramer, 2016). So gibt es in Deutschland mehr Jungen als Mädchen, die ohne Hauptschulabschluss von der Schule abgehen und mehr Mädchen als Jungen, die das Gymnasium besuchen (Statistisches Bundesamt, 2015). Überdies gibt es empirische Hinweise darauf, dass Jungen in der Grundschule schlechtere Noten erhalten als Mädchen, trotz vergleichbarer (z. B. über Schulleistungstests erfasster) Kompetenzen (siehe hierzu Hannover & Kessels, 2011). Latsch und Hannover (2014) konnten bei einer Befragung von Schülern (N = 206, 9. Klasse) feststellen, dass Jungen und Mädchen tatsächlich von einem negativen Bild in der Gesellschaft über Jungen in Bezug auf Schule ausgehen (Studie 1). In einer zweiten Studie bearbeiteten N = 124 (9. Klasse) Schüler Tests aus den Bereichen Deutsch und Mathematik. Die Experimentalgruppe las einen echten Zeitungsartikel (aus der Wochenzeitung „Der Spiegel"), der sich mit der Überlegenheit von Mädchen in schulischer Hinsicht befasste. Im Vergleich zur Kontrollgruppe schnitten die Jungen beim Lesetest *schlechter*, die Mädchen *besser* ab (beim Mathematiktest zeigten sich keine Effekte). Die Autoren interpretierten ersteres als STT-und letzteres als STL-Effekt, da durch den Zeitungsartikel negative Stereotype über Jungen aktiviert wurden und diese die Lese-, nicht aber die Mathematikkompetenz von Jungen tangierten. Es muss allerdings beachtet werden, dass in der vorliegenden Untersuchung die Ausprägung der Bedenken, schlechter bewertet zu werden, eher gering war. So lag der höchste Wert (M = 2.38 bei den Jungen in A) noch deutlich unter dem Skalenmittelpunkt von 3. Einschränkend muss weiterhin angeführt werden, dass sich beim Aspekt der Attribution („Ich habe Bedenken, dass der Versuchsleiter meine Ergebnisse auf mein Geschlecht zurückführt, wenn ich schlecht abgeschnitten habe.") keine so konsistenten Effekte zeigten (lediglich in Bedingung A berichteten Jungen einen signifikant höheren Wert als Mädchen).

5.3.2 Limitationen

Eine Schwäche der Studie ist das quasi-experimentelle Design. So wurden die Klassen geschlossen einer Bedingung zugewiesen. Demzufolge zeigten sich auch einige, für die Interpretation problematische Geschlechtsunterschiede in Bedingung A (Note und Interesse Mathematik), die bei einer experimentellen Randomisierung wahrscheinlich nicht aufgetreten wären. Zumindest in Bedingung B und C unterschieden sich die Geschlechter nicht in der Mathematiknote, was stringentere Rückschlüsse auf die Auswirkungen der Manipulation erlaubt. Offen bleibt dabei gleichwohl der Effekt weiterer, nicht gemessener Variablen, die womöglich für Unterschiede zwischen den Klassen verantwortlich waren (z. B. unterschiedliche Lehrer oder soziodemographische Zusammensetzung; siehe auch Abschnitt 5.3.1). Die Unterschiede zwischen den Bedingungen in der Mathematiknote hätten reduziert werden können durch eine frühere Erfassung der Noten (in der Vorerhebung) und eine anschließende Zuordnung der Klassen zu den Untersuchungsbedingungen.

Darüber hinaus ist die relativ kleine Stichprobengröße und die damit verbundene niedrige Power zu kritisieren. Da drei Bedingungen im Design enthalten waren und überdies der Faktor Geschlecht in den Analysen von wichtiger Bedeutung war, waren die einzelnen Zellen schwach besetzt. Auch das dürfte ein Grund dafür gewesen sein, dass einige der Effekte nicht signifikant geworden sind.

Bereits erwähnt wurde, dass es bei der Vorerhebung in einer Klasse, die sich in Bedingung A befand (Klasse 2), zu einer Störung kam. So teilte der Lehrer in der Vorerhebung (und nur dort) mit, dass es in der Studie um Prüfungsangst gehe. Es ist aber fraglich, ob dieser Vorfall einen nachdrücklichen Effekt auf die Ergebnisse hatte. Der vorgebliche Untersuchungszweck (FITAS; siehe Abschnitt 5.1.1) wurde erst in der Hauptuntersuchung explizit mitgeteilt. Überdies bestand die Vorerhebung lediglich aus der Erhebung des Pbn-Code und des TAI-G XU, womit die Relevanz des Themas Prüfungsangst für die meisten Pbn naheliegend gewesen sein dürfte. Eine besondere Auswirkung dieser Störung bei Klasse 2 ist also eher unwahrscheinlich.

Bei allen Interpretationen muss auch der Einflussgrad der Instruktionsformulierungen hinterfragt werden. So erhielten alle Pbn die Beispielitems direkt nach der Testankündigung. Diese allgemeine, für alle Pbn gleiche Instruktion des Tests war sehr ausführlich und lieferte einen klaren Eindruck vom mathematischen Charakter des Tests. Eine offene Frage ist, welchen Effekt eher „kleine“ Formulierungsunterschiede wie „mathematisches Denken“ und „schlussfolgerndes Denken“ in der manipulierten Testankündigung haben, wenn die anschließenden Beispielitems ohnehin aus zahlengebundenen Aufgaben bestehen. Auf der anderen Seite enthalten die Zahlenreihen im WIT-2 in der Originalinstruktion ebenfalls zwei Beispielitems, und laut Manual liegen in die-

sem Subtest keine geschlechtsspezifischen Leistungsunterschiede vor. Welche Instruktionselemente also notwendig und hinreichend sind (Instruktionsformulierungen und / oder Beispielitems) um einen STT auszulösen, muss somit noch geklärt werden.

Eine weitere Limitation ist, dass die domain identification (DI) nicht erhoben wurde. Zwar hätte diese Kovariate zur Klärung der Ergebnisse beitragen können. Jedoch zeigte die Metaanalyse von Nguyen und Ryan (2008), dass STT-Effekte, entgegen der theoretischen Erwartung, nicht mit zunehmender subjektiver Wichtigkeit der Domäne stärker werden (stattdessen waren die Effekte am stärksten bei moderater DI und niedriger bei hoher DI, was auf die möglichen Effekte von Stereotyp-Reaktanz zurückgeführt wurde). Flore und Wicherts (2015) prüften in ihrer Metaanalyse nicht auf eine Moderation des STT durch die DI. Die Bedeutung der DI ist also in zukünftigen Studien zu klären.

5.3.3 Implikationen

5.3.3.1 Theoretische und praktische Implikationen

Die Ergebnisse von Studie 2 sind durch eine starke Heterogenität geprägt. Während die absoluten Leistungsunterschiede in den drei Bedingungen eine klare Sprache zu sprechen scheinen, liefern die übrigen Ergebnisse eher Indizien als starke Belege für deren Erklärung. So dürfen bei Betrachtung der signifikanten Effekte die nicht signifikanten Effekte, insbesondere bei der state-TA, nicht ignoriert werden. Sie bestätigen vielmehr die Schwierigkeit, Erklärungsmechanismen für den STT zu identifizieren. Beispielsweise legen die Interaktionseffekte (bezogen auf Bedingung B und C) bei den Leistungszielen und der Belastungsfreiheit nahe, dass Jungen und Mädchen unterschiedlich auf die Instruktionen reagiert haben, während sich keine solchen Effekte bei der state-TA und dem Flow-Erleben zeigten.

Trotz – oder gerade wegen – der inkonsistenten Ergebnisse lassen sich einige Implikationen ableiten. Als erstes kann festgehalten werden, dass Studie 2 keine Hinweise darauf liefert, dass state-TA den Effekt des STT mediiert. Weder in den Mittelwertsunterschieden noch in den Korrelationen mit Leistung lassen sich Indizien dafür finden. In gewissem Sinne spiegeln die Ergebnisse bezüglich der state-TA auch die heterogene Befundlage zur Rolle von Angst bzw. Testangst beim STT wider. Die Hinweise für einen möglichen STT (bzw. STB / STL) in Bedingung B sind ebenfalls nicht einheitlich, so dass sie keine eindeutige Erklärung der Ergebnisse anbieten. Aus diesem Grund bestätigt Studie 2 die pessimistische Aussage von Pennington et al. (2016), nach der die bisherige Forschung bei der Identifikation von Erklärungsmechanismen für den STT nicht sehr erfolgreich war. Auch die Fokussierung auf die Facette Besorgtheit erbrachte keinen Mehrwert an Erkenntnis. Eine (spekulative) Erklärung ist, dass der klare mathematische Charakter der Beispielitems einen stärkeren Effekt hatte als die eigentliche Formulierung zur Manipulation, sprich dass in *allen* 3

Gruppen ein STT auftrat (i. S. eines Leistungsvorsprungs der Jungen), welcher in den evaluativen Bedingungen A und B eben besonders ausgeprägt war.

Eine mögliche Alternativerklärung bietet sich durch die Leistungsziele an. Brodish und Devine (2009) berichten eine sequenzielle Mediation über Leistungsvermeidungsziele und state-Testangst. Interessanterweise zeigte sich jedoch in Studie 2 (bei Bedingung B vs. C) ein Interaktionseffekt bei Annäherungs- *und* Vermeidungszielen. Es ist plausibel anzunehmen, dass ein STT nicht nur den Wunsch erzeugt, einen Misserfolg zu vermeiden (Vermeidung), sondern auch Reaktanz und den Wunsch, besser zu sein als andere (Annäherung) (siehe auch Kray et al., 2001). Es ist also wahrscheinlich, dass der STT selbst nicht mit erhöhter Testangst verbunden ist. Vermutlich sind spezifischere Prozesse der Besorgnis virulent, wenn ein STT auftritt. Die Items zur subjektiven Bedrohung durch das Stereotyp (Bewertung und Attribution) hätten hierfür Kandidaten sein können – jedoch erbrachten gerade diese ein unerwartetes Bild. Dies zeigt, dass die Wirkung von Stereotypen vermutlich mit komplexeren Prozessen verbunden ist als einer einfachen Steigerung von Bewertungsangst. Ein Beispiel hierfür ist die Untersuchung von Schmader und Johns (2003). Eine Stichprobe von $N = 75$ Studierenden bearbeitete eine Arbeitsgedächtnisaufgabe, welche entweder als solche (Kontrollgruppe) oder als Aufgabe zu mathematischen Fähigkeiten angekündigt wurde (Treatmentgruppe) (Experiment 1). Die Treatmentgruppe wurde überdies auf mögliche Geschlechtsunterschiede in besagter Fähigkeit aufmerksam gemacht. Nach der Aufgabe wurde die dabei erlebte Angst erfragt. Zudem wurde die Sorge erhoben, dass das eigene Ergebnis als Indiz für die Fähigkeiten des eigenen Geschlechts herangezogen wird (u. a. „I am concerned that the researcher will judge [women/men], as a whole, based on my performance on this test"). In der Treatmentgruppe, einer STT-Situation für Frauen, berichteten die Pbn beider Geschlechter erhöhte Sorgen diesbezüglich, wohingegen keine Effekte bei der Angst auftraten. Ein STT zeigte sich nur bei Frauen. Dieses Ergebnis legt nahe, dass womöglich spezifischere Bewertungssorgen induziert werden und nicht unbedingt eine „unspezifische", auf den Test gerichtete Angst.

Eine zweite Implikation bezieht sich auf die bekannten Probleme von Selbstberichtsdaten in der STT-Forschung. Hippel et al. (2005) beschreiben „stereotype denial" als eine Gegenreaktion auf eine Bedrohung durch ein Stereotyp: „In such a case, integrity of the self can be maintained either by denying the accuracy of the stereotype (a collective strategy) or by denying its self-relevance (an individualistic strategy)" (S. 23). Schmader et al. (2008) argumentieren, dass die beim STT auftretende Suppression negativer Emotionen und Gedanken – die wiederum Arbeitsgedächtnisressourcen bindet – dazu führt, dass sich die erlebte Angst nicht zwangsläufig in Selbstberichtsmaßen niederschlägt. Die Autoren geben ein Beispiel hierfür:

> *Imagine the student giving a speech who loses her train of thought because she is consciously trying to not feel anxious in front of an audience. Because she is trying*

> *to suppress or even deny that she is anxious, when asked on a questionnaire, she may not freely admit (even to herself) the anxiety she is feeling.* (S. 345)

Ein Beispiel für die Abweichung zwischen selbstberichteter Angst und nonverbalem Angstausdruck beim STT ist die Studie von Bosson, Haymovitz und Pinel (2004). In dieser Untersuchung mediierte der nonverbale, von Beobachtern eingeschätzte Angstausdruck den Effekt einer STT-Manipulation auf ein Leistungsmaß. Bei den targets zeigte sich keine erhöhte Angst im Selbstbericht. Jedoch untersuchten die Autoren eine hoch selektierte Gruppe (homosexuelle Männer) und verwendeten ein besonderes Leistungsmaß (Interaktionsqualität mit Kindern). Obwohl also die Generalisierbarkeit dieses Befunds fraglich ist, könnten derartige Designs Erkenntnisse über die Rolle von Angst bzw. Testangst liefern. Darüber hinaus ist jedoch noch eine weitere Interpretation denkbar. So folgern Schmader und Johns (2003), aufgrund der Uneinheitlichkeit bei Selbstberichten in ihrer Studie, dass STT-Effekte womöglich auch ohne das bewusste Erleben einer Bedrohung stattfinden. Die Inkonsistenz der Ergebnisse in Studie 2 – gepaart mit den beobachteten Leistungsunterschieden – ist mit dieser Vermutung durchaus vereinbar.

Eine dritte Implikation wurde bereits in Abschnitt 1.2.2.2.3 angedeutet. Dies ist der Gedanke, dass der STT-Effekt durch eine Reihe von Moderatoren beeinflusst wird (siehe z. B. Gerstenberg et al., 2012). So schlugen Shapiro und Neuberg (2007) das Multi-Threat Framework vor, das verschiedene Varianten des STT beschreibt. Diese sind einerseits determiniert durch das Ausmaß der Gruppenidentifikation und andererseits durch den Grad an Zustimmung zu einem Stereotyp (durch die eigene Person oder durch andere). Auch ist es denkbar, dass in unterschiedlichen Gruppen unterschiedliche Mediatoren vorliegen (Pennington et al., 2016). Zusätzliche Komplexität wird dadurch generiert, dass der STB nicht als „Spiegelbild" des STT verstanden werden sollte, da sich teilweise überschneidende und teilweise verschiedene Erklärungsmechanismen für STB und STT finden (Shih et al., 2011). Ebenso gibt es Hinweise darauf, dass unterschiedliche Aktivierungsformen des STT (subtil vs. offenkundig) durch unterschiedliche Prozesse zu Leistungsbeeinträchtigungen führen (Stone & McWhinnie, 2008). Dringend notwendig in der zukünftigen Forschung ist eine Differenzierung verschiedener Arten der expliziten und impliziten Aktivierung, und wie diese auf targets und nontargets wirken. Dabei muss geklärt werden, welche (vermeintlichen) Standardinstruktionen von Tests womöglich einen STT bzw. STB/STL auslösen. Selbstberichtsmaße reichen vermutlich nicht aus, um die tatsächlichen Prozesse hinreichend abzubilden. Ein Beispiel für den direkten Vergleich von expliziter und impliziter Aktivierung ist die Untersuchung von Smith und White (2002). Eine Stichprobe von $N = 70$ Studentinnen bearbeitete einen Mathematiktest unter drei verschiedenen Vorgaben (Studie 1). Der ersten Gruppe wurde der Test als ein neuer Mathematiktest vorgestellt (implizit). Die zweite und dritte Gruppe las einen Artikel, der beschrieb, dass und warum Männer in Mathematik besser sind als Frauen und erhielt dann

weitere Informationen über den eingesetzten Test. Die zweite Gruppe erhielt dann die Information, dass die eigene Forschung der Arbeitsgruppe den besagten Geschlechtseffekt beim Test bestätigte (explizit), die dritte Gruppe erfuhr hingegen, dass keine Geschlechtsunterschiede vorliegen würden („nullified"). Beide STT-Gruppen zeigten eine ähnliche Leistung, die jedoch schlechter war als die der dritten Gruppe.

Ein vierter Aspekt betrifft die Verbindung zum „traditionellen" Untersuchungsparadigma in der Testängstlichkeitsforschung, also der Kontrastierung von evaluativen und nonevaluativen Instruktionen. Offensichtlich ist die Auflösung der Konfundierung von evaluativer Instruktion mit (A) und ohne (B) stereotypem Gehalt und nichtevaluativer Instruktion (C) nicht gelungen. Diese Vermengung ist vermutlich mit für die unerwarteten Ergebnisse verantwortlich. Von besonderer Relevanz ist auch die bereits dargestellte Schlussfolgerung von Walton und Cohen (2003): „Notably, people appear to link negative stereotypes to evaluative tests more or less automatically." (S. 456). Diese Schlussfolgerung, obgleich diskussionswürdig, hätte erhebliche Konsequenzen. Jede Studie zu Testängstlichkeit, die sich einer evaluativen Instruktion bedient, wäre dann eine potenzielle Studie zum STT. Dies führt zu einer bedeutsamen praktischen Implikation, nämlich der Frage, ob bereits die diagnostische Ankündigung eines Tests einen STT auslöst. Nguyen und Ryan (2008) führen an, dass es in der praktischen Anwendung kaum möglich ist, den diagnostischen Zweck eines Tests zu verschweigen oder gar zu negieren, da dies etablierten Richtlinien zum Einsatz von Testverfahren zuwider laufen würde (zum Beispiel den internationalen Richtlinien für den Einsatz von Tests von der International Test Commission, 2001)[44]. Ein STT, der systematisch durch die Nennung des getesteten Bereichs ausgelöst würde, hätte erhebliche Konsequenzen für die Anwendung von Tests. Zukünftige Forschung sollte sich also speziell mit dieser Form der Aktivierung des STT befassen.

Als letztes soll noch auf einen kritischen Punkt eingegangen werden, der das gesamte Forschungsfeld zum STT betrifft. Als sozialpsychologische Ursache für Gruppenunterschiede in verschiedenen Leistungsbereichen ist der STT inzwischen ein Thema, das auch jenseits wissenschaftlicher Publikationen diskutiert wird (siehe z. B. Zecharia, 2015). Zuweilen ist der STT auch Ausgangspunkt für von Wissenschaftlern aufgebrachte Forderungen nach „affirmative action" (positiver Diskriminierung) (siehe z. B. Walton, Spencer & Erman, 2013). Parallel entfaltet sich in der Forschung jedoch eine kritische Diskussion über den STT, welche mehrere Aspekte betrifft. Flore und Wicherts (2015) beschränkten ihre Metaanalyse auf Studien zum STT bei Mädchen (Durchschnittsalter der Stichproben unter 18 Jahren) und ermittelten einen Effekt von $\bar{g}$ = .22 (k = 47 Effektgrößen). Mehrere der betrachteten Indikatoren sprechen für einen publication bias in der

[44] Nguyen und Ryan (2008) nennen die Teststandards der American Educational Research Association, der American Psychological Association und des National Council on Measurement in Education als Beispiel.

Literatur. Auch Untersuchungen zur Replizierbarkeit von STT-Effekten werfen Fragen auf. Stoet und Geary (2012) führten eine Metaanalyse durch zu Studien, welche die einflussreiche Untersuchung von Spencer et al. (1999) (targets: Frauen) zu replizieren versuchten. Von den 20 betrachteten Studien replizierten lediglich 11 den ursprünglichen Effekt. Die Autoren übten überdies Kritik an der in der STT-Forschung verbreiteten Praxis, verfügbare Fähigkeitsmaße (z. B. Noten in Mathematik oder SAT-Werte) als Kovariate in die Analyse zu nehmen (so z. B. bei Steele & Aronson, 1995). Von den 10 Studien, die *nicht* auf diese Weise rechneten, replizierten nur 3 den ursprünglichen Effekt. Sackett, Hardison und Cullen (2004) befassten sich mit der auf dieses Vorgehen basierenden und weit verbreiteten Fehlinterpretation der Ergebnisse von Steele und Aronson (1995): „absent stereotype threat, the African American–White difference is just what one would expect based on the African American–White difference in SAT scores, whereas in the presence of stereotype threat, the difference is larger than would be expected based on the difference in SAT scores." (S. 9). Die genannte Kritik ist also nicht als eine grundsätzliche Negation der Existenz des STT zu verstehen, jedoch als eine Infragestellung seiner umfassenden Bedeutung zur Erklärung von Gruppenunterschieden (siehe auch Stoet & Geary, 2012). Die Problematik von Kovariaten spricht andererseits für die Belastbarkeit der Vergleiche in Bedingung B und C – hier unterschieden sich Jungen und Mädchen nicht in ihrer Mathematiknote, weshalb Gruppenunterschiede eher auf die Manipulation zurückgeführt werden können.

5.3.3.2 Ausblick – weitere Forschung

Zukünftige Forschungsarbeiten sollten die zahlreichen, oben diskutierten Implikationen aufgreifen. Aus den Daten von Studie 2 lässt sich eine erklärende Relevanz von Testangst beim STT nicht ableiten. Ein wichtiger Erkenntnisbeitrag könnte geleistet werden durch den verstärkten Rückgriff auf (Test-)Angstmaße, die über den Selbstbericht hinausgehen, wie z. B. Implizite Assoziationstests oder Verhaltensbeobachtungen (vgl. Abschnitt 4.3.3.1). Die Ebene des Selbstberichts sollte dabei aber nicht völlig verlassen, sondern um genannte Methoden ergänzt werden, da gerade Diskrepanzen zwischen verschiedenen Maßen zu einem besseren Verständnis der Wirkung von Stereotypen beitragen könnten (siehe auch Pennington et al., 2016).

Parallel dazu sollte sich die Forschung systematisch mit dem Vergleich unterschiedlicher Aktivierungsformen des STT beschäftigen. Dies gilt insbesondere für die implizite Aktivierung des STT durch die Nennung des getesteten Bereichs, die für die Diagnostik ein Problem darstellt, dessen genauer Umfang und Geltungsbereich in den zitierten Metaanalysen aber nicht systematisch aufgearbeitet wurde. Spezielle Relevanz hat dieses Thema auch für die Testängstlichkeitsforschung. Hier wird „routinemäßig" mit evaluativen Instruktionen gearbeitet, um in low-stakes Situationen Bewertungsstress zu erzeugen. Neben der genannten Erfassung von Testangst mit Maßen, die

über Selbstberichte hinausgehen, sollte dabei auch – ähnlich wie es in Studie 2 versucht wurde – unterschieden werden zwischen der vergleichsweise unspezifischen Testangst und spezifischeren Sorgen, die sich auf die Bedrohung durch das Stereotyp beziehen. Ausgangspunkt können dabei auch Ansätze sein, die von verschiedenen Formen des STT ausgehen (siehe hierzu das Multi-Threat Framework; Shapiro, 2011; Shapiro & Neuberg, 2007). Auch der Einbezug anderer Emotionen, welcher im Leistungskontext von großer Bedeutung ist (z. B. Pekrun et al., 2011), könnte Mehrwert bringen. In diesem Kontext sollte berücksichtigt werden, welche Bedeutung Beispielitems haben. Generell ist es zu klären, welchen Effekt eine variierende Formulierung der Testankündigung („mathematisches Denken“ oder „schlussfolgerndes Denken“) hat, wenn danach vorgestellte Beispielitems ohnehin eindeutig aus numerischem Material bestehen (hierauf wurde bereits in Abschnitt 5.3.2 eingegangen).

Die unerwarteten Ergebnisse zur subjektiven Bedrohung durch das Stereotyp können darüber hinaus Ausgangspunkt sein, sich stärker mit der subjektiven Gestalt von Stereotypen zu befassen. Dies betrifft insbesondere Annahmen und subjektive Theorien über die Konsequenzen, die mit einem negativen Stereotyp verbunden sind. Dabei könnte der zitierte Ansatz zur Unterscheidung unterschiedlicher Formen des STT einen theoretischen Rahmen bilden. Auch dürfte die Annahme, dass Stereotype gemeinhin bekannt sind und somit per se bei targets einen STT (oder bei nontargets einen STB/STL) auslösen, in Zeiten des sozialen Wandels immer schwieriger aufrecht zu erhalten sein. Möglicherweise wird die STT-Forschung mehr Aufmerksamkeit auf die tatsächliche Existenz und Ausprägung von Stereotypen innerhalb der untersuchten Stichproben verwenden müssen. Die zitierte Studie von Latsch und Hannover (2014) ist hierfür ein positives Beispiel.

Schließlich sollten die neueren, kritischen Beiträge zum STT – auch im Kontext der Replikationskrise innerhalb der Psychologie (Open Science Collaboration, 2015) – Anlass sein, den Fokus in der STT-Forschung zu verändern. Letztere sollte sich vorrangig der Replikation des Effekts mit größeren Stichproben widmen und sich dabei auch neueren Entwicklungen in der Forschungslandschaft, wie z. B. der Präregistrierung von Studien, zuwenden (Flore & Wicherts, 2015). Dies gilt nicht nur vor dem Hintergrund der Relevanz des STT für den Einsatz von Tests insgesamt, sondern auch bezüglich der gesellschaftlichen Brisanz von vermeintlichen und tatsächlichen Unterschieden zwischen den Geschlechtern.

6. Studie 3

6.1 Methode

6.1.1 Kurzüberblick

Ausgangspunkt für Studie 3 war die Betrachtung der potenziell motivierenden Aspekte von Testangst. Ein Ziel der Studie war es, Erkenntnisse über das nomologische Netzwerk des Konstrukts anxiety motivation (AM) zu gewinnen. Darüber hinaus war Ziel der Studie, eine Kurzintervention (konkret eine reappraisal-Manipulation) zu erproben, die darauf basiert, Testteilnehmern die leistungsförderlichen Aspekte von Angst bewusst zu machen. Außerdem sollte untersucht werden, inwiefern anxiety motiation (als trait und als state) den Zusammenhang von Testängstlichkeit bzw. Testangst und Leistungskriterien moderiert.

Studie 3 wurde als Papier-Bleistift-Erhebung durchgeführt. Die Pbn beantworteten zunächst einige globale Fragen zur Selbsteinschätzung in verschiedenen Fähigkeiten und Eigenschaften (diese werden in dieser Arbeit nicht weiter ausgewertet). Anschließend wurde Testängstlichkeit (trait-TÄ) sowie anxiety motivation (AM-Energie und AM-Info) erhoben. Danach wurde der Leistungstest (GkKT) angekündigt, wobei unmittelbar vor dem ersten Aufgabenblock die reappraisal-Manipulation administriert wurde (Bedingung B) oder nicht (Bedingung A). Nach dem Test wurde in beiden Bedingungen die Akzeptanz des Verfahrens sowie die während des Tests erlebte state-Testangst und state-anxiety motivation erfasst (state-TA bzw. state-AM). Im Anschluss daran wurde ein weiterer (kristalliner) Intelligenztest eingesetzt, wobei in Bedingung A in zwei Subgruppen unterschiedliche Tests durchgeführt wurden. Ein Teil der Gruppe bearbeitete den BEFKI-GC-K (Schipolowski et al., 2013), der andere Teil bearbeitet den MWT-B (Lehrl, 2005)[45]. Diese beiden Varianten des Fragebogens in Bedingung A werden fortan als Version A_{BEFKI} und Version $A_{MWT\text{-}B}$ bezeichnet. Die Probanden in Bedingung B bearbeiteten ebenfalls den BEFKI-GC-K (hier abgekürzt als BEFKI). Nach diesen Tests wurden den Pbn einige Persönlichkeitsskalen vorgelegt. Hierzu gehörte ein Fragebogen zu den Eigenschaften nach dem Fünf-Faktoren-Modell (FFM) und zur Allgemeinen Selbstwirksamkeitserwartung. Danach wurden einige Variablen zum Erleben des eigenen Studiums (u. a. allgemeine Studienzufriedenheit) erhoben. Zum Schluss wurden demographische Variablen sowie ein Pbn-Code erhoben. Die Subgruppen mit den Fragebogenversionen A_{BEFKI} und $A_{MWT\text{-}B}$ wurden als *eine* Kontrollbedingung A behandelt. Bedingung B war die Treatmentgruppe.

[45] Die Ergebnisse aus diesen Verfahren werden in dieser Arbeit nicht näher beschrieben.

6.1.2 Herleitung der explorativen Analysen und Hypothesen

Die Hypothesen von Studie 3 leiten sich aus den Fragestellungen ab, die in Abschnitt 2.3 dargelegt wurden.

Eine explorative Analyse zielte darauf ab, das nomologische Netzwerk von AM zu sondieren. Ein Ansatz hierzu war die Betrachtung korrelativer Zusammenhänge zum Fünf-Faktoren-Modell nach Costa und McCrae (1992). Von zentralem Interesse waren darüber hinaus die Zusammenhänge mit Testängstlichkeit. Da die Zusammenhänge von Persönlichkeitseigenschaften zu Testängstlichkeit von der jeweiligen Facette letzterer abhängen (siehe Abschnitt 1.1.1.3), war zu klären, wie die Relationen von AM-Info und AM-Energie zu entsprechenden Facetten aussehen. Betrachtet man AM als eine mehr oder minder stabile Eigenschaft, würde dies bedeuten, dass Personen eine *ähnliche* Intensität an Angsterleben unter Umständen *unterschiedlich* bewerten (funktional oder nicht funktional). Diese Eigenschaft könnte Ausdruck der individuellen Überzeugung sein, mit den Auslösern der Angst erfolgreich umgehen und eine drohende Gefahr (d. h. einen Misserfolg) abwenden zu können. Die Untersuchung des Zusammenhangs zur Selbstwirksamkeit sollte hierbei Aufschluss geben. In diesen Kontext fügte sich auch Hypothese 1. Innerhalb Hypothese 1 sollte die Relation zwischen dem situativen Erleben von AM (state-AM) einerseits und den Dispositionen AM (AM-Energie und AM-Info) und Testängstlichkeit andererseits verglichen werden.

> **Explorative Analyse:** Zur Klärung des nomologischen Netzwerks werden die beiden Skalen von AM (AM-Energie und AM-Info) in Relation gesetzt zu den Facetten der Testängstlichkeit, zu den Persönlichkeitseigenschaften nach dem FFM sowie zur Selbstwirksamkeit. Es werden moderate Zusammenhänge erwartet, deren Richtung offen ist.
>
> **Hypothese 1:** Es wird erwartet, dass AM-Energie und AM-Info die state-anxiety motivation besser statistisch vorhersagen können als die Facetten der Testängstlichkeit.

Hypothese 2a und 2b befassten sich mit der Wirkung einer kurzen reappraisal-Manipulation. In einem experimentellen Design sollte dabei ein Intelligenztest bearbeitet werden, entweder mit (Treatmentgruppe) oder ohne vorheriger reappraisal-Instruktion (Kontrollgruppe). Die Effekte wurden dabei in zweierlei Hinsicht analysiert. Zum einen wurde betrachtet, welche Gruppenunterschiede sich in Form von Mittelwertsunterschieden bestimmter abhängiger Variablen zeigen. In Anlehnung an die zitierten Befunde zu reappraisal-Manipulationen (siehe Abschnitt 1.3.2) sollten dabei sowohl Effekte auf das absolute Leistungsniveau als auch auf die erlebte Testangst geprüft werden. Auf Basis der Befunde aus einem Prüfungskontext (Jamieson et al., 2010; 2016) wurde erwartet, dass sich durch die reappraisal-Manipulation ein positiver Effekt auf die Testleistung ergibt. Ob die reappraisal-Manipulation einen Effekt auf die erlebte Testangst haben würde, ließ sich aufgrund der heterogenen Befunde nicht mit Bestimmtheit vorhersagen. Da das re-

appraisal jedoch auf eine kognitive Umbewertung der Angst abzielte und nicht per se die Intensität der erlebten Angst senken sollte, wurde erwartet, dass es keine Auswirkung auf die Höhe der erlebten Testangst hat. Demgegenüber wurde erwartet, dass die Probanden, ähnlich wie bei Jamieson et al. (2010), durch die Manipulation in einem höheren Maße angeben, dass ihnen Angst bei der Aufgabe hilfreich war (state-AM). Ferner sollte untersucht werden, ob die Manipulation mit weiteren Gruppenunterschieden im Belastungserleben und der subjektiven Leistung einhergeht. Insofern das reappraisal zu einer subjektiv besseren „Beherrschbarkeit" oder auch Meisterung der Aufgabe führt, sollte die erlebte Belastung bei der reappraisal-Gruppe niedriger, die subjektive Leistung höher sein. Zweitens stand die Relation von Testangst und Leistung im Fokus. Ähnlich wie bei anderen Kurzinterventionen wurde vermutet, dass sich eine in der Kontrollgruppe erwartete negative Relation von Testangst und Testleistung in der Treatmentgruppe nicht zeigt oder schwächer ist. Das würde bedeuten, dass durch die funktionale Interpretation von Testangst die negative Wirkung von Testangst auf Leistung abgeschwächt oder gar aufgehoben wird.

> **Hypothese 2a:** In Bedingung B zeigen die Pbn eine bessere Leistung als in Bedingung A. Im Niveau an state-Testangst (Besorgtheit) zeigt sich kein Unterschied zwischen Bedingung B und A. In Bedingung B sind überdies die state-anxiety motivation, die Belastungsfreiheit und die subjektive Leistung höher als in Bedingung A.
>
> **Hypothese 2b:** In Bedingung B zeigt sich eine schwächere oder nicht signifikante Korrelation zwischen state-Testangst (Besorgtheit) und Leistung verglichen mit der entsprechenden Korrelation in Bedingung A.

Die Hypothesen 3a-c befassten sich mit den moderierenden Effekten von AM auf den Zusammenhang von Testangst bzw. Testängstlichkeit und Leistung. Grundannahme war, dass sich ungünstige Effekte von Testangst bzw. Testängstlichkeit *nur* zeigen, wenn Angst *nicht* als hilfreich (also als schädlich oder hemmend) empfunden wird. Diese Effekte wurden auf zwei Ebenen betrachtet. Ebene 1 bezog sich auf den Zusammenhang von Testängstlichkeit mit teilweise stabilen oder zumindest nicht extrem fluktuierenden Leistungskriterien. Das erste Leistungskriterium war die Abiturnote. Als zweites und drittes Kriterium wurden distale Maße für den Studienerfolg herangezogen (Erfolgszuversicht im Studium sowie Gedanken an einen Studienabbruch, beide Begriffe abgekürzt als „Erfolgszuversicht" und „Abbruchtendenz"). Auf Basis der Ergebnisse von Strack et al. (2014) wurde erwartet, dass mit höherer AM die negative Korrelation zwischen Testängstlichkeit und Abiturnote (bzw. Erfolgszuversicht) sowie die positive Korrelation zwischen Testängstlichkeit und Abbruchtendenz im Betrag *kleiner* (und ggf. nicht signifikant) wird. Inhaltlich würde dies bedeuten, dass AM einen „Puffer" gegen die ungünstigen Effekte von Testängstlichkeit bildet. Umgekehrt sollten „ungünstige" Effekte von Testängstlichkeit umso ausgeprägter ausfallen, je weniger Testängstlichkeit als hilfreich empfunden wird – was gleichbedeutend damit wäre, dass sie als hemmend oder beeinträchtigend wahrgenommen wird. Um die jeweilige Bedeutung von AM-

Info und AM-Energie in diesem Kontext zu prüfen, sollten beide Variablen simultan als Moderatoren inkludiert werden. Ebene 2 betrachtete das aktuelle Erleben in der vorliegenden Testsituation. Hierbei wurde die erlebte Testangst in Bezug zur Testleistung gesetzt. Analog zu Ebene 1 wurde erwartet, dass die negative Relation von Testangst und Testleistung *schwächer* wird, wenn die in der Situation berichtete state-AM *zunimmt*. Auf beiden Betrachtungsebenen (Disposition und Situation) wurde also angenommen, dass der Effekt von Testängstlichkeit und Testangst auf Leistung bzw. Leistungsmaße durch AM (AM-Energie und AM-Info, Ebene 1) bzw. state-AM (Ebene 2) moderiert wird. Diese Konstellation an Beziehungen ist in Tabelle 63 veranschaulicht.

Tabelle 63: Übersicht zu Hypothese 3a und 3b

Ebene	Prädiktoren	Moderatoren	Kriterien
1 – Disposition	Testängstlichkeit (trait)	AM-Energie AM-Info	Abiturnote Erfolgszuversicht Abbruchtendenz
2 – Situation	Testangst (state)	State-AM	Testleistung

Hypothese 3c bezieht sich auf die Effekte der Moderatoren auf die Kriterien

Negative Effekte von Testängstlichkeit bzw. Testangst sollten auf beiden Ebenen geringer werden, wenn AM-Tendenzen zunehmen – wobei auch denkbar war, dass Effekte von Testängstlichkeit bzw. Testangst sich statistisch und inhaltlich *umkehren* und positiv werden. In diesem Fall hätte Testangst bzw. Testängstlichkeit – da sie als hilfreich erlebt wird – positive Effekte auf Leistung bzw. Leistungskriterien. Alle Moderationsanalysen beinhalteten auch die Betrachtung des eigenständigen Effekts von AM-Energie, AM-Info und state-AM. Da sie leistungsförderliche Prozesse beschreiben, ist es plausibel, dass diese Variablen einen positiven Effekt auf Leistung haben würden (vgl. auch Strack & Esteves, 2014). Um bei der Analyse auf Ebene 2 für das Vorhandensein von Aufgeregtheit zu kontrollieren, wurde die Aufgeregtheit (state) als Kovariate in das Modell inkludiert. Dadurch sind besagte Prozesse bei einem mittleren Niveau an Aufgeregtheit abgebildet (aufgrund der Mittelwertszentrierung).

> **Hypothese 3a:** Auf Ebene 1 hat Testängstlichkeit (Besorgtheit) jeweils einen negativen Effekt auf (erstens) die Abiturnote sowie (zweitens) die Erfolgszuversicht im Studium und einen positiven Effekt auf (drittens) Gedanken an einen Studienabbruch. Alle drei Zusammenhänge werden umso geringer, je höher AM-Energie und AM-Info ausgeprägt sind.
>
> **Hypothese 3b:** Auf Ebene 2 wird erwartet, dass state-Testangst (Besorgtheit) einen negativen Effekt auf die Testleistung hat. Dieser Effekt ist umso schwächer, je höher die state-anxiety motivation ist. Möglich ist, dass sich die besagten Effekte bei hoher AM-Energie und AM-Info (3a) bzw. hoher state-anxiety motivation (3b) nicht nur abschwächen, sondern sich in ihrem Vorzeichen ändern.

Hypothese 3c: Darüber hinaus wird angenommen, dass AM-Energie und AM-Info (Ebene 1) und state-anxiety motivation (Ebene 2) jeweils positive Effekte auf die Leistung haben.

Bei den Zusammenhängen mit Leistung bzw. Leistungskriterien wurde von den Facetten lediglich die Besorgtheit berücksichtigt. Dies gilt gleichermaßen für die Analysen mit dem trait und dem state. Bei der Klärung des nomologischen Netzwerks von AM wurden hingegen sämtliche Facetten der Testängstlichkeit einbezogen.

6.1.3 Stichprobe

Teilnehmer der Studie waren Erstsemester der Agrarwissenschaften, Ernährungswissenschaften, Ökotrophologie und Umweltmanagement. Den Teilnehmern wurde mitgeteilt, dass sie die Tests im Rahmen der Normierung und Evaluierung der im Self-Assessment ihres Fachbereichs (Fachbereich 09 der JLU Gießen) eingesetzten Verfahren bearbeiten. Die Erhebung fand im November 2014 während einer für alle Erstsemester-Studierenden des Fachbereichs vorgesehenen Vorlesung („Mathematik und Statistik“) statt. Die Vorlesung wurde aufgrund der hohen Teilnehmerzahl in zwei Hörsälen durchgeführt, weswegen auch die Erhebung parallel in zwei Horsälen durch Mitglieder des Self-Assessment-Teams (darunter auch der Verfasser dieser Arbeit) administriert wurde. Insgesamt nahmen N = 504 Pbn an der Erhebung teil. Davon waren N = 338 in Bedingung A (N = 168 für A_{BEFKI}, sowie N = 170 für $A_{MWT\text{-}B}$) und N = 166 in Bedingung B. Die Aufteilung der Teilnehmer auf die Bedingungen erfolgte per Zufall. Bei der Berechnung der Skalenwerte wurde ein konservatives Vorgehen gewählt, so dass nur dann Skalenwerte berechnet wurden, wenn eine Skala vollständig beantwortet wurde. Gleichzeitig wurde bei den Analysen standardmäßig ein paarweiser und kein listenweiser Fallausschluss vorgenommen, um eine starke Reduktion der Analysestichprobe zu vermeiden. Die Stichprobengröße variierte bei den Hypothesenprüfungen zwischen N = 378-496. Es befanden sich N = 347 Frauen und N = 113 Männer (75 % weiblich) in der Stichprobe, 44 weitere Personen trafen keine Angabe zum Geschlecht. Die Pbn waren im Durchschnitt 20.90 (Md = 20.00; SD = 2.96) Jahre alt (72 % der Pbn waren 21 Jahre alt oder jünger). Die Geschlechts- sowie Altersverteilung waren für Studiengänge dieses Fachs bzw. für Erstsemester erwartungsgemäß. Aus der freien Nennung des Studiengangs wurde eine Kategorisierung vorgenommen, wobei vier der fünf resultierenden Kategorien den vier B. Sc.-Studiengängen des Fachbereichs 09 entsprachen. Die fünfte Kategorie bestand aus Studierenden der Beruflichen und Betrieblichen Bildung (BBB), damit verbundenen Kombinationsstudiengängen sowie mit Ernährung verbundenen Kombinationsstudiengängen, welche nicht mit dem Studiengang B. Sc. Ernährungswissenschaften identisch sind. Eine genaue Aufschlüsselung der Häufigkeiten ist in Tabelle 64 enthalten. Das Gros der Stichprobe stammte also aus den vier B. Sc.-Studiengängen des FB 09.

