JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

Wie hängen intraindividuelle Variabilität und Extreme Response Style zusammen und welche Rolle spielen sie in Persönlichkeitsfragebogen?

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Philosophie des Fachbereiches 06

der Justus-Liebig-Universität Gießen

vorgelegt von
Dennis Beermann

aus Frankfurt am Main

2015

Dekan:	Prof. Dr. Marco Ennemoser
1. Berichterstatter/in:	Prof. Dr. Martin Kersting
2. Berichterstatter/in:	Prof. Ute-Christine Klehe, PhD
Tag der Disputation:	28.09.2015

Meiner Frau Svenja und unserem Sohn Samuel

Danksagung

Mein erster Dank geht an Prof. Dr. Martin Kersting für seine Bereitschaft, mich als Doktoranden anzunehmen, für die hervorragende Betreuung, für sehr viele für diese Arbeit sehr wertvolle Anregungen und dafür, dass er mir als Doktorvater das richtige Maß an Freiraum und Orientierung geboten hat.

Bei Prof. Dr. Ute-Christine Klehe möchte ich mich herzlich für ihre Zweitgutachtertätigkeit sowie für hilfreiche Diskussionen und Anregungen bedanken. Herzlicher Dank für Diskussionsbeiträge und einen hilfreichen Austausch geht auch an meine „Mit-Doktorandinnen" Anna-Sophie Ulfert, Carolin Palmer und Lilith Michaelis. Meinem „Mit-Doktoranden" Michael Ott danke ich für wertvolle Diskussionspunkte und für seine ausführliche und schnelle Rückmeldung zu dem eher trockenen Kapitel zur Alpha-Adjustierung.

Ein großes Dankeschön für einen anregenden Austausch, für das gründliche Gegenlesen mehrerer Textteile, für sehr viele sehr hilfreiche Kommentare, Ideen und Rückmeldungen sowie für viele motivierende Worte geht an Dr. Katharina Lochner, die über mehrere Monate zeitgleich an ihrer Dissertation gearbeitet hat und mit der ich unzählige Stunden über unsere Arbeiten diskutiert habe. An Dr. Adrian Hoffmann geht ebenfalls ein besonderer Dank für sein sehr gründliches Review meiner Arbeit und für seine schnellen und vor allem wertvollen Rückmeldungen. Für weiteres hilfreiches Feedback bedanke ich mich bei Luisa Bergholz und Lioba Peters, die jeweils ein Kapitel dieser Arbeit gegengelesen haben.

Ferner danke ich Dr. Alexander Zimmerhofer, Dr. Timo Heydasch und Prof. Dr. Karl-Heinz Renner für die Unterstützung bei der Stichprobengewinnung und Datenerhebung. Für die Unterstützung bei der Vorbereitung von Studie 1 danke ich Leander Troll ganz herzlich!

Bedanken möchte ich auch bei meiner Mutter Rosemarie, die mir mit großem Rückhalt zur Seite stand und mir dadurch mein Studium und diesen Bildungsweg ermöglicht hat. Meiner Frau Svenja danke ich von ganzem Herzen für ihre bedingungslose Unterstützung, ihre Liebe und ihre Geduld – ihr ist diese Arbeit gewidmet. Ebenso gewidmet ist diese Arbeit meinem Sohn Samuel, der in der Endphase der Erstellung dieser Arbeit geboren wurde.

Vielen Dank!

„We have argued […] that it is not possible, in principle, to do any better than predicting some of the people some of the time."

(D. J. Bem & Allen, 1974)

Inhaltsverzeichnis

Zusammenfassung

Im Fokus der vorliegenden Arbeit stehen (intraindividuelle) Variabilität, die Variabilität innerhalb von Traits, und Extreme Response Style (ERS), die Tendenz, in Fragebogen extrem zu antworten. Zur Beschreibung von Variabilität liegen viele Forschungsarbeiten vor, zur Erfassung und Struktur fehlen jedoch klare Ergebnisse. ERS wurde bislang lediglich operational definiert; eine Beschreibung über extremes Antworten in Fragebogen hinaus findet sich nicht. Beiden Phänomenen gemein ist, dass sie als Moderatoren der Validität von Persönlichkeitsfragebogen diskutiert werden; doch auch diesbezüglich ist die Befundlage uneinheitlich.

Mit zwei empirischen Studien wurden entsprechend die Erfassung und Struktur von Variabilität, der Zusammenhang zwischen Variabilität und ERS sowie der Einfluss von Variabilität auf die Testgütekriterien untersucht. In Studie 1 bearbeiteten 405 Studierende das ITB Personality Structure Assessment (ITB-PESA), einen Facetten-Fragebogen, sowie die deutschsprachige revidierte Fassung des HEXACO-Persönlichkeitsinventars, mit der weitgehend unabhängige Dimensionen erfasst werden. In Studie 2 bearbeiteten 367 Bewerber im Rahmen eines Personalauswahlprozesses eine kürzere Version des ITB-PESA, deren Facetten mehrheitlich auf Extraversion laden; zum Vergleich wurde mit Nicht-Auswahl-Daten aus Studie 1 eine identische Fragebogenversion gebildet. Zur statistischen Analyse dienten u. a. Vergleiche von Korrelationen, Strukturgleichungsmodelle und moderierte multiple Regressionen.

Hinsichtlich der Erfassung und Struktur von Variabilität bestätigten beide Studien, dass Variabilität eine eindimensionale Eigenschaft ist, die valide mit einem Aggregat von um Skalenausprägung und -extremität korrigierten intraindividuellen Standardabweichungen der Skalen erfasst werden kann. Bezüglich des Zusammenhangs von ERS und Variabilität zeigte sich bereits in Studie 1, dass Variabilität und ERS sehr hoch korrelieren. Die Ergebnisse beider Studien ließen darauf schließen, dass extreme Antworten, die nicht auf hohe Ausprägungen auf den erfassten Merkmalen zurückzuführen sind, von Variabilität abhängen. Von den Testgütekriterien wurde lediglich die Kriteriumsvalidität in Studie 1 eindeutig von Variabilität beeinflusst: je höher die Variabilität, desto niedriger die Kriteriumsvalidität. Nicht oder sehr gering beeinflusst wurden in beiden Studien die Reliabilität und die Konstruktvalidität.

In Studie 2 (Auswahl) wurde der Effekt auf die Kriteriumsvalidität nicht repliziert, was möglicherweise auf geringe Passung von Prädiktor und Kriterium zurückzuführen ist.

Implikationen der Ergebnisse sind, dass Variabilität eindimensional ist, dass ERS kein Antwortbias ist, sondern Indikator von Variabilität, und dass Variabilität – zumindest im Nicht-Auswahl-Setting – die Zusammenhänge zwischen Persönlichkeitseigenschaften und Kriterien moderiert.

Schlüsselwörter: intraindividuelle Variabilität, Konsistenzskalen, Extreme Response Style, Persönlichkeitsfragebogen, Eignungsdiagnostik

Abstract

The present dissertation focusses on (intraindividual) variability, the variability within traits, and extreme response style (ERS), the tendency to respond extremely in questionnaires. The description of variability has been addressed in many papers, but clear results on its measurement and structure are lacking. Regarding ERS, there is no comprehensive explanation in place, it has only been described as extreme responding. Variability and ERS both are regarded as moderators of validity in personality questionnaires. Still, findings are inconsistent.

Two empirical studies were conducted to examine the measurement and structure of variability, the relationship between variability and ERS, and the effects of variability on reliability and validity. In study 1—405 students completed the ITB Personality Structure Assessment (ITB-PESA), which captures personality facets, and the German revised HEXACO personality inventory, which captures widely independent personality dimensions. In study 2—367 applicants completed a shorter ITB-PESA version as part of a personnel selection procedure. This version mainly comprised facets of extraversion. In order to compare results to non-selection context, an identical version was formed with data from study 1. Data were analysed by comparing correlations, applying structure equation models and performing moderated multiple regressions.

Both studies confirmed that variability is a unidimensional trait. This trait can be captured by aggregating intraindividual within-scale standard deviations and controlling those for trait mean and extremity. Regarding the relationship between variability and ERS, a strong correlation was shown in study 1. Results of both studies led to the conclusion that extreme responses which are not due to extreme trait manifestation are caused by variability. Study 1 yielded a strong influence of variability on the criterion validity: the higher the variability, the lower the criterion validity. Reliability and construct validity were not affected or only to a negligible extent. In study 2, the moderating effect of variability on the criterion validity was not replicated. This might be attributed to a bad predictor-criterion fit.

The results imply that variability is unidimensional, that ERS is not a response bias but an indicator of variability, and that variability moderates the relationship between traits and criteria—at least for non-selection settings.

Keywords: intraindividual variability, consistency scales, extreme response style, personality test, personnel diagnostics

1 Einleitung

Die Forschung zu Persönlichkeitsfragebogen hat in den letzten 25 Jahren stark zugenommen (Alonso-Arbiol & van de Vijver, 2010; Kersting, 2005; Morgeson et al., 2007). Persönlichkeitsfragebogen umfassen in den meisten Fällen eine Reihe von Aussagen, die vom Bearbeiter[1] nach dem Grad der Zustimmung bzw. des Zutreffens beurteilt werden (Schmidt-Atzert & Amelang, 2012). Erfasst werden üblicherweise mehrere Persönlichkeitseigenschaften, die für bestimmte Lebensbereiche relevant sind. Der Aufschwung der Forschung zu Persönlichkeitsfragebogen betrifft insbesondere berufsbezogene Persönlichkeitsfragebogen. Er ist vor allem auf die Erkenntnis zurückzuführen, dass bestimmte Persönlichkeitseigenschaften mit Verhalten am Arbeitsplatz zusammenhängen, dass sich diese Zusammenhänge nutzen lassen können (Robertson & Callinan, 1998), und dass Persönlichkeitsfragebogen zu bestimmten Eigenschaften treffsichere Prognosen für berufsrelevante Kriterien ermöglichen: Beispielsweise lassen sich Berufserfolg durch (Fragebogen-)Maße von Gewissenhaftigkeit (Barrick & Mount, 1991; Salgado, 1997), Trainingserfolg durch Maße von Offenheit für Erfahrungen (Barrick & Mount, 1991; Salgado, 1997), Arbeitszufriedenheit durch Maße von emotionaler Stabilität (Judge, Heller & Mount, 2002) und kontraproduktives Arbeitsverhalten durch Maße von Ehrlichkeit-Bescheidenheit (Marcus, Lee & Ashton, 2007; Zettler & Hilbig, 2010) vorhersagen. Zwar liegen die Validitätskoeffizienten üblicherweise im mittleren Bereich ($.20 \leq r_{tc} \leq .40$), bei der Vorhersage von Berufserfolg haben z. B. Fragebogen zu Gewissenhaftigkeit jedoch inkrementelle Validität und somit zusätzlichen Nutzen über Intelligenztests hinaus (Schmidt & Hunter, 1998). Fragebogen zu Leistungsstreben, Dominanzstreben und dem Bedürfnis nach Beachtung führen kombiniert mit Assessment Centern (AC) zu einer besseren Vorhersage von Berufserfolg als AC allein (Goffin, Rothstein & Johnston, 1996).

Trotz dieser positiven Befundlage, trotz der Objektivität und der ökonomischen Durchführung und Auswertung (Fisseni, 2004; Schmidt-Atzert & Amelang, 2012) werden Persönlichkeitsfragebogen im deutschsprachigen Raum vergleichsweise selten zur Personalauswahl eingesetzt (König, Klehe, Berchtold & Kleinmann, 2010; Schuler, Hell, Trapmann, Schaar & Boramir, 2007). Verringert wird die Bereitschaft, Persönlichkeitsfragebogen einzusetzen,

[1] Aus Gründen der Leserlichkeit wird im Folgenden stets die männliche Form verwendet. Gemeint sind jeweils Frauen und Männer.

durch Vorurteile über ihre Akzeptanz (Beermann, Kersting, Stegt & Zimmerhofer, 2013), durch ihre im Vergleich zu anderen Verfahren augenscheinlich niedrigere Prognosekraft für Berufserfolg (Rothstein & Goffin, 2006) sowie durch ihre mutmaßliche Anfälligkeit für Verfälschungen (Bott, O'Connell, Ramakrishnan & Doverspike, 2007; Rothstein & Goffin, 2006; Schmidt-Atzert & Amelang, 2012). Was die Akzeptanz von Persönlichkeitsfragebogen betrifft, liegen nur wenige Forschungsarbeiten vor. Ersten Ergebnissen zufolge lässt sich das Vorurteil mangelnder Akzeptanz nicht halten (Beermann et al., 2013). Zur Prognosekraft und zur Verfälschbarkeit von Persönlichkeitsfragebogen haben sich weitreichende Forschungstrends ausgebildet (Rothstein & Goffin, 2006).

Hinsichtlich der Prognosekraft von Persönlichkeitsfragebogen für Berufserfolg liegen positive Forschungsbefunde vor (siehe auch Beermann & Heilmann, 2014): Sie lässt sich steigern, indem zur Auswahl einzusetzender Persönlichkeitsskalen Anforderungsanalysen verwendet werden (Tett, Jackson & Rothstein, 1991), indem schmale Facetten statt breiter Dimensionen erfasst werden (Beermann, 2011; Dudley, Orvis, Lebiecki & Cortina, 2006; Paunonen & Ashton, 2001; Vinchur, Shippmann, Switzer & Roth, 1998) und indem Items bzw. Instruktionen in einen kriterienrelevanten Kontext eingebettet werden, d. h. indem mit ihnen ein Bezug zur Berufswelt hergestellt wird (Bowling & Burns, 2010; Lievens, De Corte & Schollaert, 2008; Reddock, Biderman & Nguyen, 2011). Ein weiteres aktuelles Forschungsfeld ist die Suche nach Moderatoren der Validität bzw. die Untersuchung der differenziellen Validität von Persönlichkeitsfragebogen (Rothstein & Goffin, 2006): Untersucht wird, ob Persönlichkeitsfragebogen für bestimmte Personen(gruppen) bessere Prognosen von Verhalten (am Arbeitsplatz) erlauben als für andere, also ob die Validität für unterschiedliche Personengruppen (z. B. Berufsgruppen, Geschlechter) unterschiedlich ausfällt oder von kontinuierlichen Variablen moderiert wird. Neben demografischen Merkmalen wie Geschlecht, ethnische Zugehörigkeit oder Bildungsniveau können auch Persönlichkeitsmerkmale die Validität moderieren (Rothstein & Goffin, 2006): So wurden zum Beispiel Gewissenhaftigkeit als Moderator des Zusammenhangs zwischen Extraversion und Berufserfolg (Witt, 2002) und Verträglichkeit als Moderator des Zusammenhangs zwischen Gewissenhaftigkeit und Berufserfolg (Witt, Burke, Barrick & Mount, 2002) identifiziert. Als weitere Moderatoren der Validität von Fragebogen im Allgemeinen und von Persönlichkeitsfragebogen im Speziellen gelten auch Antwortstile. Unter Antwortstilen bei der Bearbeitung von Fragebogen versteht man

Antwortverhalten, das nicht aufgrund des zu erfassenden Merkmals zustande kommt und das sich folglich auf die Validität des Fragebogens auswirkt (Cronbach, 1946; McGrath, Mitchell, Kim & Hough, 2010; Van Vaerenbergh & Thomas, 2013). Ein potenzieller Antwortstil, Extreme Response Style (ERS), und ein Persönlichkeitsmerkmal, intraindividuelle Variabilität, stehen im Fokus der vorliegenden Arbeit.

Die Forschung zu Variabilität[2] hat eine längere Tradition und umfasst mehrere Forschungszweige (Baird, Le & Lucas, 2006; D. J. Bem & Allen, 1974; Block, 1961; Britt, 1993; Donahue, Robins, Roberts & John, 1993; Fiske & Rice, 1955). Erst kürzlich wurden neue Hinweise darauf berichtet, dass die Reliabilität und die Validität von Persönlichkeitsfragebogen mit intraindividueller Variabilität zusammenhängen (Biderman & Reddock, 2012; Fleisher, Woehr, Edwards & Cullen, 2011; Reddock et al., 2011). Die Forschungslage ist zum derzeitigen Stand jedoch nicht eindeutig. In Kapitel 2 werden das Phänomen Variabilität beschrieben und der Forschungsstand referiert. Insbesondere werden offenen Fragen im Hinblick auf die Beschreibung von Variabilität und den Zusammenhang von Variabilität mit der Reliabilität und Validität von Persönlichkeitsfragebogen diskutiert.

ERS wird aktuell in verschiedenen Disziplinen der empirischen Sozialforschung untersucht (Baumgartner & Steenkamp, 2001; M. Johnson, 2013; Kieruj & Moors, 2012; Weijters, Geuens & Schillewaert, 2010b, 2010c; Wetzel, Cartensen & Böhnke, 2013b); in Kapitel 3 wird eine Übersicht über die Forschung zu ERS gegeben: Diese konzentriert sich überwiegend darauf, ERS mittels statistischer Modelle zu schätzen. Die Ursachen von ERS wurden mit nur wenigen Studien untersucht (z. B. Naemi, Beal & Payne, 2009) und gelten als weitgehend ungeklärt. Dies ist insofern verwunderlich, als dass einige Autoren einen Zusammenhang zwischen ERS und Variabilität vermuten (Greenleaf, 1992a) bzw. in ihren Daten finden (Baumgartner & Steenkamp, 2001; Biderman & Reddock, 2012). Dieser Zusammenhang wird jedoch nicht *inhaltlich* beschrieben bzw. erklärt.

[2] In dieser Arbeit wird aus Gründen der Übersichtlichkeit der Begriff „Variabilität“ verwendet. Sofern nicht anders gekennzeichnet, ist stets „intraindividuelle Variabilität“ gemeint. Der Begriff „Variabilität“ wird auch dann genutzt, wenn dasselbe Phänomen in den referierten Forschungsarbeiten mit „Inkonsistenz“ bzw. „Konsistenz“ bezeichnet wird.

In Kapitel 4 werden aus den referierten Befunden zu Variabilität und ERS Hypothesen abgeleitet. Diese beziehen sich auf

- die Struktur des Konstrukts Variabilität und auf dessen Erfassung,
- die Ursachen von ERS und den Zusammenhang zwischen Variabilität und ERS sowie
- den Einfluss von Variabilität und ERS auf die Reliabilität und die Validität von Persönlichkeitsfragebogen.

Die Hypothesen werden im Rahmen von zwei empirischen Studien untersucht. In Kapitel 5 werden die Methode und die Ergebnisse von Studie 1 berichtet. Erhoben werden ein Fragebogen, mit dem Persönlichkeitsfacetten erfasst werden, ein Fragebogen, mit dem breite weitgehend unabhängige Dimensionen erfasst werden, sowie mehrere Kriterien. In Kapitel 6 werden die Methode und die Ergebnisse von Studie 2 berichtet. Studie 2 basiert auf den Daten des Ernstfalleinsatzes eines Persönlichkeitsfragebogens zur Personalauswahl sowie eines Fragebogens zum sozialen Umfeld, mit dem ein Kriterium erfasst wird. In Kapitel 7 werden schließlich die Befunde im Hinblick auf die Hypothesen zusammengefasst, diskutiert und eingeordnet.