Tabelle 64: Häufigkeiten der Studiengangskategorien der Stichprobe in Studie 3

Agrar-wissensch.	Ernährungs-wissensch.	Öko-trophologie	Umwelt-management	BBB & Kom-binationen	Gesamt
95	97	149	107	26	474

BBB = Berufliche und Betriebliche Bildung

Die durchschnittliche Abiturnote der Stichprobe lag bei M = 2.29 (SD = .50; N = 459). Da die Erhebung früh im ersten Semester stattfand, lagen noch keine Studiennoten vor. Eine Übersicht über die deskriptiven Statistiken von Alter und Abiturnote liefert Tabelle 65.

Tabelle 65: Deskriptive Statistiken zu Alter und Abiturnote

	N	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*
Alter	462	18	49	20.90	2.96	8.74	3.35	20.57
Abiturnote	459	1.0	3.7	2.29	.50	.25	.11	.12

Abiturnote in Schulnoten von 1 bis 6

Die Männer (M = 21.76, SD = 3.06) waren etwas älter als die Frauen (M = 20.60, SD = 2.90), t (442) = 3.59, p < .001. Erwartungsgemäß hatten Frauen (M = 2.26, SD = .49) signifikant bessere Abiturnoten als Männer (M = 2.42, SD = .53), t (437) = 2.84, p = .005. Die Abiturnoten unterschieden sich signifikant zwischen den Studiengangskategorien, F (4, 450) = 18.83, p < .001 – Ernährungswissenschaftler hatten im Schnitt die relativ besten Abiturnoten (2.13), Agrarwissenschaftler die schlechtesten (2.67).

6.1.4 Beschreibung des Untersuchungsablaufs

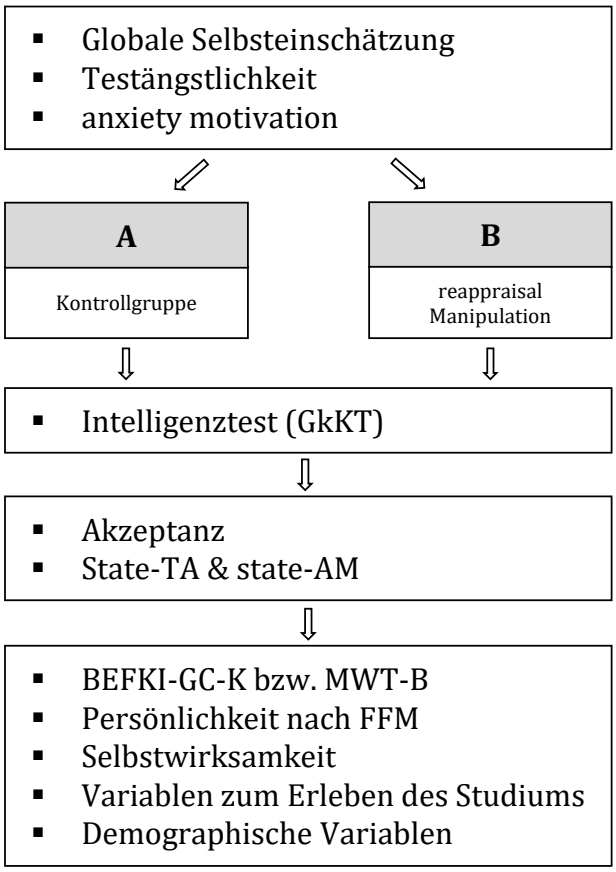

Abbildung 19: Untersuchungsablauf von Studie 3

Die Erhebung fand während einer Vorlesung statt und wurde parallel in zwei Hörsälen durchgeführt. Nach einer Einführung durch die Dozentin wurden Zweck und Ablauf der Erhebung durch Mitarbeiter des Self-Assessment-Teams vorgestellt. Nach dem Austeilen der Fragebogen- und Testbatterie (in einem Bogen) wurden die Pbn mit Hilfe von Präsentationsfolien durch die Erhebung geleitet. Ein geregelter und standardisierter Ablauf der Untersuchung sollte so gewährleistet werden. Das Untersuchungsmaterial war zu diesem Zweck zusätzlich mit Hinweisen versehen (z. B. „Bitte erst nach Aufforderung weiterblättern."). Die Randomisierung wurde dadurch erreicht, dass die Bögen beim Austeilen fortlaufend nach Bedingung sortiert waren (A_{BEFKI}-$A_{MWT\text{-}B}$-B usw.). Die Unterschiedlichkeit der Versionen bzw. Bedingungen war für die Pbn nicht unmittelbar zu erkennen. Nach der Erhebung der globalen Selbsteinschätzung, der Testängstlichkeit und der anxiety motivation wurde der Intelligenztest (GkKT) angekündigt. Nach dem GkKT wurde die Akzeptanz des Verfahrens, die state-TA sowie die state-AM erhoben. Um Kontexteffekte auf die letzten beiden Maße zu verhindern, wurden die im Akzeptanzfragebogen enthaltenen Items zur subjektiven Leistung und zur Gesamtbewertung des Tests *nach* der state-TA und der state-AM positioniert. Anschließend wurden die Pbn instruiert, dass nun ein Teil den BEFKI, der andere Teil den MWT-B bearbeiten würde. Für beide Tests wurde eine Zeitbegrenzung von 5 Minuten vorgegeben. Nach Ende der Bearbeitungszeit des BEFKI bzw. MWT-B wurde die state-TA und state-AM bezogen auf das jeweilige Verfahren erhoben (dies wird in dieser Arbeit nicht berichtet). Anschließend wurden die Pbn angewiesen, den weiteren Teil des Fragebogens nun selbständig zu bearbeiten. Dies beinhaltete den IPIP-Fragebogen zum FFM, die Items zur Selbstwirksamkeit, (nicht ausgewertete) Items zu Dominanz, demographische Variablen und Variablen zum Erleben des eigenen Studiums (siehe auch Abbildung 19). Die Pbn konnten zum Schluss angeben, ob sie eine Rückmeldung zu Ihren Ergebnissen im Test erhalten wollen. Die gesamte Erhebung dauerte etwa 60 Minuten.

6.1.5 Darstellung der Erhebungsdetails und Operationalisierung der Konstrukte

Testängstlichkeit

Analog zu Studie 1und 2 wurde der TAI-G XU (Wacker et al., 2008) zur Erfassung der dispositionellen Testängstlichkeit eingesetzt. Die deskriptiven Statistiken und Reliabilitäten sind Tabelle 66 zu entnehmen.

Tabelle 66: Deskriptive Statistiken und Reliabilitätswerte für die Skala Testängstlichkeit

	N	*Skala*	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Besorgtheit	470	5-20	5	20	13.22	3.58	12.78	.03	-.75	.86
Aufgeregtheit	468	4-16	4	16	7.37	2.65	7.01	.77	.20	.83
Interferenz	484	3-12	3	12	6.17	2.01	4.06	.43	-.09	.76
Mangel an Zuversicht	472	3-12	3	12	7.43	1.88	3.55	-.18	-.57	.76
Gesamt	429	15-60	16	55	34.17	7.50	56.31	.13	-.41	.87

Angegeben sind Skalensummenwerte; bei Umrechnung in Skalenmittelwerte ergeben sich folgende Werte: Besorgtheit = 2.64; Aufgeregtheit = 1.84; Interferenz = 2.06; Mangel an Zuversicht = 2.48; Gesamt = 2.28
Md: Aufgeregtheit = 7.00

Bei allen Subskalen war die komplette Skalenbreite ausgeschöpft. Ein tendenzieller Bodeneffekt zeigte sich bei Interferenz und insbesondere bei Aufgeregtheit. Die Reliabilitäten waren durchweg gut. Der Mittelwert bei den Frauen (M = 34.78, SD = 7.53) lag signifikant unter jenem von Wacker et al. (2008) für eine weibliche studentische Stichprobe berichteten Wert (M = 36.2), t (300) = -3.26, p = .001. Der Mittelwert bei den Männern (M = 32.25, SD = 7.03) wich hingegen nicht von dem entsprechend bei Wacker et al. (2008) berichteten Wert für Männer (M = 32.7) ab t (98) = -.63, p = .528. Frauen berichteten höhere Werte als Männer bei Besorgtheit (M = 13.59, SD = 3.46 bzw. M = 12.07, SD = 3.66), t (431) = -3.90, p < .001, bei Aufgeregtheit (M = 7.62, SD = 2.69 bzw. M = 6.67, SD = 2.41), t (432) = -3.26, p = .001, sowie bei Mangel an Zuversicht (M = 7.64, SD = 1.85 bzw. M = 6.76, SD = 1.80), t (434) = -4.27, p < .001. Unerwartet ist nur der höhere Wert für Interferenz bei Männern (M = 6.51, SD = 2.17 bzw. M = 6.05, SD = 1.98), t (445) = 2.05, p = .041. Folglich ergab sich auch ein höherer Gesamtwert bei Frauen als bei Männern t (398) = -2.95, p = .003. Die Facetteninterkorrelationen sowie Zusammenhänge mit Abiturnote sind in Tabelle 67 dargestellt.

Tabelle 67: Facetteninterkorrelationen bezüglich Testängstlichkeit und Korrelationen mit Abiturnote

	Trait-TÄ			Abiturnote
	BE	AU	IN	
Besorgtheit				.07
Aufgeregtheit	.52**			.02
Interferenz	.25**	.20**		.22**
Mangel an Zuversicht	.53**	.43**	.22**	.13**
Gesamt				.13**

N = 431-461 (beim Gesamtwert Testängstlichkeit N = 396)

Es zeigten sich erwartungsgemäß moderate bis hohe Zusammenhänge zwischen den Subskalen. Von diesen zeigte Interferenz den stärksten Zusammenhang zur Abiturnote.

Anxiety Motivation

Grundlage der Skala sind die sieben Items von Strack et al. (2014), welche die beiden Faktoren AM-Energie und -Info erfassen. Dabei wurden wesentliche Modifikationen vorgenommen. Zunächst wurde der Fokus der Items, der auf dem Arbeitskontext lag (z. B. „Wenn ich etwas Angst verspüre in einer Arbeitssituation … kanalisiere ich meine Energie in meine Arbeit“) auf einen Prüfungs- und Testkontext übertragen („Wenn ich während einer Prüfung oder einem Test Angst empfinde…“). Erfasst werden sollte das dispositionale Erleben während einer Prüfungssituation, wodurch eine analoge Interpretation zu den Items des TAI-G-XU gewährleistet werden sollte. Darüber hinaus wurden bei sechs der sieben Items Änderungen in der Itemformulierung vorgenommen. Bei den Items zu AM-Energie wurde in der Originalskala in drei der vier Items der Begriff „Energie“ verwandt – in der neuen Skala wurden die Formulierungen hingegen heterogener gestaltet, um eine größere Spanne an Merkmalsaspekten abzubilden. Ein Item („… kanalisiere ich meine Energie in meine Arbeit“) wurde semantisch einfacher formuliert („… richte ich meine Energie auf die Prüfung / den Test“). In ähnlicher Weise wurden die Items zu AM-Info modifiziert. Der Fokus der Items lag nun darauf, inwieweit Angst zu dem Bewusstsein führt, dass eine höhere Anstrengung bzw. Konzentration erforderlich ist. Bei der Umformulierung wurde darauf geachtet, dass sich die Items sowohl auf mündliche wie auf schriftliche Prüfungssituationen übertragen lassen, weshalb die Formulierungen „arbeiten“ bzw. „härter arbeiten“ durch „anstrengen“ bzw. „Mühe geben“ ersetzt wurden. Das fünfstufige Antwortformat (1 = „stimme nicht zu“ bis 5 = „stimme völlig zu“) wurde beibehalten.

Aufgrund der erheblichen Veränderung der Skala wurde eine Hauptkomponentenanalyse durchgeführt. Erwartet wurde, dass sich die Items trotz der Veränderung wieder in die beiden Komponenten Energie und Information separieren lassen würden. Das Kaiser-Meyer-Olkin-Maß der Stichprobeneignung lag bei .816, der Barltett-Test wurde mit $p < .001$ signifikant, was die Eignung der Itemauswahl für eine Faktorenanalyse bestätigte (Bühner, 2006). In Anlehnung an Strack et al. (2014) wurde oblique rotiert (Promax). Eine Parallelanalyse nach Horn ergab zwei zu extrahierende Komponenten. Der Minimum-Average-Partial-Test hingegen empfahl eine Komponente (Bühner, 2006). Bei einer Durchführung der Hauptkomponentenanalyse ohne vorher festgelegte Anzahl an Faktoren resultierten zwei Komponenten mit einem Eigenwert > 1 (Screeplot siehe Anhang E). Diese Faktorlösung klärte insgesamt 63.86% der Gesamtvarianz auf (erste Komponente 47.30 %, zweite Komponente 16.56 %). Tabelle 68 gibt die Faktorladungen sowie die Kommunalitäten wieder. Die Items sind dabei nach der erwarteten Subskalenzugehörigkeit sortiert.

Tabelle 68: Mustermatrix der Hauptkomponentenanalyse mit den Items zu anxiety motivation

		Komponente		h^2
Nr.	Wenn ich während einer Prüfung oder einem Test Angst empfinde…	1	2	
1	… verleiht mir das mehr Schwung	**.914**	-.272	.699
3	… richte ich meine Energie auf die Prüfung / den Test	**.559**	.301	.545
5	… spornt mich das an	**.816**	.075	.724
6	… bin ich aktiver bei der Problemlösung	**.807**	.021	.666
2	… erinnert mich das daran, dass ich mich konzentrieren muss	.366	**.445**	.470
4	… macht mich das darauf aufmerksam, dass ich mir Mühe geben muss	.154	**.756**	.694
7	… gibt mir das ein klares Signal dafür, dass ich mich mehr anstrengen muss	-.254	**.894**	.672

N = 432; Hauptkomponentenanalyse mit Promax-Rotation; KMO = .816, Bartlett Test $p < .001$; Komponentenkorrelation: .42; Varianzaufklärung kumuliert: 63.86%; Items nach postulierter Skalenzugehörigkeit sortiert, Hauptladungen fett dargestellt

Item 1, 5, 6, 4 und 7 ließen sich eindeutig einer Komponente zuordnen, da sie eine bedeutsame Haupt- und keine substantielle Nebenladung aufwiesen. Die Ergebnisse der Analyse unterstützten die Interpretation zweier Faktoren. Es fand sich eine annähernde Einfachstruktur, wobei Item 3 und insbesondere Item 2 schwächere Hauptladungen sowie Nebenladungen aufwiesen. Es fand sich keine so klare Einfachstruktur wie bei Strack et al. (2014), was auf die deutlich heterogeneren Items zurückgeht. Für die Zuordnung dieser beiden Items zu den Komponenten wurden eine Reihe statistischer Kriterien sowie theoretische Argumente herangezogen. Erstes Kriterium war die Zuordnung eines Items zu dem Faktor, auf dem es am höchsten lädt. Zweites Kriterium war die Überschreitung der praktischen Bedeutsamkeit der Ladung auf Basis von in der Literatur genannten Heuristiken. Diese liegt nach Bühner (2006) sowie Eid et al. (2013) bei .30, nach Tabachnick und Fidell (2010) bei .32. Peterson (2000) berichtet in einer Metaanalyse exploratorischer Faktorenanalysen, dass die am häufigsten gewählte Untergrenze der praktischen Bedeutsamkeit eine Ladungshöhe von .40 ist. Bezüglich der Differenz von Haupt- und Nebenladung findet sich als Heuristik auch ein Mindestwert von .20 (Ferguson & Cox, 1993). Als drittes Kriterium wurde das Fürntratt-Kriterium herangezogen ($a^2/h^2 > .50$) (Bühner, 2006). Eine vollständige Exklusion eines Items (viertens) sollte bei einer Kommunalität unter .40 erfolgen (Costello & Osborne, 2005). Nach allen vier Kriterien ließ sich Item 3 eindeutig der ersten Komponente zuordnen. Das erste und letzte Kriterium war für Item 2 erfüllt, das zweite Kriterium teilweise und das dritte nicht. Die schwächere Hauptladung von Item 2 lässt sich inhaltlich begründen: so fokussiert Item 2 eher auf mentale Anstrengung („dass ich mich konzentrieren muss“), während Item 4 und 7 unspezifischer sind („Mühe“ und „anstrengen“). Da Item 2 jedoch anders als die Items 1, 3, 5 und 6 eine

klare Informationskomponente besitzt („erinnert mich das daran…“), wurde es der zweiten Komponente zugeordnet. Die erste Komponente bildete die Subskala AM-Energie, die zweite Komponente die Subskala AM-Info. Die deskriptiven Statistiken und Reliabilitäten sind in Tabelle 69 aufgeführt.

Tabelle 69: Deskriptive Statistiken und Reliabilitätswerte für die Skala anxiety motivation

	N	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Energie	449	1.00	5.00	2.67	.83	.69	.20	-.34	.81
Information	456	1.00	5.00	3.33	.81	.66	-.40	.28	.63

AM-Energie: 4 Items; AM-Info: 3 Items; Skaleninterkorrelation: $r = .44^{**}$

Beide Skalen wiesen keine starken Boden- oder Deckeneffekte auf. Bei beiden Skalen wurde die gesamte Skalenbreite ausgeschöpft, wobei im Mittel höhere Werte bei AM-Info zu finden waren als bei AM-Energie. Die Reliabilität war für AM-Energie gut, für AM-Info ausreichend. Die relativ hohe Interkorrelation der Subskalen von $r = .44$ ($p < .001$) ist plausibel, da alle Items eine „aktivierende“ Wirkung von Angst umschreiben. Bei AM-Energie fand sich ein marginal signifikanter Unterschied zwischen Männern ($M = 2.83$, $SD = .96$) und Frauen ($M = 2.63$, $SD = .78$), $t\,(144.02) = 1.87$, $p = .064$, bei AM-Info lag kein bedeutsamer Geschlechtsunterschied vor (Männer: $M = 3.29$, $SD = .87$; Frauen: $M = 3.34$, $SD = .79$), $t\,(420) = -.61$, $p = .540$.

Ankündigung des Intelligenztests

Die Ankündigung des Tests (des GkKT, siehe entsprechende Ausführungen in diesem Abschnitt) diente einerseits dazu, einen kontrollierten und standardisierten Testablauf zu gewährleisten. Zum anderen sollte durch die Ankündigung der evaluative Charakter der Erhebung gesteigert werden (vgl. auch die Testankündigungen in Studie 1 und 2). Zu diesem Zweck wurde der Inhaltsbereich des Tests genannt („Aufgaben zum schlussfolgernden Denken“) und darüber hinaus mitgeteilt, dass die Pbn auf Wunsch eine Rückmeldung zu ihrem Ergebnis erhalten können, welche ihnen einen Vergleich mit der Leistung anderer Teilnehmer ermögliche. Auch auf die Zeitbegrenzung beim Test wurde hingewiesen.

Reappraisal-Manipulation

Die Formulierung der reappraisal-Manipulation wurde aus bereits publizierten Studien abgeleitet (Beltzer et al., 2014; Jamieson et al., 2010; Jamieson et al., 2013; Strack et al., 2014). Dabei wurden insgesamt vier inhaltliche Elemente extrahiert:

1. Die Aussage, dass Aufregung bzw. Angst nicht schädlich ist.

2. Die Aussage, dass Aufregung bzw. Angst hilfreich (funktional) ist.
3. Die Ermutigung, die eigene Aufregung bzw. Angst funktional zu interpretieren.
4. Die Information, dass die vorher genannten Aussagen wissenschaftlich belegt sind.

Eine besondere Herausforderung war die Begriffswahl. Es wurde bereits darauf eingegangen, dass reappraisal-Manipulationen (in der Formulierung) auf unterschiedliche Prozesse anspielen, sprich die Uminterpretation körperlicher Erregungsprozesse und / oder von tatsächlichem Angsterleben (siehe Abschnitt 1.3.2). Es finden sich also die Begriffe „arousal" und „anxiety". Der Begriff „Erregung" als Übersetzung für „arousal" schien unpassend, da der deutsche Begriff nicht denselben Bedeutungshof hat wie der englische (die gängigsten Assoziationen von „Erregung" dürften sexuelle Erregung sowie Verärgerung sein). Passender schien der Begriff der „Aufgeregtheit" oder „Aufregung" zu sein, da dieser stark mit Prüfungssituationen verbunden ist (vgl. auch „Lampenfieber"). Der Begriff „ängstlich" („anxious") schien ebenfalls zu extrem, da die Untersuchung keine high-stakes Testung war. Passender schien der Begriff „nervös". Diese Überlegungen führten zu der in Tabelle 70 dargestellten Formulierung. Der Kontrollgruppe wurde ein Textabschnitt vorgelegt, der lediglich eine Wiederholung bereits gegebener Informationen zum Testablauf beinhaltete.

Tabelle 70: Instruktionstexte in der Kontrollgruppe (Bedingung A) sowie Treatmentgruppe (Bedingung B)

	Bedingung	Instruktionstext
A	Kontroll-bedingung	Dieser und die weiteren Blöcke haben jeweils eine kurze Erläuterung und danach folgen Aufgaben. Wir werden Ihnen jeweils mitteilen, wenn die Zeit um ist. Bitte arbeiten Sie dann nicht mehr weiter.
B	Reappraisal-Bedingung = Treatment	**Bitte beachten Sie:** die Forschung zeigt: es schadet der Leistung nicht, wenn man bei einem Test etwas aufgeregt ist. Sollten Sie also während des Tests etwas nervös sein: denken Sie daran, dass das für Ihre Leistungsfähigkeit positiv sein kann.

Anm.: beide Textabschnitte standen direkt über der Instruktion des ersten Aufgabenblocks des GkKT

Gießener kognitiver Kompetenztest

Der Gießener kognitive Kompetenztest GkKT (Ulfert et al., 2014a) wurde als Instrument zur Erfassung allgemeiner Intelligenz für den Einsatz im Self-Assessment der JLU Gießen („Ready for Justus") entwickelt. Der GkKT basiert theoretisch auf dem Berliner Intelligenzstrukturmodell, welches dem Berliner Intelligenzstruktur-Test zugrunde liegt (Jäger, Süß & Beauducel, 1997). Der Annahme der mulit-trait Determination von Leistungen folgend umfasst der GkKT Aufgaben zur Verarbeitungskapazität mit verbalem, numerischem und figuralem Inhalt. In der Version 1.0 des GkKT ist allerdings keine differenzierte Auswertung der Daten nach Inhaltsbereichen (Zellebene im BIS-Modell) vorgesehen, es wird lediglich ein Wert für Verarbeitungskapazität ermittelt. Die drei weiteren Operationen Bearbeitungsgeschwindigkeit, Merkfähigkeit und Einfallsreichtum

sind im GkKT nicht abgebildet. Im Sinne der Systematik von Intelligenztests nach Schmidt-Atzert und Amelang (2012) lässt sich der GkKT einordnen als Test zur Erfassung der allgemeinen Intelligenz (g). Dabei handelt es sich um einen „breiten" Intelligenztest, da mehrere Intelligenzkomponenten (hier die drei Inhaltsbereiche gemäß des BIS-Modells) erfasst werden. Jedoch ist der GkKT kein Strukturtest, da die verschiedenen Komponenten nicht separat interpretiert werden (Schmidt-Atzert & Amelang, 2012). Der GkKT hat als konzeptuelles Vorbild den Wonderlic Personnel Test WPT (Wonderlic Inc., 1996), welcher die Erfassung allgemeiner Intelligenz (g) in einer ökonomischen Form ermöglicht (50 Items, 12 Minuten Bearbeitungszeit). In Anlehnung an den WPT wurde eine Heterogenität der Aufgabentypen im GkKT angestrebt. Während im WPT Items verschiedenen Typs vollständig durchmischt sind, wurde im GkKT eine Mischung auf Basis von Aufgabenblöcken realisiert. So enthält der GkKT elf Aufgabenblöcke, die jeweils aus einer kleinen Anzahl typgleicher Items bestehen. Die Bündelung der typgleichen Items in einem Aufgabenblock wurde vorgenommen, da bei einigen Itemtypen eine separate Instruktion für das Verständnis des Aufgabentyps unabdingbar ist. Entsprechend dieses Aufbaus wurde jeder Aufgabenblock mit einer eigenen Zeitbegrenzung versehen. Die komplette Bearbeitungszeit des GkKT betrug 18:50 Minuten. Allen Berechnungen mit dem GkKT ging eine sorgfältige Analyse auffälliger Fälle voraus, wobei Pbn exkludiert wurden, die den GkKT nicht oder nur ausgesprochen lückenhaft bearbeiteten. Fälle, die mehr als 5 der 11 Aufgabenblöcke nicht bearbeitet hatten, wurden ausgeschlossen. Tabelle 71 zeigt die deskriptiven Statistiken sowie die Reliabilität des GkKT.

Tabelle 71: Deskriptive Statistiken und Reliabilitätswert des GkKT

	N	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
GkKT	496	6	41	23.02	6.31	39.78	-.12	-.10	.81

45 Items

Die Verteilung der Gesamtpunktzahl ist in Abbildung 20 zu sehen. Die Reliabilität des GkKT war gut. Analog zu den Ausführungen zum GkKT-K (siehe Abschnitt 4.1.5) soll kurz auf die Validität des GkKT eingegangen werden. Für den GkKT wurde in vorherigen Untersuchungen eine hohe Korrelation mit dem WPT von r_p = .78 (p < .001) (*N* = 170; Studierende der Physik und Wirtschaftswissenschaften) ermittelt. Mit den in der vorliegenden Untersuchungen administrierten Verfahren BEFKI GC-K (Schipolowski et al., 2013) sowie MWT-B (Lehrl, 2005) ergaben sich

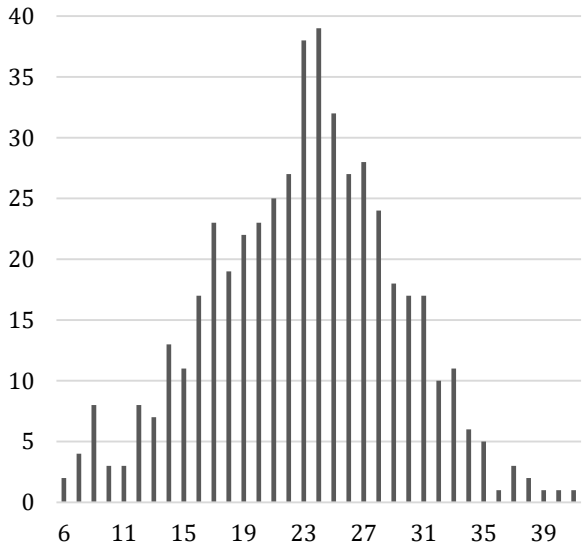

Abbildung 20: Histogramm der Gesamtpunktzahl beim GkKT

moderate Korrelationen von r = .44 bzw. .37 (N = 327 bzw. 167, jeweils p < .001). Diese Befunde sprechen für die Konstruktvalidität des GkKT, wobei der höhere Zusammenhang mit dem WPT plausibel ist, da dieser allgemeine Intelligenz erfasst, während der BEFKI GC-K und der MWT-B kristalline Intelligenz abbilden. Im GkKT hingegen sind eher fluide Intelligenzleistungen repräsentiert. Mit der Abiturnote zeigte sich ein schwacher, negativer Zusammenhang von r = -.11 (p = .025).

In der Gesamtleistung zeigte sich kein Geschlechtsunterschied zwischen Männern (M = 24.04, SD = 5.90) und Frauen (M = 23.13, SD = 6.41), t (457) = 1.34, p = .180.

Akzeptanz

Die Akzeptanz des GkKT wurde wie in Studie 1 und 2 mit dem Akzept!-L von Kersting (2008) erfasst. Die deskriptiven Statistiken und Reliabilitäten sind in Tabelle 72 aufgeführt.

Tabelle 72: Deskriptive Statistiken und Reliabilitätswerte für die Skala Akzeptanz

	N	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Kontrollierbarkeit	420	1-6	1.25	6.00	4.21	1.05	1.10	-.24	-.60	.81
Messqualität	428	1-6	1.00	5.50	3.25	.89	.79	-.08	-.33	.69
Augenschein-validität	416	1-6	1.00	5.50	2.38	.89	.79	.41	-.18	.72
Belastungsfreiheit	429	1-6	1.00	6.00	3.68	1.06	1.11	.00	-.41	.79
Note Verfahren	462	1-6	1	6	3.49	1.04	1.07	.20	-.50	-
Subjektive Leistung	459	1-6	1	6	3.41	1.04	1.08	.46	-.18	-

Itemahl: 4 Items, Note & subjektive Leistung jeweils mit einem Item erfasst

Die Reliabilitäten waren ausreichend bis gut. Die Kontrollierbarkeit des GkKT wurde relativ hoch eingeschätzt, das Instruktionsverständnis seitens der Pbn kann also angenommen werden.

Tabelle 73: Skaleninterkorrelationen bezüglich Akzeptanz sowie Korrelationen mit der Note für das Verfahren und subjektiver Leistung

	KB	MQ	AV	BF	Note
Kontrollierbarkeit					
Messqualität	.37**				
Augenscheinvalidität	.14**	.41**			
Belastungsfreiheit	.31**	-.01	.11*		
Note Verfahren	-.42**	-.46**	-.34**	-.32**	
Subjektive Leistung	-.32**	-.20**	-.13**	-.35**	.45**

N = 367-445; Note & subjektive Leistung jeweils mit einem Item erfasst (1 = „sehr gut“, 6 = „ungenügend“)

Zwischen den Skalen lagen überwiegend moderate Zusammenhänge vor (siehe Tabelle 73). Von den Skalen hing Belastungsfreiheit am stärksten mit der subjektiven Leistung zusammen mit r =

-.35 (p < .001). Die eigene Leistung wurde umso schlechter eingeschätzt, je belastender der Test empfunden wurde. Die subjektive Leistung hing überdies deutlich mit der Note für das Verfahren zusammen zu r = .45 (p < .001). Mit der Note für das Verfahren hing von den Skalen die Messqualität am stärksten zusammen, r = -.46 (p < .001). Das bedeutet, dass der Test umso besser bewertet wurde, je höher dessen Messqualität erlebt wurde.

State Testangst & state Anxiety Motivation

Ebenfalls nach dem GkKT wurde die während dieses Tests erlebte Testangst erfasst. Im Unterschied zu Studie 1 wurden hierfür nur die Items des STAI-SKD (Englert et al., 2011) zur Erfassung von Besorgtheit und Aufgeregtheit eingesetzt. Wie bei der state-TA post in Studie 1 wurden die Itemformulierungen ins Imperfekt gesetzt, das Antwortformat jedoch beibehalten. Mit einem unmittelbar danach platzierten, weiteren Item sollte erfasst werden, inwiefern die während des Tests erlebte Angst als hilfreich empfunden wurde (state-anxiety motivation bzw. state-AM). Um die Konsistenz zur reappraisal-Manipulation zu halten, wurde folgende Formulierung gewählt: „Hat Ihnen diese Aufregung beim Lösen der Aufgaben geholfen?“. Es wurde dasselbe Antwortformat wie bei den Items zur Erfassung der state-TA vorgegeben (1 = „überhaupt nicht“; 4 = „sehr“). Tabelle 74 zeigt die deskriptiven Statistiken der state-TA sowie der state-AM.

Tabelle 74: Deskriptive Statistiken und Reliabilitätswerte für die Skala state-Testangst und das Item zu state-anxiety motivation

	N	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
State-TA BE	455	1-4	1.00	4.00	1.60	.71	.51	1.11	.61	.74
State-TA AU	452	1-4	1.00	4.00	1.96	.74	.55	.49	-.49	.79
State-AM	465	1-4	1.00	4.00	1.54	.78	.61	1.16	.19	-

state-TA Besorgtheit bzw. BE: 2 Items; state-TA Aufgeregtheit bzw. AU: 3 Items
Skaleinterkorrelationen state-TA: .62**
Md: state-TA Besorgtheit = 1.50; state-AM = 1.00

Insgesamt war bei der state-TA ein Bodeneffekt zu beobachten. Die Reliabilitäten der Subskalen waren gut. Für alle fünf Items der state-TA (Besorgtheit und Aufgeregtheit) ergab sich ein Mittelwert von 1.81 (SD = .66). Wenngleich auch mit diesem Wert noch ein Bodeneffekt zu verzeichnen ist, lag er über dem Wert, der von Englert et al. (2011) nach einer bedrohlichen Aufgabenankündigung (M = 1.73, SD = .47) ermittelt wurde. Ein gewisses Maß an Bewertungsstress dürfte bei der Bearbeitung des GkKT also vorhanden gewesen sein. Vor dem Hintergrund einer „künstlichen“ Prüfungssituation ist die Ausprägung der state-TA also erwartungsgemäß. Ein deutlicher Bodeneffekt war bei der state-AM feststellbar.

IPIP-Fragebogen zum Fünf-Faktoren-Modell

Die Persönlichkeitseigenschaften nach dem Fünf-Faktoren-Modell nach Costa und McCrae (1992) (kurz: FFM) wurden mit einem Fragebogen von Michaelis et al. (2014) erhoben und dienten maßgeblich der Konstruktvalidierung der anxiety motivation-Skalen. Der Fragebogen basiert auf einem englischen 300-Item Fragebogen zur Erfassung des FFM auf Facettenebene aus dem International Personality Item Pool (IPIP; www.ipip.ori.org) (Goldberg et al., 2006), von dem eine deutsche Übersetzung vorlag (Treiber, 2013). Im Rahmen des Self-Assessment-Projekts an der JLU Gießen wurde eine Kurzversion zur Erfassung des FFM auf Dimensionsebene nach dem Vorbild des 50-Item-IPIP-Fragebogens zum FFM erstellt. Dabei wurde sowohl auf Itemübersetzungen von Treiber (2013) zurückgegriffen als auch Items vom Projektteam übersetzt. Die Fragebogenentwicklung umfasste mehrere Erhebungswellen. Auf Basis von Itemselektionen, Faktorenanalysen und mehrfachen Revisionen resultierte ein Fragebogen aus ingesamt 29 Items, welcher auch in Studie 3 eingesetzt wurde (kurz: IPIP-FFM). Der Fragebogen erfasst Extraversion mit 6 Items (z. B. „Ich bin geschickt in sozialen Interaktionen."), Offenheit für Erfahrungen mit 5 Items (z. B. „Ungewöhnlichen Ideen stehe ich offen gegenüber."), Verträglichkeit mit 6 Items (z. B. „Ich lege für jeden ein gutes Wort ein."), Gewissenhaftigkeit mit 6 Items (z. B. „Aufgaben bearbeite ich bis ins letzte Detail.") und Neurotizismus mit 6 Items (z. B. „Ich lasse mich leicht stressen."). Die Items sind jeweils mit einem fünfstufigen Antwortformat versehen (1 = „trifft überhaupt nicht zu"; 5 = „trifft vollständig zu"). Tabelle 75 gibt die deskriptiven Statistiken sowie die internen Konsistenzen der fünf Skalen wieder.

Tabelle 75: Deskriptive Statistiken und Reliabilitätswerte der Skala Persönlichkeit nach dem FFM

	N	*Skala*	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
Extraversion	432	1-5	1.17	5.00	3.22	.60	.36	-.21	.51	.76
Offenheit	443	1-5	1.40	5.00	3.17	.62	.38	.02	-.17	.61
Verträglichkeit	429	1-5	2.33	5.00	3.97	.50	.25	-.57	.46	.68
Gewissenhaftigkeit	439	1-5	1.50	5.00	3.16	.63	.40	-.03	-.15	.77
Neurotizismus	440	1-5	1.00	5.00	2.97	.74	.55	.23	-.04	.79

Extraversion: 6 Items; Offenheit: 5 Items; Verträglichkeit: 6 Items; Gewissenhaftigkeit: 6 Items; Neurotizismus: 6 Items
Md: Verträglichkeit = 4.00

Die Reliabilitäten der Skalen waren gut bis ausreichend. Bei Extraversion (Frauen: M = 3.20, SD = .57; Männer: M = 3.21, SD = .66) zeigte sich kein Geschlechtsunterschied, t (401) = .04, p = .966, ebenso bei Offenheit (Frauen: M = 3.16, SD = .59; Männer: M = 3.15, SD = .67), t (415) = -.10, p = .917. Bei Verträglichkeit (Frauen: M = 4.06, SD = .45; Männer: M = 3.74, SD = .55), t (405) = -5.76, p < .001, Gewissenhaftigkeit (Frauen: M = 3.21, SD = .64; Männer: M = 3.01, SD = .60), t (412) = -2.78, p = .006 und Neurotizismus (Frauen: M = 3.10, SD = .73; Männer: M = 2.65, SD = .69), t (410) = -5.46, p < .001, zeigten Frauen eine höhere Ausprägung.

Allgemeine Selbstwirksamkeit

Die allgemeine Selbstwirksamkeitserwartung (kurz: Selbstwirksamkeit bzw. general self-efficacy, kurz: GSE) wurde mit dem Fragebogen von Jerusalem und Schwarzer (1999) erfasst. Dieser besteht aus insgesamt 10 Items (z. B. „In unerwarteten Situationen weiß ich immer, wie ich mich verhalten soll."). Statt des vorgesehenen vierstufigen wurde das fünfstufige Antwortformat der IPIP-FFM-Items vorgegeben, da die Items zur Selbstwirksamkeit gemeinsam mit den IPIP-FFM-Items in einem Fragenblock platziert wurden und ein Wechsel des Antwortformats vermieden werden sollte. Tabelle 76 gibt die deskriptiven Statistiken sowie die Reliabilität der Skala wieder.

Tabelle 76: Deskriptive Statistiken und Reliabilitätswert für die Skala Selbstwirksamkeit

	N	Skala	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*	*α*
GSE	414	10-50	15	50	33.37	5.41	29.22	-.13	.72	.86

Skala: 10 Items; Summenscore mit Itemwerten von 1 bis 5

Die Reliabilität der Skala war gut. Für die ursprüngliche Skala mit einem vierstufigen Antwortformat (Summenwert von 10 bis 40) geben Hinz, Schumacher, Albani, Schmid und Brähler (2006) an, dass in einem Großteil der Untersuchungen der mittlere Summenwert bei etwa 29 liegt. Bei Umrechnung in die hier vorgenommene Kodierung von 1 bis 5 ergab sich ein Referenzwert von 36.25[46], der etwas über dem in der hier vorliegenden Stichprobe ermittelten Wert von 33.37 liegt. Frauen (M = 32.78, SD = 5.21) berichteten – in Übereinstimmung mit den Erkenntnissen aus anderen Studien – niedrigere Werte als Männer (M = 34.60, SD = 5.69), t (388) = 2.91, p = .004.

Variablen zum Erleben des Studiums

Mit insgesamt sieben Items wurden unterschiedliche Aspekte zum Erleben des eigenen Studiums erfasst. Drei Items zur Erfassung der Zufriedenheit mit den Studieninhalten wurden einem Fragebogen von Hiemisch, Westermann und Michael (2005) entnommen (z. B. „Ich habe richtig Freude an dem, was ich studiere."). Vier weitere Items wurden vom Verfasser dieser Arbeit formuliert. Je ein Item befasste sich mit dem bisher gefundenen Anschluss an Kommilitoninnen und Kommilitonen, der Erfolgszuversicht bezüglich der kommenden Prüfungen („Ich komme gut mit dem Stoff mit und bin optimistisch, dass ich die Prüfungen schaffe.", abgekürzt als „Erfolgszuversicht"), Gedanken an einen Fachwechsel sowie Gedanken an einen Studienabbruch („Ich habe mir schon einmal Gedanken darüber gemacht, das Studium ganz abzubrechen.", abgekürzt als „Abbruchtendenz"). Diese sieben Items hatten ein fünfstufiges Antwortformat (1 = „trifft nicht zu"; 5 = „trifft

[46] 29 / 4 = 7.52; 7.25 x 5 = 36.25

vollständig zu“). Da die Pbn sich im ersten Semester befanden und somit keine Studiennoten verfügbar waren, wurden letztere Variablen auch als distale Maße für Studienerfolg erhoben. Da im Fachbereich 09 ein früher Fachwechsel innerhalb des Fachbereichs möglich ist und auch als sinnvoll erachtet wird, sind Gedanken an einen Fachwechsel qualitativ zu unterscheiden von Abbruchgedanken. Beide Variablen wurden also getrennt behandelt. Die deskriptiven Statistiken sind in Tabelle 77 zu finden.

Tabelle 77: Deskriptive Statistiken der Items zum Erleben des Studiums

	N	*Min*	*Max*	*M*	*SD*	*Var*	*Sch*	*Kurt*
Anschluss an Kommilitonen	467	1	5	4.10	.87	.75	-.85	.56
Erfolgszuversicht	462	1	5	3.13	.91	.83	-.16	-.08
Gedanken an Fachwechsel	462	1	5	2.09	1.28	1.63	.96	-.26
Abbruchtendenz	465	1	5	1.72	1.14	1.30	1.49	1.15
Zufriedenheit mit Inhalten[1]	447	1	5	3.76	.77	.59	-.71	.77

Skala von 1 = „trifft nicht zu“ bis 5 = „trifft vollständig zu“; [1] α = .85
Md: Anschluss an Kommilitonen = 4.00; Gedanken an Fachwechsel = 2.00; Abbruchtendenz = 1.00; Zufriedenheit mit Inhalten = 4.00

Die Reliabilität der Skala „Zufriedenheit mit Studieninhalten“ war gut. Insgesamt waren die Probanden relativ zufrieden mit ihrem Studium: die Zufriedenheit mit Studieninhalten lag über dem Skalenmittelpunkt, Gedanken an Fachwechsel und Abbruch deutlich unter diesem.

6.1.6 Statistische Verfahren

Die Wahl der Verfahren und die Voraussetzungsprüfungen entsprechen den Vorgehensweisen in Studie 1 und 2 (siehe auch Abschnitt 3). Dies betrifft die Normalverteilungsprüfungen und die Voraussetzungsprüfungen der Moderationsanalysen (siehe Anhang F). Die Moderationsanalysen wurden mit der SPSS-Macro PROCESS von Hayes (2013) berechnet. Die vorhandene Berechnung von Modellen mit zwei Moderatoren wurde ebenfalls mit PROCESS durchgeführt (analog wurde gerechnet in Studie 1, Hypothese 2b-d). Dieses Vorgehen unterscheidet sich in zwei Punkten von der einfachen Moderation des Effekts von *X* auf *Y* durch *M*. Erstens wird zum ersten Moderator *M* ein zweiter Moderator *W* hinzugefügt (im Falle von Hypothese 3a sind das AM-Energie und AM-Info). Zweitens wird eine Interaktion zwischen *M* und *W* eingefügt, wodurch eine moderierte Moderation vorliegt. Dies war notwendig, da bei Hypothese 3a nicht klar war, ob zusätzlich eine Interaktion zwischen AM-Energie und AM-Info vorliegt. Abbildung 21 zeigt das konzeptuelle und statistische Modell einer moderierten Moderation mit zwei Moderatoren (Model 3; Hayes, 2013). Verglichen mit einer Moderation mit zwei Moderatoren ohne Interaktion der Moderatoren kommen zu den Interaktionstermen *XM* und *XW* noch der Interaktionsterm *MW* sowie die Dreifachinteraktion *XWM*.

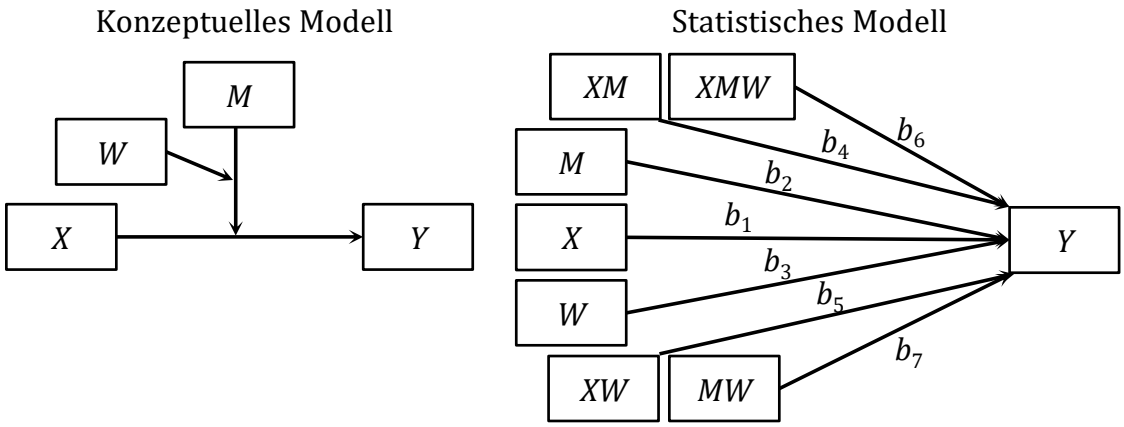

Abbildung 21: Konzeptuelles und statistisches Modell einer moderierten Moderation (Hayes, 2013, S. 444)

6.2 Ergebnisse

6.2.1 Vorbereitende Analysen

Die vorbereitenden Analysen für Studie 3 konzentrierten sich auf die Frage, ob es zwischen den verschiedenen Untersuchungsbedingungen Unterschiede in demographischen Variablen oder weiteren erfassten Merkmalen gegeben hat. Dies diente auch dazu, den Erfolg der Randomisierung zu prüfen. Die Analysen wurden dabei gegliedert in den Vergleich der Bedingungen A und B sowie in den Vergleich der Pbn in den beiden Hörsälen. Auf etwaige Geschlechtsunterschiede in den betrachteten Variablen wurde bereits im Methodenteil eingegangen.

6.2.1.1 Vergleich der Gruppen in den verschiedenen Bedingungen

In der Stichprobe befanden sich insgesamt mehr Frauen als Männer (siehe Abschnitt 6.1.3). Der relative Anteil an Männern unterschied sich jedoch in Bedingung A nicht von B, χ^2 (1, N = 460) = .02, p = .881. Bezüglich des Alters lag zwischen Bedingung A (M = 20.84, SD = 2.63) und B (M = 21.00, SD = 3.53) kein signifikanter Unterschied vor, t (460) = -.53, p = .594. Bezüglich der Abiturnote zeigte sich zwischen Bedingung A (M = 2.30, SD = .52) und Bedingung B (M = 2.27, SD = .47) kein signifikanter Unterschied, t (457) = .53, p = .594. Tabelle 78 zeigt die Mittelwertsunterschiede in den erhobenen dispositionellen Variablen. Es fanden sich keine Gruppenunterschiede. Auch Mittelwertsvergleiche zwischen den drei Fragebogenversionen A_{BEFKI}, $A_{MWT\text{-}B}$ und B ergaben keine signifikanten Unterschiede in den hier betrachteten Variablen.

Tabelle 78: Unterschiede zwischen den Gruppen in den Bedingungen A und B bezüglich dispositioneller Variablen (geprüft via t-Tests für unabhängige Stichproben)

		Bedingung A		Bedingung B				
	N	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Trait-TÄ								
Besorgtheit	470	13.12	3.42	13.44	3.88	-.88	.381	.09
Aufgeregtheit	468	7.42	2.64	7.26	2.67	.62	.539	-.06
Interferenz	484	6.20	2.07	6.12	1.91	.43	.667	-.04
Mangel an Zuversicht	472	7.44	1.84	7.40	1.98	.21	.835	-.02
Gesamt	429	34.16	7.40	34.17	7.75	-.02	.987	.00
anxiety motivation								
Energie	449	2.70	.83	2.61	.84	1.07	.284	-.11
Information	456	3.36	.80	3.25	.83	1.33	.186	-.14
Selbstwirksamkeit	414	33.46	5.41	33.20	5.40	.48	.635	-.05
Extraversion	432	3.22	.59	3.22	.62	-.12	.902	.00
Offenheit	443	3.18	.62	3.17	.63	.16	.872	-.02
Verträglichkeit	429	3.98	.50	3.96	.50	.33	.739	-.04
Gewissenhaftigkeit	439	3.18	.63	3.11	.64	.99	.325	-.11
Neurotizismus	440	3.01	.72	2.91	.79	1.33	.185	-.13

Es lagen keine überzufälligen Unterschiede in der relativen Häufigkeit der Studiengangskategorien pro Bedingung bzw. Fragebogenversion vor, χ^2 (4, N = 474) = .75, p = .946. Entsprechend waren die Häufigkeiten der jeweiligen Studiengangskategorie in den drei Fragebogenversionen A_{BEFKI}, $A_{MWT\text{-}B}$ und B nahezu identisch (z. B. Agrarwissenschaften jeweils N = 31, 33 und 31).

6.2.1.2 Vergleich der Gruppen in den verschiedenen Untersuchungsräumen

Die Untersuchung fand parallel in einem großen Hörsaal (HS 01, N = 419) und einem kleinen Hörsaal (HS 02, N = 85) statt. Da die Dozentin nur im HS 01 physisch anwesend war (HS 02 ist im Vorlesungsbetrieb durch die parallel ablaufende Präsentation und Tonübertragung zugeschaltet), waren dispositionelle Unterschiede zwischen den Pbn denkbar. Der Anteil an Männern unterschied sich nicht zwischen den beiden Räumen, χ^2 (1, N = 460) = 2.26, p = .133. Probanden in HS 01 (M = 20.78, SD = 2.95) waren etwas jünger als jene in HS 02 (M = 21.47, SD = 2.92), t (460) = -1.87, p = .062, wobei der Unterschied knapp nicht signifikant ist. In der Abiturnote zeigte sich kein signifikanter Unterschied zwischen HS 01 (M = 2.28, SD = .49) und HS 02 (M = 2.35, SD = .55), t (457) = -1.25, p = .211. Tabelle 79 zeigt die Gruppenunterschiede in dispositionellen Variablen.

Tabelle 79: Unterschiede zwischen den Gruppen in den beiden Untersuchungsräumen bezüglich dispositioneller Variablen (geprüft via t-Tests für unabhängige Stichproben)

		HS 01		HS 02				
	N	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
Trait-TÄ								
Besorgtheit	470	13.20	3.60	13.32	3.50	-.27	.787	.03
Aufgeregtheit	468	7.42	2.65	7.15	2.67	.82	.411	-.10
Interferenz	484	6.18	2.00	6.16	2.08	.07	.941	-.01
Mangel an Zuversicht	472	7.49	1.89	7.12	1.84	1.61	.108	-.20
Gesamt	429	34.28	7.50	33.65	7.56	.67	.505	-.08
anxiety motivation								
Energie	449	2.67	.83	2.67	.84	-.05	.958	.00
Information	456	3.34	.82	3.27	.76	.68	.500	-.09
Selbstwirksamkeit	414	33.15	5.41	34.55	5.26	-1.96	.051	.26
Extraversion	432	3.20	.61	3.30	.53	-1.30	.194	.17
Offenheit	443	3.16	.63	3.23	.57	-.85	.397	.11
Verträglichkeit	429	3.99	.50	3.91	.50	1.19	.233	-.16
Gewissenhaftigkeit	439	3.16	.63	3.14	.64	.27	.784	-.03
Neurotizismus	440	2.98	.75	2.96	.70	.15	.884	-.03

Es zeigten sich keine Unterschiede zwischen den Räumen, mit Ausnahme eines marginal signifikanten Unterschieds in der Selbstwirksamkeit. Die Wahl des Raums schien also kein Ausdruck der Persönlichkeit zu sein. Wahrscheinlich waren „alltagspraktische“ Argumente maßgeblich für die

Raumwahl, wie z. B. Gruppenbildung, Zeit des Eintreffens und insbesondere Zufall (Anfahrtssituation, Parkplatzsuche, etc.).

Die vorbereitenden Analysen weisen darauf hin, dass es keinerlei Unterschiede in demographischen Merkmalen sowie in Persönlichkeitseigenschaften der Probanden in den zwei Bedingungen sowie den beiden Untersuchungsräumen gab. Die einzigen Subgruppenunterschiede fanden sich beim Vergleich der Geschlechter, wobei diese Unterschiede, insbesondere bei der Testängstlichkeit, nicht überraschend sind und bekannte empirische Trends widerspiegeln. Alle Ergebnisse der Voranalysen sprechen also für eine erfolgreiche Randomisierung.

6.2.2 Ergebnisse der explorativen Analysen und Hypothesenprüfung

6.2.2.1 Explorative Analyse und Hypothese 1

Für die explorative Analyse wurden bivariate Korrelationen berechnet zwischen den beiden Skalen von anxiety motivation (AM-Energie sowie AM-Info) und den weiteren erhobenen Eigenschaften. Zuerst wurden hierzu Zusammenhänge zur Testängstlichkeit betrachtet. Da AM-Energie und AM-Info „Facetten" von anxiety motivation darstellen, wurde auch die Testängstlichkeit in diesem Fall in ihre Facetten untergliedert. Die Ergebnisse sind Tabelle 80 zu entnehmen.

Tabelle 80: Bivariate Korrelationen von anxiety motivation Energie und Information mit Testängstlichkeit

	Testängstlichkeit				
	Besorgtheit	Aufgeregtheit	Interferenz	Mangel an Zuversicht	Gesamt
AM-Energie	.01	-.02	-.10*	-.19**	-.09
AM-Information	.17**	.11*	.02	.00	.13*

N = 396-444

Insgesamt zeigten sich eher schwache Zusammenhänge zwischen AM und Testängstlichkeit. Dabei sind jedoch für die Skalen von AM Unterschiede zu erkennen. So korrelierte AM-Energie negativ mit Interferenz (r = -.10, p = .032) und Mangel an Zuversicht (r = -.19, p < .001). Demgegenüber korrelierte AM-Info positiv mit Besorgtheit (r = .17, p = .001) und Aufgeregtheit (r = .11, p = .020). Als nächstes wurden die Zusammenhänge zu den Persönlichkeitseigenschaften nach dem FFM und zur Selbstwirksamkeit berechnet. Die Ergebnisse sind Tabelle 81 zu entnehmen.

Tabelle 81: Bivariate Korrelationen von anxiety motivation Energie und Information mit den Persönlichkeitseigenschaften nach dem FFM und Selbstwirksamkeit

	Weitere traits					
	Extra-version	Offenheit	Verträg-lichkeit	Gewissen-haftigkeit	Neuroti-zismus	Selbstwirk-samkeit
AM-Energie	.07	.02	-.04	.21**	-.08	.14**
AM-Information	.03	.07	-.01	.07	.07	.04

N = 378-407

Es zeigten sich insgesamt keine oder aber schwache Zusammenhänge zwischen AM und Persönlichkeit sowie Selbstwirksamkeit. AM-Information wies zu keiner der sechs betrachteten Dispositionen substanzielle Korrelation auf. AM-Energie hingegen korrelierte signifikant positiv mit Gewissenhaftigkeit (r = .21, p < .001) und Selbstwirksamkeit (r = .14, p = .008).

Hypothese 1 ging der Vermutung nach, dass AM-Energie und -Information die state-AM besser vorhersagen können als die Facetten der Testängstlichkeit. Hierzu wurden bivariate Korrelationen berechnet, die in Tabelle 82 zusammengefasst sind.

Tabelle 82: Bivariate Korrelatoinen von der state-anxiety motivation mit anxiety motivation (trait) und Testängstlichkeit

	anxiety motivation		Testängstlichkeit				
	Energie	Informa-tion	Besorgt-heit	Aufge-regtheit	Interfe-renz	Mangel an Zuversicht	Gesamt
State-AM	.34**	.13**	-.03	.03	-.06	-.07	-.04

N = 404-452

Es zeigten sich keine signifikanten Zusammenhänge zwischen state-AM und Testängstlichkeit. Demgegenüber korrelierte state-AM schwach positiv mit AM-Information (r = .13, p = .008) sowie substanziell positiv mit AM-Energie (r = .34, p < .001) (zum Korrelationsvergleich siehe Anhang H). Hypothese 1 wurde somit bestätigt.

6.2.2.2 Hypothese 2a und 2b

Zur Prüfung von Hypothese 2a wurden Mittelwertsvergleiche in den relevanten abhängigen Variablen via t-Tests für unabhängige Stichproben vorgenommen. Diese sind in Tabelle 83 aufgeführt.

Tabelle 83: Vergleich der Kontrollgruppe (Bedingung A) sowie der Treatmentgruppe (Bedingung B) bezüglich der abhängigen Variablen

		Bedingung A		Bedingung B				
	N	*M*	*SD*	*M*	*SD*	*t*	*p*	*d*
State-AM	464	1.57	.80	1.48	.73	1.15	.250	-.12
Leistung GkKT	496	22.93	6.39	23.20	6.16	-.44	.657	.04
State-Testangst	455	1.59	.72	1.61	.71	-.37	.710	.03
Belastungsfreiheit	429	3.73	1.05	3.56	1.06	1.62	.107	-.16
Subjektive Leistung	459	3.37	1.04	3.48	1.04	-1.03	.303	.11

State-Testangst = Facette Besorgtheit

Es traten zwischen Kontrollgruppe und Treatmentgruppe keinerlei signifikante Unterschiede in den abhängigen Variablen auf. Während dieses Ergebnis für die state-TA erwartungsgemäß ist, widersprechen die fehlenden Unterschiede in der state-AM, der Leistung, der Belastungsfreiheit und der subjektiven Leistung der Vorannahme. Hypothese 2a wurde daher abgelehnt.

Zur Prüfung von Hypothese 2b wurde die Korrelation zwischen der objektiven Leistung im GkKT und der state-TA berechnet. Für die Kontrollgruppe resultierte ein signifikant negativer Zusammenhang von r = -.21 (p < .001) und für die Treatmentgruppe ein nicht signifikanter Zusammenhang von r = -.09 (p = .273). Dieses Ergebnis entspricht der Erwartung, wobei jedoch der Unterschied in der Höhe der Korrelationen nicht signifikant war (Fisher's z = -1.22, p = .222). Der negative Zusammenhang in der Kontrollgruppe und der abgeschwächte (negative) Zusammenhang in der Treatmentgruppe sind in Abbildung 22 sichtbar. Hypothese 2b ist somit anzunehmen.

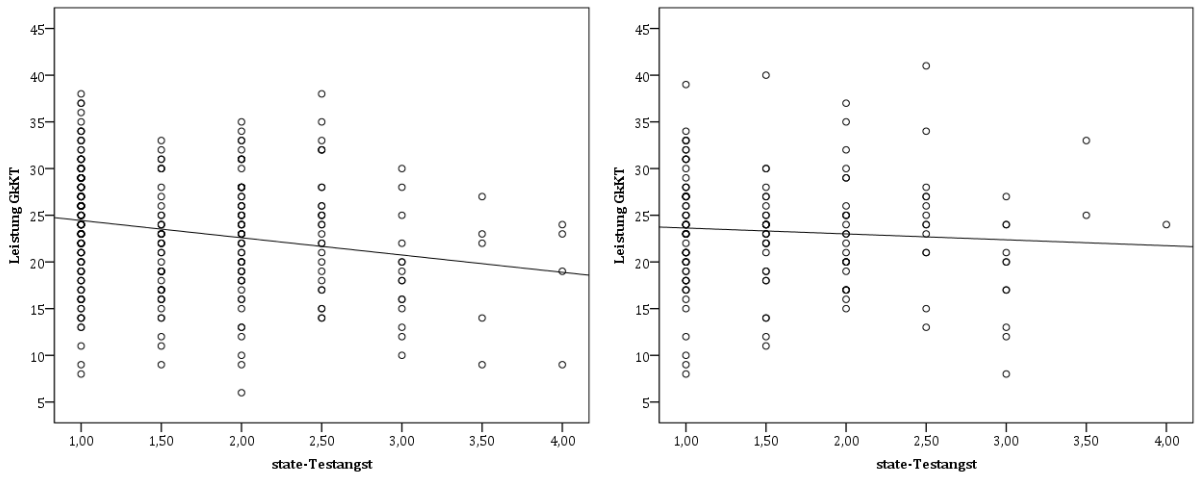

Abbildung 22: Streudiagramme zum Zusammenhang von state-Testangst (Besorgtheit) und Leistung im GkKT in der Kontrollgruppe (links; Bedingung A) sowie der Treatmentgruppe (rechts; Bedingung B)

6.2.2.3 Hypothese 3a-c

Zur Prüfung der Hypothese 3a wurde für jede der abhängigen Variablen (Abiturnote, Erfolgszuversicht im Studium, Abbruchtendenz) eine moderierte Moderation berechnet (siehe Abschnitt

6.1.6), wobei die Testängstlichkeit (hier nur die Facette Besorgtheit) als unabhängige Variable fungierte und AM-Energie und AM-Information Moderatoren waren (PROCESS Model 3). Die Ergebnisse zur abhängigen Variable Abiturnote sind in Tabelle 84 aufgeführt.

Tabelle 84: Zusammenfassung des Regressionsmodells zur Vorhersage der Abiturnote durch Testängstlichkeit (Besorgtheit), AM-Energie und AM-Info sowie die Interaktionen bzw. Moderationen

	Kriterium (Y): Abiturnote			
	Koeff.	*SE*	*t*	*p*
Konstante	2.29	.03	83.65	.000
X: Trait-TÄ	.02	.01	2.19	.029
M: AM-Energie	-.09	.04	-2.55	.011
W: AM-Info	.05	.04	1.36	.176
Int 1: trait-TÄ x AM-Energie	-.02	.01	-1.78	.075
Int 2: trait-TÄ x AM-Info	.02	.01	1.99	.047
Int 3: AM-Energie x AM-Info	-.02	.03	-.62	.539
Int 4: trait-TÄ x AM-Energie x AM-Info	-.01	.01	-1.06	.292
		R^2 = .05		
		F (7, 372) = 2.69, p = .010		

N = 380; X, M und W jeweils mittelwertszentriert; Abiturnote kodiert von 1 = „sehr gut“ bis 6 = „ungenügend“

Die Ergebnisse zeigen, dass trait-TÄ (Koeff. = .02, *p* = .029) einen (inhaltlich) negativen und AM-Energie (Koeff. = -.09, *p* = .011) einen (inhaltlich) positiven Effekte auf die Abiturnote hatte. Zwei auffällige Ergebnisse zeigten sich auch bei den Interaktionen. So war die Interaktion von trait-TÄ und AM-Energie marginal signifikant (Koeff. = -.02, *p* = .075) und die Interaktion von trait-TÄ und AM-Info signifikant (Koeff = .02, *p* = .047). Die Dreifachinteraktion war nicht signifikant. Zunächst bedeutet dies, dass mit *zunehmender* AM-Info der Effekt der trait-TÄ *stärker* wurde. Eine darüber hinaus gehende Analyse der Signifikanzregionen nach Johnson-Neyman (Hayes, 2013) zeigte, dass die Interaktion von trait-TÄ und AM-Energie ab einer Ausprägung von AM-Info knapp über dem Mittelwert signifikant wurde (zentrierter Wert von AM-Info = .28, *p* = .046), wenngleich die Dreifachinteraktion insgesamt nicht signifikant war. Dieses Bild ist auch bei Betrachtung der Interaktion von trait-TÄ und AM-Energie bei den drei Ausprägungen von AM-Info erkennbar: bei niedriger AM-Info war die Interaktion nicht signifikant (Effekt = -.011, *p* = .340), bei mittlerer marginal signifikant (Effekt = -.017, *p* = .074) und bei hoher signifikant (Effekt = -.024, *p* = .031) (Konstellationen für die Ausprägungen von AM-Info: *M* = 0 sowie *M* = ± 1 *SD*). Bei hoher AM-Info wurde also der negative Effekt von trait-TÄ auf die Abiturnote bedeutsam *schwächer*, wenn AM-Energie *zunahm*. Diese Konstellation ist in Abbildung 23 graphisch dargestellt. Während diese Richtung der Interaktion von trait-TÄ und AM-Energie erwartet wurde, war jene der Interaktion von trait-TÄ und AM-Info unerwartet.

Als Zweites wurde dasselbe Modell für die abhängige Variable Erfolgszuversicht im Studium berechnet. Die Ergebnisse sind in Tabelle 85 aufgeführt.

Tabelle 85: Zusammenfassung des Regressionsmodells zur Vorhersage der Erfolgszuversicht im Studium durch Testängstlichkeit (Besorgtheit), AM-Energie und AM-Info sowie die Interaktionen bzw. Moderationen

	Kriterium (Y): Erfolgszuversicht im Studium			
	Koeff.	*SE*	*t*	*p*
Konstante	3.12	.05	68.01	.000
X: Trait-TÄ	-.10	.01	-7.44	.000
M: AM-Energie	.21	.06	3.46	.001
W: AM-Info	-.05	.06	-.77	.440
Int 1: trait-TÄ x AM-Energie	.02	.02	.90	.370
Int 2: trait-TÄ x AM-Info	.01	.02	.36	.720
Int 3: AM-Energie x AM-Info	.01	.05	.11	.910
Int 4: trait-TÄ x AM-Energie x AM-Info	-.01	.01	-.48	.629
		R^2 = .20		
		F (7, 380) = 13.41, p < .001		

N = 388; X, M und W jeweils mittelwertszentriert

Die Analysen zeigen, dass trait-TÄ einen negativen (Koeff. = -.10, p < .001), AM-Energie hingegen einen positiven (Koeff. = .21, p = .001) Effekt auf die Erfolgszuversicht im Studium hatte. Dieser Befund ähnelt den Ergebnissen bezüglich der Abiturnote. Jedoch war keine der Interaktionen signifikant, was der Erwartung widerspricht.

Als Drittes wurde dieses Modell für die abhängige Variable Abbruchtendenz berechnet. Die Ergebnisse sind in Tabelle 86 aufgeführt.

Tabelle 86: Zusammenfassung des Regressionsmodells zur Vorhersage der Abbruchtendenz durch Testängstlichkeit (Besorgtheit), AM-Energie und AM-Info sowie die Interaktionen bzw. Moderationen

	Kriterium (Y): Abbruchtendenz			
	Koeff.	*SE*	*t*	*p*
Konstante	1.74	.06	28.47	.000
X: trait-TÄ	.10	.02	5.38	.000
M: AM-Energie	.09	.08	1.15	.253
W: AM-Info	-.11	.08	-1.26	.208
Int 1: trait-TÄ x AM-Energie	-.03	.02	-1.23	.218
Int 2: trait-TÄ x AM-Info	.05	.02	2.05	.041
Int 3: AM-Energie x AM-Info	-.06	.07	-.89	.374
Int 4: trait-TÄ x AM-Energie x AM-Info	.01	.02	.65	.514
		R^2 = .10		
		F (7, 379) = 6.21, p < .001		

N = 387; X, M und W jeweils mittelwertszentriert

Lediglich trait-TÄ hatte einen positiven Effekt auf die Abbruchtendenz (Koeff. = .10, p < .001). Darüber hinaus war die Interaktion von trait-TÄ und AM-Info signifikant (Koeff. = .05, p = .041). Dies

bedeutet, dass mit *zunehmender* Ausprägung von AM-Info der Effekt der trait-TÄ auf die Abbruchtendenz *stärker* wurde. Die Interaktion ist in Abbildung 23 schematisch dargestellt. Hypothese 3a ist also lediglich teilweise bestätigt.

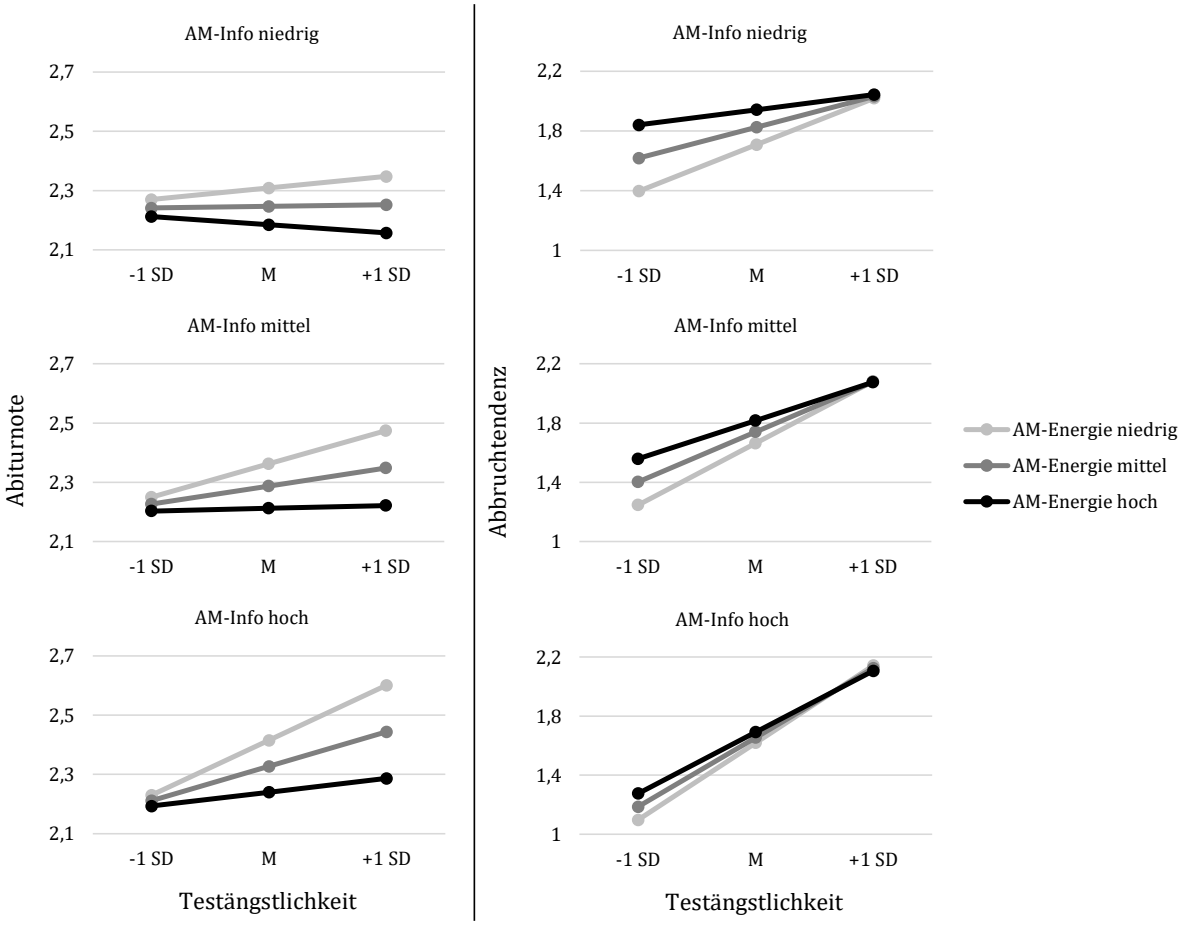

Abbildung 23: Schematische Darstellung der Interaktion von Testängstlichkeit (Besorgtheit), AM-Energie und AM-Info für die abhängigen Variablen Abiturnote und Abbruchtendenz; alle unabhängigen Variablen sind mittelwertszentriert; niedrig = -1 SD, mittel = M = 0, hoch = + 1 SD.

Die Prüfung von Hypothese 3b bezog sich auf Ebene 2, also die Zusammenhänge von state-TA, state-AM und Leistung im GkKT. Hierzu wurde wie auf Ebene 1 (Hypothese 3a) das Model 3 in PROCESS berechnet. Die beiden Moderatoren waren die state-AM und die Bedingung, um mögliche Effekte der Manipulation berücksichtigen zu können. Da die primäre unabhängige Variable state-TA mit der Subskala Besorgtheit operationalisiert wurde, wurde die Subskala Aufgeregtheit als Kovariate ins Modell inkludiert. Die Zusammenfassung der Ergebnisse ist Tabelle 87 zu entnehmen.

Tabelle 87: Zusammenfassung des Regressionsmodells zur Vorhersage der objektiven Leistung durch state-Testangst (Besorgtheit), state-anxiety motivation, Bedingung, die Interaktionen bzw. Moderationen sowie die Kovariate state-Testangst (Aufgeregtheit)

	Kriterium (Y): Leistung GkKT			
	Koeff.	*SE*	*t*	*p*
Konstante	23.24	.37	63.55	.000
X: state-TA (Besorgtheit)	-2.17	.62	-3.49	.001
M: state-AM	.96	.47	2.03	.043
W: Bedingung	.21	.64	.33	.742
Int 1: state-TA x state-AM	1.52	.66	2.33	.021
Int 2: state-TA x Beding.	1.05	.90	1.18	.241
Int 3: state-AM x Beding.	-.02	.89	-.02	.984
Int 4: state-TA x state-AM x Beding.	-2.08	1.12	-1.86	.064
Kovariate: state-TA (Aufgeregtheit)	.31	.57	.54	.587
		R^2 = .06		
		F (8, 409) = 3.29, p = .001		

N = 418; Bedingung = A vs. B; state-TA (Besorgtheit & Aufgeregtheit) sowie state-AM jeweils mittelwertszentriert; die Interaktionen sind mit der state-TA Facette Besorgtheit berechnet (X)

Die Analyse zeigt, dass state-TA (Besorgtheit) einen negativen Effekt auf die Leistung hatte (Koeff. = -2.17, p = .001), state-AM hingegen einen positiven (Koeff. = .96, p = .043). Überdies zeigte sich eine signifikante Interaktion von state-TA und state-AM (Koeff. = 1.52, p = .021) und eine marginal signifikante Dreifachinteraktion zwischen state-TA, state-AM und Bedingung (Koeff. = -2.08, p = .064). Letztere lässt sich damit erklären, dass nur in Bedingung A eine bedeutsame Interaktion zwischen state-TA und state-AM bestand[47]. In Bedingung B war die Interaktion nicht signifikant (Effekt = -.55, p = .546). Dieser Befund lässt sich in Abbildung 24 nachvollziehen. Dabei ist zu berücksichtigen, dass aufgrund des Bodeneffekts bei der state-AM eine niedrige Ausprägung (-1 *SD*) zugleich dem Skalenminimum (1.00) entsprach. Inhaltlich bedeutet dies, dass *nur* in Bedingung A die erwartete Interaktion auftrat: mit *zunehmender* state-AM wurde der negative Effekt der state-TA auf die Leistung *kleiner*. Bei Betrachtung der simple slopes (bezogen auf Bedingung A) ist erkennbar, dass state-TA einen signifikanten Effekt bei niedriger (Effekt = -3.00, p < .001) und mittlerer (Effekt = -2.15, p = .001) Ausprägung der state-AM hatte, nicht aber bei hoher Ausprägung der state-AM (Effekt = -.96, p = .221). In Bedingung B hingegen zeigten sich auf keiner Ausprägung der state-AM signifikante Effekte der state-TA. Hypothese 3b ist somit teilweise anzunehmen.

[47] Die in Tabelle 87 verzeichnete Interaktion entspricht der Interaktion in Bedingung A, da die Manipulation (Bedingung A vs. Bedingung B) dummy-kodiert ist, so dass Bedingung A mit 0 und Bedingung B mit 1 kodiert ist.

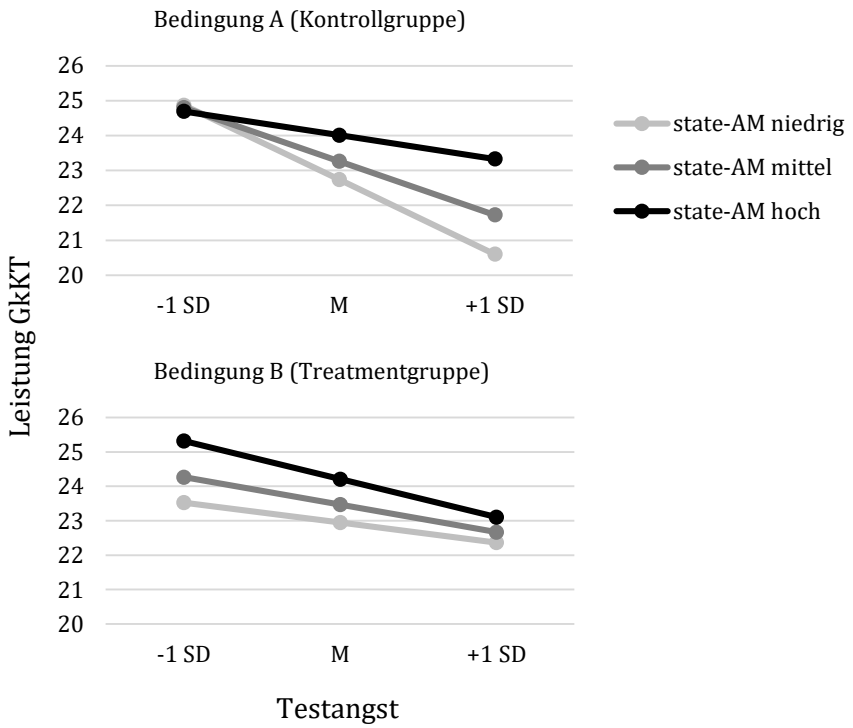

Abbildung 24: Schematische Darstellung der Interaktion von state-Testangst (Besorgtheit) und state-anxiety motivation in Bedingung A und B (Kovariate state-Testangst Aufgeregtheit); alle unabhängigen Variablen mittelwertszentriert; niedrig = -1 SD, mittel = M = 0, hoch = + 1 SD.

Zu Hypothese 3c, deren Prüfung sich aus den Analysen zu Hypothese 3a und 3b ableitet, fanden sich gemischte Ergebnisse. Einerseits wies AM-Energie positive Effekte auf die Abiturnote und auf die Erfolgszuversicht auf, ebenso wie state-AM einen positiven Effekt auf die Leistung im GkKT zeigte, was erwartet wurde. Andererseits zeigten sich diese Effekte nicht für die AM-Info, weshalb Hypothese 3c vorläufig abgelehnt wurde.

6.2.3 Weiterführende Analysen

Wie in Abschnitt 6.1.1 dargelegt, wurden die Fragebogenversionen A_{BEFKI} sowie $A_{MWT\text{-}B}$ zu Bedingung A zusammengelegt und in der Hypothesenprüfung der Bedingung B gegenübergestellt. Da sich auch bei Differenzierung von Version A_{BEFKI} sowie $A_{MWT\text{-}B}$ zwischen den drei Fragebogenversionen keinerlei Unterschiede in den erfassten Dispositionen, in der Abiturnote, in der Geschlechterverteilung, im Alter sowie den vertretenen Studiengängen zeigten, scheint dies gerechtfertigt. Dennoch sei auf eine Besonderheit eingegangen, ergänzend zu Hypothese 2b. So unterschieden sich in Bedingung A die Subgruppen A_{BEFKI} (r = -.13, p = .106) und $A_{MWT\text{-}B}$ (r = -.29, p < .001) deutlich in der Korrelation zwischen state-TA und Leistung, woraus die Korrelation von r = -.21, p < .001, für Bedingung A resultiert, welche bereits berichtet wurde. Dass dies auf die Unterschiede in der Testbatterie zurückgeht ist hochgradig unwahrscheinlich. So wussten die Pbn nicht, in welcher Bedingung sie waren bzw. dass es überhaupt Bedingungen gab. Die Ankündigung, dass ein Teil

den BEFKI und der andere den MWT-B bearbeiten würde, wurde erst unmittelbar vor Beginn beider Tests gegeben. Selbst wenn einige Pbn (durch Vorblättern) entdeckt hätten, dass sie später den BEFKI oder aber den MWT-B bearbeiten würden, gibt es keine plausible Erklärung, warum dies einen systematischen Effekt hätte produzieren sollen. Eine Erklärung lässt sich auch auf statistischem Wege nicht finden, weder durch den Vergleich von Mittelwerten und Standardabweichungen im GkKT sowie der state-TA bei den Pbn mit Version A_{BEFKI} sowie A_{MWT-B}, noch durch Exklusion von Fällen mit den höchsten Hebelwerten. Auch gibt es keine inhaltliche Begründung für die weitere Exklusion von Ausreißern (Eid et al., 2013), weshalb es sich bei den unterschiedlichen Korrelationen in A_{BEFKI} sowie A_{MWT-B} wahrscheinlich um einen zufälligen Befund handelt.

6.2.4 Zusammenfassung

Die Ergebnisse zur explorativen Analyse und zu Hypothese 1 sind, sofern Erwartungen formuliert wurden, überwiegend erwartungsgemäß. Die explorative Analyse diente in erster Linie der Sondierung des nomologischen Netzwerks von anxiety motivation. Zu Testängstlichkeit zeigten sich nur schwache Zusammenhänge, wobei AM-Energie zu zwei Facetten (Interferenz und Mangel an Zuversicht) negative, AM-Info zu den zwei weiteren Facetten (Besorgtheit und Aufgeregtheit) positive Zusammenhänge aufwies. Während AM-Info keinerlei bedeutsame Korrelationen zu den Persönlichkeitseigenschaften des FFM und Selbstwirksamkeit aufwies, korrelierte AM-Energie schwach positiv mit Gewissenhaftigkeit und Selbstwirksamkeit. Erwartungsgemäß konnten AM-Energie und -Info jeweils die state-AM besser vorhersagen als sämtlichen Facetten der Testängstlichkeit, die alle nicht signifikant mit state-AM korrelierten. Hierbei wies AM-Energie einen höheren Zusammenhang mit der state-AM auf als AM-Info. Hypothese 1 wurde somit bestätigt.

Die Ergebnisse zu den Hypothesen 2a und 2b widersprechen überwiegend der Erwartung. Bezüglich Hypothese 2a zeigte sich zwar erwartungsgemäß kein Effekt der Manipulation auf die state-TA. Jedoch wiesen auch die anderen abhängigen Variablen objektive Leistung, Belastungsfreiheit und subjektive Leistung keinerlei Gruppenunterschiede auf. Dies gilt auch für die state-AM. Hypothese 2a war somit abzulehnen. Demgegenüber fanden sich Hinweise, die Hypothese 2b unterstützen. So war der Zusammenhang zwischen state-TA und Leistung in der Kontrollgruppe mit (r = -.21) schwächer als in der Treatmentgruppe (r = -.09). Obzwar dieser Unterschied nicht signifikant war, weist er in die vermutete Richtung.