2 Intraindividuelle Variabilität

Variabilität bezeichnet das Ausmaß, in dem das Erleben und Verhalten von Personen über Situationen, über Rollen und über die Zeit hinweg variiert, sowie die Variation zwischen verschiedenen Indikatoren eines Traits (Baird et al., 2006; Baumeister & Tice, 1988; D. J. Bem & Allen, 1974; Fleeson, 2001, 2007; Reddock et al. 2011). Auch in Persönlichkeitsfragebogen, mit denen Eigenschaften situations-, rollen- und zeitpunktunabhängig gemessen werden, zeigt sich Variabilität, und zwar als Variation zwischen den Indikatoren der jeweiligen Eigenschaften – also als Variabilität der Antworten auf die Items der Skalen. Dies lässt sich sehr gut anhand von Antwortmustern veranschaulichen: Abbildung 1 zeigt das Antwortmuster zweier Personen auf einer (Adjektiv-)Persönlichkeitsskala. Beide Personen haben den gleichen Mittelwert auf der Skala, sie unterscheiden sich in der Variabilität ihres Antwortverhaltens. Die Person links hat eine hohe Variabilität und die Person rechts eine niedrige. Intuitiv ergibt sich (aus der Abbildung) die Beziehung zwischen der Variabilität und der Vorhersagbarkeit des Verhaltens bzw. der Ausprägung weiterer Indikatoren derselben Eigenschaft: Würde man ein weiteres geeignetes Item (z. B. „gutmütig“) zur abgebildeten Skala hinzufügen, ließe sich die Antwort des konsistent Bearbeitenden (rechts) wahrscheinlich besser bzw. genauer vorhersagen als die des variabel Bearbeitenden (links).

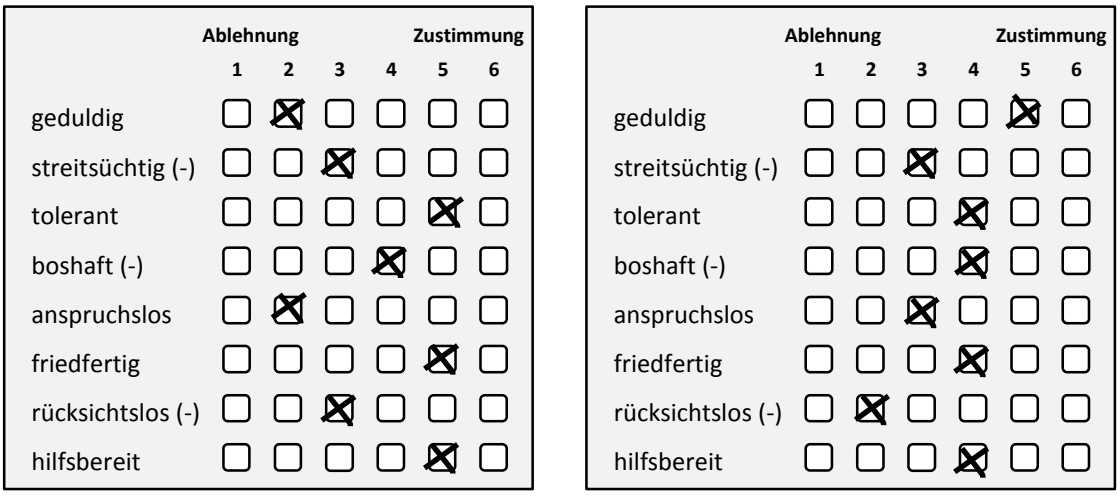

Abbildung 1: Selbstbeschreibung einer Person mit hoher (links) und einer Person mit niedriger Variabilität (rechts) auf einer Adjektiv-Persönlichkeitsskala

Ein „(-)“ kennzeichnet, dass das Adjektiv-Item den Gegenpol der abgebildeten Skala erfasst: Hohe Werte stehen hier für Ablehnung des Items und niedrige Werte für Zustimmung.

Die Hypothesen, dass Variabilität im Verhalten bedeutsam ist, dass Variabilität für verschiedene Messungen konvergiert und dass Variabilität eine Auswirkung auf die Validität einer Messung hat, berichteten Fiske und Rice bereits 1955 im ersten systematischen Review zum Thema. Abgesehen von einzelnen Studien wurde Variabilität erst wesentlich später umfassend untersucht. Die entsprechenden Forschungsarbeiten und -befunde werden in diesem Kapitel vorgestellt und diskutiert: Mittlerweile liegen Befunde zur Struktur, zur Stabilität und zur Universalität von Variabilität (Abschnitt 2.1) sowie Konzepte zur Einordnung in das Situations-Eigenschafts-Paradigma (Abschnitt 2.2) vor. Auch erste Ergebnisse zur Abhängigkeit der Reliabilität und Validität von Persönlichkeitsfragebogen von Variabilität wurden berichtet (Abschnitt 2.3).

2.1 Die Beschreibung und Erfassung intraindividueller Variabilität

Wenige Jahre nach Veröffentlichung des Reviews von Fiske und Rice (1955) zur Variabilität in Messungen untersuchte Block (1961) Variabilität aus einer anderen Perspektive, und zwar als (In)Konsistenz des Selbstbildes über unterschiedliche soziale Rollen hinweg. U. a. infolge der beiden Arbeiten entwickelten sich zwei Forschungszweige: Auf das Review von Fiske und Rice folgte die Forschung zu *Metatraits*. Die Studie von Block war ein Ausgangspunkt für die Forschung zu *Self-Concept-Differentiation* (SCD). Im Folgenden werden die Befunde der Metatraits-Forschung (Abschnitt 2.1.1) und die Befunde zu SCD (Abschnitt 2.1.2) dargestellt. Daran anschließend wird auf die methodischen Probleme, mit denen diese beiden Forschungsstränge konfrontiert waren, eingegangen (Abschnitt 2.1.3). Und schließlich wird Evidenz für die Universalität und Stabilität von Variabilität berichtet (Abschnitt 2.1.4).

2.1.1 Metatraits

Im Rahmen der Metatraits-Forschung wurde Variabilität lediglich für Verhalten, das einen gemeinsamen Trait indiziert, bzw. für einzelne Persönlichkeitsskalen untersucht (vgl. D. J. Bem & Allen, 1974; Baumeister & Tice, 1988; Britt, 1993). Angenommen wurde, dass für jeden Trait ein Metatrait existiert, der die Relevanz des Traits beschreibt, und dass Variabilität auf den Indikatoren eines Traits die Ausprägung des zugehörigen Metatraits widerspiegelt. Personen mit niedriger Variabilität wurde eine hohe Ausprägung des jeweiligen Metatraits zugeschrieben, sie wurden der Terminologie der meisten Studien zufolge als *traited* auf dem

jeweiligen Trait beschrieben. Personen mit hoher Variabilität wurde eine niedrige Ausprägung des jeweiligen Metatraits zugeschrieben, sie galten als *untraited* (Baumeister & Tice, 1988; Britt, 1993; Dwight, Wolf & Golden, 2002). Sind Personen *traited* auf einem Trait, dann hat dieser Trait Relevanz für ihr Verhalten; sind sie *untraited*, spielt die Trait-Ausprägung keine (bzw. eine kleinere) Rolle für das Verhalten dieser Personen. Veranschaulichen lassen sich Metatraits sehr gut durch die Ausführung von Cucina und Vasilopoulos (2005):

> For example, consider two individuals with average scores on an extraversion scale. One individual could be traited on extraversion (i.e., consistently average in extraversion across situations) and the other individual could be untraited on extraversion (i.e., extraverted in some situations and introverted in others). (S. 228)

Besonders prominent im Forschungsfeld „Metatraits“ ist die Studie von D. J. Bem und Allen (1974). Teilnehmer in ihrer Studie bearbeiteten einen Fragebogen, mit dem Extraversion und Gewissenhaftigkeit erfasst wurden, und gaben eine globale Selbsteinschätzung für die Dimensionen Extraversion und Gewissenhaftigkeit. In einer weiteren Selbsteinschätzung gaben sie an, wie stark sie zwischen Situationen variieren, was ihre Extraversion und ihre Gewissenhaftigkeit betrifft („*How much do you vary from one situation to another in how friendly and outgoing [conscientious] you are?“*). Mit diesen Fragen wurden die Metatraits für Extraversion und Gewissenhaftigkeit explizit erfasst. Zusätzlich zu den Selbstauskünften erhoben D. J. Bem und Allen die Traits via Fremdeinschätzungen (durch Eltern und Peers) und Verhaltensbeobachtungen. Um zu prüfen, ob der Metatrait jeweils den Zusammenhang zwischen dem Fragebogenmaß für den Trait und Fremdeinschätzungen für den Trait bzw. Beobachtungen für den jeweiligen Trait moderiert, teilten D. J. Bem und Allen ihre Untersuchungsteilnehmer für jeden Trait zwei Gruppen ein, je eine mit hoher *Traitedness* und eine mit niedriger *Traitedness*. Geschlecht und Trait-Ausprägung (globale Selbsteinschätzung) waren jeweils ausbalanciert. Anhand von Mediansplits zeigte sich, dass der Metatrait für Extraversion die Zusammenhänge zwischen den Messungen von Extraversion moderiert und der Metatrait für Gewissenhaftigkeit die Zusammenhänge zwischen den Messungen der Gewissenhaftigkeit moderiert; hohe *Traitedness* (d. h. niedrige Variabilität auf dem Trait) ging jeweils mit höheren Zusammenhängen einher als niedrige *Traitedness*.