Die Ergebnisse zu den Hypothesen 3a-c sind gemischt. Auf Ebene 1 (Dispositionen und Zusammenhang zu Leistungskriterien) ist festzustellen, dass trait-TÄ bzw. AM-Energie einen (inhaltlich) negativen bzw. positiven Effekt auf die Abiturnote hatten. Eine signifikante Interaktion von trait-TÄ und AM-Info indiziert, dass der Effekt von trait-TÄ auf die Abiturnote mit steigender AM-Info

zunahm. Die Interaktion von trait-TÄ und AM-Energie war lediglich marginal signifikant. Bei differenzierter Betrachtung der Effekte zeigte sich aber, dass bei hoher AM-Info eine signifikante Interaktion von trait-TÄ und AM-Energie vorlag: je höher AM-Energie, desto schwächer war der Effekt der trait-TÄ. Bezüglich der Erfolgszuversicht im Studium als abhängige Variable ist das Befundbild weniger komplex. Hier wies trait-TÄ einen negativen, AM-Energie einen positiven Effekt auf. Keine der Interaktionen war signifikant. Die dritte betrachtete Variable in diesem Kontext war die Tendenz zum Studienabbruch. Auch hier zeigte sich ein positiver Effekt der trait-TÄ. Darüber hinaus lag eine signifikante Interaktion von trait-TÄ und AM-Info vor: mit zunehmender AM-Info wurde der Effekt der trait-TÄ stärker. Insgesamt stützt dieses Befundmuster Hypothese 3a nur teilweise. Auf Ebene 2 (situatives Erleben und Testleistung) zeigten sich ein negativer Effekt der state-TA und ein positiver Effekt der state-AM auf die Leistung im GkKT. Überdies lag eine signifikante Interaktion von state-TA und state-AM vor sowie eine marginal signifikante Dreifachinteraktion zwischen state-TA, state-AM und der Manipulation. Eine differenzierte Betrachtung zeigte, dass lediglich in Bedingung A eine bedeutsame Interaktion von state-TA und state-AM vorlag, die in diesem Fall jedoch erwartungsgemäß ist: mit zunehmender state-AM schwächte sich der (negative) Effekt von state-TA auf die Leistung ab. Demgegenüber zeigte sich in Bedingung B kein signifikanter Effekt der state-TA auf die Leistung, was bereits aus der nicht signifikanten Korrelation von state-TA und Leistung in Bedingung B folgt. Hypothese 3b ist somit nur teilweise anzunehmen. Hypothese 3c ist vorläufig abzulehnen. Zwar zeigte AM-Energie einen positiven Effekt auf die Abiturnote und die Erfolgszuversicht. Auch zeigte state-AM einen positiven Effekt auf die Testleistung. Entgegen der Erwartung sind solche Effekte jedoch für AM-Info nicht beobachtbar.

6.3 Diskussion

Die Koppelung mehrerer Fragestellungen schlug sich im Design von Studie 3 nieder. So sollte einerseits, wie bei neueren Konstrukten erforderlich, das nomologische Netzwerk und auch der theoretische „Mehrwert“ des Konstrukts eruiert werden. Darüber hinaus sollte die Frage nach der moderierenden Wirkung von anxiety motivation sowohl innerhalb der Situation (d. h. der Testung) als auch jenseits der Testung (durch Betrachtung von Dispositionen und anderen Leistungskriterien) beantwortet werden. Schließlich wurde mit der reappraisal-Manipulation gleichzeitig eine Kurzintervention getestet.

6.3.1 Bewertung der explorativen Analysen und Hypothesen

Zunächst sollen die Ergebnisse der explorativen Analyse bewertet werden. Die Korrelationen zwischen Testängstlichkeit und anxiety motivation sind aufschlussreich. Der schwach positive Zusammenhang von AM-Info und Besorgtheit ist theoretisch plausibel. So ist es wahrscheinlich, dass der Gedanke, sich mehr anstrengen zu müssen (z. B. „Wenn ich während einer Prüfung oder einem Test Angst empfinde … erinnert mich das daran, dass ich mich konzentrieren muss“), Ursache von verstärkter Besorgtheit sein kann und gleichzeitig dessen Folge. AM-Info beinhaltet theoretisch eine Signalwirkung der Angst, also die Signalisierung einer Ist-Soll-Diskrepanz in einem bestimmten Bereich (siehe Abschnitt 1.3.2.1). Interessant ist vor diesem Hintergrund das teilweise gegensätzliche Korrelationsmuster von AM-Energie und AM-Info: so korrelierte AM-Energie negativ mit Interferenz und Mangel an Zuversicht und nicht signifikant mit Besorgtheit und Aufgeregtheit. Das bedeutet, dass AM-Energie einherging mit einem größeren Vertrauen in die eigenen Fähigkeiten, was sich auch in der schwach positiven Korrelation mit Selbstwirksamkeit zeigte. Während AM-Info gänzlich unabhängig von den Persönlichkeitseigenschaften nach dem FFM war, korrelierte AM-Energie schwach positiv mit Gewissenhaftigkeit. Dies lässt sich auf Gemeinsamkeiten beider Konstrukte zurückführen. Während Gewissenhaftigkeit bedeutet, die eigenen Energien beständig auf ein Ziel zu richten (z. B. „Aufgaben bearbeite ich bis ins letzte Detail.“), kann AM-Energie verstanden werden als die Mobilisierung der eigenen Energie zur Erreichung eines Ziels, welche durch das Erleben von Testangst motiviert wird. Die gefundenen Zusammenhänge waren insgesamt relativ schwach.

Drei Schlüsse können daraus gezogen werden. Erstens sprechen die Resultate dafür, dass AM ein von der Testängstlichkeit und anderen Persönlichkeitseigenschaften zu unterscheidendes Konstrukt ist. Zweitens stützen die Befunde die empirische Eigenständigkeit von AM-Energie und -Info, was für die Konstruktvalidität der facettentheoretischen Interpretation spricht. Drittens (und den zweiten Punkt wiederum stützend) kann interpretiert werden, dass AM-Energie eher funktionale oder adaptive Bedeutung hat, da es negativ mit zwei Facetten von Testängstlichkeit

und positiv mit Gewissenhaftigkeit und Selbstwirksamkeit korrelierte. Demgegenüber scheint AM-Info eher dysfunktional zu sein bzw. stärker mit dem eigentlichen Angsterleben zu korrespondieren. Diese Ergebnisse lassen sich gut mit denen von Strack et al. (2014) vereinbaren. Sie berichten eine schwach positive (nicht signifikante) Korrelation zwischen AM-Info und trait-Ängstlichkeit (r = .12), wohingegen AM-Energie mit letzterer deutlich negativ korreliert (r = -.37). Die von Strack et al. (2014) gefundene negative Korrelation von AM-Info mit emotionaler Erschöpfung (r = -.23) scheint tendenziell im Widerspruch zu den Ergebnissen der vorliegenden Studie zu stehen. Auf der anderen Seite zeigte AM-Energie auch in den Daten von Strack et al. (2014) ein adaptiveres Zusammenhangsmuster: AM-Energie hing noch deutlicher negativ mit emotionaler Erschöpfung zusammen (r = -.49).

Die Ergebnisse zu Hypothese 1 untermauern den theoretischen Nutzen von AM. State-AM als situatives Maß für die funktionale Wirkung von Angst ließ sich durch AM-Energie und -Info wesentlich besser prädizieren als durch die Facetten von Testängstlichkeit. Mit anderen Worten ließ sich state-AM durch die korrespondierende „trait"-AM (AM-Energie & -Info) besser vorhersagen als durch Testängstlichkeit. Es muss dabei berücksichtigt werden, dass – anders als im Fragebogen zu AM – im TAI-G XU nicht nach motivierenden Wirkungen von Angst gefragt wird. Dennoch verdeutlicht dieses Ergebnis, dass eine Unterscheidung zwischen dem *Erleben* und der motivationalen *Wirkung* von Angst sinnvoll und möglich ist.

Ein positiver Effekt der reappraisal-Manipulation, der gemäß Hypothese 2a erwartet wurde, konnte nicht bestätigt werden. Weder bei der objektiven Leistung, der Belastungsfreiheit, noch bei der subjektiven Leistung zeigten sich Unterschiede zwischen Kontroll- und Treatmentgruppe. Der erwartete, nicht signifikante Gruppenunterschied bei der state-TA sollte daher nicht als Argument für die Gültigkeit der Hypothese herangezogen werden. In dieses Bild passt, dass es keine Gruppenunterschiede bei der state-AM gab. Diese Befunde sprechen gegen die Wirksamkeit der Kurzintervention. Da sich die Formulierungsbestandteile der Kurzintervention stark an bereits in der Literatur eingesetzten orientierten, ist dieses Ergebnis klärungsbedürftig.

Mindestens drei Erklärungen sind denkbar. Erstens könnte die Intervention von zu geringer Intensität gewesen sein, um eine Wirkung zu entfalten. Die Intervention war explizit kurz gestaltet worden, um deren Einsetzbarkeit in standardisierten Testsituationen zu erleichtern. Andere Kurzinterventionen waren deutlich umfangreicher. Beispielsweise gaben Lang und Lang (2010) beim Kompetenzpriming eine maximale Bearbeitungszeit von 10 Minuten vor. Park et al. (2014) gewährten den Pbn beim expressiven Schreiben 7 Minuten. Jamieson und Nock et al. (2012) berichten einen Zeitaufwand von 10 bis 15 Minuten für das Lesen von Zusammenfassungen von Zeitschriftenartikeln, in denen die Manipulation transportiert wurde. Weitere Untersuchungen

von Jamieson und Kollegen setzten dasselbe Prozedere ein (Beltzer et al., 2014; Jamieson et al., 2013). Selbst die kürzlich von Jamieson et al. (2016) berichtete Bearbeitungsdauer von 5 bis 8 Minuten überschreitet den zeitlichem Umfang der in dieser Studie administrierten Intervention deutlich. Johns et al. (2008) berichten hingegen nur eine kurze, vom Versuchsleiter verbal mitgeteilte reappraisal-Instruktion. Auch Jamieson et al. (2010) berichten, nur eine kurze Instruktion verwendet zu haben. Allerdings ist in beiden Fällen nicht angegeben, wie lange und umfangreich die Manipulation tatsächlich war. Es ist daher offen, wie kurz eine reappraisal-Manipulation sein darf oder wie intensiv sie mindestens sein muss (ob Dauer mit Intensität gleichgesetzt werden kann, ist zusätzlich klärungsbedürftig). Zweitens ist möglich, dass die Wirkung der Intervention durch ihre Position direkt zu Beginn des ersten Aufgabenblocks neutralisiert wurde. Möglicherweise haben die meisten Pbn unverzüglich mit der Bearbeitung der Aufgabe begonnen. Dass dies eine Rolle gespielt hat ist nicht von der Hand zu weisen, da die Pbn wussten, dass der GkKT einer Zeitbegrenzung unterliegt. Eine unbestimmte Anzahl von Pbn dürfte also die Instruktion „überlesen“ haben. Hierbei muss kritisch erwähnt werden, dass keine Prüfung erfolgen konnte, ob die Teilnehmer die reappraisal-Manipulation tatsächlich gelesen haben. Ein expliziter Manipulationscheck hat also gefehlt. Damit zusammenhängend und drittens gab es keine klare Trennung zwischen der reappraisal-Manipulation und der Aufgabenbearbeitung, was wahrscheinlich zu einer oberflächlicheren Verarbeitung des gelesenen Textabschnitts oder gar zu einem Abbruch während des Lesens desselbigen geführt hat. Aus Sicht eines Befragten in einer Umfrage ist es wahrscheinlich, dass nur jene Informationen gelesen werden, die subjektiv einen Nutzen für die weitere Bearbeitung eines Fragebogens haben. Es ist anzunehmen, dass dies ebenso für Leistungstests gilt. Hierbei kann des Weiteren davon ausgegangen werden, dass ein unbestimmter Teil von Testanden die Instruktion nur dann aufmerksam liest, wenn das Aufgabenprinzip nicht unmittelbar und intuitiv verständlich ist. Dass nun das Lesen des anfänglichen Hinweises zum reappraisal nicht unmittelbar zur Lösung der Aufgaben im ersten Aufgabenblock nötig war, dürfte zum „Nichtlesen“ des Textabschnitts beigetragen haben. Dieser Schwachpunkt im Design kann für folgende Erkenntnis genutzt werden: eine eher subtile und vor allem kurze reappraisal-Manipulation, die in das Testmaterial hineingemischt wird, genügt offenkundig nicht, um einen Effekt auf die Leistung zu bewirken.

Die Ergebnisse zu Hypothese 2b scheinen dem Misserfolg der Manipulation zu widersprechen: in der Treatmentgruppe zeigte sich eine abgeschwächte Korrelation zwischen Testangst und Leistung, verglichen mit der Kontrollgruppe. Ähnliche Befunde wurden beim Kompetenzpriming (Lang & Lang, 2010) sowie beim expressiven Schreiben (Ramirez & Beilock, 2011) festgestellt. Die Ergebnisse könnten dazu verleiten, den Effekt auf die Manipulation zurückzuführen. Wie jedoch gerade ausgeführt wurde, gab es keine weiteren Hinweise für eine erfolgreiche Manipulation. Zwei Erklärungen sind möglich. Entweder handelt es sich um einen Zufallsbefund – in diese

Richtung weist auch der nicht näher erklärbare Unterschied in den Korrelationen zwischen state-TA und Leistung in den Versionen A_{BEFKI} sowie A_{MWT-B}. Alternativ ist denkbar, dass die Manipulation erfolgreich *war*, ohne dass sich in den betrachteten abhängigen Variablen Effekte zeigten. Einer solchen, spekulativen Interpretation soll an dieser Stelle aber nicht gefolgt werden.

Hinter dem komplexen Befundbild zu Hypothese 3a ist eine unerwartete, aber plausible Systematik erkennbar. Die Interaktion von trait-TÄ mit AM-Info bei der Prädiktion der Abiturnote bedeutet inhaltlich, dass sich der negative Effekt von trait-TÄ mit zunehmender AM-Info *verstärkte*. Bei hoher AM-Info war der Effekt von trait-TÄ am stärksten. AM-Info hatte also einen moderierenden Effekt, jedoch nicht in der erwarteten Richtung. AM-Info beinhaltet die Wahrnehmung, dass die bisherige Investition in ein Ziel nicht ausreicht und zusätzliche Anstrengung investiert werden muss (z. B. „Wenn ich während einer Prüfung oder einem Test Angst empfinde … gibt mir das ein klares Signal dafür, dass ich mich mehr anstrengen muss"). Entgegen der Erwartung, dass dies die negative Auswirkung von Testängstlichkeit reduziert, verstärkte sie diese. Unter Rückgriff auf die Theorie zur Selbstwirksamkeit ergibt dies durchaus Sinn: „Stressful and taxing situations generally elicit emotional arousal that, depending on the circumstances, might have informative value concerning personal competency." (Bandura, 1977, S. 198). Emotionale und physiologische Erregung ist eine von vier Antezedenzien von Selbstwirksamkeit (Bandura, 1977). Beispielsweise kann sich die Selbstwirksamkeit verringern, wenn eine Person bei einer Anforderung starke Beanspruchung erlebt, wohingegen ihre Erfolgszuversicht hoch ist, solange sie aufgrund einer souveränen Beherrschung einer Aufgabe keine Anspannung erlebt (Bandura, 1994). Ähnlich ist dies bei der informierenden Wirkung von Angst denkbar: interpretiert eine testängstliche Person ihre Testangst als Signal für Ihren Abstand zu einem angestrebten Ziel, könnte dies eine Verstärkung der Testangst und ihrer negativen Konsequenzen bewirken. Anschaulich beschreibt Bandura (1977) eine Situation, die der hier skizzierten sehr ähnlich ist: „By conjuring up fear-provoking thoughts about their ineptitude, individuals can rouse themselves to elevated levels of anxiety that far exceed the fear experienced during the actual threatening situation." (S. 199). Die moderierende Bedeutung von AM-Info verdeutlicht auch die Notwendigkeit, das Erleben von Testängstlichkeit von deren Bewertung zu trennen. Die Interpretation von Angst als Signal für eine „Ist-Soll-Diskrepanz" könnte eine conditio sine qua non für negative Effekte von Testangst sein.

In diesem Kontext sollte auch die marginal signifikante Interaktion von trait-TÄ und AM-Energie nicht ignoriert werden. Es zeigte sich, dass bei hoher Ausprägung der AM-Info der Zusammenhang von trait-TÄ mit der Abiturnote durch AM-Energie moderiert wurde, und zwar in erwarteter Richtung: mit zunehmender AM-Energie schwächte sich der Effekt der trait-TÄ ab. Das bedeutet: je *mehr* Angst als motivierend bewertet wurde, umso *schwächer* war der ungünstige Effekt der trait-

TÄ – allerdings nur bei hoch ausgeprägter AM-Info. Zum einen unterstreicht dies die Notwendigkeit der theoretischen Trennung von informierender und „energetisierender" Funktion von Angst, auch wenn beide Facetten miteinander korrelierten (r = .44). Zum anderen scheint AM-Energie eine Art „Pufferfunktion" zu besitzen: in einer eigentlich ungünstigen Situation hoch ausgeprägter AM-Info verringerte sich mit zunehmender AM-Energie der negative Effekt der trait-TÄ. Die positive Rolle von AM-Energie wird zusätzlich untermauert durch deren positiven Effekt auf die Abiturnote, bei Kontrolle (konkret: bei mittlerer Ausprägung) der trait-TÄ. Dies stützt eine Interpretation von AM-Energie als adaptiver Eigenschaft. Das Zusammenspiel von AM-Info und AM-Energie kann also durchaus als komplex bezeichnet werden. Womöglich sind die Bewertungsprozesse, die in AM-Info abgebildet sind, in der Emotionsentstehung chronologisch früher angesiedelt als jene in AM-Energie: wird Testangst empfunden und als handlungsrelevant wahrgenommen (AM-Info) stellt sich als nächstes die Frage, ob die Angst als förderlich erlebt wird oder nicht (AM-Energie). Ist dies der Fall, werden negative Effekte der trait-TÄ abgepuffert. Wichtig scheint hierbei der Befund, dass der negative Effekt der trait-TÄ abhing von deren Interpretation – mit zunehmender AM-Info wurde der negative Effekt deutlicher. Festgehalten werden muss dabei auch, dass der Effekt von trait-TÄ auf die Abiturnote in keinem Fall durch die AM-Energie „umgekehrt" wurde: selbst wenn also Angst als motivierend empfunden wurde, zeigte sie keine positive Relation mit diesem Leistungskriterium. Vereinfacht ließe sich sagen, dass Testängstlichkeit als maladaptive Eigenschaft durch eine funktionale Interpretation nicht zu einer adaptiven Eigenschaft wird.

Aus theoretischer Sicht stellt sich die Frage, ob die Relevanz von AM-Energie nicht auf Personen mit hoher trait-TÄ beschränkt ist. Dies ist aus zweierlei Gründen zu verneinen. Aus statistischer Sicht ist der Effekt von AM-Energie (M) auf die Abiturnote (Y) im berechneten Modell (Tabelle 84) als Effekt von M auf Y bei *mittlerer* Ausprägung von X (der trait-TÄ) zu interpretieren. Dies ist durch die Mittelwertszentrierung der unabhängigen Variablen (X, M und W) bedingt. Zudem sind die Items im Fragebogen zu AM nicht spezifisch auf Situationen hin formuliert, in denen „starke" oder „hohe" Testangst vorliegt, sondern nur daraufhin, *dass* Angst empfunden wird.

Es findet sich partielle Bestätigung für diese Vermutungen bei der Analyse der abhängigen Variable Erfolgszuversicht. Auch hier zeigte sich neben dem erwarteten negativen Effekt der trait-TÄ ein eigenständiger, positiver Effekt der AM-Energie. Jedoch wurde keine der Interaktionen signifikant. Auch bei der Abbruchtendenz spiegelt sich das Ergebnis aus der diskutierten Analyse zur Abiturnote nur partiell. Auch hier hatte trait-TÄ einen positiven Effekt, jedoch zeigte sich eine signifikante Interaktion von trait-TÄ und AM-Info, deren Richtung mit jener in der Analyse zur Abiturnote übereinstimmte: mit zunehmender AM-Info wurde der Effekt der trait-TÄ auf die Abbruchtendenz stärker. Dass sich bei der Erfolgszuversicht und der Abbruchtendenz die Ergebnisse bezüglich der Abiturnote nur teilweise replizieren ließen, ist erklärungsbedürftig. Sicherlich

weisen Abiturnote, Erfolgszuversicht und Abbruchtendenz gemeinsame, aber auch voneinander verschiedene Determinanten auf. Allgemein sind Erfolgszuversicht und Abbruchtendenz jedoch weniger belastbare Leistungskriterien als die Abiturnote, die ein vergleichsweise „hartes" Kriterium darstellt. Die teilweise divergierenden Ergebnisse deuten aber an, dass die obige Interpretation als vorläufig betrachtet und in Zukunft weiter empirisch überprüft werden sollte.

Schließlich wurde geprüft, ob sich der vermutete Moderationseffekt auch bezüglich state-TA, state-AM und objektiver Leistung im GkKT zeigt. Insgesamt ähneln die Befunde am deutlichsten jenen auf der dispositionellen Ebene bezüglich der abhängigen Variable Abiturnote. So zeigte state-TA einen negativen und state-AM einen positiven Effekt auf die Testleistung. Auch wurde die Interaktion von state-TA und state-AM signifikant. Die nähere Betrachtung der Ergebnisse zeigte, dass sich diese Interaktion jedoch auf Bedingung A beschränkte, also jene Bedingung, in der state-TA auch einen signifikanten Effekt auf die Leistung vorwies. Unter Berücksichtigung dieser Einschränkung stützen diese Ergebnisse die Analysen zur Abiturnote. Der negative Effekt der state-TA wurde durch state-AM abgepuffert und war bei hoher state-AM nicht mehr signifikant.

Ein zentrales Ergebnis ist, dass sich positive Effekte *sowie* die moderierende Funktion von AM auf zwei Ebenen, nämlich bei Betrachtung von traits *und* states, feststellen ließen. Die von Strack und Esteves (2014) sowie Strack et al. (2014) berichteten Befunde zu derartigen Moderationseffekten konnten also mit Leistungskriterien bestätigt werden. Dieser Befund verweist auf komplexe Mechanismen im Erleben und der Bewertung von Testangst, die noch präziser zu entschlüsseln sind. In jedem Fall unterstützen die Ergebnisse die Forderung, Erleben und Bewertung von Testangst zu trennen. Diese Ergebnisse können ein Impuls sein, den (nicht allzu neuen) theoretischen Gedanken von Carver und Scheier (1988) in Zukunft stärker zu berücksichtigen: „The person who expects to be able to cope, who is sufficiently confident of being able to complete the action, responds to anxiety arousal with renewed effort. When this person's attention is self-directed, the result is enhanced persistence and even enhanced performance." (S. 18).

6.3.2 Limitationen

Die dargestellten Schlussfolgerungen unterliegen einigen Limitationen. Beide Erhebungen fanden in Papier-Bleistift-Form im Rahmen einer Vorlesung statt, wodurch nur eingeschränkte Kontrolle über die Testbedingungen herrschte. Suboptimal war dabei insbesondere der geringe Sitzabstand zwischen den Pbn. Auch herrschte trotz einer insgesamt konzentrierten Arbeitsatmosphäre ein gewisser Pegel an Störgeräuschen (Umblättern, Gespräche zwischen Pbn). Überdies wurde die Erhebung parallel in zwei verschiedenen Räumen durchgeführt, was per se nicht ideal ist. Aufgrund der in den Voranalysen berichteten Befunde gibt es jedoch keinen Anlass dafür, einen kon-

fundierenden Effekt durch die Variable Untersuchungsraum zu erwarten. Es kann nicht ausgeschlossen werden, dass es durch den geringen Sitzabstandes zwischen den Pbn einen Transfer zwischen Treatment- und Kontrollgruppe gab. Aufgrund der Kürze der reappraisal-Manipulation ist dies aber eher unwahrscheinlich.

Die Interpretation von anxiety motivation unterliegt einigen Einschränkungen. Aus methodischer Sicht kann kritisiert werden, dass die state-AM mit nur einem Item gemessen wurde und somit lediglich ein grober Indikator für die erleichternde Wirkung von Angst war. Auch lässt der Wortlaut des Items („Hat Ihnen diese Aufregung beim Lösen der Aufgaben geholfen?“) keine Rückschlüsse darauf zu, wie genau die erlebte Angst tatsächlich geholfen hat. Aus dem Wortlaut ergibt sich auch eine weitere Frage: so spielte das Item zu state-AM scheinbar direkt auf die erlebte Aufgeregtheit, nicht aber auf die Besorgtheit, an. Es mag als Widerspruch erscheinen, dass die state-TA (Besorgtheit) als unabhängige Variable, die state-TA (Aufgeregtheit) hingegen nur als Kovariate in das Modell inkludiert wurde (Hypothese 3b). Jedoch ist es unwahrscheinlich, dass die Pbn sich bei der Beantwortung des Items ausschließlich auf die in der Aufgeregtheit implizierten affektiven Angstkomponenten bezogen, da das Item zu state-AM unmittelbar nach den fünf Items zur Erfassung der state-TA positioniert war. Ein anderer Kritikpunkt erwächst aus der interpretativen Verbindung von Ebene 1 und 2 (Analyse zu Abiturnote und Testleistung), wonach AM in beiden Kontexten den Zusammenhang von Testängstlichkeit bzw. Testangst und einem Leistungskriterium moderiert. Diese unterstellt implizit, dass state-AM das situative Analogon zur Eigenschaft AM-Energie ist. Diese Position ist sicherlich kritisierbar. Evident ist gleichwohl, dass state-AM phänomenologisch näher an AM-Energie als an AM-Info ist, da die leistungsförderliche Bewertung von Angst in state-AM und AM-Energie unmittelbarer erfasst ist als in AM-Info. Eine bedeutendere Einschränkung erwächst daraus, dass sich die Moderation durch state-AM auf Bedingung A beschränkte. Aufgrund dieser Einschränkungen ist die vorgeschlagene Interpretation sicherlich vorläufig.

Eine weitere Limitation stellen die Bodeneffekte insbesondere bei der state-TA und state-AM dar (vgl. Abschnitt 6.1.5). Zwar ist der gefundene Mittelwert beim STAI-SKD (M = 1.81) vergleichbar mit dem, den Englert et al. (2011) nach einer evaluativen Instruktion berichten (M = 1.73), dennoch liegt der Wert deutlich unter dem möglichen Skalenmittelpunkt von 2.5. Der Versuch, eine Bewertungssituation zu simulieren, ist also gelungen, jedoch ist die Generalisierung der Befunde auf eine echte high-stakes Testung auf Basis der erhobenen Daten nicht zulässig. Relevant ist dies hinsichtlich der Frage, ob ein bestimmtes Niveau an Angst erlebt werden *muss*, bevor letztere als motivierend erlebt werden kann.

Die bedeutsamste Limitation ist die fehlende Prüfung, ob der Textabschnitt zur reappraisal-Manipulation gelesen und bewusst zur Kenntnis genommen wurde. Die nicht vorhandenen Mittelwertsunterschiede und der Unterschied in der Korrelation von Testangst und Leistung zwischen Treatment- und Kontrollgruppe können somit nicht zweifelsfrei auf ein Funktionieren oder Scheitern der Manipulation zurückgeführt werden.

Eine methodische Limitation ist der Einsatz von Kurzskalen, die nur aus wenigen Items bestehen. Insbesondere die Erfassung der Testangst mit dem STAI-SKD in Studie 3 (dies gilt auch für Studie 1) und die von den Autoren nicht vorgesehene, getrennte Interpretation von Besorgtheit (2 Items) und Aufgeregtheit (3 Items) seien hier genannt. Kurzskalen erfreuen sich jedoch aufgrund ihrer Ökonomie zunehmender Beliebtheit und können grundsätzlich die Anforderungen an die Testgüte erfüllen (siehe Rammstedt & Beierlein, 2014). Verbesserungswürdig ist die Skala AM-Info hinsichtlich ihrer Reliabilität.

Für die Interpretation der Ergebnisse muss ferner berücksichtigt werden, dass die gefundenen Effekte insgesamt relativ klein sind (insbesondere bei der Abiturnote; für eine Betrachtung standardisierter Regressionsgewichte aus den multiplen Regressionen siehe Anhang F). Diese Ergebnisse sind Ausdruck der Tatsache, dass Testängstlichkeit (und deren Bewertung) nur eine von mehreren Variablen ist, die mit Leistungsergebnissen in Zusammenhang stehen.

6.3.3 Implikationen

6.3.3.1 Theoretische und praktische Implikationen

Die zentrale theoretische Implikation der Ergebnisse ist, dass in der Operationalisierung von Testängstlichkeit bzw. Testangst und in der Erforschung der damit verbundenen Konsequenzen in stärkerem Maße die Trennung von Emotion bzw. Emotionserleben (Testängstlichkeit und Testangst) und Bewertung (anxiety motivation) etabliert werden sollte (siehe Abschnitt 1.3.2.1). Hierbei sollte berücksichtigt werden, dass Bewertungsprozesse komplex sind: „appraisals appear as consequences of emotions as well as antecedents“ (Frijda & Zeelenberg, 2001, S. 141). Ob Testangst wirklich leistungsschädlich ist, könnte also davon abhängen, wie sie bewertet wird. So wie sich in den Jahrzehnten der Forschung eine mehrdimensionale Auffassung des Erlebens von Testangst etabliert hat, könnte auch eine mehrdimensionale Betrachtung der Interpretation von Testangst (z. B. anregend, motivierend, lähmend) Mehrwert dabei generieren, ungünstige Effekte von Testangst zu erklären und ggf. zu modifizieren. Eine eindimensionale Perspektive, dass Testangst per se maladaptiv ist und stets negative Konsequenzen hat, sollte durch ein komplexeres Verständnis der Wirkungen von Testangst abgelöst werden.

Die Betrachtung des nomologischen Netzwerks erweitert das theoretische Verständnis von anxiety motivation. Allerdings erhebt die Differenzierung in informierende und energetisierende Wirkungen oder Bewertungen von Angst keineswegs Anspruch darauf, erschöpfend zu sein. Ebenso bedürfen beide Konzepte vor dem Hintergrund der teilweise unerwarteten Ergebnisse weiterer theoretischer Schärfung. AM-Info fußt auf der grundlegenden Bedeutung der Emotion Angst und ergibt sich auch aus einer Regulationsperspektive: Angst kann demnach Handlungsbedarf signalisieren (siehe Abschnitt 1.3.2.1). Dennoch weisen die Ergebnisse dieser Studie darauf hin, dass AM-Info eher maladaptiv ist, was zunächst überrascht. Aufschluss geben könnten die schwachen positiven Korrelationen zu den Facetten Besorgtheit und Aufgeregtheit. Was den maladaptiven Charakter von AM-Info ausmachen könnte ist die unangemessene oder verzerrte Definition von Ist- und Soll-Zuständen. Testängstliche Personen weisen beispielsweise einen niedrigeren Selbstwert, eine niedrigere Selbstwirksamkeit, eine ausgeprägte externale Kontrollüberzeugung und ein negatives Selbstkonzept auf (siehe Abschnitt 1.1.1.3). Eine Folge hiervon könnte sein, dass diese Personen den Ist-Zustand – ihre eigenen Fähigkeiten, ihren Lernfortschritt oder ihr Abschneiden in einer Prüfung – habituell als *zu niedrig* einschätzen. Darüber hinaus könnte sozial orientierter (bzw. maladaptiver) Perfektionismus, der mit Testängstlichkeit positiv korreliert (siehe Abschnitt 1.1.1.3), mit einem überhöhten Anspruch an die eigene Leistung einhergehen, was zu einer unangemessen hohen Setzung des Soll-Niveaus führen kann. Im Ergebnis zeigt sich bei testängstlichen Personen vermutlich häufig eine unangemessene Diskrepanz von aktuellem und angestrebtem Zustand, was auf eine verzerrte Einschätzung *beider* Parameter zurückgeht. Dies passt auch zu den Befunden, dass sich bei Testängstlichen kognitive Verzerrungen finden (siehe Abschnitt 1.2.1.2). Vereinfacht lautet die hier beschriebene Vermutung: Testängstliche schätzen das Ist zu niedrig und das Soll zu hoch ein.

Nichtsdestotrotz ist die wichtige Frage, was genau eigentlich erleichternde oder „motivierende" Aspekte oder Wirkungen von Angst phänomenologisch ausmacht, noch zu beantworten. Mindestens drei theoretische Erklärungen liegen nahe. Die erste Variante ist, dass leistungsförderliche Testangst schlicht und ergreifend ein niedriges Ausmaß, also *wenig* Testangst bedeutet. Raffety, Smith und Ptacek (1997) erhoben in einer Tagebuchstudie bei einer studentischen Stichprobe (N = 158) unter anderem die erlebte Testangst (erfasst wurden die drei Facetten Worry, Distraction und Tension mit einer eigens entwickelten Skala) sowie die selbst eingeschätzten erleichternden und hemmenden Wirkungen von Testangst beim Lernen sowie bei der Bearbeitung von Prüfungen (diese Konzeption war kongruent mit der Unterscheidung von AAT- und AAT+ im AAT, siehe Abschnitt 1.1.1.2.3). Die Autoren bildeten anhand von Mediansplits Subgruppen von Probanden und unterschieden „debilitators" (niedrige erleichternde und hohe hemmende Testängstlichkeit) und „facilitators" (hohe erleichternde und niedrige hemmende Testängstlichkeit). Insgesamt zeig-

ten sich bei den „debilitators“ auf allen drei Facetten der Testängstlichkeit höhere Werte. Die Autoren erklärten dies folgendermaßen: „This finding suggests that one aspect of viewing anxiety as facilitating may be lower overall levels of anxiety.“ (Raffety et al., 1997, S. 903). In diesem Kontext besonders interessant ist die von den Autoren weiter geführte Überlegung, dass die Bewertung der Angst an deren Intensität gekoppelt ist: „On the other hand, anxiety level may influence how one perceives an experience, with low-to-moderate anxiety being seen as facilitative and higher levels of anxiety being seen as debilitative.“ (Raffety et al., 1997, S. 903). Um dies zu klären ist es erforderlich, die leistungsförderlichen explizit von den leistungshinderlichen Prozessen zu trennen, da das „Fehlen“ leistungsförderlicher Prozesse (z. B. in Form einer niedrigen AM-Energie) nicht unbedingt gleichbedeutend mit einer leistungsbeeinträchtigenden Interpretation von Testangst sein muss. Dass erleichternde und hemmende Angst nicht vollständig orthogonal sind, zeigt die von Raffety et al. (1997) berichtete Interkorrelation von r = -.35 zwischen den beiden Kurzskalen für erleichternde und hemmende Testängstlichkeit. Wie von den Autoren betont, liegt dies in einem ähnlichen Bereich wie die in Abschnitt 1.1.1.2.3 berichteten Skaleninterkorrelationen des AAT.

Eine zweite Erklärung folgt einer anderen Argumentation. Wine (1980) schlug ein (hypothetisches) bidirektionales Modell der Testängstlichkeit vor, nach der die Unterschiede zwischen niedrig und hoch Testängstlichen über bloße Ausprägungsunterschiede in bestimmten Variablen hinausgehen. Zwar unterscheiden sich beide Gruppen durchaus anhand der Ausprägung bestimmter Merkmale (z. B. niedrige vs. hohe Selbstwirksamkeit), jedoch gibt es demnach auch phänomenologisch unterschiedliche Erlebensweisen (z. B. „Arousal interpreted as distress“ vs. „Arousal interpreted as energy“, S. 378). Zusammengefasst bedeutet dies: „The cognitive structures and self-statements of the low-test-anxious individual are not simply the opposite of that of the highs but rather differ qualitatively.“ (Wine, 1980, S. 376). Verwandt mit diesem Modell ist die jüngere, in der Sportpsychologie diskutierte Unterscheidung von Intensität und Richtung von Ängstlichkeit bzw. Angst, welche bereits in Abschnitt 1.3.2.1 erwähnt wurde. Jones (1995) stützt dabei die Trennung von Intensität und Richtung auf ein Modell von Carver und Scheier (1988) zur Wirkung von Angst. Diesem zufolge hängt eine erleichternde oder hemmende Wirkung von Angst von der – positiven oder negativen – Erwartung ab, die Angst bewältigen und ein gesetztes Ziel erreichen zu können. Ein Beispiel für die Differenzierung von Intensität und Richtung ist eine Untersuchung von Jones, Hanton und Swain (1994) mit N = 211 Wettkampfschwimmern (Elite- und Nicht-Eliteschwimmer die unter bzw. über einer vorgegebenen Qualifikationszeit lagen). Sie setzten das Competitive State Anxiety Inventory-2 (Martens, Burton, Vealey, Bump & Smith, 1990) ein, welches körperliche und kognitive Angstsymptome sowie Zuversicht erfasst (Intensität). Eine zusätzliche Skala erfasste für jedes Angstsymptom, ob dieses erleichternd oder beeinträchtigend erlebt wird (Richtung). Elite- und Nichteliteschwimmer unterschieden sich nicht in der Intensität der

kognitiven und körperlichen Angstsymptome. Jedoch war die leistungsförderliche Bewertung beider Angstkomponenten bei den Eliteschwimmern höher ausgeprägt als bei den Nicht-Eliteschwimmern.

Eine dritte Erklärung basiert auf der von Jones (1995) diskutierten Frage, ob erleichternde Angst noch Angst im eigentlichen Sinne darstellt: „It is likely, of course, that a state in which cognitive and physiological symptoms, however intense, are perceived as being facilitative to performance does not represent 'anxiety' at all. Instead, it will probably be labelled by the performer as 'anticipatory excitement' or being 'psyched up '." (Jones, 1995, S. 464). Jones (1995) argumentiert, dass sich die Operationalisierung im Fragebogen – und damit die Vorstellung vom Konstrukt seitens der Fragebogenentwickler – nicht unbedingt mit dem Erleben der befragten Personen decken muss. Jones und Hanton (2001) pointieren diese Kritik in einem späteren Artikel: „we believe that if a negative score on the direction scale is revealed, then this signifies a state of anxiety. If a positive direction score is found, this points to another state previously mislabelled as anxiety." (S. 393). Dies bedeutet, dass bei der Operationalisierung von Testängstlichkeit bzw. Testangst stets funktionale und dysfunktionale Bewertungen mit zu berücksichtigen sind, wobei auch ein qualitativer Zugang zum Erleben hilfreich sein kann, um die angesprochenen Probleme einer möglichen Fehlinterpretation zu umgehen.

Somit finden sich verschiedene Vorstellungen, was unter „motivierender Angst" oder motivierenden Wirkungen von Angst zu verstehen ist. Theoretisch stellt sich die Frage, ob es sich um zwei Formen von Angst (also zwei Konstrukte) oder unterschiedliche Ausprägungen eines Konstrukts, handelt. Die diskutierte Literatur liefert hierfür unterschiedliche Antworten. Auf der einen Seite gibt es Konzeptionen zur Trennung erleichternder und hemmender Angst (siehe Alpert & Haber, 1960). Zu dieser Perspektive könnte auch die theoretische Trennung zwischen Ängstlichkeit und anxiety motivation gezählt werden (siehe Strack et al., 2014). Dem stehen Vorstellungen gegenüber, dass erleichternde Angst schlichtweg ein niedriges Angstniveau bedeutet (siehe Raffety et al., 1997) oder die extreme Position, dass „erleichternde Angst" nicht mehr als „Angst" bezeichnet werden sollte (siehe Jones & Hanton, 2001).

Bezogen sich diese Ausführungen auf die Disposition, Angst als motivierend zu erleben, drängen sich ähnliche Fragen auch bei reappraisal-Manipulationen auf. Hier gilt zu klären, was für Prozesse eigentlich bei derartigen Manipulationen konkret in Gang gesetzt werden (siehe Abschnitt 1.3.2). Dies bezieht sich darauf, dass sich reappraisal-Manipulationen auf die Umbewertung von körperlichen Erregungsprozessen, aber auch von spezifischem Angsterleben beziehen können. Ein weiterer Gedanke hierzu wurde noch nicht betrachtet: In Abschnitt 1.3.2 wurde beschrieben, dass reappraisal-Manipulationen „adaptiv" sind, da sie nur von Relevanz sein sollten, wenn Personen tatsächlich Testangst erleben (und irrelevant wenn dies nicht der Fall ist). Ein Hinweis, dass

die Nervosität oder Aufregung bei der Testbearbeitung *nicht* schädlich ist, könnte jedoch selbst ein Auslöser dafür sein, Testangst zu erleben. Wird ein Proband darauf hingewiesen, dass er möglicherweise aufgeregt sein könnte, regt ihn das unter Umständen zu der Überlegung an, dass andere Probanden höchstwahrscheinlich aufgeregt *sind*, was im Kontrast zum eigenen Erleben steht. Dieses Missverhältnis zwischen subjektivem Empfinden und einem in der Situation als „angemessen" dargestellten Erleben könnte manche Personen aufgrund informativen sozialen Einflusses (Deutsch & Gerard, 1955; Hewstone & Martin, 2014) zu der Überlegung verleiten, dass man die Situation nicht adäquat eingeschätzt hat. Dementsprechend könnte das Nichtauftreten von Testangst als unpassend empfunden und – wenn nicht zu Testangst selbst – zu Unbehagen und der gedanklichen Beschäftigung mit den eigenen Emotionen führen. Noch subtilere Effekte sind möglich bei einer genauen Sezierung der Formulierungen einer reappraisal-Manipulation: im Sinne eines Primings könnte es einen Unterschied ausmachen, ob Aufregung als „nicht schädlich" oder aber als „förderlich" beschrieben wird oder beides.

Diese Überlegungen sollten in weiterer Forschung untersucht werden, auf die nun ein Ausblick gegeben wird.