Problematisch an dem expliziten Maß von D. J. Bem und Allen (1974) war, dass es als Single-Item nicht immer hohe Reliabilität aufwies (Rushton, Jackson & Paunonen, 1981) und dass

die Anforderungen an den Bearbeiter sehr hoch waren: Personen mussten zum Beantworten nämlich gleichzeitig Informationen über ihr Verhalten zusammenführen, integrieren und bewerten (Baumeister & Tice, 1988). Baumeister und Tice zufolge könnte geringe Variabilität auch sozial erwünscht sein, was die Validität des expliziten Maßes weiter beeinträchtigte. Möglicherweise war es diesen Schwächen der Messung von Metatraits geschuldet, dass Chaplin und Goldberg (1984) die Ergebnisse von D. J. Bem und Allen mit gleicher Methode nicht replizieren konnten. Infolge der Studie von Chaplin und Goldberg wurden Metatraits vorwiegend implizit erfasst, jeweils als Standardabweichungen der Antworten einer Person auf die Items einer Skala (Inter-Item-SD). Mit diesem Maß untersuchten Baumeister und Tice (1988), ob Metatraits – der Theorie entsprechend – den Zusammenhang zwischen Traits und Verhalten moderieren. Teilnehmer an ihrer Studie mussten eine Videospiel-Aufgabe bewältigen und hatten vorher Gelegenheit, das Videospiel zu üben. Sie bearbeiteten auch eine Skala zur Kontrollüberzeugung. Bei Personen, die eine niedrige Inter-Item-SD auf der Skala hatten, konnte die Übungszeit für die Videospiel-Aufgabe besser durch Kontrollüberzeugung vorhergesagt werden als bei Personen mit hoher Inter-Item-SD. Mit anderen Worten hatte der Trait unter Personen, die als *traited* beschrieben werden konnten, mehr Einfluss auf das Verhalten als bei Personen, auf die das Attribut *untraited* zutraf. Allerdings war die Stichprobe sehr klein ($N = 33$). Als Methode verwendeten Baumeister und Tice wie D. J. Bem und Allen (1974) einen Mediansplit. Mit einer größeren Stichprobe ($N = 125$) ermittelte Baumeister (1991) die Retestreliabilität der Inter-Item-SD mehrerer Skalen. Diese liegt in einem Zeitraum von zwei Wochen für Dimensionsskalen im mittleren bis hohen Bereich ($.66 \leq r_{tt} \leq .74$). Für homogene und kurze Skalen fällt sie etwas niedriger aus ($r_{tt} \approx .50$). Die Zusammenhänge zwischen Metatraits für verschiedene Traits waren in Baumeisters Studie vergleichbar mit denen zwischen den untersuchten Traits ($.25 \leq r \leq .45$). Die Annahme, dass Metatraits die Retestreliabilität der Trait-Maße moderieren, wurde nicht hinreichend durch die Daten bestätigt.

Britt (1993) interpretiert die Metatraits-Theorie umfassender. Seiner Auffassung nach bedeutet *Traitedness*, dass eine Person einen Trait repräsentiert. Entsprechend sollte die Korrelation zwischen zwei verwandten Traits höher sein, wenn Personen hohe Traitedness für beide Traits haben. Für Personen, die niedrige Traitedness haben, sollten sich geringere Zusammenhänge zeigen. Anders als die Autoren der Vorgängerstudien verwendete Britt mode-

rierte multiple Regressionen[3] statt Mediansplits und fand damit die erwarteten Moderatoreffekte. Die Zusammenhänge zwischen den Metatraits waren gering. Metatraits korrelierten nicht mit der Selbsteinschätzung der Relevanz des entsprechenden Traits.

Aufbauend auf den Befunden von Baumeister und Tice (1988), Baumeister (1991) und Britt (1993) führten Dwight et al. (2002) eine praxisnahe Studie durch: Bei einer Stichprobe von Ablesepersonal eines Energiekonzerns wurde untersucht, ob der Metatrait die Beziehung zwischen Traits und Kriterien moderiert. Prädiktor war ein Fragebogen zu sicherheitsrelevantem Verhalten am Arbeitsplatz, Kriterien waren ein subjektives (Vorgesetztenbeurteilung) und ein objektives Leistungsmaß (zusammengesetzt u. a. aus der Anzahl abgelesener Geräte und der Fehlerrate). Während das subjektive Maß nicht durch den Fragebogen vorhergesagt wurde, wurde ein substanzieller Varianzanteil des objektiven Kriteriums durch den Fragebogen aufgeklärt. Moderierte multiple Regressionen zeigten, dass der Anteil durch den Fragebogen aufgeklärter Kriteriumsvarianz unter Personen mit hoher *Traitedness* (mit niedriger Inter-Item-SD im Fragebogen) größer war als unter Personen mit geringer *Traitedness*.

Zusammenfassend lässt sich für die Metatraits-Forschung bis um die Jahrtausendwende festhalten, dass Metatraits als Inter-Item-SD einer Skala erfasst werden können. Metatraits sind zeitstabil und für jeden Trait wird ein eigener Metatrait angenommen. Hinsichtlich der Konvergenz von verschiedenen Metatraits lassen sich auf Basis der Studien zu Metatraits keine Aussagen treffen. Die Studien zeigen jedoch, dass Metatraits den Zusammenhang zwischen Variablen sowie zwischen Traits und Verhalten moderieren. Die Moderation zeigt sich in einigen Studien auch bei der Kriteriumsvalidität von Trait-Fragebogen: Bei Personen mit geringer Inter-Item-SD sind zum Teil bessere Vorhersagen möglich als bei Personen mit hoher Inter-Item-SD. Dies lässt sich jedoch auf Basis der Metatraits-Forschung nicht verallgemeinern.

2.1.2 Self-Concept-Differentiation

Unter SCD wird eine Inkonsistenz des Selbstkonzepts verstanden (Baird et al., 2006; Block, 1961; Donahue et al., 1993): Geringe SCD deutet auf ein über Rollen hinweg konsistentes Selbstkonzept hin, hohe SCD darauf, dass das Selbstkonzept ausdifferenziert und variabel ist.

[3] Dieses Verfahren wird auch für die in dieser Arbeit berichteten empirischen Studien verwendet und wird im Ergebnisteil von Studie 1 skizziert (Abschnitt 5.2.4).

Block (1961) vermutete, dass hohe ebenso wie geringe SCD maladaptiv sei: Starke Ausdifferenzierung sei Ausdruck von Rollen-Diffusion, hohe Konsistenz im Selbstkonzept sei Ausdruck von Rollen-Rigidität. Beide Zustände verhindern Blocks Hypothese zufolge das Einstellen auf und das erfolgreiche Bewältigen von verschiedenen Situationen bzw. sozialen Problemen. Diese Hypothese erschien zunächst plausibel, wurde jedoch in der Studie von Block durch die Daten nicht bestätigt. Allerdings fand Block einen linearen Zusammenhang zwischen SCD und der sozialen Anpassung: Personen mit konsistentem Selbstbild sind den Ergebnissen zufolge besser angepasst als Personen mit hoher SCD. Diesen Befund replizierten Donahue et al. (1993). In ihrer Studie korrelierte SCD mit schlechter Angepasstheit: So berichten sie Zusammenhänge von SCD mit Depressivität ($r = .44$) und mit geringem Selbstwertgefühl ($-.39$). Ein weiteres Ergebnis der Arbeit von Donahue et al. waren Zusammenhänge von SCD mit Neurotizismus ($.30$), mit Gewissenhaftigkeit ($-.45$) und mit Verträglichkeit ($-.27$). Ferner fanden Donahue et al. heraus, dass sich die Zusammenhänge zwischen SCD und den Persönlichkeitseigenschaften auch dann nachweisen lassen, wenn diese erst 30 Jahre später erfasst werden. Insgesamt legten die Befunde nahe, dass Personen mit weniger konsistentem Selbstbild schlechter angepasst sind, und sie lassen vermuten, dass SCD möglicherweise mit der Entstehung von psychischen Erkrankungen in Verbindung steht.

Besonders interessant an den Studien von Donahue et al. (1993) und Block (1961) ist die Methode, mit der SCD gemessen wurde. Personen beurteilten, wie gut Adjektive sie in unterschiedlichen sozialen Rollen beschreiben. Dazu beantworteten sie einen Fragebogen mehrere Male, jeweils mit der Instruktion, die Selbsteinschätzungen für eine bestimmte Rolle (z. B. Freund, Partner, Kind, Fremder) vorzunehmen. Entsprechend lag für jede Person eine Matrix mit Selbsteinschätzungen auf Adjektiven (Zeilen) für mehrere Rollen (Spalten) vor. Um herauszufinden, ob Personen sich in allen Rollen ähnlich beschreiben oder ob ihre Selbstbeschreibungen stark auseinander gehen, berechneten die Autoren für jede Person eine Hauptkomponentenanalyse (*Principal Component Analysis*, PCA) über alle Rollen hinweg. Die erste Hauptkomponente dieser PCA gab Auskunft darüber, wie viel Varianz zwischen den Items die Rollen gemeinsam hatten, d. h. wie stabil das Selbstkonzept über verschiedene Rollen hinweg war. Als *PCA-Index* für SCD wurde dieser Anteil von 1 subtrahiert.

In einer Fußnote berichteten Donahue et al. (1993), dass sie als Alternative zum PCA-Index auch einen anderen Index für SCD berechnet hatten: Sie addierten für jede Person die Stan-

dardabweichungen der Adjektiv-Items über die Rollen hinweg. Da die Autoren jedoch für diesen Index die gleichen Ergebnisse erhielten wie für den PCA-Index, berichteten sie diese nicht zusätzlich. Was sie dagegen berichteten, war die Reliabilität dieser Summe aus 60 Standardabweichungen, die mit $\alpha = .95$ hoch ausfiel. Die hohe Korrelation zum PCA-Index ($r \approx .80$) deutete darauf hin, dass SCD mit beiden Indizes reliabel erfasst wurde.

2.1.3 Methodische Probleme in der SCD- und Metatraits-Forschung

Der PCA-Index der SCD-Forschung (Block, 1961; Donahue et al., 1993) wurde von Baird et al. (2006) genauer analysiert: Die Hauptkomponente wird für eine Person über die Rollen hinweg gebildet, d. h. sie erklärt die gemeinsame Varianz zwischen den Rollen in der Matrix von Rollen und Adjektiven. Grundlage für die PCA sind jeweils die Korrelationen der Adjektiv-Paare für zwei Rollen. Diese Korrelationen fallen höher aus, je höher die Varianz zwischen den Adjektiv-Items innerhalb der Rollen ausfällt: Entsprechend wird die Hauptkomponente größer für Personen mit (in beiden Richtungen) extremerer Trait-Ausprägung. Laut Baird et al. handelt es sich dabei um eine Quelle für SCD irrelevanter Varianz. Sie illustrieren dies anhand eines Beispiels, das in Tabelle 1 referiert wird.

Tabelle 1: Antwortmuster auf Adjektiv-Items in verschiedenen Rollen von einer Person mit hoher Varianz und einer Person mit niedriger Varianz zwischen den Adjektiven (nach Baird, Le & Lucas, 2006; eigene Übersetzung)

	Person 1 (hohe Varianz zw. Adjektiven)			Person 2 (niedrige Varianz zw. Adjektiven)		
Adjektiv-Item	Rolle 1	Rolle 2	Rolle 3	Rolle 1	Rolle 2	Rolle 3
gesprächig	5	4	4	4	3	3
nervös	1	1	2	2	2	3
durchsetzungsfähig	4	5	4	3	4	3
organisiert	2	1	1	3	2	2
einfühlsam	5	5	4	4	4	3
fleißig	2	1	1	3	2	2
fürsorglich	4	5	5	3	4	4
reizbar	1	2	1	2	3	2
Standardabweichung zwischen den Items	1.69	1.93	1.67	0.76	0.93	0.71
Anteil der durch PCA aufgeklärten Varianz			92.69			63.59

Beide Personen, deren Antwortprofil abgebildet ist, haben für jedes Item die gleiche Streuung zwischen den Rollen ($SD_{x-role} = 0.58$). Die Ausprägungen der Adjektiv-Items innerhalb der Rollen streuen jedoch bei Person 1 stärker als bei Person 2, daher ist der durch die PCA

aufgeklärte Varianzanteil von Person 1 größer. Der PCA-Index ist entsprechend kleiner und führt zu dem Fehlschluss, dass Person 1 ein konsistenteres Selbstbild hat als Person 2. Baird et al. (2006) belegen dies anhand ihrer Daten: Die Varianz innerhalb der Rollen korreliert zu $r = -.24$ mit dem PCA-Index. Die Autoren schließen, dass der PCA-Index nicht nur die Konsistenz des Persönlichkeitsprofils über Rollen hinweg erfasst, sondern auch, wie ähnlich eine Person verschiedene Adjektiv-Items innerhalb einer Rolle beantwortet.

Die von Donahue et al. (1993) vorgeschlagene Methode zur Messung von SCD, die Summe der Standardabweichungen der Item-Antworten über die Rollen hinweg, ist eine Alternative zum PCA-Index. Standardabweichungen über Items hinweg dominierten auch in der Metatraits-Forschung als Operationalisierung, mit dem Unterschied, dass die Standardabweichung nicht für ein Item über Rollen hinweg, sondern für Items, die dieselbe Eigenschaft indizieren, berechnet wurde (Inter-Item-SD, vgl. Abschnitt 2.1.1). Unabhängig davon, ob sie für ein Item in verschiedenen Rollen oder für die Items einer Skala berechnet wird, stellt die Inter-Item-SD ebenfalls keine optimale bzw. valide Operationalisierung von SCD oder Variabilität dar: Sie ist mit dem Messwert (Item-Mittelwert oder Summe) der Skala bzw. des Items konfundiert, für die bzw. das sie bestimmt wird (Baird et al., 2006; Paunonen & Jackson, 1985): Hohe und niedrige Ausprägungen auf einer Trait-Skala oder einem (für verschiedene Rollen zu beantwortenden) Item können nämlich nur dann zustande kommen, wenn eine Person alle Items in der Nähe des Endpunktes der Likert-Skala beantwortet hat. In diesem Fall ist die Inter-Item-SD klein. Bei Personen mit mittlerer Ausprägung kann die Inter-Item-SD dagegen entweder hoch oder niedrig sein (Baird et al., 2006; Paunonen & Jackson, 1985). Abbildung 2 gibt eine Übersicht über alle möglichen Kombinationen von Mittelwerten[4] und Standardabweichungen einer Skala aus vier sechs-stufigen Likert-Items (vgl. Baird et al., 2006)[5].

[4] Bei Messungen von Variabilität beziehen sich die Begriffe „Mittelwert“ und „Quadrat des (z-standardisierten) Mittelwerts“ stets auf die individuellen Werte, d. h. auf den Item-Mittelwert einer Person auf einer Skala bzw. auf das Quadrat des (z-standardisierten) Item-Mittelwerts einer Person auf einer Skala. Gruppen-Mittelwerte werden explizit als solche bezeichnet.

[5] Baird, Le und Lucas (2006) veranschaulichen den Zusammenhang mit einer Grafik für sechs fünf-stufige Likert-Items; er wird hier mit Blick auf die empirischen Studien (Kapitel 5 und 6) für sechs-stufige Likert-Items illustriert. Die Darstellung ist zur besseren Übersichtlichkeit auf vier Items begrenzt.

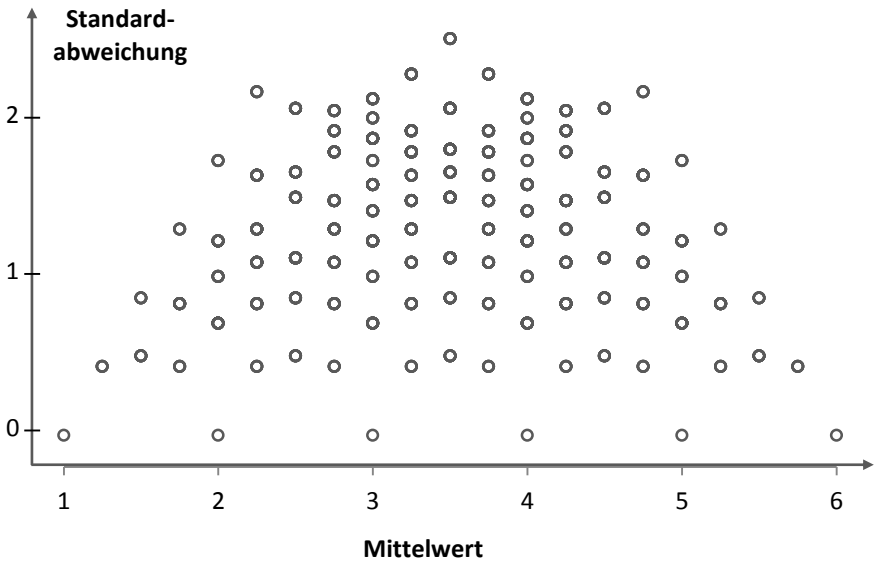

Abbildung 2: Überblick über die theoretisch möglichen Kombinationen an Mittelwerten und Standardabweichungen für vier sechs-stufige Likert-Items (vgl. Baird, Le & Lucas, 2006)

In Fällen, in denen der Gruppen-Mittelwert der Item-Mittelwerte nahe der Mitte der Likert-Skala liegt ($M = 3.5$), korreliert die Inter-Item-SD daher tendenziell negativ mit dem Abstand des Item-Mittelwerts vom Gruppen-Mittelwert. Personen an den Extremen haben i. d. R. eine geringere Inter-Item-SD als Personen im mittleren Bereich; Personen mit hoher Inter-Item-SD können keinen extremen Item-Mittelwert haben. Folglich ist die Streuung der Item-Mittelwerte unter Personen, die eine hohe Inter-Item-SD aufweisen, kleiner als unter Personen mit niedriger Inter-Item-SD (Baird et al., 2006; Paunonen & Jackson, 1985). Wird die Inter-Item-SD mittels Mediansplit als Moderator betrachtet – wie bei vielen Studien der Metatraits-Forschung – dann zeigt sich vermutlich allein aufgrund dieser unterschiedlichen Streuungen ein Moderatoreffekt bei der Vorhersage eines Kriteriums auf Basis des Skalen-Mittelwerts. Abbildung 3 veranschaulicht diesen Sachverhalt (vgl. Paunonen & Jackson, 1985)[6]. Der Index für Variabilität kann um diesen Einfluss bereinigt werden, indem die Inter-Item-SD um das Quadrat des (z-standardisierten) Item-Mittelwerts korrigiert wird bzw. werden (Baird et al., 2006; Reddock et al., 2011).

[6] Paunonen und Jackson (1985) führen dieses Beispiel für die Vorhersage von Fremdbeurteilungen auf der Basis von Selbsteinschätzungen an.

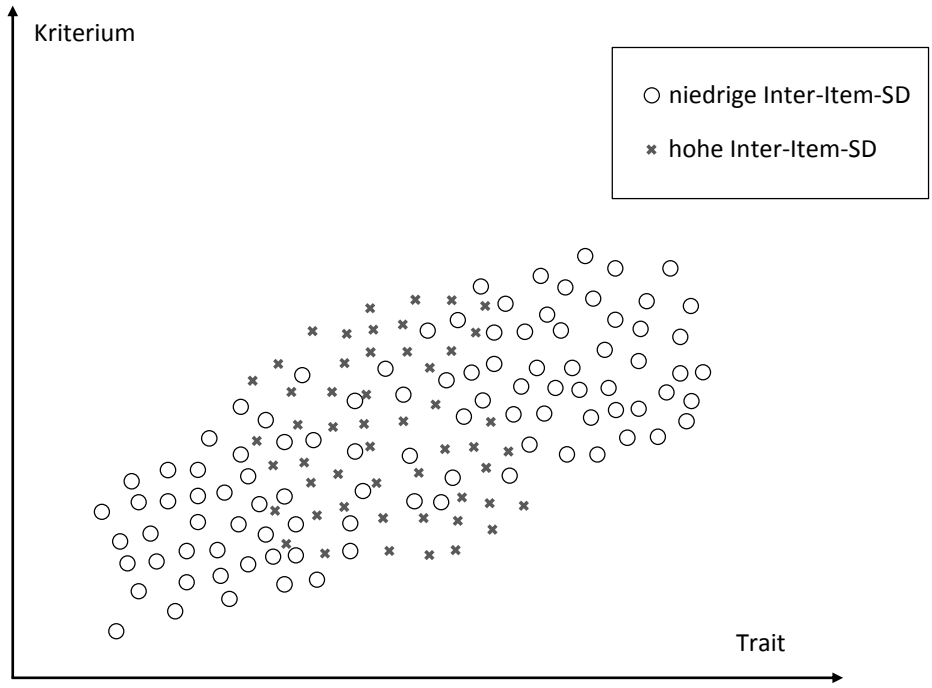

Abbildung 3: Streudiagramm für eine hypothetische Trait-Kriteriums-Beziehung bei Personen mit auf der Trait-Skala niedriger oder hoher Inter-Item-SD (vgl. Paunonen & Jackson, 1985)