6.3.3.2 Ausblick – weitere Forschung

Entsprechend der gerade aufgeführten theoretischen Überlegungen sollte bei weiteren Studien zu reappraisal-Manipulationen genau untersucht werden, welche Textelemente tatsächlich gelesen und bewusst verarbeitet werden und welche davon einen Effekt auf welche abhängigen Variablen zeigen. Eine systematische Variation von reappraisal-Manipulationen mit genauer Kontrolle des gelesenen Materials (z. B. indirekt durch Kontrollfragen oder direkt durch Eye-Tracking) wäre hier sinnvoll. Eine theoretisch vielversprechende abhängige Variable sind die subjektiv eingeschätzten Bewältigungsressourcen, welche möglicherweise den positiven Effekt von reppraisal-Manipulationen erklären (siehe Jamieson et al., 2016).

Damit verbunden sollte untersucht werden, was genau leistungsförderliche Wirkungen von Testangst ausmacht, also wie genau sich etwa eine energetisierende Bewertung von Angst positiv auf die Leistung auswirkt. In der Tradition der Testängstlichkeitsforschung dürfte die Arbeitsgedächtniskapazität eine bedeutsame Variable sein, aber auch Annäherungs- und Vermeidungszielorientierung könnten eine Rolle spielen.

Sowohl was die spezifischen Effekte von reappraisal-Manipulationen als auch die funktionale Bewertung von Testangst angeht bedarf es Studien, die die Aufgabenbearbeitung auf einer elementaren Ebene betrachten und dabei beispielsweise Bearbeitungszeiten, Lesezeiten, Verweildauer bei einzelnen Items und auch Bearbeitungsstrategien analysieren. Diese Variablen liefern unter

Umständen genauere Erkenntnisse über die stattfindenden Prozesse als ein bloßer Vergleich von Mittelwerten in der erreichten Leistung. Hierbei können Untersuchungsparadigmen wie jene zur Erforschung der Verarbeitungseffizienz (siehe Abschnitt 1.2.1.1.3) als Ausgangspunkt dienen.

Insbesondere bezüglich der Operationalisierung der leistungsförderlichen Bewertung von Testangst bedarf es weiterer Forschung. Die in dieser Studie eingesetzte Skala bietet noch Raum für weitere Optimierung. Die von Jones (1995) formulierte Problematik der Konzeption erleichternder Angst sollte in der zukünftigen Operationalisierung von erleichternder bzw. hemmender Testangst Berücksichtigung finden. Zur Vermeidung einer unangemessenen Vereinfachung der Phänomene wäre auch der stärkere Einbezug qualitativer Methodik denkbar.

Weitere Forschung sollte sich auch damit beschäftigen, ob die Differenzierung von anxiety motivation in Information und Energie hinreichend ist. Übertragen auf den Kontext von Prüfungen und Tests wäre es notwendig, das Erleben und die Bewertung von Testangst in unterschiedlichen Situationen zu unterscheiden, so z. B. während einer Prüfungsvorbereitung und bei der Bearbeitung einer Prüfung (vgl. z. B. das Vorgehen bei Raffety et al., 1997). Insbesondere die Definition von AM-Info sollte weiter präzisiert werden. Die angesprochene, durch Angst signalisierte Diskrepanz zwischen einem aktuell erlebten und einem angestrebten Zustand kann sich auf verschiedenste Aspekte beziehen, z. B. die eigenen Fähigkeiten, den eigenen Wissensstand oder auch den Vergleich mit anderen Personen. Je nach auslösender Situation könnten auch unterschiedliche Wechselwirkungen mit der energetisierenden Wirkung von Angst auftreten. Bezüglich beider Aspekte sollten auch die Relationen zu motivationalen Variablen geklärt werden. So ist denkbar, dass das Gegenstück zu einer motivierenden Wirkung von Angst nicht das *Fehlen* einer motivierenden Wirkung, sondern eine explizit demotivierende Wirkung ist. Beispielsweise könnte eine erleichternde bzw. hemmende Wirkung von Testangst mit einer erhöhten Annäherungs- bzw. Vermeidungszielorientierung einhergehen.

7. Schlusswort

In der Einleitung wurde die bereits von Mandler und Sarason (1952) angestoßene Frage aufgegriffen: führt Testangst dazu, dass eine Person in einer Testsituation ihre wahren Fähigkeiten nicht zeigen kann? Letztlich betrifft dies auch die Kriteriumsvalidität von Tests, die in dieser Arbeit nicht im Fokus stand. Diesbezüglich ist es fraglich, ob der negative Einfluss von Testangst auf Leistung tatsächlich ein Problem für die Interpretation von Testwerten darstellt. So ist es plausibel anzunehmen, dass Personen, die in einem Test schlecht abschneiden, auch im Kriterium eine schlechte Leistung erbringen. Ein Beispiel ist eine Person, die in einer Präsentationsübung bei einem Assessment Center schlecht abschneidet und in einer späteren Stresssituation im Beruf ebenso unzureichende Leistung erbringt – in beiden Fällen aufgrund von Testangst (Zeidner, 1995). Neuere empirische Untersuchungen ergaben bislang uneinheitliche Ergebnisse zu der Frage, inwieweit die Kriteriumsvalidität von besagter Thematik betroffen ist (Reeve, Heggestad & Lievens, 2009; Wicherts & Zand Scholten, 2010). Diese Problematik verdeutlicht, dass es wenig sinnvoll ist, Prüfungs- und Testsituationen gänzlich stress- und belastungsfrei zu gestalten (Zeidner, 1998). Zielführender ist die Untersuchung von Bedingungen, unter denen Testangst ungünstig ist und ob Testangst zwangsläufig ungünstig sein muss. Der negative Zusammenhang von Angst und Leistung sollte also nicht als „gegeben“ hingenommen werden, angesichts seiner Abhängigkeit von zahlreichen Situationsparametern (siehe Abschnitt 1.2.2) und seiner Veränderbarkeit durch gezielte Manipulationen (siehe Abschnitt 1.3.1). Auch die Tatsache, dass Prüfungen und Tests eine der Hauptquellen von Stress im Leben von Schülern und Studierenden darstellen, gibt Anlass für weitere Forschung zu Interventionen.

Diese Arbeit hatte nicht zuletzt zum Ziel, einige „blinde Flecken“ in der mehr als umfangreichen Forschung zu Testängstlichkeit bzw. Testangst zu beleuchten und somit Impulse für zukünftige Forschung zu geben. Erstens sollten zukünftige Arbeiten stärker darauf eingehen, welche Auswirkungen kleine und mitunter ungewollte Manipulationen der Testsituation auf das Erleben von Testangst und den Zusammenhang von Testangst und Leistung haben. Insbesondere gehören hierzu Elemente von Testinstruktionen (Studie 2 & 3), aber auch Skalen, die vor einem Test erhoben werden, gerade solche zu Testangst (Studie 1). Die Literatur zu Primingeffekten zeigt, dass auch „kleine“ Veränderungen einer Untersuchungssituation bedeutsame Effekte nach sich ziehen können und ernst genommen werden sollten. Die Forschung wird dabei nicht umhinkommen, Selbstberichtsmaße um weitere Erhebungsmethoden zu ergänzen und das Testerleben auf elementarer Ebene (z. B. Reaktionen auf einzelne Items) zu betrachten. Zweitens machen die Ergebnisse von Studie 3 deutlich, dass eine alleinige Betrachtung der leistungsmindernden Funktion von Testangst unzureichend ist. Statt der Frage, ob Testangst schädlich ist, muss untersucht wer-

den, wann und unter welchen Bedingungen Testangst schädlich ist und wann nicht. Die theoretischen Ausführungen gaben einen Einblick in die große „Tradition“ der Erforschung situativer Determinanten von Testangst und des Zusammenhangs von Angst und Leistung. Dieses Forschungsfeld sollte ergänzt werden um die Analyse von weiteren Moderatoren, wie etwa anxiety motivation. Die dritte Schlussfolgerung betrifft einen Aspekt, der in dieser Arbeit nicht untersucht wurde, aber das Forschungsfeld bereichern könnte. Zeidner (1998) beschrieb unterschiedliche Typen testängstlicher Personen (das Kapitel trägt den bezeichnenden Titel: „Debunking the Uniformity Myth: Different Types of Test-Anxious Students“, S. 52). Die dargestellten Typen (u. a. „failure-accepting examinees“, „perfectionistic overstrivers“ oder „self-handicappers”) lassen sich anhand von Einzelbefunden gut nachvollziehen. Untersuchungen zur empirischen Bildung solcher Typen, z. B. in Form von Clusteranalysen, sind jedoch äußerst rar (siehe z. B. Putwain & Daly, 2013). Die Zugehörigkeit zu einem bestimmten Typ könnte erklären, wie eine Person Testangst berichtet (Studie 1) oder wie sie auf bestimmte Elemente der Testinstruktion reagiert (Studie 2). Insbesondere die in Studie 3 untersuchte anxiety motivation könnte eine Persönlichkeitseigenschaft sein, die einen Typus von Personen ausmacht – solche, die ausgeprägte Angst erleben, aber davon nicht blockiert, sondern motiviert werden. Viertens steht die zukünftige Forschung vor der Herausforderung, die enorme und stetig wachsende Menge an Einzelbefunden zu aggregieren, dabei „blinde Flecken“ zu vermeiden und unterschiedliche Theorien zu integrieren. In dem Forschungsfeld besteht ein enormer Bedarf an Metaanalysen, welche die Befundlage aggregieren. Die beiden wegweisenden Metaanalysen von Hembree (1988) und Seipp (1991) liegen bereits über 25 Jahre zurück. Fünftens und letztens ist es nötig, die verschiedenen Theorien, die jeweils in der Lage sind, bestimmte Effekte zu erklären, in ein Gesamtmodell zu integrieren und dieses auch einer umfassenden empirischen Prüfung zuzuführen. Das S-REF Modell ist ein solches Modell, da es komplexe Wechselwirkungen von Prozessen, insbesondere auch von Defizit- und Interferenzerklärungen, integrieren kann (Zeidner & Matthews, 2010). Auf diese Weise könnten nicht nur die Effekte unterschiedlicher Instruktionen (Studie 1) und die Bedeutung der Bewertung der Angst (Studie 3), sondern auch die in Studie 2 untersuchte Rolle von Testangst bei der Selbst(wert-)regulation besser verstanden werden.

Diese Arbeit konnte einerseits Fragen beantworten, wirft aber andererseits neue auf. Jede der drei durchgeführten Studien konnte sich jeweils nur einem einzelnen Teilbereich eines großen Forschungsgebiets widmen. Diese Annäherung an das Phänomen aus unterschiedlichen „Richtungen“ kann aber ein Weg zu einem ganzheitlichen Verständnis der Entstehung und Wirkung von Testangst sein. Dass dies nur schrittweise geht, bringen Zeidner und Matthews (2010) auf den Punkt:

> *Anxiety is easy to experience but hard to understand, at least for psychologists. The conclusion we feel most confident in advancing is that anxiety must be understood at multiple levels.* (S. 238)

Literaturverzeichnis

Abbott, A. (2013). *Disputed results a fresh blow for social psychology. Failure to replicate intelligence-priming effects ignites row in research community.* Zugriff am 18.04.2016. Verfügbar unter www.nature.com/news/disputed-results-a-fresh-blow-for-social-psychology-1.12902

Ackerman, P. L. & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121* (2), 219-245.

Alpert, R. & Haber, R. N. (1960). Anxiety in academic achievement situations. *The Journal of Abnormal and Social Psychology, 61* (2), 207-215.

Alter, A. L., Aronson, J., Darley, J. M., Rodriguez, C. & Ruble, D. N. (2010). Rising to the threat: Reducing stereotype threat by reframing the threat as a challenge. *Journal of Experimental Social Psychology, 46* (1), 166-171.

American Psychiatric Association. (2013a). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)* (5th ed.). Washington: American Psychiatric Publishing.

American Psychiatric Association. (2013b). *Highlight of Changes from DSM-IV-TR to DSM-5:* American Psychiatric Publishing.

Anderson, S. B. & Sauser, W. I. (1995). Measurement of Test Anxiety: An Overview. In C. D. Spielberger & P. R. Vagg (Hrsg.), *Test Anxiety. Theory, Assessment, and Treatment* (S. 15-33). Washington, DC: Taylor & Francis.

Appel, M., Weber, S. & Kronberger, N. (2015). The influence of stereotype threat on immigrants. Review and meta-analysis. *Frontiers in Psychology, 6*.

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M. & Brown, J. (1999). When White Men Can't Do Math. Necessary and Sufficient Factors in Stereotype Threat. *Journal of Experimental Social Psychology, 35* (1), 29-46.

Arvey, R. D., Strickland, W., Drauden, G. & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43* (4), 695-716.

Asendorpf, J. B. & Neyer, F. J. (2012). *Psychologie der Persönlichkeit* (Springer-Lehrbuch, 5., vollst. überarb. Aufl.). Berlin, Heidelberg: Springer.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64* (6, Pt.1), 359-372.

Atkinson, J. W. & Litwin, G. H. (1960). Achievement motive and test anxiety conceived as motive to approach success and motive to avoid failure. *The Journal of Abnormal and Social Psychology, 60* (1), 52-63.

Augustine, A. A. & Hemenover, S. H. (2009). On the relative effectiveness of affect regulation strategies: A meta-analysis. *Cognition & Emotion, 23* (6), 1181-1220.

Bachmann, G. (2009). *Zielorientierungen und aktuelle Motivation: Eine Integration im Kontext des selbstregulierten Lernens*. Dissertation, Goethe-Universität Frankfurt a. M. Frankfurt a. M.

Baddeley, A. (2012). Working Memory. Theories, Models, and Controversies. *Annual Review of Psychology, 63* (1), 1-29.

Bandura, A. (1977). Self-efficacy. Toward a unifying theory of behavioral change. *Psychological Review, 84* (2), 191-215.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Hrsg.), *Encyclopedia of human behavior* (Bd. 4, S. 71-81). New York: Academic Press (Reprinted in H. Friedman [Ed.], Encyclopedia of mental health. San Diego: Academic Press, 1998).

Bargh, J. A. (2012). *Priming Effects Replicate Just Fine, Thanks. In response to a ScienceNews article on priming effects in social psychology.* Zugriff am 18.04.2016. Verfügbar unter https://www.psychologytoday.com/blog/the-natural-unconscious/201205/priming-effects-replicate-just-fine-thanks

Bargh, J. A., Chen, M. & Burrows, L. (1996). Automaticity of social behavior. Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71* (2), 230-244.

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J. & van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin, 133* (1), 1-24.

Baron, R. M. & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research. Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51* (6), 1173-1182.

Beck, A. T. & Clark, D. A. (1997). An information processing model of anxiety. Automatic and strategic processes. *Behaviour Research and Therapy, 35* (1), 49-58.

Beilock, S. L., Rydell, R. J. & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General, 136* (2), 256-276.

Beltzer, M. L., Nock, M. K., Peters, B. J. & Jamieson, J. P. (2014). Rethinking butterflies: The affective, physiological, and performance effects of reappraising arousal during social evaluation. *Emotion, 14* (4), 761-768.

Benjamin, M., McKeachie, W. J., Lin, Y.-g. & Holinger, D. P. (1981). Test anxiety: Deficits in information processing. *Journal of Educational Psychology, 73* (6), 816-824.

Benson, J., Bandalos, D. & Hutchinson, S. (1994). Modeling test anxiety among men and women. *Anxiety, Stress & Coping, 7* (2), 131-148.

Benson, J. & El-Zahhar, N. (1994). Further refinement and validation of the revised test anxiety scale. *Structural Equation Modeling: A Multidisciplinary Journal, 1* (3), 203-221.

Benson, J., Moulin-Julian, M., Schwarzer, C., Seipp, B. & El-Zahhar, N. (1992). Cross-validation of a revised test anxiety scale using multi-national samples. In K. A. Hagtvet & T. B. Johnson (Hrsg.), *Advances in test anxiety research* (Vol. 7, S. 62-83). Lisse: Swets & Zeitlinger.

Berglas, S. & Jones, E. E. (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology, 36* (4), 405-417.

Blankstein, K. R., Toner, B. B. & Flett, G. L. (1989). Test anxiety and the contents of consciousness. Thought-listing and endorsement measures. *Journal of Research in Personality, 23* (3), 269-286.

Borkenau, P. & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae*. Göttingen: Hogrefe.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (Springer-Lehrbuch, 7., vollständig überarbeitete und erweiterte Auflage). Berlin, Heidelberg: Springer-Verlag.

Bosson, J. K., Haymovitz, E. L. & Pinel, E. C. (2004). When saying and doing diverge: The effects of stereotype threat on self-reported versus non-verbal anxiety. *Journal of Experimental Social Psychology, 40* (2), 247-255.

Brodish, A. B. & Devine, P. G. (2009). The role of performance–avoidance goals and worry in mediating the relationship between stereotype threat and performance. *Journal of Experimental Social Psychology, 45* (1), 180-185.

Bruch, M. A. (1981). Relationship of test-taking strategies to test anxiety and performance. Toward a task analysis of examination behavior. *Cognitive Therapy and Research, 5* (1), 41-56.

Brunstein, J. & Heckhausen, H. (2010). Leistungsmotivation. In J. Heckhausen & H. Heckhausen (Hrsg.), *Motivation und Handeln* (S. 145-192). Berlin, Heidelberg: Springer.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2., aktualisierte und erw. Aufl.). München: Pearson Studium.

Cadinu, M., Maass, A., Rosabianca, A. & Kiesner, J. (2005). Why Do Women Underperform Under Stereotype Threat?: Evidence for the Role of Negative Thinking. *Psychological Science, 16* (7), 572-578.

Calvo, M. G. & Eysenck, M. W. (1996). Phonological working memory and reading in test anxiety. *Memory (Hove, England), 4* (3), 289-305.

Calvo, M. G., Eysenck, M. W. & Estevez, A. (1994). Ego-threat interpretive bias in test anxiety. Online inferences. *Cognition & Emotion, 8* (2), 127-146.

Calvo, M. G., Eysenck, M. W., Ramos, P. M. & Jiménez, A. (1994). Compensatory reading strategies in test anxiety. *Anxiety, Stress & Coping, 7* (2), 99-116.

Calvo, M. G., Ramos, P. M. & Estevez, A. (1992). Test anxiety and comprehension efficiency: The role of prior knowledge and working memory deficits. *Anxiety, Stress & Coping, 5* (2), 125-138.

Cannon, W. B. (1929). *Bodily changes in pain, hunger, fear, and rage*. Boston, MA: Branford.

Carroll, J. B. (1993). *Human cognitive abilities. A survey of factor-analytic studies*. Cambridge: Cambridge Univ. Press.

Cartwright-Hatton, S. & Wells, A. (1997). Beliefs about Worry and Intrusions: The Meta-Cognitions Questionnaire and its Correlates. *Journal of Anxiety Disorders, 11* (3), 279-296.

Carver, C. S. & Scheier, M. F. (1988). A control-process perspective on anxiety. *Anxiety Research, 1* (1), 17-22.

Carver, C. S. & Scheier, M. F. (1990). Origins and functions of positive and negative affect. A control-process view. *Psychological Review, 97* (1), 19-35.

Carver, C. S. & Scheier, M. F. (2001). *On the self-regulation of behavior* (1. pbk. ed.). Cambridge: Cambridge University Press.

Cassady, J. C. & Johnson, R. E. (2002). Cognitive Test Anxiety and Academic Performance. *Contemporary Educational Psychology, 27* (2), 270-295.

Cattell, R. B., Eber, H. W. & Tatsuoka, M. M. (1985). *Handbook for the sixteen personality factor questionnaire (16PF)*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Chamorro-Premuzic, T., Ahmetoglu, G. & Furnham, A. (2008). Little more than personality: Dispositional determinants of test anxiety (the Big Five, core self-evaluations, and self-assessed intelligence). *Learning and Individual Differences, 18* (2), 258-263.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A. et al. (2005). Test Anxiety and Academic Performance in Undergraduate and Graduate Students. *Journal of Educational Psychology, 97* (2), 268-274.

Chung, B. G., Ehrhart, M. G., Holcombe Ehrhart, K., Hattrup, K. & Solamon, J. (2010). Stereotype Threat, State Anxiety, and Specific Self-Efficacy as Predictors of Promotion Exam Performance. *Group & Organization Management, 35* (1), 77-107.

Cisler, J. M., Bacon, A. K. & Williams, N. L. (2009). Phenomenological Characteristics of Attentional Biases Towards Threat: A Critical Review. *Cognitive Therapy and Research, 33* (2), 221-234.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hoboken: Taylor and Francis.

Cohen, J. (2010). *Applied multiple regression/correlation analysis for the behavioral sciences* (3. ed., [Nachdr.]. Mahwah, NJ: Erlbaum.

Collani, G. von & Herzberg, P. Y. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie, 24* (1), 3-7.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation. Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Corbetta, M. & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3* (3), 201-215.

Costa, P. T. & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI)*. Professional Manual Odessa: Psychological Assessment Resources.

Costa, Paul, Jr., Terracciano, A. & McCrae, R. R. (2001). Gender differences in personality traits across cultures. Robust and surprising findings. *Journal of Personality and Social Psychology, 81* (2), 322-331.

Costello, A. B. & Osborne, J. W. (2005). Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. *Practical Assessment, Research & Evaluation, 10* (7), 1-9.

Covington, M. V. (1984a). The motive for self-worth. In R. E. Ames & C. Ames (Hrsg.), *Research on motivation in education* (S. 77-105). New York: Academic Press.

Covington, M. V. (1984b). The Self-Worth Theory of Achievement Motivation: Findings and Implications. *The Elementary School Journal, 85* (1), 5-20.

Covington, M. V. & Omelich, C. L. (1987). "I knew it cold before the exam": A test of the anxiety-blockage hypothesis. *Journal of Educational Psychology, 79* (4), 393-400.

Covington, M. V. & Omelich, C. L. (1988). Achievement dynamics. The interaction of motives, cognitions, and emotions over time. *Anxiety Research, 1* (3), 165-183.

Coy, B., O'Brien, W. H., Tabaczynski, T., Northern, J. & Carels, R. (2011). Associations between evaluation anxiety, cognitive interference and performance on working memory tasks. *Applied Cognitive Psychology, 25* (5), 823-832.

Crocker, J. & Knight, K. M. (2005). Contingencies of Self-Worth. *Current Directions in Psychological Science, 14* (4), 200-203.

Crocker, J., Luhtanen, R. K., Cooper, M. L. & Bouvrette, A. (2003). Contingencies of Self-Worth in College Students: Theory and Measurement. *Journal of Personality and Social Psychology, 85* (5), 894-908.

Crocker, J. & Wolfe, C. T. (2001). Contingencies of self-worth. *Psychological Review, 108* (3), 593-623.

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24* (4), 349-354.

Csikszentmihalyi, M. (2014). *Applications of Flow in Human Development and Education: The Collected Works of Mihaly Csikszentmihalyi*. s.l.: Springer Netherlands.

Culler, R. E. & Holahan, C. J. (1980). Test anxiety and academic performance: The effects of study-related behaviors. *Journal of Educational Psychology, 72* (1), 16-20.

Darke, S. (1988a). Anxiety and working memory capacity. *Cognition & Emotion, 2* (2), 145-154.

Darke, S. (1988b). Effects of anxiety on inferential reasoning task performance. *Journal of Personality and Social Psychology, 55* (3), 499-505.

Davey, G. C., Hampton, J., Farrell, J. & Davidson, S. (1992). Some characteristics of worrying: Evidence for worrying and anxiety as separate constructs. *Personality and Individual Differences, 13* (2), 133-147.

Deci, E. L. & Ryan, R. M. (1995). Human autonomy: The basis for true self-esteem. In M. H. Kernis (Hrsg.), *Efficacy, agency, and self-esteem* (S. 31-49). New York: Plenum Press.

Deffenbacher, J. L. (1978). Worry, emotionality, and task-generated interference in test anxiety: An empirical test of attentional theory. *Journal of Educational Psychology, 70* (2), 248-254.

Deffenbacher, J. L. (1980). Worry and emotionality in test anxiety. In I. G. Sarason (Ed.), *Test anxiety. Theory, research, and applications* (pp. 111-128). Hillsdale, N.J.: Erlbaum.

Deffenbacher, J. L. (1986). Cognitive and physiological components of test anxiety in real-life exams. *Cognitive Therapy and Research, 10* (6), 635-644.

Deffenbacher, J. L. & Deitz, S. R. (1978). Effects of test anxiety on performance, worry, and emotionality in naturally occurring exams. *Psychology in the Schools, 15* (3), 446-450.

Deffenbacher, J. L. & Hazaleus, S. L. (1985). Cognitive, emotional, and physiological components of Test Anxiety. *Cognitive Therapy and Research, 9* (2), 169-180.

Derakshan, N. & Eysenck, M. W. (2009). Anxiety, Processing Efficiency, and Cognitive Performance. *European Psychologist, 14* (2), 168-176.

Deutsch, M. & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology, 51* (3), 629-636.

Dickhäuser, O., Schöne, C., Spinath, B. & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept. *Zeitschrift für Differentielle und Diagnostische Psychologie, 23* (4), 393-405.

Diedenhofen, B. & Musch, J. (2015). cocor: a comprehensive solution for the statistical comparison of correlations. *PLoS ONE, 10* (3), e0121945.

Dienstbier, R. A. (1989). Arousal and physiological toughness. Implications for mental and physical health. *Psychological Review, 96* (1), 84-100.

Döpfner, M., Schnabel, M., Goletz, H. & Ollendick, T. H. (2006). *PHOKI. Phobiefragebogen für Kinder und Jugendliche*. Göttingen: Hogrefe.

Doyen, S., Klein, O., Pichon, C.-L., Cleeremans, A. & Lauwereyns, J. (2012). Behavioral Priming. It's All in the Mind, but Whose Mind? *PLoS ONE, 7* (1), e29081.

Eckert, C., Schilling, D. & Stiensmeier-Pelster, J. (2006). Einfluss des Fähigkeitsselbstkonzepts auf die Intelligenzund Konzentrationsleistung. *Zeitschrift für Pädagogische Psychologie, 20* (1/2), 41-48.

Egloff, B. & Schmukle, S. C. (2002). Predictive validity of an implicit association test for assessing anxiety. *Journal of Personality and Social Psychology, 83* (6), 1441-1455.

Eid, M., Gollwitzer, M. & Schmitt, M. (2013). *Statistik und Forschungsmethoden. Lehrbuch* (3., korrigierte Aufl.). Weinheim: Beltz.

Ekman, P. & Cordaro, D. (2011). What is Meant by Calling Emotions Basic. *Emotion Review, 3* (4), 364-370.

Elliot, A. J. (2006). The Hierarchical Model of Approach-Avoidance Motivation. *Motivation and Emotion, 30* (2), 111-116.

Elliot, A. J. & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72* (1), 218-232.

Elliot, A. J. & McGregor, H. A. (1999). Test anxiety and the hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 76* (4), 628-644.

Elliot, A. J. & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology, 80* (3), 501-519.

Elliot, A. J. & Pekrun, R. (2007). Emotion in the hierarchical model of approach-avoidance achievement motivation. In P. A. Schutz & R. Pekrun (Hrsg.), *Emotion in education* (Educational psychology series, S. 57-73). Amsterdam: Academic Press.

Englert, C., Bertrams, A. & Dickhäuser, O. (2011). Entwicklung der Fünf-Item-Kurzskala STAI-SKD zur Messung von Zustandsangst. *Zeitschrift für Gesundheitspsychologie, 19* (4), 173-180.

Ergene, T. (2003). Effective Interventions on Test Anxiety Reduction: A Meta-Analysis. *School Psychology International, 24* (3), 313-328.

Eum, K. & Rice, K. G. (2011). Test anxiety, perfectionism, goal orientation, and academic performance. *Anxiety, Stress & Coping, 24* (2), 167-178.

Evers, A. (2001). The Revised Dutch Rating System for Test Quality. *International Journal of Testing, 1* (2), 155-182.

Everson, H. T., Millsap, R. E. & Rodriguez, C. M. (1991). Isolating Gender Differences in Test Anxiety. A Confirmatory Factor Analysis of the Test Anxiety Inventory. *Educational and Psychological Measurement, 51* (1), 243-251.

Eysenck, M. W. (1982). *Attention and Arousal. Cognition and Performance*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Eysenck, M. W. (1992). *Anxiety. The cognitive perspective* (Essays in cognitive psychology). Hove: Erlbaum.

Eysenck, M. W. (1997). *Anxiety and Cognition. A Unified Theory* (Essays in cognitive psychology). Hoboken: Taylor and Francis.

Eysenck, M. W. & Calvo, M. G. (1992). Anxiety and Performance: The Processing Efficiency Theory. *Cognition & Emotion, 6* (6), 409-434.

Eysenck, M. W., Derakshan, N., Santos, R. & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7* (2), 336-353.

Fehm, L. & Fydrich, T. (2011). *Prüfungsangst* (Fortschritte der Psychotherapie, Bd. 44). Göttingen [u.a.]: Hogrefe.

Ferguson, E. & Cox, T. (1993). Exploratory Factor Analysis. A Users' Guide. *International Journal of Selection and Assessment, 1* (2), 84-94.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics. And sex and drugs and rock 'n' roll* (MobileStudy, 4th edition). Los Angeles: Sage.

Flore, P. C. & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of school psychology, 53* (1), 25-44.

Folkman, S. (1984). Personal control and stress and coping processes. A theoretical analysis. *Journal of Personality and Social Psychology, 46* (4), 839-852.

Folkman, S. & Lazarus, R. S. (1985). If it changes it must be a process. Study of emotion and coping during three stages of a college examination. *Journal of Personality and Social Psychology, 48* (1), 150-170.

Folkman, S., Lazarus, R. S., Dunkel-Schetter, C., DeLongis, A. & Gruen, R. J. (1986). Dynamics of a stressful encounter. Cognitive appraisal, coping, and encounter outcomes. *Journal of Personality and Social Psychology, 50* (5), 992-1003.

Fox, E., Russo, R., Bowles, R. & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General, 130* (4), 681-700.

French, J. W. (1962). Effect of Anxiety on Verbal and Mathematical Examination Scores. *Educational and Psychological Measurement, 22* (3), 553-564.

Frenzel, A. C., Götz, T. & Pekrun, R. (2009). Emotionen. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (Springer-Lehrbuch). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Frenzel, A. C., Pekrun, R. & Goetz, T. (2007). Girls and mathematics —A "hopeless" issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education, 22* (4), 497-514.

Freund, P. A. & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138* (2), 296-321.

Friedman, I. A. & Bendas-Jacob, O. (1997). Measuring Perceived Test Anxiety in Adolescents: A Self-Report Scale. *Educational and Psychological Measurement, 57* (6), 1035-1046.

Frijda, N. H. & Zeelenberg, M. (2001). Appraisal: What is the dependent? In K. R. Scherer, A. Schorr & T. Johnstone (Hrsg.), *Appraisal processes in emotion: Theory, methods, research. Series in affective science.* (S. 141-155). New York: Oxford University Press.

Fritz, C. O., Morris, P. E. & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology. General, 141* (1), 2-18.

Frost, R. O., Marten, P., Lahart, C. & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research, 14* (5), 449-468.

Galassi, J. P., Frierson, H. T. & Sharer, R. (1981). Concurrent versus retrospective assessment in test anxiety research. *Journal of Consulting and Clinical Psychology, 49* (4), 614-615.

Ganzer, V. J. (1968). Effects of audience presence and test anxiety on learning and retention in a serial learning situation. *Journal of Personality and Social Psychology, 8* (2, Pt.1), 194-199.

Gerstenberg, F. X. R., Imhoff, R. & Schmitt, M. (2012). 'Women are Bad at Math, but I'm Not, am I? ' Fragile Mathematical Self-concept Predicts Vulnerability to a Stereotype Threat Effect on Mathematical Performance. *European Journal of Personality, 26* (6), 588-599.

Gilliland, S. W. (1993). The Perceived Fairness of Selection Systems. An Organizational Justice Perspective. *The Academy of Management Review, 18* (4), 694.

Glaesmer, H., Hoyer, J., Klotsche, J. & Herzberg, P. Y. (2008). Die deutsche Version des Life-Orientation-Tests (LOT-R) zum dispositionellen Optimismus und Pessimismus. *Zeitschrift für Gesundheitspsychologie, 16* (1), 26-31.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R. et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40* (1), 84-96.

Gorsuch, R. L. (1966). The general factor in the test anxiety questionnaire. *Psychological Reports, 19* (1), 308.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A. & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109* (1), 3-25.

Greenwald, A. G. & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology, 79* (6), 1022-1038.

Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition. The implicit association test. *Journal of Personality and Social Psychology, 74* (6), 1464-1480.

Greiner, L. & Padtberg-Kruse, C. (2015, 11. Dezember). *Prüfungsangst: Acht Tipps gegen Blackout,* Spiegel Online. Zugriff am 04.01.2016. Verfügbar unter http://www.spiegel.de/schulspiegel/blackout-tipps-gegen-aussetzer-in-der-pruefung-a-1066860.html

Gross, J. J. (1998). The emerging field of emotion regulation. An integrative review. *Review of General Psychology, 2* (3), 271-299.

Gross, J. J. & Thompson, R. A. (2007). Emotion regulation: Conceptual foundations. In J. J. Gross (Hrsg.), *Handbook of emotion regulation* (S. 3-24). New York: Guilford Press.

Hadwin, J. A., Brogan, J. & Stevenson, J. (2005). State anxiety and working memory in children: A test of processing efficiency theory. *Educational Psychology, 25* (4), 379-393.

Hagtvet, K. A. & Benson, J. (1997). The motive to avoid failure and test anxiety responses: Empirical support for integration of two research traditions. *Anxiety, Stress & Coping, 10* (1), 35-57.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S. & Gernsbacher, M. A. (2007). The Science of Sex Differences in Science and Mathematics. *Psychological Science in the Public Interest, 8* (1), 1-51.

Hancock, D. R. (2001). Effects of Test Anxiety and Evaluative Threat on Students' Achievement and Motivation. *The Journal of Educational Research, 94* (5), 284-290.

Hannover, B. & Kessels, U. (2011). Sind Jungen die neuen Bildungsverlierer? Empirische Evidenz für Geschlechterdisparitäten zuungunsten von Jungen und Erklärungsansätze. *Zeitschrift für Pädagogische Psychologie, 25* (2), 89-103.

Hansen, J. & Wänke, M. (2009). Think of Capable Others and You Can Make It! Self-Efficacy Mediates the Effect of Stereotype Activation on Behavior. *Social Cognition, 27* (1), 76-88.

Harrison, L. A., Stevens, C. M., Monty, A. N. & Coakley, C. A. (2006). The consequences of stereotype threat on the academic performance of White and non-White lower income college students. *Social Psychology of Education, 9* (3), 341-357.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis. A regression-based approach* (Methodology in the social sciences). New York, NY: Guilford Press.

Heckhausen, J. (1991). Adults' expectancies about development and its controllability: Enhancing self-efficacy by social comparison. In R. Schwarzer (Ed.), *Self-Efficacy. Thought Control Of Action* (pp. 107-126). Washington, DC: Hemisphere.

Heinrich, D. L. & Spielberger, C. D. (1982). Anxiety and complex learning. In H. W. Krohne & L. Laux (Hrsg.), *Achievement, stress and anxiety* (S. 145-165). Washington, DC: Hemishpere.

Hembree, R. (1988). Correlates, Causes, Effects, and Treatment of Test Anxiety. *Review of Educational Research, 58* (1), 47-77.

Herbst, U., Voeth, M., Eidhoff, A. T., Müller, M. & Stief, S. (2016). *Studierendenstress in Deutschland – eine empirische Untersuchung. Herausgeber: AOK-Bundesverband,* Potsdam & Stuttgart.

Hermans, H., Petermann, F. & Zielinski, W. (1978). *Leistungs-Motivations-Test LMT*. Amsterdam: Swets & Zeitliner.

Herzer, F., Wendt, J. & Hamm, A. O. (2014). Discriminating Clinical From Nonclinical Manifestations of Test Anxiety: A Validation Study. *Behavior Therapy, 45* (2), 222-231.

Hess, T. M., Auman, C., Colcombe, S. J. & Rahhal, T. A. (2003). The Impact of Stereotype Threat on Age Differences in Memory Performance. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 58* (1), P3-P11.

Hewitt, P. L. & Flett, G. L. (1991). Perfectionism in the self and social contexts: Conceptualization, assessment, and association with psychopathology. *Journal of Personality and Social Psychology, 60* (3), 456-470.

Hewstone, M. & Martin, R. (2014). Sozialer Einfluss. In K. Jonas, W. Stroebe & M. Hewstone (Hrsg.), *Sozialpsychologie* (Springer-Lehrbuch, S. 269-313). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hiemisch, A., Westermann, R. & Michael, A. (2005). Die Abhängigkeit der Zufriedenheit mit dem Medizinstudium von Studienzielen und ihrer Realisierbarkeit. *Zeitschrift für Psychologie / Journal of Psychology, 213* (2), 97-108.

Hill, K. T. & Sarason, S. B. (1966). The relation of test anxiety and defensiveness to test and school performance over the elementary school years: A longitudinal study. *Monographs of the Society for Research in Child Development, 31* (2, Serial No. 104).

Hinz, A., Schumacher, J., Albani, C., Schmid, G. & Brähler, E. (2006). Bevölkerungsrepräsentative Normierung der Skala zur Allgemeinen Selbstwirksamkeitserwartung. *Diagnostica, 52* (1), 26-32.

Hippel, W. von, Hippel, C. von, Conway, L., Preacher, K. J., Schooler, J. W. & Radvansky, G. A. (2005). Coping with stereotype threat: denial as an impression management strategy. *Journal of Personality and Social Psychology, 89* (1), 22-35.

Hirt, E. R., Deppe, R. K. & Gordon, L. J. (1991). Self-reported versus behavioral self-handicapping. Empirical evidence for a theoretical distinction. *Journal of Personality and Social Psychology, 61* (6), 981-991.

Hittner, J. B., May, K. & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *The Journal of general psychology, 130* (2), 149-168.

Hodapp, V. (1989). Anxiety, fear of failure, and achievement: Two path-analytical models. *Anxiety Research, 1* (4), 301-312.

Hodapp, V. (1991). Das Prüfungsängstlichkeitsinventar TAI-G: Eine erweiterte und modifizierte Version mit vier Komponenten. *Zeitschrift für Pädagogische Psychologie, 5* (2), 121-130.

Hodapp, V. & Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress & Coping, 10* (3), 219-244.

Hodapp, V., Laux, L. & Spielberger, C. D. (1982). Theorie und Messung der emotionalen und kognitiven Komponente der Prüfungsangst. *Zeitschrift für Differentielle und Diagnostische Psychologie, 13,* 169-184.

Hodapp, V., Rohrmann, S. & Ringeisen, T. (2011). *Der Prüfungsangstfragebogen (PAF)*. Göttingen: Hogrefe.

Hong, E. (1998). Differential stability of individual differences in state and trait test anxiety. *Learning and Individual Differences, 10* (1), 51-69.

Hong, E. & Karstensson, L. (2002). Antecedents of State Test Anxiety. *Contemporary Educational Psychology, 27* (2), 348-367.

Hopko, D. R. (2003). Confirmatory Factor Analysis Of The Math Anxiety Rating Scale-Revised. *Educational and Psychological Measurement, 63* (2), 336-351.

Hopko, D. R., Crittendon, J. A., Grant, E. & Wilson, S. A. (2005). The impact of anxiety on performance IQ. *Anxiety, Stress & Coping, 18* (1), 17-35.

Hopko, D. R., Hunt, M. K. & Armento, M. E. (2005). Attentional Task Aptitude and Performance Anxiety. *International Journal of Stress Management, 12* (4), 389-408.

Hornke, L. F., Küppers, A. & Etzel, S. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests. *Diagnostica, 46* (4), 182-188.

Hyde, J. S., Fennema, E. & Lamon, S. J. (1990). Gender differences in mathematics performance. A meta-analysis. *Psychological Bulletin, 107* (2), 139-155.

International Test Commission. (2001). International Guidelines for Test Use. *International Journal of Testing, 1* (2), 93-114.

Irwin, J. R. & McClelland, G. H. (2003). Negative Consequences of Dichotomizing Continuous Predictor Variables. *Journal of Marketing Research, 40* (3), 366-371.

Isserstedt, W., Middendorff, E., Kandulla, M., Borchert, L. & Leszczensky, M. (2010). *Die wirtschaftliche und soziale Lage der Studierenden in der Bundesrepublik Deutschland 2009. 19. Sozialerhebung des Deutschen Studentenwerks*. Bonn, Berlin: HIS Hochschul-Informations-System.

Izard, C. E. (2009). Emotion theory and research: highlights, unanswered questions, and emerging issues. *Annual review of psychology, 60,* 1-25.

Izard, C. E. & Ackerman, B. P. (2000). Motivational, Organizational, and Regulatory Functions of Discrete Emotions. In M. Lewis & J. M. Haviland-Jones (Hrsg.), *Handbook of emotions* (S. 253-264). New York, NY: Guilford Press.

Jäger, A. O., Süß, H.-M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test (Form 4; BIS-4)*. Göttingen: Hogrefe.

Jamieson, J. P., Mendes, W. B., Blackstock, E. & Schmader, T. (2010). Turning the knots in your stomach into bows: Reappraising arousal improves performance on the GRE. *Journal of Experimental Social Psychology, 46* (1), 208-212.

Jamieson, J. P., Mendes, W. B. & Nock, M. K. (2012). Improving Acute Stress Responses: The Power of Reappraisal. *Current Directions in Psychological Science, 22* (1), 51-56.

Jamieson, J. P., Nock, M. K. & Mendes, W. B. (2012). Mind over matter: Reappraising arousal improves cardiovascular and cognitive responses to stress. *Journal of Experimental Psychology: General, 141* (3), 417-422.

Jamieson, J. P., Nock, M. K. & Mendes, W. B. (2013). Changing the Conceptualization of Stress in Social Anxiety Disorder: Affective and Physiological Consequences. *Clinical Psychological Science, 1* (4), 363-374.

Jamieson, J. P., Peters, B. J., Greenwood, E. J. & Altose, A. J. (2016). Reappraising Stress Arousal Improves Performance and Reduces Evaluation Anxiety in Classroom Exam Situations. *Social Psychological and Personality Science,* 1-9.

Jerusalem, M. & Schwarzer, R. (Hrsg.). (1986). *Skalen zur Befindlichkeit und Persönlichkeit*. Forschungsbericht Bd. 5. Berlin: Freie Universität, Institut für Psychologie.

Jerusalem, M. & Schwarzer, R. (Hrsg.). (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Berlin: Freie Universität Berlin.

Johns, M., Inzlicht, M. & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General, 137* (4), 691-705.

Jones, G. (1995). More than just a game. Research developments and issues in competitive anxiety in sport. *British Journal of Psychology, 86* (4), 449-478.

Jones, G. & Hanton, S. (2001). Pre-competitive feeling states and directional anxiety interpretations. *Journal of sports sciences, 19* (6), 385-395.

Jones, G., Hanton, S. & Swain, A. (1994). Intensity and interpretation of anxiety symptoms in elite and non-elite sports performers. *Personality and Individual Differences, 17* (5), 657-663.

Jones, M. V. & Uphill, M. (2004). Responses to the Competitive State Anxiety Inventory-2(d) by athletes in anxious and excited scenarios. *Psychology of Sport and Exercise, 5* (2), 201-212.

Kausch, T. J. (2013). *Geschlechtsunterschiede in der Selbstwertkontingenz*. Unveröffentlichte Bachelor-Thesis, Justus-Liebig-Universität Gießen. Gießen.

Keinan, G. & Zeidner, M. (1987). Effects of decisional control on state anxiety and achievement. *Personality and Individual Differences, 8* (6), 973-975.

Keith, N., Hodapp, V., Schermelleh-Engel, K. & Moosbrugger, H. (2003). Cross-sectional and longitudinal confirmatory factor models for the german test anxiety inventory: A construct validation. *Anxiety, Stress & Coping, 16* (3), 251-270.

Keller, J. (2007). When Negative Stereotypic Expectancies Turn Into Challenge or Threat: The Moderating Role of Regulatory Focus. *Swiss Journal of Psychology, 66* (3), 163-168.

Keller, J. & Dauenheimer, D. (2003). Stereotype Threat in the Classroom: Dejection Mediates the Disrupting Threat Effect on Women's Math Performance. *Personality and Social Psychology Bulletin, 29* (3), 371-381.

Kellogg, J. S., Hopko, D. R. & Ashcraft, M. H. (1999). The Effects of Time Pressure on Arithmetic Performance. *Journal of Anxiety Disorders, 13* (6), 591-600.

Keogh, E., Bond, F. W., French, C. C., Richards, A. & Davis, R. E. (2004). Test anxiety, susceptibility to distraction and examination performance. *Anxiety, Stress & Coping, 17* (3), 241-252.

Kernis, M. H. (2003). Toward a Conceptualization of Optimal Self-Esteem. *Psychological Inquiry, 14* (1), 1-26.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests. *Report Psychologie, 33,* 420-433.

Kersting, M., Althoff, K. & Jäger, A. O. (2008). *Wilde-Intelligenz-Test 2 (WIT-2)*. Göttingen: Hogrefe.

Kirkland, K. & Hollandsworth, J. G. (1980). Effective test taking. Skills-acquisition versus anxiety-reduction techniques. *Journal of Consulting and Clinical Psychology, 48* (4), 431-439.

Kirschbaum, C., Pirke, K.-M. & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' - A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting. *Neuropsychobiology, 28* (1-2), 76-81.

Klein, K. & Boals, A. (2001). Expressive writing can increase working memory capacity. *Journal of Experimental Psychology: General, 130* (3), 520-533.

Klinger, E. (1984). A consciousness-sampling analysis of test anxiety and performance. *Journal of Personality and Social Psychology, 47* (6), 1376-1390.

Köhler, C. (2015). *Der Einfluss des Stereotype Threat bei Schülerinnen auf die Leistungsangst und Testleistung bei einer numerischen Intelligenztestaufgabe*. Unveröffentlichte Master-Thesis, Justus-Liebig-Universität Gießen. Gießen.

Kramer, B. (2016). *Mythos und Wahrheit: Sind Jungen die neuen Verlierer?,* Spiegel Online. Zugriff am 24.08.2016. Verfügbar unter http://www.spiegel.de/schulspiegel/schlechtere-noten-als-maedchen-sind-jungen-schulverlierer-a-1059134.html

Krampen, G. (1988). Competence and control orientations as predictors of test anxiety in students: Longitudinal results. *Anxiety Research, 1* (3), 185-197.

Krampen, G. (1991). *Fragebogen zu Kompetenz- und Kontrollüberzeugungen (FKK)*. Göttingen: Hogrefe.

Kray, L. J., Thompson, L. & Galinsky, A. (2001). Battle of the sexes. Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology, 80* (6), 942-958.

Krinzinger, H., Kaufmann, L., Dowker, A., Thomas, G., Graf, M., Nuerk, H.-C. et al. (2007). Deutschsprachige Version des Fragebogens für Rechenangst (FRA) für 6- bis 9-jährige Kinder. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 35* (5), 341-351.

Krohne, H. W. & Hock, M. (2015). *Psychologische Diagnostik. Grundlagen und Anwendungsfelder* (Kohlhammer Standards Psychologie, 2., überarb. und aktualisierte Aufl.). Stuttgart: Kohlhammer.

Lamont, R. A., Swift, H. J. & Abrams, D. (2015). A review and meta-analysis of age-based stereotype threat. Negative stereotypes, not facts, do the damage. *Psychology and Aging, 30* (1), 180-193.

Landy, F. J. & Conte, J. M. (2013). *Work in the 21st century. An introduction to industrial and organizational psychology* (4. ed.). Hoboken, NJ: Wiley.

Lang, J. W. & Fries, S. (2006). A Revised 10-Item Version of the Achievement Motives Scale. *European Journal of Psychological Assessment, 22* (3), 216-224.

Lang, J. W. B. & Lang, J. (2010). Priming Competence Diminishes the Link Between Cognitive Test Anxiety and Test Performance: Implications for the Interpretation of Test Scores. *Psychological Science, 21* (6), 811-819.

Latsch, M. & Hannover, B. (2014). Smart Girls, Dumb Boys! ? *Social Psychology, 45* (2), 112-126.

Laux, L., Glanzmann, P., Schaffner, P. & Spielberger, C. D. (1981). *Das State-Trait-Angstinventar*. Göttingen: Beltz.

Lawrence, J. S. & Williams, A. (2013). Anxiety explains why people with domain-contingent self-worth underperform on ability-diagnostic tests. *Journal of Research in Personality, 47* (3), 227-232.

Lazarus, R. S. (1991a). *Emotion and adaptation*. New York: Oxford University Press.

Lazarus, R. S. (1991b). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist, 46* (8), 819-834.

Lazarus, R. S. (2001). Relational meaning and discrete emotions. In K. R. Scherer, A. Schorr & T. Johnstone (Hrsg.), *Appraisal processes in emotion. Theory, methods, research* (Series in affective science, S. 37-67).

Lazarus, R. S. (2006). *Stress and Emotion. A New Synthesis*. New York: Springer Publishing Company.

Lazarus, R. S. & Averill, J. R. (1972). Emotion and cognition: Wth special reference to anxiety. In C. D. Spielberger (Hrsg.), *Anxiety: Current trends in theory and research* (Vol. 2, S. 241-283). New York: Academic Press.

Lazarus, R. S. & Folkman, S. (1984). *Stress, appraisal, and coping*. New York: Springer Publ.

Lazarus, R. S. & Folkman, S. (1987). Transactional theory and research on emotions and coping. *European Journal of Personality, 1* (3), 141-169.

Lazarus, R. S. & Smith, C. A. (1988). Knowledge and Appraisal in the Cognition—Emotion Relationship. *Cognition & Emotion, 2* (4), 281-300.

Leary, M. R. & Shepperd, J. A. (1986). Behavioral self-handicaps versus self-reported handicaps. A conceptual note. *Journal of Personality and Social Psychology, 51* (6), 1265-1268.

Lehrl, S. (2005). *Manual zum MWT-B. Mehrfachwahl-Wortschatz-Intelligenztest* (5., unveränd. Aufl.). Balingen: Spitta-Verl.

Leininger, S. & Skeel, R. (2012). Cortisol and Self-report Measures of Anxiety as Predictors of Neuropsychological Performance. *Archives of Clinical Neuropsychology, 27* (3), 318-328.

Lenhard, W. & Lenhard, A. (2016). *Berechnung von Effektstärken*. verfügbar unter: https://www.psychometrica.de/effektstaerke.html. Bibergau: Psychometrica.

Leszczensky, M., Cordes, A., Kerst, C., Meister, T. & Wespel, J. (2013). *Bildung und Qualifikation als Grundlage der technologischen Leistungsfähigkeit Deutschlands. Bericht des Konsortiums „Bildungsindikatoren und technologische Leistungsfähigkeit“*. (HIS: Forum Hochschule 11 | 2013). Hannover: HIS Hochschul-Informations-System.

Liebert, R. M. & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports, 20* (3), 975-978.

Liepmann, D., Beauducel, A., Brocke, B. & Nettelnstroth, W. (2012). *IST-Screening*. Göttingen: Hogrefe.

MacLeod, C., Mathews, A. & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology, 95* (1), 15-20.

Mandler, G. & Cowen, J. E. (1958). Test Anxiety Questionnaires. *Journal of Consulting Psychology, 22* (3), 228-229.

Mandler, G. & Sarason, S. B. (1952). A study of anxiety and learning. *The Journal of Abnormal and Social Psychology, 47* (2), 166-173.

Manley, M. J. & Rosemier, R. A. (1972). Developmental trends in general and test anxiety among junior and senior high school students. *The Journal of genetic psychology, 120* (2d Half), 219-226.

Martens, R., Burton, D., Vealey, R. S., Bump, L. A. & Smith, D. E. (1990). Development and validation of the Competitive State Anxiety Inventory-2 (CSAI-2). In R. Martens, R. S. Vealey & D. Burton (Eds.), *Competitive anxiety in sport* (pp. 117-213). Champaign, Ill.: Human Kinetics Books.

Maslach, C., Schaufeli, W. B. & Leiter, M. P. (2001). Job burnout. *Annual review of psychology, 52,* 397-422.

Matthews, G., Hillyard, E. J. & Campbell, S. E. (1999). Metacognition and maladaptive coping as components of test anxiety. *Clinical Psychology & Psychotherapy, 6* (2), 111-125.

Matthews, G., Schwean, V. L., Campbell, S. E., Saklofske, D. H. & Mohamed, A. A. (2000). Personality, Self-Regulation, and Adaptation. In M. Boekaerts, Pintrich, Paul, R. & M. Zeidner (Hrsg.), *Handbook of Self-Regulation* (S. 171-207). New York: Academic Press.

Matthews, G. & Wells, A. (1999). The Cognitive Science of Attention and Emotion. In T. Dalgleish & M. J. Power (Hrsg.), *Handbook of Cognition and Emotion* (S. 171-192). Chichester, UK: John Wiley & Sons, Ltd.

Mavilidi, M.-F., Hoogerheide, V. & Paas, F. (2014). A Quick and Easy Strategy to Reduce Test Anxiety and Enhance Test Performance. *Applied Cognitive Psychology, 28* (5), 720-726.

Mayer, J. D., Roberts, R. D. & Barsade, S. G. (2008). Human abilities: emotional intelligence. *Annual review of psychology, 59,* 507-536.

Mayer, J. D., Salovey, P., Caruso, D. R. & Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion, 1* (3), 232-242.

McDonald, A. S. (2001). The Prevalence and Effects of Test Anxiety in School Children. *Educational Psychology, 21* (1), 89-101.

McIlroy, D. & Bunting, B. (2002). Personality, Behavior, and Academic Achievement. Principles for Educators to Inculcate and Students to Model. *Contemporary Educational Psychology, 27* (2), 326-337.

Meijer, J. & Oostdam, R. (2007). Test anxiety and intelligence testing: A closer examination of the stage-fright hypothesis and the influence of stressful instruction. *Anxiety, Stress & Coping, 20* (1), 77-91.

Meijer, J. & Oostdam, R. (2011). Effects of instruction and stage-fright on intelligence testing. *European Journal of Psychology of Education, 26* (1), 143-161.

Michaelis, L., Ott, M., Palmer, C., Ulfert, A.-S. & Kersting, M. (2013). *Gießener anforderungsanalytischer Fragebogen (GaF)*. Unveröffentlichter Fragebogen, Justus-Liebig-Universität Gießen. Gießen.

Michaelis, L., Ott, M., Palmer, C., Ulfert, A.-S., Kersting, M., Treiber, L. et al. (2014). *Skalendokumentation: IPIP Kurztest zur Erfassung der Persönlichkeit nach dem Fünf-Faktoren-Modell. Modifikation der dt. Version der 300-Item IPIP-Skala von Treiber, Thunsdorff, Schmitt & Schreiber, 2013,* Justus-Liebig-Universität Gießen. Gießen.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks. A Latent Variable Analysis. *Cognitive Psychology, 41* (1), 49-100.

Morris, L. W., Davis, M. A. & Hutchings, C. H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry-emotionality scale. *Journal of Educational Psychology, 73* (4), 541-555.

Morris, L. W. & Fulmer, R. S. (1976). Test anxiety (worry and emotionality) changes during academic testing as a function of feedback and test importance. *Journal of Educational Psychology, 68* (6), 817-824.

Morris, L. W. & Liebert, R. M. (1969). Effects of anxiety on timed and untimed intelligence tests. Another look. *Journal of Consulting and Clinical Psychology, 33* (2), 240-244.

Morris, L. W. & Liebert, R. M. (1973). Effects of negative feedback, threat of shock, and level of trait anxiety on the arousal of two components of anxiety. *Journal of Counseling Psychology, 20* (4), 321-326.

Mrazek, M. D., Chin, J. M., Schmader, T., Hartson, K. A., Smallwood, J. & Schooler, J. W. (2011). Threatened to distraction. Mind-wandering as a consequence of stereotype threat. *Journal of Experimental Social Psychology, 47* (6), 1243-1248.

Musch, J. & Bröder, A. (1999a). Psychometrische Eigenschaften und Validität des multidimensionalen Prüfungsängstlichkeitsinventars TAI-G. *Zeitschrift für Pädagogische Psychologie, 13* (1-2), 100-105.

Musch, J. & Bröder, A. (1999b). Test anxiety versus academic skills: A comparison of two alternative models for predicting performance in a statistics exam. *British Journal of Educational Psychology, 69* (1), 105-116.

National Council on Measurement in Education. (2015). *Glossary of Important Assessment and Measurement*. Verfügbar unter https://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061

Naveh-Benjamin, M., McKeachie, W. J. & Lin, Y.-g. (1987). Two types of test-anxious students. Support for an information processing model. *Journal of Educational Psychology, 79* (2), 131-136.

Nelson, D. W. & Knight, A. E. (2010). The Power of Positive Recollections: Reducing Test Anxiety and Enhancing College Student Efficacy and Performance. *Journal of Applied Social Psychology, 40* (3), 732-745.

Nguyen, H.-H. D. & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93* (6), 1314-1334.

Nie, Y., Lau, S. & Liau, A. K. (2011). Role of academic self-efficacy in moderating the relation between task importance and test anxiety. *Learning and Individual Differences, 21* (6), 736-741.

Nygård, R. & Gjesme, T. (2006). Assessment of Achievement Motives. Comments and Suggestions. *Scandinavian Journal of Educational Research, 17* (1), 39-46.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science (New York, N.Y.), 349* (6251).

Orne, M. T. (1962). On the social psychology of the psychological experiment. With particular reference to demand characteristics and their implications. *American Psychologist, 17* (11), 776-783.

Ortner, T. M. & Caspers, J. (2011). Consequences of Test Anxiety on Adaptive Versus Fixed Item Testing. *European Journal of Psychological Assessment, 27* (3), 157-163.

Osborne, J. W. (2001). Testing Stereotype Threat: Does Anxiety Explain Race and Sex Differences in Achievement? *Contemporary Educational Psychology, 26* (3), 291-310.

Papadogiannis, P. K., Logan, D. & Sitarenios, G. (2009). An Ability Model of Emotional Intelligence: A Rationale, Description, and Application of the Mayer Salovey Caruso Emotional Intelligence Test (MSCEIT). In C. Stough, D. H. Saklofske & J. D. A. Parker (Hrsg.), *Assessing Emotional Intelligence: Theory, Research, and Applications.* Boston, MA: Springer US.

Park, D., Ramirez, G. & Beilock, S. L. (2014). The role of expressive writing in math anxiety. *Journal of Experimental Psychology: Applied,* 103-111.

Parrot, W. G. (2002). The Functional Utility of Negative Emotions. In L. F. Barrett & P. Salovey (Eds.), *The wisdom in feeling. Psychological processes in emotional intelligence* (Emotions and social behavior, pp. 341-359). New York: Guilford Press.

Paulman, R. G. & Kennelly, K. J. (1984). Test anxiety and ineffective test taking: Different names, same construct? *Journal of Educational Psychology, 76* (2), 279-288.

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions. Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review, 18* (4), 315-341.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P. & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology, 36* (1), 36-48.

Pendry, L. (2014). Soziale Kognition. In K. Jonas, W. Stroebe & M. Hewstone (Hrsg.), *Sozialpsychologie* (Springer-Lehrbuch, S. 107-140). Berlin, Heidelberg: Springer Berlin Heidelberg.

Pennebaker, J. W. (1997). Writing about emotional experiencesas a therapeutic process. *Psychological Science, 8,* 162-166.

Pennebaker, J. W. & Chung, C. K. (2011). Expressive writing and its links to mental and physical health. In H. S. Friedman (Hrsg.), *Oxford handbook of health psychology.* New York, NY: Oxford University Press.

Pennington, C. R., Heim, D., Levy, A. R., Larkin, D. T. & Pavlova, M. A. (2016). Twenty Years of Stereotype Threat Research. A Review of Psychological Mediators. *PLOS ONE, 11* (1), e0146487.

Perry, J. D. & Williams, J. M. (1998). Relationship of Intensity and Direction of Competitive Trait Anxiety to Skill Level and Gender in Tennis. *The Sport Psychologist, 12,* 169-179.

Peterson, R. A. (2000). A Meta-Analysis of Variance Accounted for and Factor Loadings in Exploratory Factor Analysis. *Marketing Letters, 11* (3), 261-275.

Picho, K., Rodriguez, A. & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of females under stereotype threat: a meta-analysis. *The Journal of social psychology, 153* (3), 299-333.

Plake, B. S. & Parker, C. S. (1982). The Development and Validation of a Revised Version of the Mathematics Anxiety Rating Scale. *Educational and Psychological Measurement, 42* (2), 551-557.

Pohl, C. (2006). *Eine Validierungsstudie des Prüfungsängstlichkeitsinventars TAI-G an Realschülern mit Hilfe einer vorgestellten Situation*. Unveröffentlichte Diplomarbeit, Goethe-Universität Frankfurt a. M.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32* (1), 3-25.

Prins, P. J. M. & Hanewald, G. J. F. P. (1997). Self-statements of test-anxious children: Thought-listing and questionnaire approaches. *Journal of Consulting and Clinical Psychology, 65* (3), 440-447.

Putwain, D. W. (2008a). Deconstructing test anxiety. *Emotional and Behavioural Difficulties, 13* (2), 141-155.

Putwain, D. W. (2008b). Test anxiety and GCSE performance: the effect of gender and socio-economic background. *Educational Psychology in Practice, 24* (4), 319-334.

Putwain, D. W., Connors, L. & Symes, W. (2010). Do cognitive distortions mediate the test anxiety–examination performance relationship? *Educational Psychology, 30* (1), 11-26.

Putwain, D. W. & Daly, A. L. (2013). Do clusters of test anxiety and academic buoyancy differentially predict academic performance? *Learning and Individual Differences, 27,* 157-162.

Putwain, D. W. & Daniels, R. A. (2010). Is the relationship between competence beliefs and test anxiety influenced by goal orientation? *Learning and Individual Differences, 20* (1), 8-13.

Putwain, D. W. & Symes, W. (2012). Achievement goals as mediators of the relationship between competence beliefs and test anxiety. *British Journal of Educational Psychology, 82* (2), 207-224.

Raffety, B. D., Smith, R. E. & Ptacek, J. T. (1997). Facilitating and debilitating trait anxiety, situational anxiety, and coping with an anticipated stressor: A process analysis. *Journal of Personality and Social Psychology, 72* (4), 892-906.

Ramirez, G. & Beilock, S. L. (2011). Writing About Testing Worries Boosts Exam Performance in the Classroom. *Science, 331* (6014), 211-213.

Rammstedt, B. & Beierlein, C. (2014). Can't We Make It Any Shorter? *Journal of Individual Differences, 35* (4), 212-220.

Rapee, R. M. (1993). The utilisation of working memory by worry. *Behaviour Research and Therapy, 31* (6), 617-620.

Reeve, C. L. & Bonaccio, S. (2008). Does test anxiety induce measurement bias in cognitive ability tests? *Intelligence, 36* (6), 526-538.

Reeve, C. L., Heggestad, E. D. & Lievens, F. (2009). Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests. *Intelligence, 37* (1), 34-41.

Reips, U.-D. (2002). Theory and techniques of conducting Web experiments. In M. Batinic, U.-D. Reips & M. Bosnjak (Hrsg.), *Online Social Sciences* (S. 229-250). Seattle: Hogrefe & Huber.

Rheinberg, F. & Vollmeyer, R. (2003). Flow-Erleben in einem Computerspiel unter experimentell variierten Bedingungen. *Zeitschrift für Psychologie, 211* (4), 161-170.

Rheinberg, F., Vollmeyer, R. & Engeser, S. (2003). Die Erfassung des Flow-Erlebens. In J. Stiensmeier-Pelster & F. Rheinberg (Hrsg.), *Diagnostik von Selbstkonzept, Lernmotivation und Selbstregulation* (S. 261-279) [Tests und Trends, Jahrbuch der pädagogisch-psychologischen Diagnostik N.F., Band 2]. Göttingen: Hogrefe.

Richardson, M., Abraham, C. & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138* (2), 353-387.

Ringeisen, T. & Buchwald, P. (2010). Test anxiety and positive and negative emotional states during an examination. *Cognition, Brain, Behavior. An Interdisciplinary Journal, 14* (4), 431-447.

Ringeisen, T., Buchwald, P. & Hodapp, V. (2010). Capturing the multidimensionality of test anxiety in cross-cultural research: An english adaptation of the German Test Anxiety Inventory. *Cognition, Brain, Behavior. An Interdisciplinary Journal, 14* (4), 347-364.

Rohrmann, S., Bechtoldt, M., Schnell, K. & Hodapp, V. (2010). Validation of the German Test Anxiety Inventory by self-concept scales. *Cognition, Brain, Behavior. An Interdisciplinary Journal, 14* (4), 401-412.

Rosenberg, M. J. (1965). When dissonance fails. On eliminating evaluation apprehension from attitude measurement. *Journal of Personality and Social Psychology, 1* (1), 28-42.

Rost, D. H. & Schermer, F. J. (1989). »Reaktionsweisen gegenüber Tests« (RTT) und »Manifestationen von Leistungsangst« (DAI-MAN): una eademque res?, *10* (3), 169-179.

Rost, D. H. & Schermer, F. J. (2007). *Differentielles Leistungsangst Inventar DAI* (2., erw. Aufl.). Frankfurt a. M.: Harcourt Test Services.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied, 80* (1), 1-28.

Rotter, J. B. (1975). Some problems and misconceptions related to the construct of internal versus external control of reinforcement. *Journal of Consulting and Clinical Psychology, 43* (1), 56-67.

Rydell, R. J., McConnell, A. R. & Beilock, S. L. (2009). Multiple social identities and stereotype threat: imbalance, accessibility, and working memory. *Journal of Personality and Social Psychology, 96* (5), 949-966.

Sackett, P. R., Hardison, C. M. & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *The American psychologist, 59* (1), 7-13.

Sarason, I. (1956). Effect of anxiety, motivational instructions, and failure on serial learning. *Journal of Experimental Psychology, 51* (4), 253-260.

Sarason, I. G. (1958a). Effects on verbal learning of anxiety, reassurance, and meaningfulness of material. *Journal of Experimental Psychology, 56* (6), 472-477.

Sarason, I. G. (1958b). Interrelationships among individual difference variables, behavior in psychotherapy, and verbal conditioning. *The Journal of Abnormal and Social Psychology, 56* (3), 339-344.

Sarason, I. G. (1961). The effects of anxiety and threat on the solution of a difficult task. *The Journal of Abnormal and Social Psychology, 62* (1), 165-168.

Sarason, I. G. (1972). Experimental approaches to test anxiety: Attention and the uses of information. In C. D. Spielberger (Hrsg.), *Anxiety: Current trends in theory and research* (Vol. 2). New York: Academic Press.

Sarason, I. G. (1973). Test anxiety and social influence. *Journal of Personality, 41* (2), 261-271.

Sarason, I. G. (1978). The Test Anxiety Scale: Concept and research. In C. D. Spielberger & I. G. Sarason (Hrsg.), *Stress and anxiety* (Bd. 5). New York: Hemisphere/Wiley.

Sarason, I. G. (1980). Introduction to the Study of Test Anxiety. In I. G. Sarason (Ed.), *Test anxiety. Theory, research, and applications* (pp. 3-14). Hillsdale, N.J.: Erlbaum.

Sarason, I. G. (1981). Test anxiety, stress, and social support. *Journal of Personality, 49* (1), 101-114.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46* (4), 929-938.

Sarason, I. G. & Sarason, B. R. (1990). Test Anxiety. In H. Leitenberg (Ed.), *Handbook of social and evaluation anxiety* (pp. 475-495). New York: Plenum Pr.

Sarason, I. G., Sarason, B. R., Keefe, D. E., Hayes, B. E. & Shearin, E. N. (1986). Cognitive interference: Situational determinants and traitlike characteristics. *Journal of Personality and Social Psychology, 51* (1), 215-226.

Sarason, I. G., Sarason, B. R. & Pierce, G. R. (1990). Anxiety, Cognitive Interference, and Performance. *Journal of Social Behavior and Personality, 5* (1), 1-18.

Sarason, I. G. & Stoops, R. (1978). Test anxiety and the passage of time. *Journal of Consulting and Clinical Psychology, 46* (1), 102-109.

Sassenrath, J. M. (1964). A factor analysis of rating-scale item on the test anxiety questionnaire. *Journal of Consulting Psychology, 28* (4), 371-377.

Schachter, S. & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69* (5), 379-399.

Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C. J. & Rammstedt, B. (2013). BEFKI GC-K: Eine Kurzskala zur Messung kristalliner Intelligenz. *methoden, daten, analysen, 7,* 153-181.

Schmader, T. & Beilock, S. L. (2012). An Integration of Processes that Underlie Stereotype Threat. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat. Theory, process, and application* (pp. 34-50). New York: N.Y.; Oxford University Press.

Schmader, T. & Johns, M. (2003). Converging Evidence That Stereotype Threat Reduces Working Memory Capacity. *Journal of Personality and Social Psychology, 85* (3), 440-452.

Schmader, T., Johns, M. & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review, 115* (2), 336-356.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (5., vollständig überarbeitete und erweiterte Auflage). Berlin: Springer.

Schmidt-Atzert, L., Peper, M. & Stemmler, G. (2014). *Emotionspsychologie. Ein Lehrbuch* (Standards Psychologie, 2. Aufl.). Stuttgart: Kohlhammer Verlag.

Schmitt, D. P. & Allik, J. (2005). Simultaneous administration of the Rosenberg Self-Esteem Scale in 53 nations: exploring the universal and culture-specific features of global self-esteem. *Journal of personality and social psychology, 89* (4), 623-642.

Schmitt, D. P., Realo, A., Voracek, M. & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94* (1), 168-182.

Schnell, K., Tibubos, A. N., Rohrmann, S. & Hodapp, V. (2013). Test and Math Anxiety: A Validation of the German Test Anxiety Questionnaire. *Polish Psychological Bulletin, 44* (2).

Schöne, C. (2007). *Zielorientierung und Bezugsnormpräferenz in Lern- und Leistungssituationen*. Dissertation, Justus-Liebig-Universität Gießen. Gießen.

Schöne, C., Hermann, J. & Stiensmeier-Pelster, J. (in Vorb.). *Entwicklung und Überprüfung einer Skala zur Erfassung der Selbstwertkontingenz bei Studierenden (SESKON-ST)*. Manuskript in Vorbereitung.

Schwarz, N. (2012). Feelings-as-information theory. In P. van Lange, A. W. Kruglanski & E. T. Higgins (Eds.), *Theories of social psychology* (pp. 289-308). Los Angeles, Calif.: Sage.

Schwarzer, R. (2000). *Streß, Angst und Handlungsregulation* (4., überarb. Aufl.). Stuttgart: Kohlhammer.

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4* (1), 27-41.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z. et al. (2013). Priming Intelligent Behavior. An Elusive Phenomenon. *PLoS ONE, 8* (4), e56515.

Shapiro, J. R. (2011). Different groups, different threats: a multi-threat approach to the experience of stereotype threats. *Personality & social psychology bulletin, 37* (4), 464-480.

Shapiro, J. R. & Neuberg, S. L. (2007). From Stereotype Threat to Stereotype Threats: Implications of a Multi-Threat Framework for Causes, Moderators, Mediators, Consequences, and Interventions. *Personality and Social Psychology Review, 11* (2), 107-130.

Shih, M., Ambady, N., Richeson, J. A., Fujita, K. & Gray, H. M. (2002). Stereotype performance boosts: The impact of self-relevance and the manner of stereotype activation. *Journal of Personality and Social Psychology, 83* (3), 638-647.

Shih, M. J., Pittinsky, T. L. & Ho, G. C. (2011). Stereotype Boost Positive Outcomes from the Activation of Positive Stereotypes. In M. Inzlicht & T. Schmader (Eds.), *Stereotype ThreatTheory, Process, and Application* (pp. 141-156). Oxford University Press.

Sieber, J. E., O'Neil, H. F. & Tobias, S. (1977). *Anxiety Learning and Instruction*. Hillsdale, N.J.: Erlbaum.

Slaney, R. B., Rice, K. G., Mobley, M., Trippi, J. & Ashby, J. S. (2001). The Revised Almost Perfect Scale. *Measurement and Evaluation in Counseling and Development, 34* (3), 130-145.

Smith, J. L. (2006). The Interplay among Stereotypes, Performance-Avoidance Goals, and Women's Math Performance Expectations. *Sex Roles, 54* (3-4), 287-296.

Smith, J. L. & Johnson, C. S. (2006). A Stereotype Boost or Choking Under Pressure? Positive Gender Stereotypes and Men Who Are Low in Domain Identification. *Basic and Applied Social Psychology, 28* (1), 51-63.

Smith, J. L. & White, P. H. (2002). An Examination of Implicitly Activated, Explicitly Activated, and Nullified Stereotypes on Mathematical Performance: It's Not Just a Woman's Issue. *Sex Roles, 47* (3/4), 179-191.

Smith, R. J., Arnkoff, D. B. & Wright, T. L. (1990). Test anxiety and academic competence: A comparison of alternative models. *Journal of Counseling Psychology, 37* (3), 313-321.

Smith, T. W., Snyder, C. R. & Handelsman, M. M. (1982). On the self-serving function of an academic wooden leg: Test anxiety as a self-handicapping strategy. *Journal of Personality and Social Psychology, 42* (2), 314-321.

Snyder, C. R., Smith, T. W., Augelli, R. W. & Ingram, R. E. (1985). On the self-serving function of social anxiety. Shyness as a self-handicapping strategy. *Journal of Personality and Social Psychology, 48* (4), 970-980.

Sommer, M. & Arendasy, M. E. (2014). Comparing different explanations of the effect of test anxiety on respondents' test scores. *Intelligence, 42,* 115-127.

Sommer, M. & Arendasy, M. E. (2015). Further evidence for the deficit account of the test anxiety–test performance relationship from a high-stakes admission testing setting. *Intelligence, 53,* 72-80.

Sparfeldt, J. R., Rost, D. H., Baumeister, U. M. & Christ, O. (2013). Test anxiety in written and oral examinations. *Learning and Individual Differences, 24,* 198-203.

Sparfeldt, J. R., Schilling, S. R., Rost, D. H., Stelzl, I. & Peipert, D. (2005). Leistungsängstlichkeit: Facetten, Fächer, Fachfacetten? *Zeitschrift für Pädagogische Psychologie, 19* (4), 225-236.

Spencer, S. J., Steele, C. M. & Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology, 35* (1), 4-28.

Spielberger, C. D. (1972a). Anxiety as an emotional state. In C. D. Spielberger (Hrsg.), *Anxiety: Current trends in theory and research* (Vol. 1, S. 23-49). New York: Academic Press.

Spielberger, C. D. (1972b). Conceptual and methodological issues in anxiety research. In C. D. Spielberger (Hrsg.), *Anxiety: Current trends in theory and research* (Vol. 2, S. 481-493). New York: Academic Press.

Spielberger, C. D. (1980). *Preliminary professional manual for the Test Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., Gonzalez, H. P., Taylor, C. J., Algaze, B. & Anton, W. D. (1978). Examination stress and test anxiety. In C. D. Spielberger & I. G. Sarason (Hrsg.), *Stress and anxiety* (Bd. 5, S. 167-191). New York: Hemisphere/Wiley.

Spielberger, C. D. & Vagg, P. R. (Hrsg.). (1995a). *Test Anxiety. Theory, Assessment, and Treatment*. Washington, DC: Taylor & Francis.

Spielberger, C. D. & Vagg, P. R. (1995b). Test anxiety: A transactional process model. In C. D. Spielberger & P. R. Vagg (Hrsg.), *Test Anxiety. Theory, Assessment, and Treatment* (S. 3-14). Washington, DC: Taylor & Francis.

Stangor, C., Carr, C. & Kiang, L. (1998). Activating stereotypes undermines task performance expectations. *Journal of Personality and Social Psychology, 75* (5), 1191-1197.

Statistisches Bundesamt. (2015). *Allgemeinbildende Schulen - Fachserie 11 Reihe 1 - Schuljahr 2014/2015*. Wiesbaden.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52* (6), 613-629.

Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69* (5), 797-811.

Stingel, A. (2009). *Die Validierung des Prüfungsängstlichkeitsinventars TAI-G mit Hilfe des Differentiellen Leistungsangstinventars (DAI).* Unveröffentlichte Diplomarbeit, Goethe-Universität Frankfurt a. M.

Stöber, J. & Esser, K. B. (2001). Test anxiety and metamemory: general preference for external over internal information storage. *Personality and Individual Differences, 30* (5), 775-781.

Stoeber, J., Feast, A. R. & Hayward, J. A. (2009). Self-oriented and socially prescribed perfectionism: Differential relationships with intrinsic and extrinsic motivation and test anxiety. *Personality and Individual Differences, 47* (5), 423-428.

Stoet, G. & Geary, D. C. (2012). Can stereotype threat explain the gender gap in mathematics performance and achievement? *Review of General Psychology, 16* (1), 93-102.

Stone, J. & McWhinnie, C. (2008). Evidence that blatant versus subtle stereotype threat cues impact performance through dual processes. *Journal of Experimental Social Psychology, 44* (2), 445-452.

Stowell, J. R. & Bennett, D. (2010). Effects of Online Testing on Student Exam Performance and Test Anxiety. *Journal of Educational Computing Research, 42* (2), 161-171.

Strack, J. & Esteves, F. (2014). Exams? Why worry? Interpreting anxiety as facilitative and stress appraisals. *Anxiety, stress, and coping,* 1-10.

Strack, J., Lopes, P. N. & Esteves, F. (2014). Will you thrive under pressure or burn out? Linking anxiety motivation and emotional exhaustion. *Cognition and Emotion,* 1-14.

Strohbeck-Kühner, P. (1999). Testangst bei Fahreignungsbegutachtungen: Die Angst-Leistung-Relation. *Zeitschrift für Differentielle und Diagnostische Psychologie, 20* (1), 39-57.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18* (6), 643-662.

Suhr, L. & Döpfner, M. (2000). Leistungs- und Prüfungsängste bei Kindern und Jugendlichen - Ein multimodales Therapiekonzept. *Kindheit und Entwicklung, 9* (3), 171-186.

Tabachnick, B. G. & Fidell, L. S. (2010). *Using multivariate statistics* (5. ed., Pearson internat. ed.). Boston, Mass.: Pearson/Allyn and Bacon.

Tamir, M. & Ford, B. Q. (2012). When feeling bad is expected to be good: emotion regulation and outcome expectancies in social conflicts. *Emotion (Washington, D.C.), 12* (4), 807-816.

Tandler, S., Schwinger, M., Kaminski, K. & Stiensmeier-Pelster, J. (2014). Self-Affirmation Buffers Claimed Self-Handicapping? A Test of Contextual and Individual Moderators. *Psychology, 05* (05), 321-327.

Techniker Krankenkasse. (2015). *TK-CampusKompass. TK-Studie zur Gesundheit und Mediennutzung von Studierenden*. Hamburg.

Thompson, T. (1996). Self-worth Protection in Achievement Behaviour. A Review and Implications for Counselling. *Australian Psychologist, 31* (1), 41-47.

Thompson, T. & Dinnel, D. L. (2003). Construction and initial validation of the self-worth protection scale. *The British journal of educational psychology, 73* (Pt 1), 89-107.

Tobias, S. (1985). Test Anxiety: Interference, Defective Skills, and Cognitive Capacity. *Educational Psychologist, 20* (3), 135-142.

Treiber, L. (2013). *Entwicklung, Psychometrische Überprüfung und konvergente Validierung der deutschsprachigen 30-Facetten-IPIP-Skala,* Universität Koblenz-Landau. Koblenz-Landau.

Tuschen, E. (2014). *Effekte evaluativer und nicht-evaluativer Instruktion auf die Leistungsangst und Testleistung bei einer numerischen Intelligenztestaufgabe*. Unveröffentlichte Master-Thesis, Justus-Liebig-Universität Gießen. Gießen.

Ulfert, A.-S., Ott, M., Michaelis, L. & Kersting, M. (2014a). *Gießener kognitiver Kompetenztest (GkKT). Verfahrenshinweise. Version 1.0: Stand 01. Dez. 2014*. Unveröffentlichtes Manuskript, Justus-Liebig-Universität Gießen. Gießen.

Ulfert, A.-S., Ott, M., Michaelis, L. & Kersting, M. (2014b). *Gießener kognitiver Kompetenztest-Kurzversion (GkKT-K)*. Unveröffentlichtes Manuskript, Justus-Liebig-Universität Gießen. Gießen.

Urban, D. & Mayerl, J. (2011). *Regressionsanalyse: Theorie, Technik und Anwendung* (Studienskripten zur Soziologie, 4., überabeitete und erweiterte Auflage). Wiesbaden: VS Verl. für Sozialwiss.

Van Yperen, N. W. (2007). Performing well in an evaluative situation: The roles of perceived competence and task-irrelevant interfering thoughts. *Anxiety, Stress & Coping, 20* (4), 409-419.

Vancouver, J. B., More, K. M. & Yoder, R. J. (2008). Self-efficacy and resource allocation: Support for a nonmonotonic, discontinuous model. *Journal of Applied Psychology, 93* (1), 35-47.

Wacker, A., Jaunzeme, J. & Jaksztat, S. (2008). Eine Kurzform des Prüfungsängstlichkeitsinventars TAI-G. *Zeitschrift für Pädagogische Psychologie, 22* (1), 73-81.

Walton, G. M. & Cohen, G. L. (2003). Stereotype Lift. *Journal of Experimental Social Psychology, 39* (5), 456-467.

Walton, G. M. & Spencer, S. J. (2009). Latent ability: grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science, 20* (9), 1132-1139.

Walton, G. M., Spencer, S. J. & Erman, S. (2013). Affirmative Meritocracy. *Social Issues and Policy Review, 7* (1), 1-35.

Ware, W. B., Galassi, J. P. & Dew, K. M. H. (1990). The test anxiety inventory: A confirmatory factor analysis. *Anxiety Research, 3* (3), 205-212.

Weiner, B. A. & Carton, J. S. (2012). Avoidant coping: A mediator of maladaptive perfectionism and test anxiety. *Personality and Individual Differences, 52* (5), 632-636.

Wells, A. & Matthews, G. (1996). Modelling cognition in emotional disorder. The S-REF model. *Behaviour Research and Therapy, 34* (11-12), 881-888.

Wheeler, S. C. & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127* (6), 797-826.

Wicherts, J. M. & Zand Scholten, A. (2010). Test anxiety and the validity of cognitive tests: A confirmatory factor analysis perspective and some empirical findings. *Intelligence, 38* (1), 169-178.

Wieczerkowski, W., Nickel, H., Janowski, A., Fittkau, B. & Rauer, W. (1975). *Angsfragebogen für Schüler (AFS)*. Braunschweig: FRG: Westermann.

Wigfield, A. & Eccles, J. S. (1989). Test Anxiety in Elementary and Secondary School Students. *Educational Psychologist, 24* (2), 159-183.

Wigfield, A. & Meece, J. L. (1988). Math anxiety in elementary and secondary school students. *Journal of Educational Psychology, 80* (2), 210-216.

Williams, J. M. G., Mathews, A. & MacLeod, C. (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin, 120* (1), 3-24.