In Fällen, in denen der Gruppen-Mittelwert des Item-Mittelwerts von der Mitte der Likert-Skala abweicht und die Item-Mittelwerte mehrheitlich entweder unterhalb oder oberhalb der Mitte verteilt sind, sind moderat lineare Zusammenhänge zwischen dem Item-Mittelwert und der Inter-Item-SD plausibel. Nachvollziehen lässt sich dies, indem man sich anhand des Streudiagramms in Abbildung 2 (Seite 13) vorstellt, der Gruppen-Mittelwert des Item-Mittelwerts sei größer als $M = 3.5$: In diesem Fall wären wenige bis keine Datenpunkte am unteren Extrem der Item-Mittelwerte und die meisten Datenpunkte wären in der rechten Hälfte der Grafik; Item-Mittelwert und Inter-Item-SD wären entsprechend moderat negativ korreliert. Befände sich der Schwerpunkt der Verteilung der Item-Mittelwerte unterhalb von $M = 3.5$, fiele die Korrelation vermutlich positiv aus. Für den Zusammenhang der Inter-Item-SD und dem Item-Mittelwert führen Baird et al. (2006) Belege an: Bei Items, die sie für mehrere Rollen erheben, korrelieren diese Statistiken im Mittel $|r| = .35$ miteinander. Die Zusammenhänge lassen sich auf die Schiefen der Verteilungen der Item-Mittelwerte zurückführen: Je größer diese für ein Item ausfielen, desto stärker hingen die Item-Mittelwerte mit den Inter-Item-SD zusammen. Nach Baird et al. sollten die Inter-Item-SD zur validen Messung von Variabilität entsprechend auch um den Item-Mittelwert korrigiert werden.

In einer weiteren Studie untersuchte Paunonen (1988) die Zusammenhänge zwischen Traits, Metatraits und subjektiver Wichtigkeit des Traits anhand expliziter Maße, die er mit Hilfe von Single-Items im Likert-Format erfasste. Die Trait-Einschätzung erfolgte für ein bipolares Adjektiv-Item (z. B. *meek* vs. *arrogant*), die Variabilität dieses Traits wurde wie bei D. J. Bem und Allen (1974) erhoben („*How much do you vary …?*“) und auch nach der Wichtigkeit wurde explizit gefragt („*How important or central to your self-description is the dimension …?*“). Die Ergebnisse von Paunonen waren vereinbar mit den Befunden von Paunonen und Jackson (1985): Variabilität und subjektive Wichtigkeit hingen für die meisten der erfassten Adjektiv-Items linear und kurvilinear mit dem Trait zusammen. Während die meisten der linearen – zum Teil negativen und zum Teil positiven – Zusammenhänge auf schiefe Verteilungen zurückzuführen waren, ließen sich die kurvilinearen Zusammenhänge verallgemeinern: Je weiter der Abstand vom Gruppen-Mittelwert, desto geringer war die Variabilität und desto höher war die subjektive Wichtigkeit des Traits. Die Befunde gingen weiter als die Folgerungen von Paunonen und Jackson (1985): Da die Zusammenhänge für explizite Maße gefunden wurden, ließen sich Schlüsse auf der Konstruktebene ziehen; schließlich ließ sich der Zusammenhang zwischen der Abweichung des Traits vom Gruppen-Mittelwert und der Variabilität nicht auf die methodischen Restriktionen (vgl. Abbildung 2) zurückführen, die bei impliziter Messung (als Inter-Item-SD) bestehen. Paunonen (1988) folgerte:

> Measures of variability, importance […] are nonlinearly related to measures of trait level. The individuals most extreme on a bipolar dimension of behavior, either high or low, generally are the most consistent in those behaviors, are likely to perceive the trait as being important to self-description, frequently engage in behaviors relevant to the trait, and tend to view their trait behaviors as being highly visible to observers. (S. 638)

Inwieweit dieses Fazit Bestand hat, hängt jedoch stark davon ab, wie valide die Messungen von Variabilität mit expliziten Maßen sind. Die Inter-Item-SD als implizites Maß von Variabilität korreliert aufgrund der Messmethode mit der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert, so dass sich das Fazit nicht ohne weiteres überprüfen lässt. Überprüfbar sind dagegen die Implikationen der Befunde von Paunonen (1988): Wenn Variabilität mit der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert korreliert, ist sie spezifisch für einen Trait. D. h., für zwei Traits konvergiert die Variabilität stärker, wenn die Traits korreliert sind, als wenn diese nicht korreliert sind. Diese Implikation steht im Einklang mit der Metatraits-Theorie. Systematische Studien zur Konvergenz von Variabilität für verschiedene

Traits und zur Struktur von Variabilität bzw. Metatraits führten Eid und Diener (1999) sowie Baird et al. (2006) durch. Die Befunde werden im nächsten Abschnitt (2.1.4) vorgestellt.

2.1.4 Intraindividuelle Variabilität als globaler und universeller Trait

Zur Beschreibung des Phänomens Variabilität trugen maßgeblich die Forschungsarbeiten von Eid und Diener (1999) sowie von Baird et al. (2006) bei. Eid und Diener untersuchten die Reliabilität, die Stabilität und die Struktur von Variabilitätsmaßen und, inwieweit die Vorhersagbarkeit von States von der Variabilität abhängt. Die Studie von Baird et al. knüpft an die SCD-Forschung an und behandelt die Messung und Struktur von Variabilität sowie die Vorhersage von *Well-Being* durch Variabilität. Im Folgenden wird zunächst die Studie von Eid und Diener vorgestellt; im Anschluss werden die Ergebnisse von Baird et al. referiert und abschließend werden die Ergebnisse von Baird in Bezug auf das Konzept *Self-Pluralism* erklärt.

Eid und Diener (1999): Intraindividuelle Variabilität im Affekt

Eid und Diener (1999) analysierten Variabilität nicht für Persönlichkeitseigenschaften, sondern für Emotionen. Über 52 Tage beantworteten Teilnehmer ihrer Studie täglich einen State-Fragebogen; die Variabilität wurde als Inter-Item-SD der jeweiligen States über die Zeit berechnet. Mit Strukturgleichungsmodellen (*Structure Equation Models*, SEM) fanden Eid und Diener heraus, dass Variabilität reliabel und stabil gemessen werden kann – und multidimensional ist: Den Ergebnissen zufolge konvergiert Variabilität für verschiedene Emotionen zwar hoch ($.41 \leq r \leq .84$) – 8 von 21 bivariaten Korrelationen zwischen den Konstrukten lagen sogar bei $r > .70$ –, dennoch handelt es sich um unterscheidbare Persönlichkeitsmerkmale. Ein weiteres Ergebnis war, dass die Vorhersagbarkeit von States zu späteren Zeitpunkten von der Variabilität der jeweiligen Emotion abhängt, unabhängig von der Zeitspanne zwischen der Erhebung von Variabilität und der Messung des States: Die Abweichung (Betrag) vom durch eine Baseline vorhergesagten State geht mit Variabilität einher. Die State-Emotionen variieren somit um einen Mittelwert (Trait-Emotion) und je größer die Variabilität der States ist, desto breiter ist die Verteilung der States um diesen Mittelwert.

Eid und Diener (1999) prüften auch die Zusammenhänge der Variabilitätsmaße mit Persönlichkeitsmerkmalen. Dazu führten sie multiple Regressionen zur Vorhersage der Variabilität der Emotionen durch, in denen der jeweilige Mittelwert der Emotion, das Quadrat des (z-

standardisierten) Mittelwerts sowie die Big Five als Prädiktoren eingingen. Jeder Variabilitätsindex wurde signifikant vom Mittelwert und dem Mittelwerts-Quadrat der entsprechenden Emotion vorhergesagt, ein weiterer Teil der Varianz der Variabilität der meisten Emotionen wurde durch Neurotizismus aufgeklärt; in die Regressionsgleichungen zur Vorhersage einiger Variabilitätsindizes ging zudem Extraversion ein. Variabilität wies jedoch für jede Emotion einen großen Varianzanteil (50-90 %) auf, der weder durch die Emotion noch durch die Big Five aufgeklärt wurde. Eid und Diener schlossen, dass Variabilität hinreichend verschieden von anderen Merkmalen ist und jeweils als eigener Trait betrachtet werden kann.

An der Studie von Eid und Diener (1999) sind mehrere Aspekte hervorzuheben: Zum einen konnten mit der mehrwöchigen Erhebung die Reliabilität und zugleich die Stabilität von Variabilität für verschiedene Merkmale (in diesem Fall Emotionen) bestimmt werden. Zum anderen wurde die Faktorenstruktur von Variabilität für verschiedene Merkmale systematisch untersucht. Während Eid und Diener den Zusammenhang von Variabilität und Skalen-Ausprägung (linear und kurvilinear, vgl. Abschnitt 2.1.3) in ihren Regressionsanalysen berücksichtigten, vernachlässigten sie diese Konfundierung allerdings bei der Ermittlung der Faktorenstruktur: Eid und Diener identifizierten eine multidimensionale Struktur anhand der Inter-Item-SD. Das Ergebnis, dass Variabilität für jede Emotion ein eigenes Konstrukt ist, könnte auch durch die Varianzanteile der Variabilitätsindizes bedingt sein, die jeweils auf den Item-Mittelwert der Emotion (bzw. seine Abweichung vom Gruppen-Mittelwert) zurückzuführen sind. Dieser Varianzanteil verringert nämlich die Zusammenhänge zwischen den Variabilitätsindizes, wenn die Emotionen nicht bzw. nicht hoch miteinander korrelieren.