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76* (2), 92-104.

Wine, J. D. (1980). Cognitive-Attentional Theory of Test Anxiety. In I. G. Sarason (Ed.), *Test anxiety. Theory, research, and applications* (pp. 349-385). Hillsdale, N.J.: Erlbaum.

Wise, S. L., Roos, L. L., Plake, B. S. & Nebelsick-Gullett, L. J. (1994). The Relationship between Examinee Anxiety and preference for Self-Adapted Testing. *Applied Measurement in Education, 7* (1), 81-91.

Wonderlic Inc. (1996). *Wonderlic Personal Test (WPT - German Version)*. Libertyville, IL: Wonderlic Personnel Test, Inc.

Wong, S. S. (2008). The Relations of Cognitive Triad, Dysfunctional Attitudes, Automatic Thoughts, and Irrational Beliefs with Test Anxiety. *Current Psychology, 27* (3), 177-191.

Wood, J. V. (1996). What is Social Comparison and How Should We Study it? *Personality and Social Psychology Bulletin, 22* (5), 520-537.

Yeager, D. S. & Walton, G. M. (2011). Social-Psychological Interventions in Education: They're Not Magic. *Review of Educational Research, 81* (2), 267-301.

Yerkes, R. M. & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18* (5), 459-482.

Yogo, M. & Fujihara, S. (2008). Working memory capacity can be improved by expressive writing: A randomized experiment in a Japanese sample. *British Journal of Health Psychology, 13* (1), 77-80.

Zecharia, A. (2015). *Viewpoint: Everyone must fight sexism in science,* BBC. Zugriff am 26.08.2016. Verfügbar unter http://www.bbc.com/news/magazine-33096157

Zeidner, M. (1991). Test anxiety and aptitude test performance in an actual college admissions testing situation: Temporal considerations. *Personality and Individual Differences, 12* (2), 101-109.

Zeidner, M. (1994). Personal and contextual determinants of coping and anxiety in an evaluative situation. A prospective study. *Personality and Individual Differences, 16* (6), 899-918.

Zeidner, M. (1995). Personality trait correlates of intelligence. In D. H. Saklofske & M. Zeidner (Hrsg.), *International Handbook of Personality and Intelligence* (S. 299-319). Boston, MA: Springer US.

Zeidner, M. (1998). *Test anxiety. The state of the art* (Perspectives on individual differences). New York: Plenum Press.

Zeidner, M. & Matthews, G. (2007). Evaluation Anxiety. In A. J. Elliot (Ed.), *Handbook of competence and motivation* (pp. 141-166). New York, NY: Guilford Press.

Zeidner, M. & Matthews, G. (2010). *Anxiety 101* (Psych 101). New York, NY: Springer.

Zoske, S. (2015, 21. Januar). *„Ich dachte, ich bin nichts mehr wert". Prüfungsangst im Studium,* Frankfurter Allgemeine Zeitung. Zugriff am 04.01.2015. Verfügbar unter http://www.faz.net/aktuell/rhein-main/frankfurt/pruefungsangst-im-studium-ich-dachte-ich-bin-nichts-mehr-wert-13379965.html

Zuckerman, M. & Tsai, F.-F. (2005). Costs of Self-Handicapping. *Journal of Personality, 73* (2), 411-442.

Anhang

A. Abkürzungsverzeichnis

AM:	anxiety motivation
AU:	Aufgeregtheit
BE:	Besorgtheit
FFM:	Fünf-Faktoren-Modell
GPA:	grade point average
IN:	Interferenz
Kurt:	Kurtosis
MA:	Mangelnde Anstrengung
MQ:	Messqualität
MZ:	Mangel an Zuversicht
Pbn:	Proband
Sch:	Schiefe
SKEF:	Selbstkonzept eigener Fähigkeiten
STT:	stereotype threat
SW:	Selbstwert
SWK:	Selbstwertkontingenz
TÄ / trait-TÄ:	Testängstlichkeit (trait)
TA / state-TA:	Testangst (state)
Var:	Varianz

B. Verfahren

Tabelle B-1: Fragebogen zu Testängstlichkeit (trait)

Nr.	Item	Skala
01	Ich mache mir Sorgen, ob ich auch alles schaffe.	BE
02	Ich bin zuversichtlich, was meine Leistung betrifft.	MZ
03	Ich denke über die Konsequenzen eines möglichen Misserfolges nach.	BE
04	Ich frage mich, ob meine Leistung ausreicht.	BE
05	Ich denke an andere Dinge und werde dadurch abgelenkt.	IN
06	Ich fühle mich unbehaglich.	AU
07	Ich weiß, dass ich mich auf mich selbst verlassen kann.	MZ
08	Mir schießen plötzlich Gedanken durch den Kopf, die mich blockieren.	IN
09	Das Herz schlägt mir bis zum Hals.	AU
10	Ich mache mir Gedanken über mein Abschneiden.	BE
11	Ich fühle mich ängstlich.	AU
12	Ich werde in meinem Gedankengang unterbrochen, weil mir etwas Nebensächliches einfällt.	IN
13	Ich habe ein beklemmendes Gefühl.	AU
14	Ich denke daran, was passiert, wenn ich schlecht abschneide.	BE
15	Ich bin überzeugt, dass ich gut abschneiden werde.	MZ

Antwortskala: 1 = fast nie; 2 = manchmal; 3 = oft; 4 = fast immer
AU = Aufgeregtheit; BE = Besorgtheit; IN = Interferenz; MZ = Mangel an Zuversicht

Tabelle B-2: Fragebögen zu Testangst (state)

Nr.	Item	Skala
	Studie 1 (state-TA prä)	
01	Ich bin angespannt.	AU
02	Ich bin zuversichtlich, was meine Leistung betrifft.	MZ
03	Ich bin aufgeregt.	AU
04	Ich denke an andere Dinge und werde dadurch abgelenkt.	IN
05	Ich bin besorgt, dass etwas schiefgehen könnte.	BE
06	Ich weiß, dass ich mich auf mich selbst verlassen kann.	MZ
07	Ich bin beunruhigt.	BE
08	Mir schießen plötzlich Gedanken durch den Kopf, die mich blockieren.	IN
09	Ich bin nervös.	AU
10	Ich werde in meinem Gedankengang unterbrochen, weil mir etwas Nebensächliches einfällt.	IN
11	Ich bin überzeugt, dass ich gut abschneiden werde.	MZ
	Studie 1 (state-TA post)	
01	Ich war angespannt.	AU
02	Ich war zuversichtlich, was meine Leistung betrifft.	MZ
03	Ich war aufgeregt.	AU
04	Ich dachte an andere Dinge und wurde dadurch abgelenkt.	IN
05	Ich war besorgt, dass etwas schiefgehen könnte.	BE
06	Ich wusste, dass ich mich auf mich selbst verlassen kann.	MZ
07	Ich war beunruhigt.	BE
08	Mir schossen plötzlich Gedanken durch den Kopf, die mich blockierten.	IN
09	Ich war nervös.	AU
10	Ich wurde in meinem Gedankengang unterbrochen, weil mir etwas Nebensächliches einfiel.	IN
11	Ich war überzeugt, dass ich gut abschneiden werde.	MZ
	Studie 2	
01	Ich mache mir Sorgen, ob ich auch alles schaffe.	BE
02	Ich bin zuversichtlich, was meine Leistung betrifft.	MZ
03	Ich denke über die Konsequenzen eines möglichen Misserfolges nach.	BE
04	Ich frage mich, ob meine Leistung ausreicht.	BE
05	Ich denke an andere Dinge und werde dadurch abgelenkt.	IN
06	Ich fühle mich unbehaglich.	AU
07	Ich weiß, dass ich mich auf mich selbst verlassen kann.	MZ
08	Mir schießen plötzlich Gedanken durch den Kopf, die mich blockieren.	IN
09	Das Herz schlägt mir bis zum Hals.	AU
10	Ich mache mir Gedanken über mein Abschneiden.	BE
11	Ich fühle mich ängstlich.	AU
12	Ich werde in meinem Gedankengang unterbrochen, weil mir etwas Nebensächliches einfällt.	IN
13	Ich habe ein beklemmendes Gefühl.	AU
14	Ich denke daran, was passiert, wenn ich schlecht abschneide.	BE
15	Ich bin überzeugt, dass ich gut abschneiden werde.	MZ
	Studie 3	
01	Ich war angespannt.	AU
02	Ich war aufgeregt.	AU
03	Ich war besorgt, dass etwas schiefgehen könnte.	BE
04	Ich war beunruhigt.	BE
05	Ich war nervös.	AU

Antwortskala Studie 1: 1 = überhaupt nicht; 2 = ein wenig; 3 = ziemlich; 4 = sehr
Antwortskala Studie 2: 1 = fast nie; 2 = manchmal; 3 = oft; 4 = fast immer
Antwortskala Studie 3: 1 = überhaupt nicht; 2 = ein wenig; 3 = ziemlich; 4 = sehr
AU = Aufgeregtheit; BE = Besorgtheit; IN = Interferenz; MZ = Mangel an Zuversicht

Tabelle B-3: Fragebogen zu Leistungszielen

Nr.	Item	Skala
	Beim Bearbeiten des Tests …	
01	... ist es für mich wichtig, besser zu sein als andere.	Annäherung
02	... will ich vermeiden, im Vergleich zu Anderen schlechte Leistungen zu haben.	Vermeidung
03	... ist es für mich wichtig, im Vergleich zu den anderen gut zu sein.	Annäherung
04	... treibt mich meine Angst vor schlechter Leistung, im Vergleich zu anderen, an.	Vermeidung
05	... möchte ich nur verhindern, etwas schlechter zu machen als andere.	Vermeidung
06	... ist es mir wichtig, eine bessere Bewertung als die meisten anderen zu erhalten.	Annäherung

Antwortskala: 1 = stimmt gar nicht; 7 = stimmt ganz genau

Tabelle B-4: Fragebogen zu Selbstwert

Nr.	Item
01	Alles in allem bin ich mit mir selbst zufrieden.
02	Hin und wieder denke ich, daß ich gar nichts tauge. (R)
03	Ich besitze eine Reihe guter Eigenschaften.
04	Ich kann vieles genauso gut wie die meisten anderen Menschen auch.
05	Ich fürchte, es gibt nicht viel, worauf ich stolz sein kann. (R)
06	Ich fühle mich von Zeit zu Zeit richtig nutzlos. (R)
07	Ich halte mich für einen wertvollen Menschen, jedenfalls bin ich nicht weniger wertvoll als andere auch.
08	Ich wünschte, ich könnte vor mir selbst mehr Achtung haben. (R)
09	Alles in allem neige ich dazu, mich für einen Versager zu halten. (R)
10	Ich habe eine positive Einstellung zu mir selbst gefunden.

Antwortskala: 0 = trifft gar nicht zu; 3 = trifft voll und ganz zu

Tabelle B-5: Fragebogen zu Selbstwertkontingenz

Nr.	Item	Skala
01	Mein Selbstwertgefühl leidet stark darunter, wenn andere in der Uni bessere Beurteilungen für ihre Leistung erhalten als ich.	LK
02	Selbst wenn es auffällt, dass ich etwas schlechter beherrsche als andere, bleiben meine Selbstwertgefühle davon völlig unberührt. (R)	KK
03	Meine Selbstwertgefühle sind stark davon abhängig, wie ich meine Leistung in der Uni einschätze.	LK
04	Selbst wenn ich merke, dass andere eine schnellere Auffassungsgabe haben als ich, bleibt mein Selbstwertgefühl davon völlig unberührt. (R)	KK
05	Ich fühle mich minderwertig, wenn andere merken, dass ich etwas nicht gut beherrsche.	KK
06	Wie wertvoll ich mich fühle wird stark davon beeinflusst, wie andere meine Leistung beurteilen.	LK
07	Wenn andere in der Uni bessere Leistungen erbringen als ich, fühle ich mich schnell minderwertig.	LK
08	Wenn ich zeigen kann, dass ich etwas besser beherrsche als andere, fühle ich mich wesentlich wertvoller.	KK
09	Selbst wenn andere merken, dass ich etwas nicht gut beherrsche, bleibt mein Selbstwert davon völlig unberührt. (R)	KK
10	Mein Selbstwertgefühl verringert sich deutlich, wenn andere meine Leistung in der Uni kritisieren.	LK
11	Wenn es mir schwer fällt etwas zu lernen oder zu verstehen, fühle ich mich deswegen nicht weniger wertvoll. (R)	KK
12	Ich fühle mich minderwertig, wenn mir jemand etwas mehrfach erklären muss, weil ich es permanent nicht verstehe.	KK
13	Mein Selbstwertgefühl wird stark davon beeinflusst, wenn andere merken, dass ich mit einer Aufgabe überfordert bin.	KK

Antwortskala: 0 = trifft gar nicht zu; 3 = trifft voll und ganz zu;
LK = Leistungskontingenz; KK = Kompetenzkontingenz

Tabelle B-6: Fragebogen zu Selbstkonzept eigener Fähigkeiten

Nr.	Item
01	Ich halte meine Begabung im Bereich Intelligenz für – niedrig / hoch
02	Neues zu lernen fällt mir – schwer / leicht
03	Meiner Meinung nach bin ich – nicht intelligent / sehr intelligent
04	Meine Fähigkeiten im logischen Denken sind – niedrig / hoch
05	Aufgaben die logisches Denken erfordern fallen mir – schwer / leicht

Antwortskala jeweils von 1 bis 7

Tabelle B-7: Fragebogen zu Anxiety Motivation

Nr.	Item	Skala
	Wenn ich während einer Prüfung oder einem Test Angst empfinde	
01	…verleiht mir das mehr Schwung.	Energie
02	…erinnert mich das daran, dass ich mich konzentrieren muss.	Information
03	…richte ich meine Energie auf die Prüfung / den Test.	Energie
04	…macht mich das darauf aufmerksam, dass ich mir Mühe geben muss.	Information
05	…spornt mich das an.	Energie
06	…bin ich aktiver bei der Problemlösung.	Energie
07	…gibt mir das ein klares Signal dafür, dass ich mich mehr anstrengen muss.	Information

Antwortskala: 1 = stimme nicht zu, 2 = stimme eher nicht zu, 3 = stimme teils/teils zu, 4 = stimme eher zu, 5 = stimme völlig zu

Tabelle B-8: Fragebogen zu Self-Handicapping

Nr.	Item
	Haben die folgenden Aspekte Ihre Leistung beeinträchtigt?
01	Müdigkeit
02	Schlechte Laune
03	Nervosität
04	Ablenkung durch Umgebung
05	Krankheit/ körperliches Unwohlsein
06	Stress im Alltag
07	Mangelnde Anstrengung/ Faulheit
08	Prüfungsangst
09	Mangelhaftigkeit des Tests

Antwortskala: 1 = trifft nicht zu; 3 = trifft teilweise zu; 5 = trifft zu

Tabelle B-9: Fragebogen zu Akzeptanz

Nr.	Item	Skala
01	Die Testaufgaben waren klar und verständlich.	KO
02	Mit dem Test kann man die hinsichtlich des getesteten Merkmals bestehenden Unterschiede präzise abbilden.	MQ
03	Die Testaufgaben spiegeln Anforderungen wider, die auch im Berufsleben gefordert sind.	AV
04	Bei der Testung fühlte ich mich überfordert.	BF
05	Dass man mit dem Test geeignete Personen für einen Job herausfinden kann, ist zu bezweifeln.	AV
06	Der Test misst das, was er misst, zuverlässig.	MQ
07	Ich habe die Aufgabenstellung nicht verstanden.	KO
08	Die Bearbeitung der Testaufgaben ist belastend.	BF
09	Bei der Bearbeitung der Testaufgaben wusste ich jederzeit, was ich tun muss.	KO
10	Ob jemand bei den Testaufgaben gut abschneidet oder im Beruf gut ist, das sind zwei völlig verschiedene Dinge.	AV
11	Der Test ermöglicht es, die zwischen verschiedenen Menschen bestehenden Leistungsunterschiede in der vom Test erfassten Fähigkeit exakt zu messen.	MQ
12	Die Testaufgaben waren überwiegend zu schwer für mich.	BF
13	Die Testaufgaben haben zu wenig mit der Realität zu tun, um wirklich Berufserfolg vorherzusagen.	AV
14	Die Bearbeitung der Testaufgaben ist anstrengend.	BF
15	Ich habe die Testaufgaben nicht verstanden.	KO
16	Die Auswertung des Tests kann einen zutreffenden Eindruck von den Fähigkeiten einer Person vermitteln.	MQ
18#	Welche Schulnote würden Sie dem soeben bearbeiteten Intelligenztest geben?	
19#	Im Vergleich mit anderen Personen meiner Altersgruppe (mit gleicher Schulbildung) denke ich, dass ich im Intelligenztest ..... abgeschnitten habe.	

Antwortskala: 1 = trifft nicht zu; 6 = trifft genau zu; #Antwortskala: 1 = sehr gut, 6 = ungenügend
KO = Kontrollierbarkeit; MQ = Messqualität; AV = Augenscheinvalidität; BF = Belastungsfreiheit

Tabelle B-10: Fragebogen zu Flow

Nr.	Item	Skala
01	Ich fühlte mich optimal beansprucht.	Absorbiertheit
02	Meine Gedanken bzw. Aktivitäten liefen flüssig und glatt.	Glatter Verlauf
03	Ich merkte gar nicht, wie die Zeit vergeht.	Absorbiertheit
04	Ich hatte keine Muhe, mich zu konzentrieren.	Glatter Verlauf
05	Mein Kopf war völlig klar.	Glatter Verlauf
06	Ich war ganz vertieft in das, was ich gerade machte.	Absorbiertheit
07	Die richtigen Gedanken/Bewegungen kamen wie von selbst.	Glatter Verlauf
08	Ich wusste bei jedem Schritt, was ich zu tun hatte.	Glatter Verlauf
09	Ich hatte das Gefühl, den Ablauf unter Kontrolle zu haben.	Glatter Verlauf
10	Ich war völlig selbstvergessen.	Absorbiertheit
11	Es stand für mich Wichtiges auf dem Spiel.	Besorgnis
12	Ich durfte keine Fehler machen.	Besorgnis
13	Ich machte mir Sorgen über einen Misserfolg.	Besorgnis

Antwortskala: 1 = trifft nicht zu; 4 = teils-teils; 7 = trifft zu

Tabelle B-11: Fragebogen zu Persönlichkeitseigenschaften nach dem Fünf-Faktoren-Modell

Nr.	Item	Skala
01	Aufgaben bearbeite ich bis ins letzte Detail.	G
02	Ich bin geschickt in sozialen Interaktionen.	E
03	Ich mache mehr, als von mir verlangt wird.	G
04	Ich interessiere mich für anspruchsvolle, fachübergreifende Literatur.	O
05	Ich bin sehr selbstkritisch.	N
06	Ich bin an abstrakten Ideen nicht interessiert. (R)	O
07	Ich lasse mich leicht stressen.	N
08	Ich habe häufig Stimmungsschwankungen.	N
09	Ich halte mich eher im Hintergrund. (R)	E
10	Ich lege für jeden ein gutes Wort ein.	V
11	Für meinen eigenen Vorteil nehme ich das schlechtere Abschneiden anderer in Kauf. (R)	V
12	Philosophische Diskussionen vermeide ich. (R)	O
13	Meinen Aufgaben widme ich mich mit viel Disziplin und Ausdauer.	G
14	Ich gerate leicht in Panik.	N
15	Ich mag es nicht, Aufmerksamkeit auf mich zu ziehen. (R)	E
16	Aufgaben beginne ich stets rechtzeitig.	G
17	Ich kann Leute begeistern.	E
18	Ich respektiere andere Menschen.	V
19	Ich fühle mich oft unglücklich.	N
20	Ich biete stets meine Hilfe an, wenn jemand ein Problem hat.	V
21	Ungewöhnlichen Ideen stehe ich offen gegenüber.	O
22	Ich freunde mich leicht mit anderen an.	E
23	Ich akzeptiere andere Menschen, so wie sie sind.	V
24	Ich bin stets vorbereitet.	G
25	Ich bringe Unterhaltungen auf ein höheres Niveau.	O
26	Ich nehme mir Zeit für andere.	V
27	Ich kann andere von meinen Ideen und Auffassungen überzeugen.	E
28	Oft fühle ich mich traurig.	N
29	Meine Pläne führe ich aus.	G

Antwortskala: 1 = trifft überhaupt nicht zu; 2 = trifft eher nicht zu; 3 = trifft teils/teils zu; 4 = trifft eher zu; 5 = trifft vollständig zu

G = Gewissenhaftigkeit; E = Extraversion; O = Offenheit für Erfahrungen; N = Neurotizismus; V = Verträglichkeit

Tabelle B-12: Fragebogen zu Allgemeiner Selbstwirksamkeit

Nr.	Item
01	Wenn sich Widerstände auftun, finde ich Mittel und Wege, mich durchzusetzen.
02	Die Lösung schwieriger Probleme gelingt mir immer, wenn ich mich darum bemühe.
03	Es bereitet mir keine Schwierigkeiten, meine Absichten und Ziele zu verwirklichen.
04	In unerwarteten Situationen weiß ich immer, wie ich mich verhalten soll.
05	Auch bei überraschenden Ereignissen glaube ich, dass ich gut mit ihnen zurechtkommen kann.
06	Schwierigkeiten sehe ich gelassen entgegen, weil ich meinen Fähigkeiten immer vertrauen kann.
07	Was auch immer passiert, ich werde schon klarkommen.
08	Für jedes Problem kann ich eine Lösung finden.
09	Wenn eine neue Sache auf mich zukommt, weiß ich, wie ich damit umgehen kann.
10	Wenn ein Problem auftaucht, kann ich es aus eigener Kraft meistern.

Antwortskala: 1 = trifft überhaupt nicht zu; 2 = trifft eher nicht zu; 3 = trifft teils/teils zu; 4 = trifft eher zu; 5 = trifft vollständig zu

C. Weiterführende Informationen zum Online-Fragebogen in Studie 1

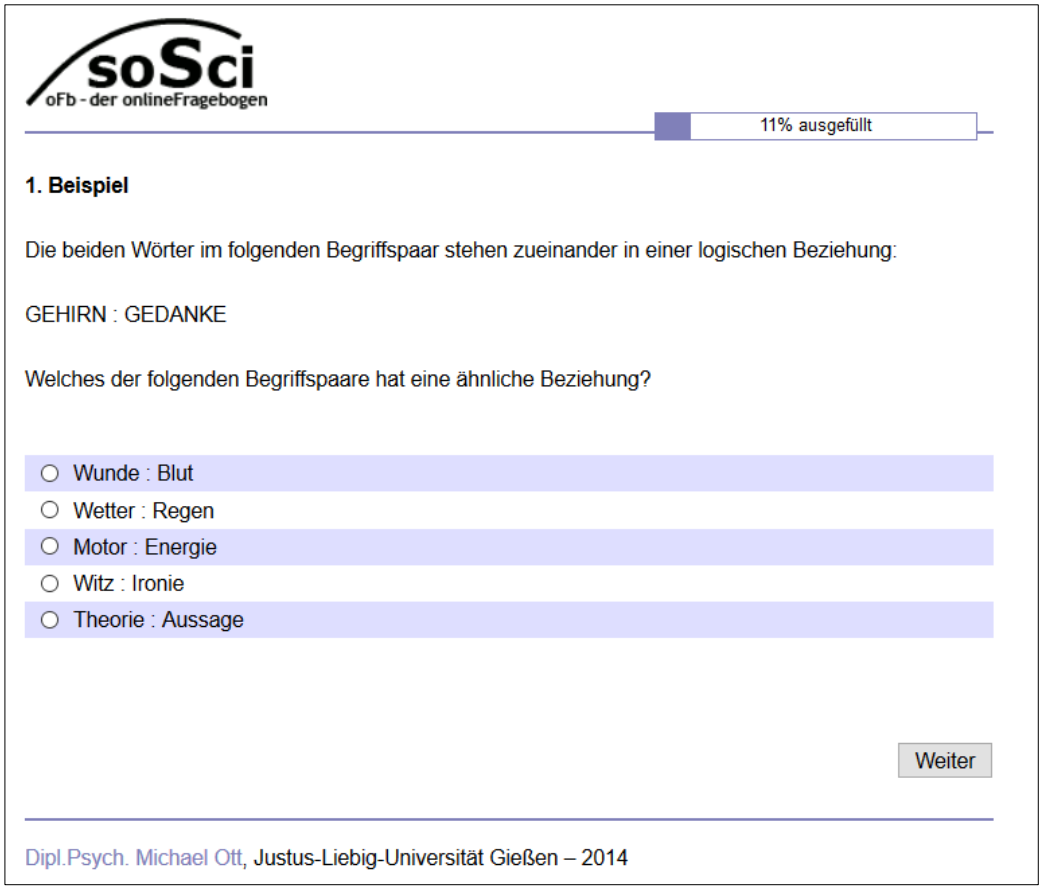

Abbildung C-1: Erstes Beispielitem für den GkKT-K in Studie 1

Tabelle C-1: E-Mail zur Rekrutierung der Probanden in Studie 1

Sehr geehrte Studierende, im Rahmen meiner Doktorarbeit in der Abteilung Psychologische Diagnostik führe ich eine Studie zum Thema "Selbstbild und Motivation" durch. Wie verhalten wir uns in Leistungssituationen? Wann sind wir motiviert? Und welche Rolle spielt unser Selbstbild dabei? Diesen Fragen möchte ich mit einer 30-minütigen Online-Studie nachgehen - dafür benötige ich Ihre Hilfe. Was haben Sie von der Teilnahme? 1. Sie unterstützen die Forschung! 2. Sie haben die Chance, einen von zwei Amazon-Gutscheinen (im Wert von 30€ und 20€) zu gewinnen Selbstverständlich werden die erhobenen Daten anonym behandelt. Sie gelangen über folgenden Link zum Fragebogen: https://www.soscisurvey.de/studie_tamp/ Vielen Dank für Ihre Unterstützung! Mit freundlichen Grüßen Michael Ott

Analyse des Dropout in Studie 1

Von den 354 Personen, die den Fragebogen aufgerufen haben, haben 140 (40%) bereits auf den ersten vier Seiten des Fragebogens abgebrochen. Auf diesen Seiten waren die Begrüßung und die ersten drei Fragebögen (die Skalen zu Selbstwert, Testängstlichkeit und Selbstwertkontingenz) enthalten. Ein hoher Anteil an Abbrüchen zu Beginn eines Fragebogens ist für Online-Erhebungen durchaus typisch (siehe hierzu Reips, 2002). Bis zur Gruppenzuweisung haben insgesamt 176 Teilnehmer abgebrochen (50%). Der während bzw. nach der Manipulation erfolgte drop-out verteilt sich relativ gleichmäßig auf die beiden Bedingungen mit 15 Abbrüchen in Bedingung A (mit prä-Messung) und 11 Abbrüchen in Bedingung B (ohne prä-Messung).

D. Testankündigungen

Studie	Ankündigungstext
Studie 1	Der folgende Test erfasst Ihre Intelligenz. Ihre individuelle Intelligenzleistung wird in der Auswertung mit der von anderen Versuchsteilnehmern verglichen. Nach dem Test erhalten Sie auf Wunsch eine ausführliche Rückmeldung über Ihre Intelligenzleistung.
Studie 3	Nun kommen einige Aufgaben zum schlussfolgernden Denken. Diese Aufgaben sind immer in Blöcken dargestellt. Es geht los bei Block A, fährt fort mit Block B usw. Jeder Block enthält eine Reihe von kurzen Aufgaben. Sie erhalten auf Wunsch eine Rückmeldung zu Ihrem Ergebnis, anhand der Sie Ihre Leistung mit der der anderen Teilnehmer/innen vergleichen können (Infos dazu am Ende des Fragebogens). Wir werden bei jedem Block die Zeit stoppen. Dies tun wir erstens damit die Ergebnisse besser interpretierbar sind und zweitens um die Erhebung für Sie kürzer zu gestalten. Wir stoppen die Zeit, sobald Sie mit einem Block beginnen. Jeder Block hat eine kurze Erläuterung und danach folgen Aufgaben. Wir werden Ihnen jeweils mitteilen, wenn die Zeit um ist. Bitte arbeiten Sie dann nicht mehr weiter.

Die Testinstruktionen für Studie 2 sind in Abschnitt 5.1.5 aufgeführt.

E. Diagramme

Dargestellt sind die Streudiagramme für die Zusammenhänge von Leistung und Testangst in Studie 1 und 2. Die Streudiagramme für Studie 3 sind in Abschnitt 6.2.2.2 abgebildet. Für Studie 3 ist darüber hinaus der Screetest für die Hauptkomponentenanalyse zu Anxiety Motivation dargestellt.

1. Studie 1

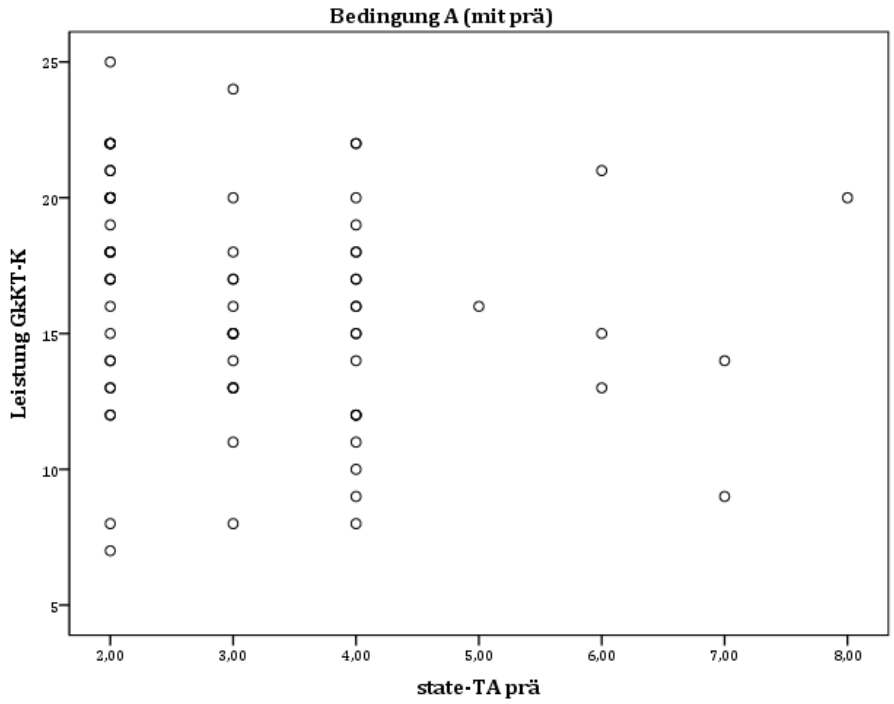

Abbildung E-1: Streudiagramm des Zusammenhangs von state-TA prä und Leistung (r = -.25); N = 74*

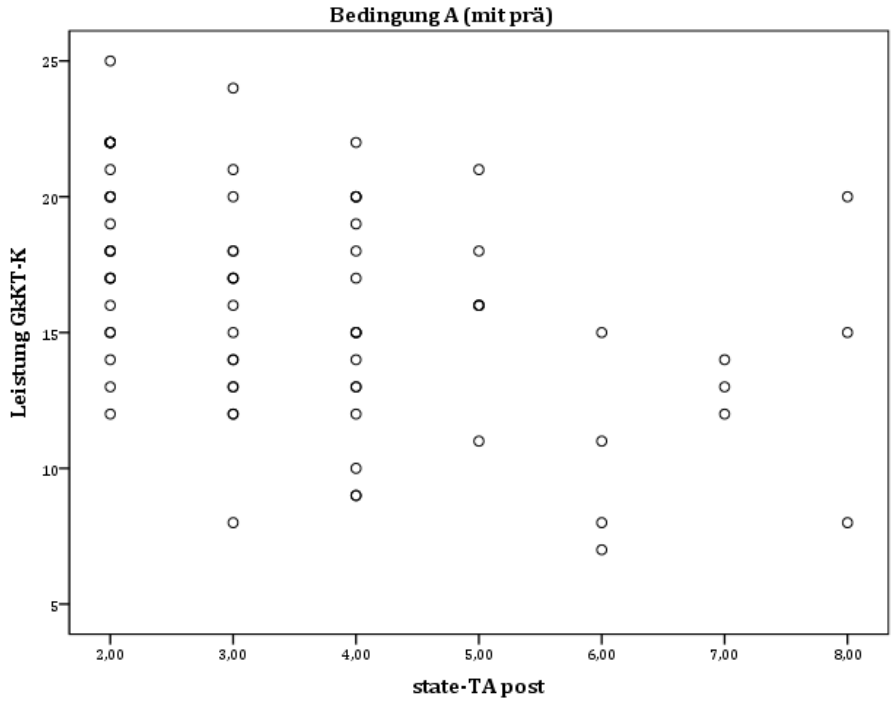

*Abbildung E-2: Streudiagramm des Zusammenhangs von state-TA post und Leistung in Bedingung A (r = -.43**); N = 78*

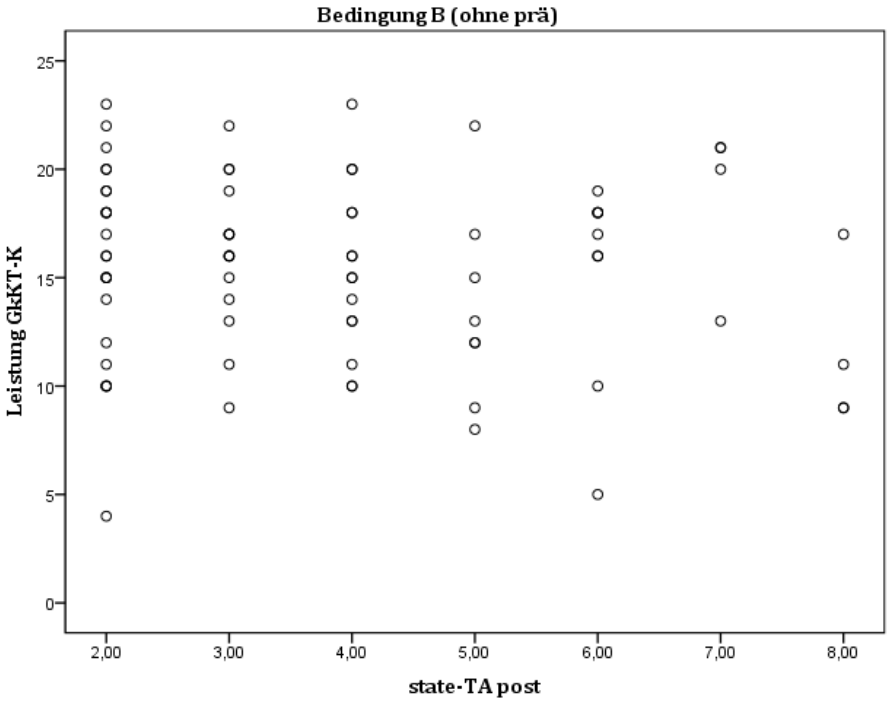

Abbildung E-3: Streudiagramm des Zusammenhangs von state-TA post und Leistung in Bedingung B (r = -.10); N = 78

2. Studie 2

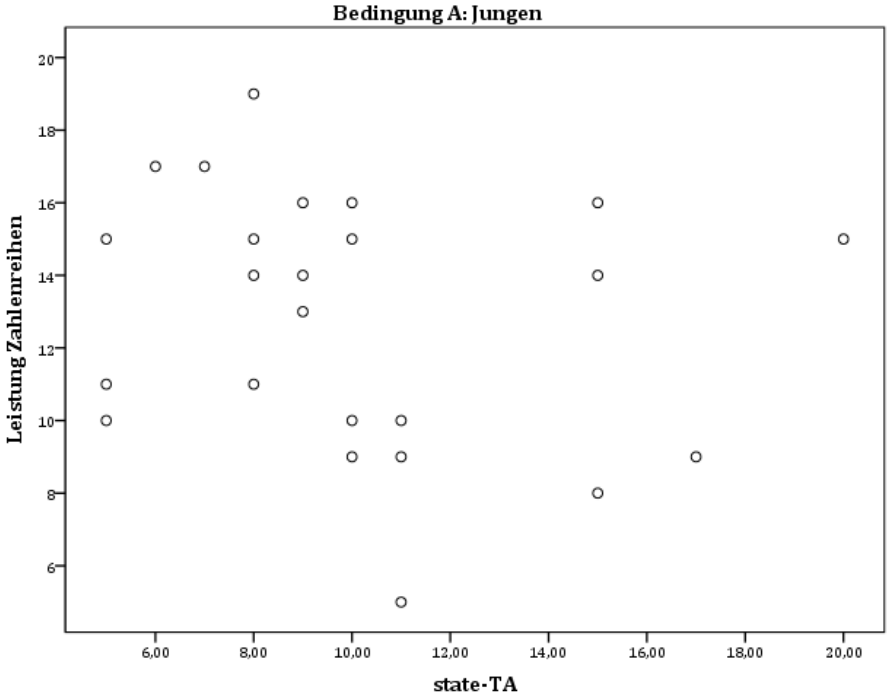

Abbildung E-4: Streudiagramm des Zusammenhangs von state-TA und Leistung bei Jungen in Bedingung A (r = -.35); N = 24

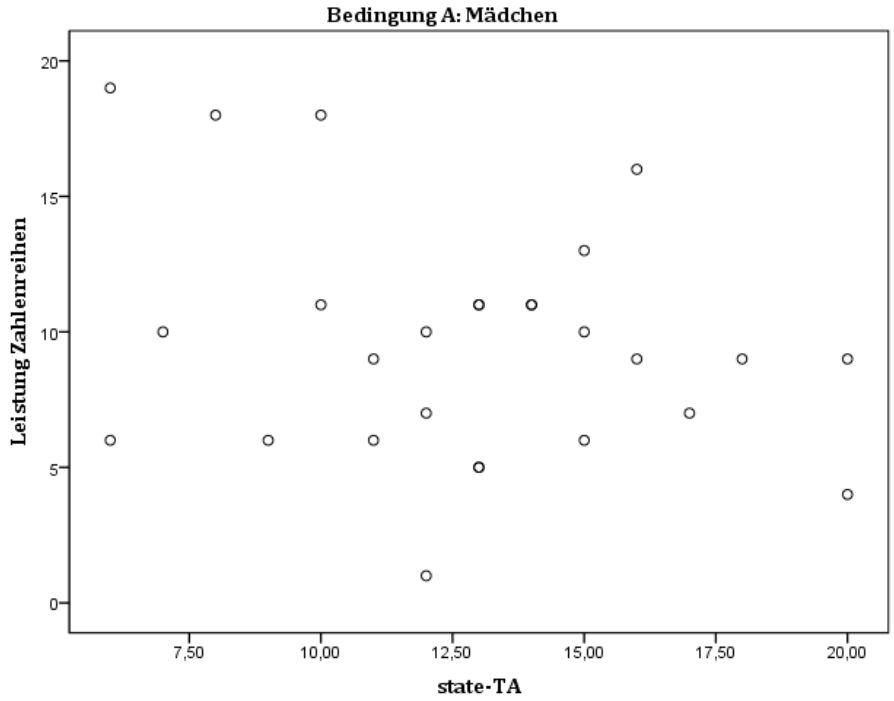

Abbildung E-5: Streudiagramm des Zusammenhangs von state-TA und Leistung bei Jungen in Bedingung A (r = -.17); N = 27

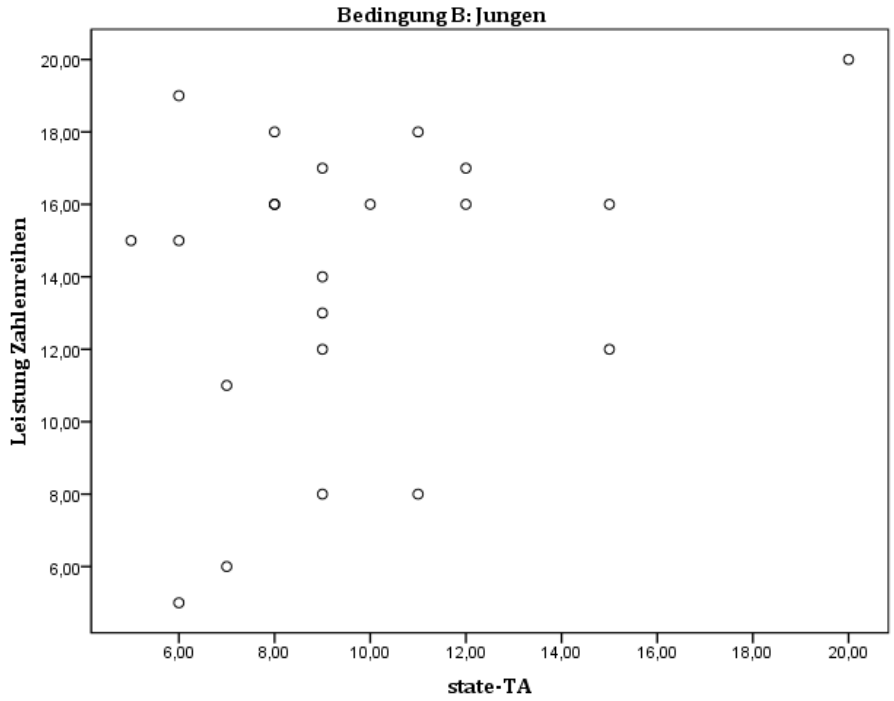

Abbildung E-6: Streudiagramm des Zusammenhangs von state-TA und Leistung bei Mädchen in Bedingung B (r = .26); N = 22