Baird, Le und Lucas (2006): Zur „Natur" intraindividueller Variabilität

Baird et al. (2006) griffen die Methoden und Ergebnisse der Forschung zu SCD sowie die Befunde der Metatraits-Forschung auf und untersuchten die psychometrischen Eigenschaften von Variabilitätsmaßen umfassend. Zudem überprüften sie die Ergebnisse von Block (1961) und Donahue et al. (1993) mit valideren Maßen von SCD bzw. Variabilität. Block sowie Donahue und Kollegen fanden, dass hohe SCD Ausdruck eines inkonsistenten Selbst ist und mit schlechter Anpassung, d. h. mit niedrigerem psychosozialem Funktionsniveau und Vulnerabilität für psychische Erkrankungen, einhergeht. Diesen Befunden entgegen stehen Theorien, denen zufolge Variabilität adaptiv, d. h. Zeichen guter Anpassung, ist (S. L. Bem, 1975; Paul-

hus & C. L. Martin, 1988). Um diese beiden entgegenstehenden Theorien zu überprüfen und die Bedeutung von intraindividueller Variabilität zu bestimmen, untersuchten Baird et al. den Zusammenhang zwischen Variabilität und Well-Being. Bemerkenswert dabei war das konzeptuell und methodisch gründliche Vorgehen (La Guardia & R. M. Ryan, 2007) und das umfassende Untersuchungsdesign: In jeder der drei Studien von Baird et al. bearbeiteten Personen 20 Adjektiv-Items zur Erfassung der Big Five, einmal allgemein und dann für sechs verschiedene Rollen. Bearbeitet wurden auch ein weiterer Fragebogen zu den Big Five sowie Fragebogen zum Affekt und zur Lebenszufriedenheit. In der zweiten und in der dritten Studie wurde zusätzlich die *Experience Sampling Methodology* (ESM) eingesetzt: Über eine Woche hinweg gaben die Teilnehmer der Studie an acht zufällig gewählten Zeitpunkten pro Tag via Pager an, wie gut die 20 Adjektiv-Items sie im Augenblick beschreiben und in welcher der sechs Rollen sie am ehesten sind. In der dritten Studie wurden auch ein Retest nach 6-9 Monaten durchgeführt sowie Fremdbeschreibungen erhoben; zur Validierung der Variabilitätsindizes wurde die *Self-Pluralism*-Skala (McReynolds, Altrocchi & House, 2000) erhoben, mit der Selbsteinschätzungen von Konsistenz im Affekt und im Verhalten erfasst werden.

Als Indikator von Variabilität berechneten Baird et al. die Inter-Item-SD jedes Items über Rollen (Fragebogen) oder Momente hinweg (ESM) und korrigierten diese um den Item-Mittelwert und um das Quadrat des (z-standardisierten) Item-Mittelwerts. Die korrigierten Inter-Item-SD addierten sie zu einem Index für intraindividuelle Variabilität. Zur Bestimmung der Reliabilität wurde die Summe der korrigierten Inter-Item-SD für jede Dimension der Big Five berechnet: Für die fünf Summen zeigte sich sowohl für die Rollen-Fragebogen als auch bei den ESM-Daten eine hohe Konsistenz ($\alpha > .70$); bei der ESM-Erhebung zeigte sich die hohe Konsistenz der fünf Summen unabhängig davon, ob die Variabilität jeweils über Rollen oder Zeitpunkte hinweg berechnet wurde. Die Autoren schlossen, dass Variabilität ein breiter, globaler und eindimensionaler Trait ist: Personen, die bezüglich eines Persönlichkeitsmerkmals variabel über verschiedene Situationen und Rollen hinweg sind, sind dies auch bezüglich anderer Persönlichkeitsmerkmale. Auch die Stabilität von Variabilität wurde durch die Daten belegt: Für den Rollen-Fragebogen zeigte sich ebenso wie für die ESM-Daten eine hohe Retestreliabilität über einen 6-bis-9-Monats-Zeitraum ($r_{tt} \geq .70$). Die Hypothese, dass Variabilität über Rollen und Situationen hinweg mit Veränderungen über die Zeit einhergeht, konnte in Teilen bestätigt werden. Variabilität hing auch – parallel zu den Befunden von D. J. Bem

und Allen (1974) – mit der Übereinstimmung von Selbst- und Fremdbeschreibungen zusammen: Je variabler sich Personen für verschiedene Rollen einschätzten, desto weniger stimmte die Fremdeinschätzung ihrer Persönlichkeit mit der Selbsteinschätzung überein.

Baird et al. (2006) berichteten nicht nur den Index der korrigierten Inter-Item-SD, sondern auch den PCA-Index von Block (1961) und Donahue et al. (1993) sowie die Summe der (nicht korrigierten) Inter-Item-SD. Während sich für die Summe der korrigierten Inter-Item-SD keine nennenswerten Zusammenhänge mit den Big Five zeigten, fielen sie für die beiden anderen Indizes stellenweise moderat ($.20 < r < .50$) aus. Diese Zusammenhänge führten Baird et al. jedoch auf die Abhängigkeit des PCA-Index von der Variabilität des Profils bzw. auf die Abhängigkeit der Inter-Item-SD von den Item-Mittelwerten zurück (vgl. Abschnitt 2.1.3). Die Autoren demonstrierten auch, dass sich der Zusammenhang zwischen Well-Being und Variabilität vollständig durch die Abhängigkeit der Variabilitäts-Messungen vom Profil bzw. von der Trait-Ausprägung erklären ließ: Variabilität und Well-Being korrelierten nur, wenn Variabilität durch den PCA-Index oder die Summe der Inter-Item-SD erfasst wurde; die Korrelation ließ sich durch die mit den Items erfassten Traits, die Big Five, aufklären.

Von den drei Maßen für Variabilität korrelierte der PCA-Index am höchsten mit der Self-Pluralism-Skala ($r = .62$), die Summe der korrigierten Inter-Item-SD am niedrigsten ($r = .20$). Die Self-Pluralism-Skala ihrerseits klärte Varianz von Well-Being auf und korrelierte mit den Big Five: Also klärte das Gefühl bzw. die Einschätzung, ein konsistentes Selbst zu haben, Well-Being auf; es hing jedoch nicht mit Variabilität über Rollen und Situationen hinweg zusammen. Unklar ist, ob die Self-Pluralism-Skala Well-Being über die Big Five hinaus vorhersagte.

Zusammengefasst gelingt es Baird et al. (2006), die Summe der korrigierten Inter-Item-SD als Maß von Variabilität zu etablieren. Ihre Ergebnisse sind ein Hinweis darauf, dass Variabilität – anders als mit der Metatraits-Theorie angenommen (Abschnitt 2.1.1) und mit den Befunden Paunonens (1988) impliziert (vgl. Abschnitt 2.1.3) – ein stabiler, globaler Trait ist und für verschiedene Traits konvergiert. Bei variablen Personen sind die Diskrepanzen zwischen Selbst- und Fremdeinschätzungen der Persönlichkeit größer und die Veränderungen der Persönlichkeit über die Zeit möglicherweise größer als bei konsistenten Personen. Well-Being ist allerdings unabhängig von Variabilität. Die in früheren Arbeiten berichteten Zusammenhän-

ge (Block, 1961; Donahue et al., 1993) sind auf methodische Artefakte zurückzuführen. Den Einwand, dass bei der Korrektur der Inter-Item-SD zu viel Varianz herauspartialisiert wird, entkräften Baird et al.: Schließlich wurden die Inter-Item-SD jeweils lediglich um den Item-Mittelwert und das Quadrat des (z-standardisierten) Item-Mittelwerts eines Items korrigiert. Da jeder Trait mit vier Items gemessen wurde, wären nach Auffassung der Autoren keine bedeutsame Varianz eliminiert worden und Zusammenhänge zwischen den globalen Traits und der Variabilität möglich gewesen.

Intraindividuelle Variabilität, Self-Pluralismus und Authentizität

Die Ergebnisse zur Self-Pluralism-Skala bei Baird et al. (2006) sind nicht in derselben Weise durch statistische Artefakte zu erklären wie die zum Zusammenhang zwischen SCD und psychosozialer Anpassung (Block, 1961; Donahue et al., 1993). Anscheinend ist „Selbst-Pluralismus" ein anderes Merkmal als die Variabilität von Traits über Situationen, über Rollen und über die Zeit hinweg. McReynolds et al. (2000) definieren Selbst-Pluralismus als das Ausmaß, in dem eine Person sich selbst zu unterschiedlichen Zeitpunkten im Erleben und Verhalten unterschiedlich wahrnimmt. Ganz ähnlich – lediglich als Gegenpol formuliert – verstehen Sheldon, R. M. Ryan, Rawsthorne und Ilardi (1997) das Merkmal Authentizität. Dabei handelt es sich um die Wahrnehmung, „wie das eigene Selbst zu handeln" oder „man selbst zu sein"[7]. Interessanterweise hängt diese Eigenschaft mit Well-Being und Zufriedenheit zusammen (Sheldon et al., 1997). Dieser Zusammenhang ist unabhängig vom PCA-Index für SCD, obwohl Authentizität stark mit SCD zusammenhängt ($r = -.61$). Diese Befunde sind vereinbar mit denen von Baird et al. (2006) und können die Ergebnisse zur Self-Pluralism-Skala erklären: In beiden Studien wurden – vorausgesetzt, die Messungen von Authentizität und Selbst-Pluralismus sind konstruktvalide – auch sehr ähnliche Merkmale gemessen. Diese sind unabhängig von der Variabilität der Persönlichkeit über Rollen oder Situationen hinweg und klären Well-Being auf. Der Zusammenhang von Selbst-Pluralismus mit dem PCA-Index (Baird et al., 2006) ist möglicherweise auf den Zusammenhang dieses Index mit den Big Five zurückzuführen. Gleiches könnte entsprechend auch auf den Zusammenhang von Authenti-

[7] In der Self-Pluralism-Skala (McReynolds, Altrocchi & House, 2000) liegt der Fokus auf Erleben und Verhalten allgemein (Beispielitem: „*I occasionally behave unlike my normal self.*"); die Items zur Erfassung von Authentizität von Sheldon, R. M. Ryan, Rawsthorne und Ilardi (1997) waren auf unterschiedliche soziale Rollen (Student, Angestellter, Kind, Freund, Partner) bezogen (Beispielitem: „*I experience this aspect of myself as an authentic part of who I am.*").

zität mit dem PCA-Index (Sheldon et al., 1997) zutreffen. Als Fazit lässt sich festhalten, dass Variabilität nicht oder nur wenig mit dem Gefühl innerer Kohärenz zusammenhängt und davon abzugrenzen ist. Erleben innerer Konsistenz lässt sich als Authentizität bzw. als Gegenpol von Selbst-Pluralismus beschreiben und ist vermutlich höher ausgeprägt bei Personen, die extravertiert, verträglich, gewissenhaft, emotional stabil und intellektuell sind.