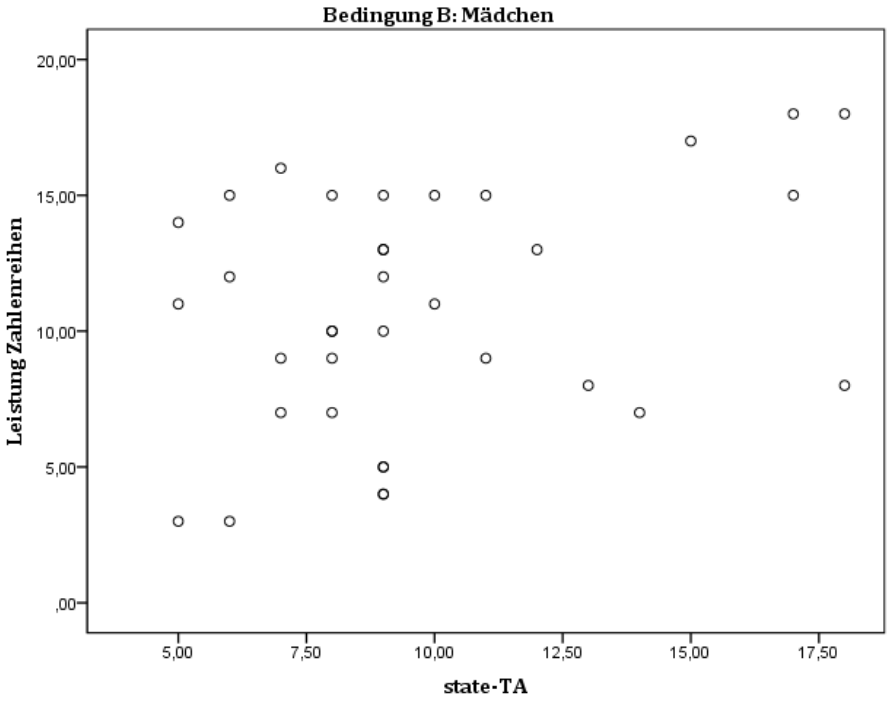

Abbildung E-7: Streudiagramm des Zusammenhangs von state-TA und Leistung bei Mädchen in Bedingung B (r = .26); N = 35

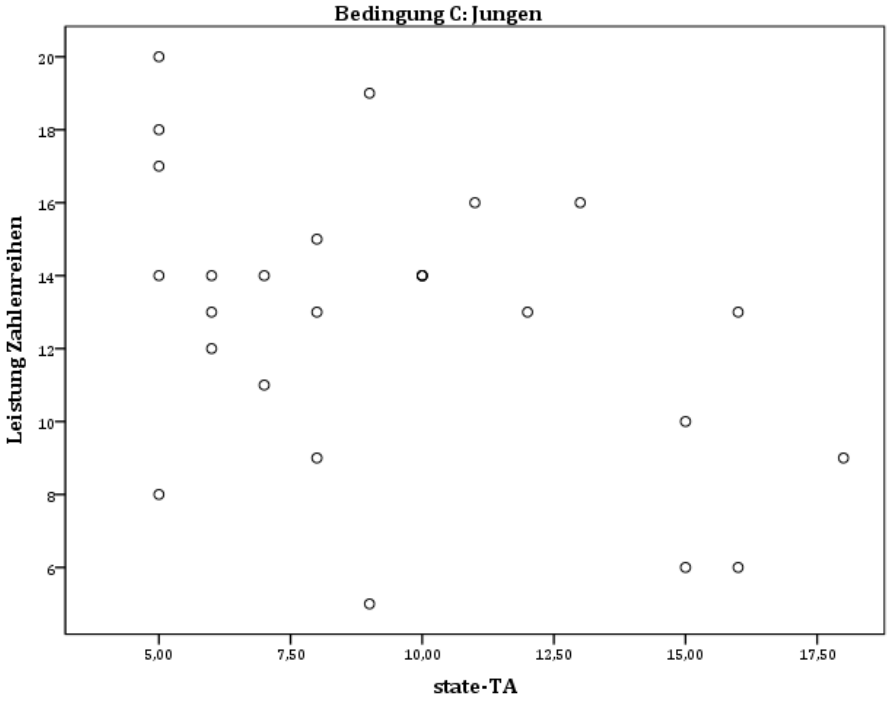

Abbildung E-8: Streudiagramm des Zusammenhangs von state-TA und Leistung bei Jungen in Bedingung C (r = -.36); N = 26

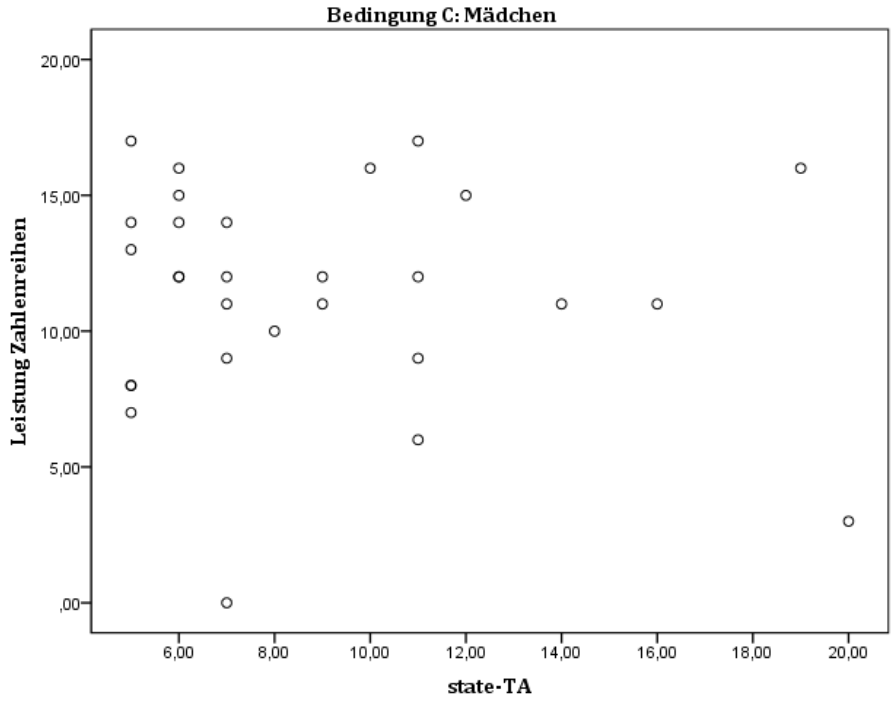

Abbildung E-9: Streudiagramm des Zusammenhangs von state-TA und Leistung bei Jungen in Bedingung C (r = -.06); N = 29

3. Studie 3

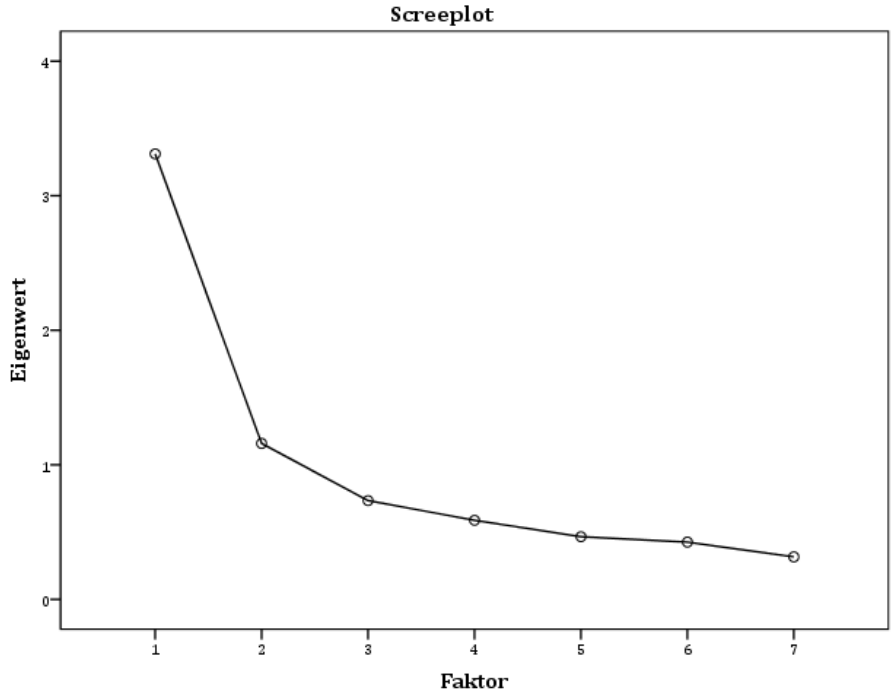

Abbildung E-10: Screeplot der Hauptkomponentenanalyse zu Anxiety Motivation

F. Voraussetzungsprüfungen

Aufgrund der Verletzung der Normalverteilungsannahme bei den meisten Variablen wurden die Korrelationsanalysen nonparametrisch durchgeführt (siehe Abschnitt 3). Im Anschluss werden für jede der drei Studien die Ergebnisse der Tests auf Normalverteilung bei den zentralen Variablen geschildert.

Wie in Abschnitt 3 berichtet, wurden bei den multiplen Regressionsanalysen Prüfungen auf Multikollinearität, Normalverteilung der Residuen und Homoskedastizität vorgenommen. Im Folgenden werden die Ergebnisse dieser Analysen (insbesondere Auffälligkeiten) behandelt. Die Multikollinearität wurde anhand der in Abschnitt 3 genannten Kennwerte geprüft, wobei für die Toleranz eine Untergrenze von 0.25 und den VIF eine Obergrenze von 5.00 gewählt wurde (Urban & Mayerl, 2011). Die Prüfung auf Normalverteilung der Residuen wurde durch eine optische Inspektion der Histogramme der standardisierten und studentisierten Residuen sowie des Probability-Probability-Plot (P-P-Plot) vorgenommen. Ebenfalls optisch überprüft wurde die Annahme der Homoskedastizität mittels Residuenplots (Eid et al., 2013). Zur Prüfung der Voraussetzungen wurden entsprechend der jeweils in PROCESS berechneten Modelle multiple Regressionen mit den Prädiktoren und den Interaktionstermen (Produkttermen) berechnet. Da PROCESS nur unstandardisierte Koeffizienten ausgibt, werden jeweils, zusätzlich zur Angabe von Toleranz und VIF, die standardisierten Regressionskoeffizienten mitberichtet. Um die Übersichtlichkeit der Darstellung zu wahren, werden jeweils nur das Histogramm der studentisierten Residuen sowie der Residuenplot je Analyse berichtet.

1. Studie 1

Tabelle F-1: Normalverteilungsprüfungen für zentrale Variablen von Studie 1

	K-S-Test	
	Z	*p*
Selbstwert	.12	.000
Testängstlichkeit (trait)		
Besorgtheit	.10	.001
Aufgeregtheit	.11	.000
Interferenz	.15	.000
Mangel an Zuversicht	.13	.000
Gesamt	.10	.002
Selbstwertkontingenz (Gesamt)	.05	.200
Selbstkonzept eigener Fähigkeiten	.10	.002
Testangst (state – prä)		
Besorgtheit	.22	.000
Aufgeregtheit	.17	.000
Interferenz	.26	.000
Mangel an Zuversicht	.18	.000
Gesamt	.13	.002
Leistung GkKT-K	.08	.030
Testangst (state – post)		
Besorgtheit	.19	.000
Aufgeregtheit	.17	.000
Interferenz	.21	.000
Mangel an Zuversicht	.15	.000
Gesamt	.09	.006
Akzeptanz		
Kontrollierbarkeit	.17	.000
Messqualität	.08	.029
Augenscheinvalidität	.08	.015
Belastungsfreiheit	.09	.005
Note Verfahren	.26	.000
Subjektive Leistung	.26	.000
Self-Handicapping-Aspekte		
Müdigkeit	.17	.000
Schlechte Laune	.32	.000
Nervosität	.22	.000
Ablenkung durch Umgebung	.24	.000
Krankheit / körperliches Unwohlsein	.37	.000
Stress	.23	.000
Mangelnde Anstrengung	.24	.000
Prüfungsangst	.35	.000
Mangelhaftigkeit des Tests	.36	.000

Z = Teststatistik des K-S-Test

Bezüglich der Voraussetzung der Mediationsanalyse (Hypothese 2a) wurden die Voraussetzung für die regressionsanalytische Vorhersage von Y (state-TA post), nicht aber von M (subjektive Leistung), geprüft, da es sich dabei um das komplexere Modell handelte. In Tabelle F-2 sind die Kennwerte zur Prüfung der Multikollinearität aufgeführt. Demnach kann Multikollinearität ausgeschlossen werden. Aus den Abbildungen F-1 bis F-5 ist zu erkennen, dass tendenziell leichte Verletzungen der Normalverteilungen bei den Residuen gegeben sind. Gravierende Abweichungen in Form von Boden- oder Deckeneffekten sind allerdings nicht beobachtbar. Außerdem ist erkennbar, dass Hinweise auf Heteroskedastizität vorliegen. Heteroskedastizität führt nicht zu einer verzerrten Schätzung der Regressionskoeffizienten, jedoch der Standardfehler (d.h. auch der

Irrtumswahrscheinlichkeiten) (Eid et al., 2013). Diese spezielle statistische Limitation ist bei Hypothese 2a-d zu berücksichtigen.

Tabelle F-2: Studie 1 – Voraussetzungsprüfungen zu Hypothese 2

	B	*SE*	*β*	*t*	*p*	Toleranz	VIF
Hypothese 3a – AV: state-TA post							
X: objektive Leistung	-.10	.05	-.24	-2.00	.047	.39	2.59
M: subjektive Leistung	.62	.20	.36	3.16	.002	.43	2.33
W: Bedingung	.19	.26	.05	.73	.466	.99	1.01
Int 1: obj. Leistung x Beding.	.08	.07	.14	1.22	.224	.41	2.41
Int 2: subj. Leistung x Beding.	-.03	.27	-.01	-.09	.925	.46	2.16
Hypothese 3b – AV: state-TA post							
X: subjektive Leistung	.73	.17	.42	4.35	.000	.49	2.02
M: Selbstwertkontingenz	.30	.14	.23	2.16	.032	.41	2.42
W: Bedingung	.04	.24	.01	.16	.873	.94	1.07
Int 1: subj. Leistung x SWK	-.06	.15	-.04	-.38	.703	.39	2.56
Int 2: subj. Leistung x Beding.	-.31	.25	-.12	-1.25	.214	.49	2.03
Int 3: SWK x Beding.	.28	.19	.16	1.48	.140	.41	2.44
Int 4: subj. Leistung x SWK x Beding.	.27	.19	.15	1.40	.164	.39	2.60
Hypothese 3c – AV: state-TA post							
X: subjektive Leistung – AV: state-TA post	.82	.18	.47	4.44	.000	.50	2.00
M: Messqualität	.23	.23	.12	1.02	.310	.41	2.45
W: Bedingung	.17	.27	.05	.61	.543	.92	1.08
Int 1: subj. Leistung x MQ	.06	.21	.03	.30	.762	.51	1.97
Int 2: subj. Leistung x Beding.	-.21	.27	-.08	-.77	.441	.49	2.05
Int 3: MQ x Beding.	-.25	.31	-.09	-.79	.430	.43	2.32
Int 4: subj. Leistung x MQ x Beding.	-.05	.32	-.02	-.15	.880	.50	2.00
Hypothese 3d – AV: state-TA post							
X: subjektive Leistung	.78	.18	.45	4.33	.000	.52	1.92
M: Mangelnde Anstrengung	.09	.14	.06	.62	.535	.55	1.81
W: Bedingung	.23	.26	.07	.87	.386	.99	1.01
Int 1: subj. Leistung x MA	-.10	.13	-.08	-.78	.437	.58	1.72
Int 2: subj. Leistung x Beding.	-.18	.27	-.07	-.67	.507	.50	2.01
Int 3: MA x Beding.	-.03	.21	-.02	-.15	.882	.55	1.82
Int 4: subj. Leistung x MA x Beding.	.11	.20	.05	.53	.595	.55	1.83
Weiterf. Analyse – AV: state-TA post							
X: Selbstwertkontingenz	.33	.15	.25	2.28	.024	.39	2.58
M: Messqualität	.14	.21	.07	.68	.499	.43	2.35
W: Bedingung	.11	.25	.03	.46	.649	.93	1.07
Int 1: SWK x MQ	.10	.15	.07	.66	.513	.39	2.59
Int 2: SWK x Beding.	.18	.19	.10	.91	.363	.39	2.56
Int 3: MQ x Beding.	-.17	.28	-.06	-.61	.543	.45	2.23
Int 4: SWX x MQ x Beding.	.11	.20	.06	.57	.572	.38	2.62
Kovariate: subjektive Leistung	.58	.12	.33	4.68	.000	.94	1.06

VIF = Varianz-Inflations-Faktor

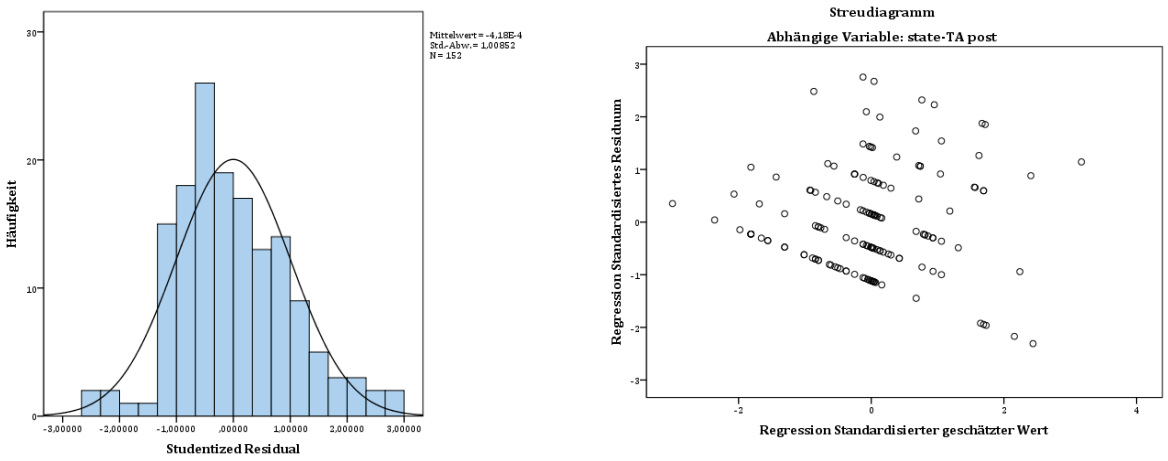

Abbildung F-1: Prüfung auf Normalverteilung und Homoskedastizität bei Hypothese 2a

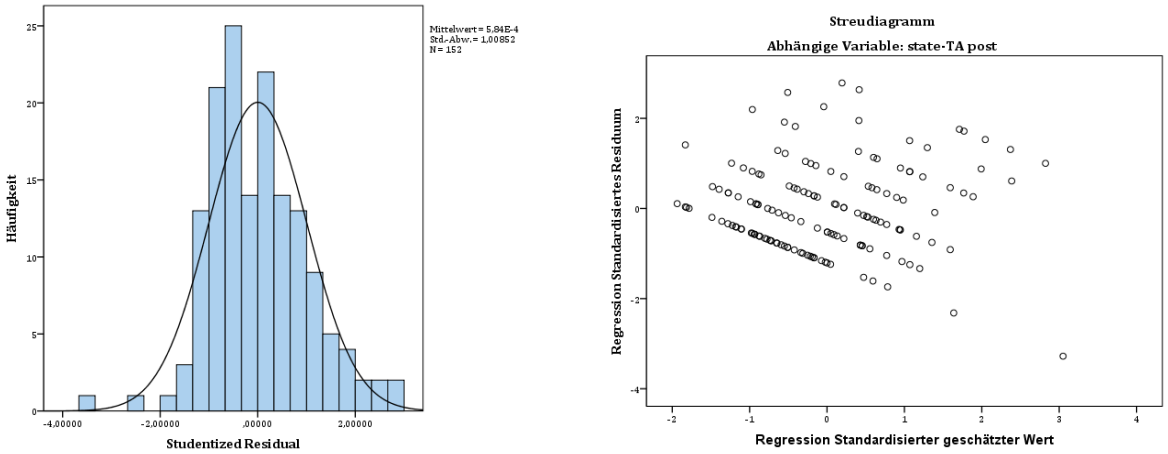

Abbildung F-2: Prüfung auf Normalverteilung und Homoskedastizität bei Hypothese 2b

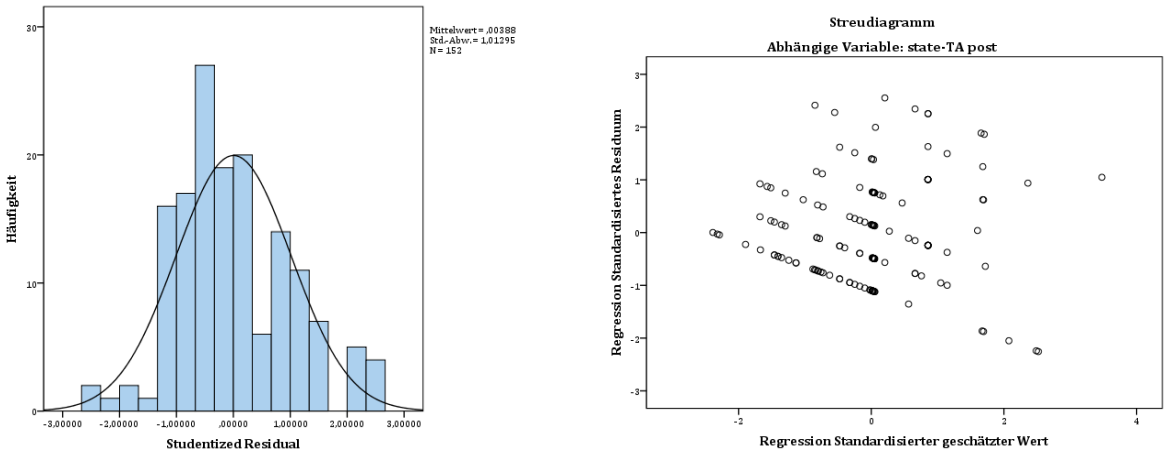

Abbildung F-3: Prüfung auf Normalverteilung und Homoskedastizität bei Hypothese 2c

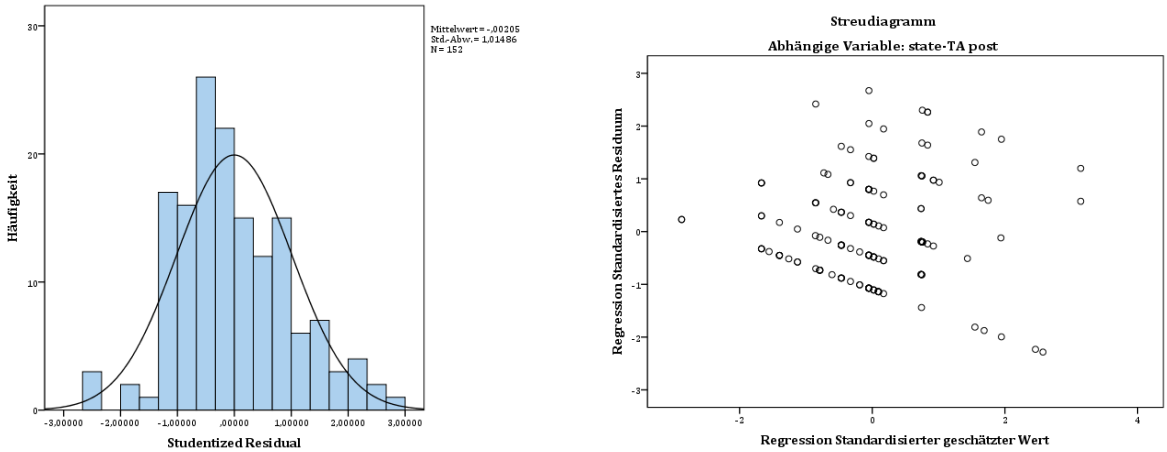

Abbildung F-4: Prüfung auf Normalverteilung und Homoskedastizität bei Hypothese 2d

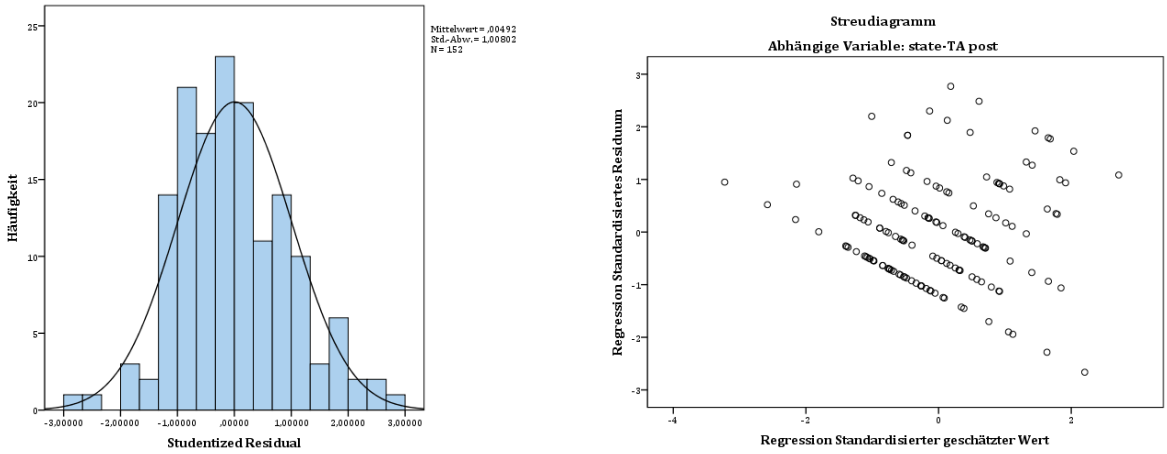

Abbildung F-5: Prüfung auf Normalverteilung und Homoskedastizität bei den weiterführenden Analysen

2. Studie 2

Tabelle F-3: Normalverteilungsprüfungen für zentrale Variablen von Studie 2

	K-S-Test	
	Z	*p*
Testängstlichkeit (trait)		
Besorgtheit	.09	.003
Aufgeregtheit	.18	.000
Interferenz	.14	.000
Mangel an Zuversicht	.12	.000
Gesamt	.10	.001
Testangst (state – prä)		
Besorgtheit	.15	.000
Aufgeregtheit	.21	.000
Interferenz	.15	.000
Mangel an Zuversicht	.17	.000
Gesamt	.14	.000
Leistungszielorientierung		
Annäherung	.08	.012
Vermeidung	.08	.010
Leistung Zahlenreihen	.09	.001
Akzeptanz		
Kontrollierbarkeit	.20	.000
Messqualität	.10	.001
Augenscheinvalidität	.12	.000
Belastungsfreiheit	.09	.003
Note Verfahren	.25	.000
Subjektive Leistung	.23	.000
Flow-Erleben		
Gesamt	.10	.000
Besorgnis	.24	.000

Z = Teststatistik des K-S-Test

3. Studie 3

Tabelle F-4: Normalverteilungsprüfungen für zentrale Variablen von Studie 3

	K-S-Test	
	Z	*p*
Testängstlichkeit (trait)		
Besorgtheit	.08	.000
Aufgeregtheit	.14	.000
Interferenz	.15	.000
Mangel an Zuversicht	.13	.000
Gesamt	.05	.016
Anxiety Motivation		
Energie	.08	.000
Information	.10	.000
Leistung GkKT	.07	.000
Akzeptanz		
Kontrollierbarkeit	.09	.000
Messqualität	.09	.000
Augenscheinvalidität	.09	.000
Belastungsfreiheit	.07	.002
Note Verfahren	.23	.000
Subjektive Leistung	.26	.000
Testangst (state)		
Besorgtheit	.27	.000
Aufgeregtheit	.13	.000
Gesamt	.12	.000
Anxiety Motivation (state)	.38	.000
Extraversion	.06	.005
Offenheit	.08	.000
Verträglichkeit	.10	.000
Gewissenhaftigkeit	.06	.006
Neurotizismus	.07	.001
Allgemeine Selbstwirksamkeit	.07	.000
Leistung BEFKI	.13	.000
Leistung MWT-B	.06	.200
Variablen zum Erleben des Studiums		
Erfolgszuversicht	.22	.000
Abbruchgedanken	.39	.000

Z = Teststatistik des K-S-Test

Wie in Tabelle F-5 zu sehen ist, kann Multikollinearität ausgeschlossen werden. Mit Ausnahme der Analyse zur abhängigen Variable Abiturnote kann auch Homoskedastizität angenommen werden, was eine Limitation dieser Analyse darstellt (siehe Abbildungen F-6 bis 9). Die deutlich nicht normalverteilten Residuen bei Analyse der abhängigen Variable Abbruchtendenz sind relativ unproblematisch, da die Stichprobe groß ist (Eid et al., 2013).

Tabelle F-5: Studie 3 – Voraussetzungsprüfungen zu Hypothese 3

	B	*SE*	*β*	*t*	*p*	Toleranz	VIF
Hypothese 3a – AV: Abiturnote							
X: Trait-Testängstlichkeit	.02	.01	.13	2.19	.029	.78	1.28
M: AM-Energie	-.09	.04	-.15	-2.54	.011	.74	1.35
W: AM-Info	.05	.04	.08	1.36	.176	.72	1.38
Int 1: trait-TÄ x Energie	-.02	.01	-.11	-1.78	.075	.64	1.57
Int 2: trait-TÄ x Info	.02	.01	.13	1.99	.047	.64	1.56
Int 3: Energie x Info	-.02	.03	-.03	-.62	.539	.92	1.09
Int 4: trait-TÄ x Energie x Info	-.01	.01	-.06	-1.06	.292	.75	1.33
Hypothese 3a – AV: Erfolgszuversicht							
X: Trait-Testängstlichkeit	-1.00	.01	-.39	-7.44	.000	.78	1.29
M: AM-Energie	.21	.06	.19	3.46	.001	.73	1.38
W: AM-Info	-.05	.06	-.04	-.77	.440	.71	1.41
Int 1: trait-TÄ x Energie	.02	.02	.05	.90	.370	.63	1.58
Int 2: trait-TÄ x Info	.01	.02	.02	.36	.720	.63	1.59
Int 3: Energie x Info	.01	.05	.01	.11	.910	.93	1.08
Int 4: trait-TÄ x Energie x Info	-.01	.01	-.03	-.48	.629	.75	1.33
Hypothese 3a – AV: Abbruchtendenz							
X: Trait-Testängstlichkeit	.10	.02	.30	5.38	.000	.78	1.28
M: AM-Energie	.09	.08	.07	1.15	.253	.73	1.37
W: AM-Info	-.11	.08	-.07	-1.26	.208	.71	1.40
Int 1: trait-TÄ x Energie	-.03	.02	-.08	-1.23	.218	.64	1.57
Int 2: trait-TÄ x Info	.05	.02	.13	2.05	.041	.63	1.58
Int 3: Energie x Info	-.06	.07	-.05	-.89	.374	.92	1.08
Int 4: trait-TÄ x Energie x Info	.01	.02	.04	.65	.514	.76	1.32
Hypothese 3b – AV: Leistung GkKT							
X: state-TA (Besorgtheit)	-2.17	.62	-.25	-3.49	.001	.45	2.25
M: state-AM	.96	.47	.12	2.03	.043	.64	1.57
W: Bedingung	.21	.64	.02	.33	.742	.96	1.04
Int 1: state-TA x state-AM	1.52	.66	.14	2.33	.021	.64	1.57
Int 2: state-TA x Beding.	1.05	.90	.07	1.18	.241	.65	1.53
Int 3: state-AM x Beding.	-.02	.89	.00	-.02	.984	.68	1.48
Int 4: state-TA x state-AM x Beding.	-2.08	1.12	-.11	-1.86	.064	.61	1.64
Kovariate: state-TA (Aufgeregtheit)	.31	.57	.04	.54	.587	.50	2.00

VIF = Varianz-Inflations-Faktor.

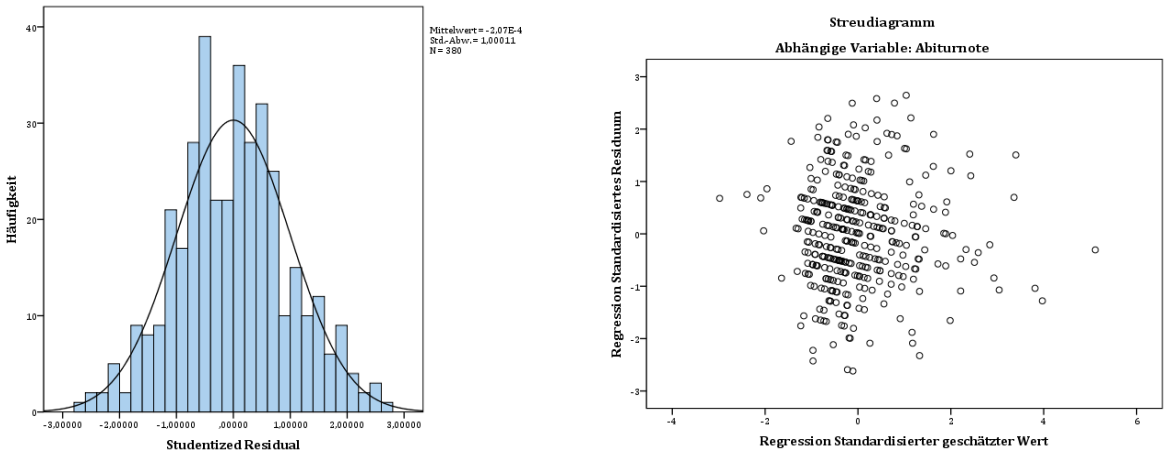

Abbildung F-6: Prüfung auf Normalverteilung und Homoskedastizität bei Hypothese 3a (AV Abiturnote)

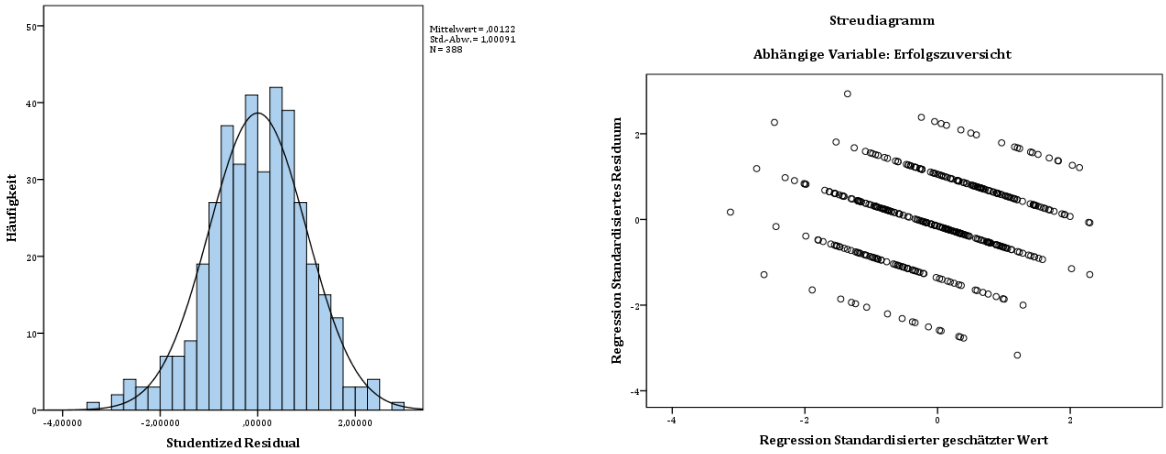

Abbildung F-7: Prüfung auf Normalverteilung und Homoskedastizität bei Hypothese 3a (AV Erfolgszuversicht)

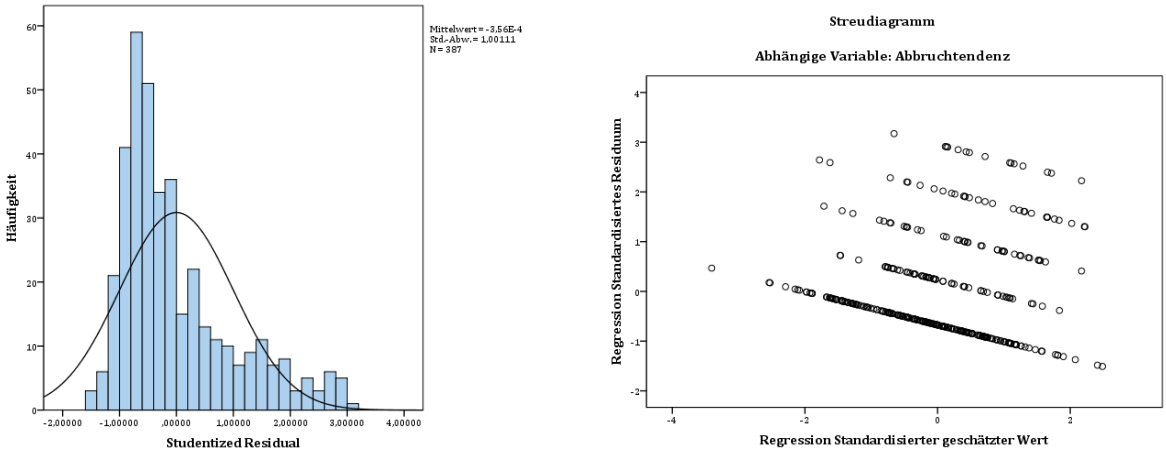

Abbildung F-8: Prüfung auf Normalverteilung und Homoskedastizität bei Hypothese 3a (AV Abbruchtendenz)

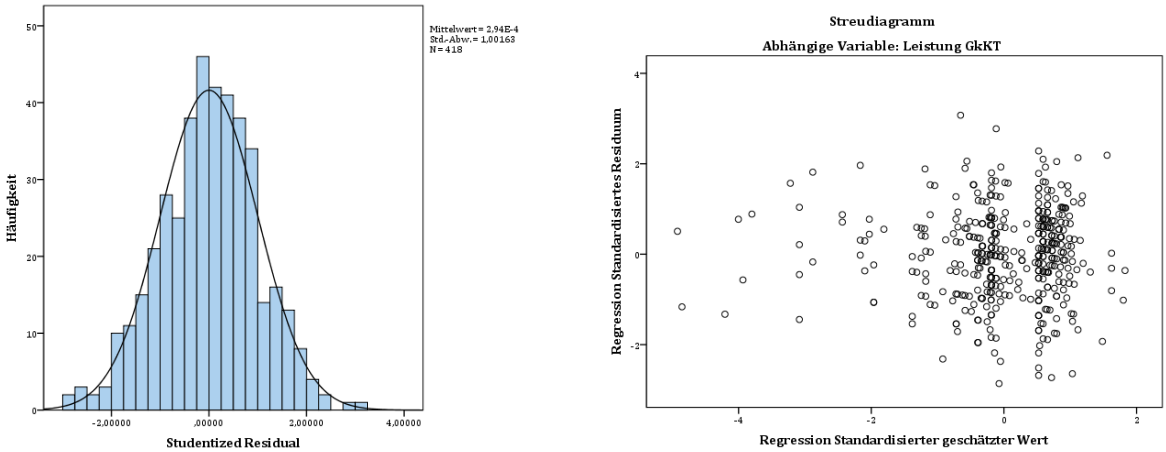

Abbildung F-9: Prüfung auf Normalverteilung und Homoskedastizität bei Hypothese 3b (AV Leistung GKKT)

G. Informationen zur Geschlechterverteilung je Klasse in Studie 2

Tabelle G-1: Geschlechterverteilung je Klasse in Studie 2

Bedingung	Klasse	Mädchen	Jungen
A	1	3	13
	2	13	7
	3	12	4
B	4	11	5
	5	14	7
	6	10	10
C	7	7	8
	8	13	9
	9	11	10
Gesamt		94	73

N = 167

H. Korrelationsvergleiche in Hypothese 1 in Studie 3

Aufgrund des paarweisen Fallausschusses in Studie 3 basieren die Korrelationskoeffizienten von AM-Information (r = .13, p = .008, N = 430) und AM-Energie (r = .34, p < .001, N = 425) und Testängstlichkeit (r = -.04, p = .478, N = 404) mit state-AM nicht auf den exakt gleichen Teilstichproben. Für einen Signifikanztest zum Korrelationsunterschied wurden beide Vergleiche daher wiederholt und jeweils mit listenweisem Fallausschuss durchgeführt. Die Ergebnisse sind in Tabelle H-1 dargestellt.

Tabelle H-1: Korrelationsvergleiche zu Hypothese 1 in Studie 3 (listenweiser Fallausschuss)

	Korrelationsvergleiche					
	AM-Energie	AM-Information	Testängstlichkeit	*N*	z[1]	*p*
State-AM	.33**	-	-.01	377	4.57	.000
State-AM	-	.08	-.04	378	1.74	.081

Korrelation AM-Energie & Testängstlichkeit: -.10 (p = .055); Korrelation AM-Info & Testängstlichkeit: .11 (p = .035)
[1] z ermittelt nach der Prozedur von Hittner, May und Silver (2003) (zitiert nach Diedenhofen & Musch, 2015)

Erkennbar ist, dass durch die schwächere Korrelation von AM-Info mit state-AM der Korrelationsunterschied zur Korrelation von Testängstlichkeit mit state-AM knapp nicht signifikant wird.

Eigenständigkeitserklärung

Ich, Michael Konrad Ott, geboren am 29.04.1986 in Fürth, erkläre: Ich habe die vorgelegte Dissertation selbständig und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Diese Arbeit wurde weder in der vorliegenden noch in einer modifizierten Form, sowie weder vollständig noch auszugsweise veröffentlicht oder einer anderen Prüfungsbehörde vorgelegt.

Frankfurt am Main, 17.01.2017