2.2 Einordnung in das Situations-Eigenschafts-Paradigma

Inwiefern lässt sich Variabilität mit Trait-Theorien vereinen und welchen Stellenwert hat Variabilität bei der Beschreibung der Zusammenhänge zwischen Traits, Situationen und Verhalten? Besonders wichtig für die Antworten auf diese Fragen sind die Ergebnisse von William Fleeson (2001, 2007) und ihre Implikationen: Fleeson fasste Variabilität als Bindeglied zwischen Eigenschaften und Situationen auf. Im Folgenden werden zunächst die Person-Situation-Debatte skizziert und situative und personale Determinanten von Verhalten erläutert (Abschnitt 2.2.1). Daran anschließend werden die Beiträge von Fleeson vorgestellt (Abschnitt 2.2.2) und schließlich wird Variabilität in das Situations-Eigenschafts-Paradigma eingeordnet (Abschnitt 2.2.3).

2.2.1 Die Person und die Situation als Determinanten von Verhalten

In der ersten Hälfte des 20. Jahrhunderts dominierte in der Persönlichkeitspsychologie das Eigenschaftsparadigma (Amelang & Bartussek, 2001; Mischel, 2004). Kern dieses Paradigmas ist die Vorstellung, dass Personen zeitlich stabile Persönlichkeitszüge aufweisen und dass diese Persönlichkeitszüge das Verhalten deutlich stärker beeinflussen als die objektive Reizstruktur (Amelang & Bartussek, 2001; Mischel, 2004); Persönlichkeitszüge entsprechen den Ausprägungen einer Person auf kontinuierlichen Eigenschaftsdimensionen (Asendorpf, 2004; Beermann, 2011). Die Tragweite des Eigenschaftsparadigmas zeigte sich in den umfassenden Bemühungen, Persönlichkeitsmerkmale – mittels dimensionaler Ansätze – zu identifizieren und zu kategorisieren (z. B. Cattell, 1944, 1945; Eysenck, 1944; J. P. Guilford & Braly, 1930; J. P. Guilford & R. B. Guilford, 1936; McCrae & Costa, 1987; Tupes & Christal, 1958, 1961). Daran, dass Verhalten maßgeblich durch Persönlichkeitseigenschaften erklärt werden kann, hegte insbesondere Mischel (1968) Zweifel. Angesichts nur geringer empirischer Konvergenz zwischen Traits und Verhalten räumte er situativen Faktoren einen deutlich größeren Stel-

lenwert bei der Erklärung von Verhalten ein (vgl. Amelang & Bartussek, 2001; Mischel, 2004; Schmitt, 2005). Zu dieser Ansicht beigetragen haben insbesondere die Befunde von Hartshorne und May (1928, zitiert nach Amelang & Bartussek, 2001; Asendorpf & Neyer, 2012; Mischel, 1968, 2004; Schmitt, 2005): Unter einer Gruppe von Schülern ermittelten Hartshorne und May nur eine durchschnittliche Korrelation von $\bar{r} = .19$ zwischen unterschiedlichen Formen ehrlichen Verhaltens. Diese Ergebnisse und die Arbeiten von Mischel (1968) haben die sogenannte Person-Situation-Debatte hervorgerufen und das Eigenschaftsparadigma in eine Krise geführt, die erst in den letzten Dekaden des 20. Jahrhunderts überwunden wurde (Amelang & Bartussek, 2001; Schmitt, 2005). Einen großen Beitrag zu Überwindung der Konsistenzkrise haben Arbeitsgruppen um Mischel selbst (Mischel & Peake, 1982; Shoda, Mischel & Wright, 1993, 1994) geleistet. So untersuchten Shoda et al. (1994) aggressives Verhalten von Kindern in einem Ferienlager und fanden nur wenig Konsistenz zwischen verschiedenen Situationen. Die Aggressions-Profile der Kinder waren allerdings stabil. Mit anderen Worten zeigt sich im Verhalten nicht unbedingt transsituative Konsistenz, aber intraindividuelle Kohärenz (= situationsspezifische Konsistenz).

Ebenfalls zur Auflösung der Person-Situation-Debatte haben interaktionistische Ansätze beigetragen. Mittlerweile ist unstrittig, dass sowohl Eigenschaften als auch Situationen das Verhalten beeinflussen (Schmitt, 2005): Verhalten kann durch (i) die Situation, durch (ii) Eigenschaften, durch (iii) die Situation und Eigenschaften sowie durch (iv) die Interaktion von Situation und Eigenschaften determiniert sein:

(i) Der Einfluss der Situation ist maßgeblich, wenn diese mächtig ist (z. B. Haney, Banks & Zimbardo, 1973; Milgram, 1963, 1974). Im Alltag bleiben zum Beispiel fast alle Menschen an einer roten Ampel stehen und alle Zuschauer klatschen nach einem Theaterstück oder nach einem Konzert, hier ist die interindividuelle Varianz im Verhalten klein. Auch soziale Rollen können – als situative Determinante – das Verhalten beeinflussen. Je stärker eine Situation ist, desto besser lässt sich das Verhalten einer beliebigen Person in dieser Situation vorhersagen (Mischel, 1968; Schmitt, 2005).

(ii) Eigenschaften bestimmen das Verhalten besonders in schwachen Situationen – zum Beispiel, wenn Menschen einen Urlaub planen, eine Wohnung einrichten oder eine lang-

fristig zu erledigende Arbeitsaufgabe beginnen. Als schwache Situation kann auch ein (Leistungstest-)Item mit mittlerer Schwierigkeit angesehen werden (Schmitt, 2005).

(iii) Es kann auch sein, dass Verhalten additiv durch ein Personenmerkmal und die Situation bestimmt wird. Ein Beispiel dafür liefern Asendorpf und Neyer (2012): Abbildung 4 zeigt, mit welcher Ausprägung an Angst vier Individuen auf unterschiedliche Situationen (von links nach rechts nach bedrohlicher) reagieren. Das Ausmaß an Angst hängt stark von der Situation ab, gleichzeitig bleiben die interindividuellen Unterschiede über alle Situationen hinweg nahezu konstant.

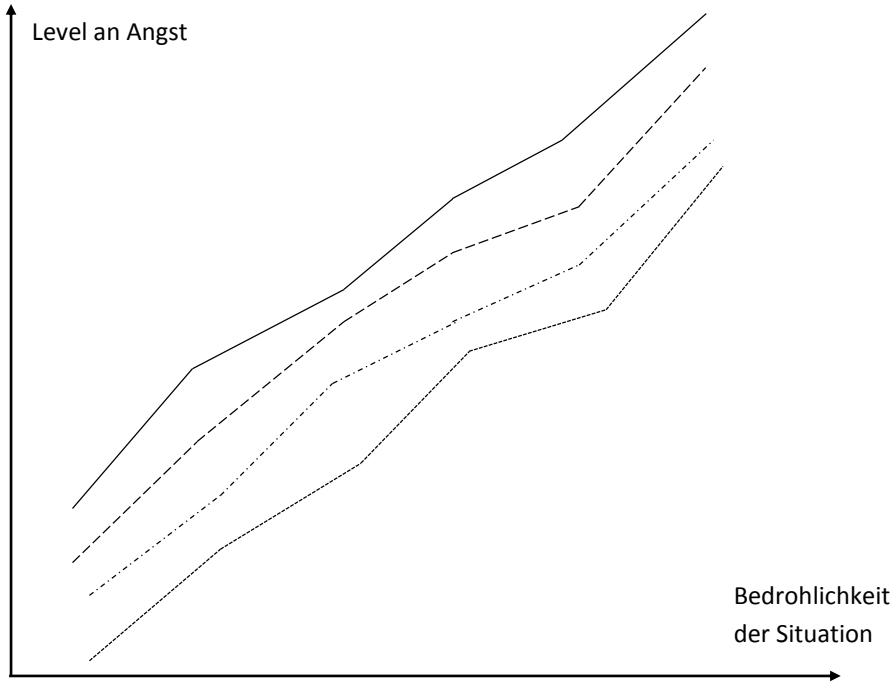

Abbildung 4: Angst-Level von vier Personen in Abhängigkeit der Bedrohlichkeit der Situation (Asendorpf & Neyer, 2012)

(iv) In vielen Fällen bestimmen nicht Eigenschaften *oder* die Situation *oder* die Addition beider Determinanten das Verhalten, sondern die Interaktion zwischen Traits und Situationen. Als Alltagsbeispiel dient der Vergleich einer Situation, in der Menschen Smalltalk führen, mit einer Situation, in der Menschen unterschiedlicher Meinung sind. Während in erster Situation Unterschiede in der beobachtbaren Konfliktbereitschaft zwischen verträglichen und wenig verträglichen Personen klein sind, sollten sie im Falle der Meinungsverschiedenheit gravierender ausfallen; Abbildung 5 veranschaulicht die Zusammenhänge. Effekte dieser Art wurden mehrfach belegt: So steigt die Aggressionsbereit-

schaft nach einem Karatefilm bei aggressiven Probanden stärker an als bei nicht aggressiven (Bushman, 1995) und Autofahrer mit Ärgerneigung werden nach frustrierenden Verkehrssituationen aggressiver als Autofahrer ohne Ärgerbereitschaft (Deffenbacher, 2003). Interaktionen von Eigenschaften und Situationen liegen auch dann vor, wenn interindividuelle Unterschiede in qualitativ unterschiedlichen Situationen unterschiedlich ausfallen wie bei Shoda et al. (1994) und in der Studie von Hartshorne und May (1928, zitiert nach Amelang & Bartussek, 2001; Asendorpf & Neyer, 2012; Mischel, 1968, 2004; Schmitt, 2005)

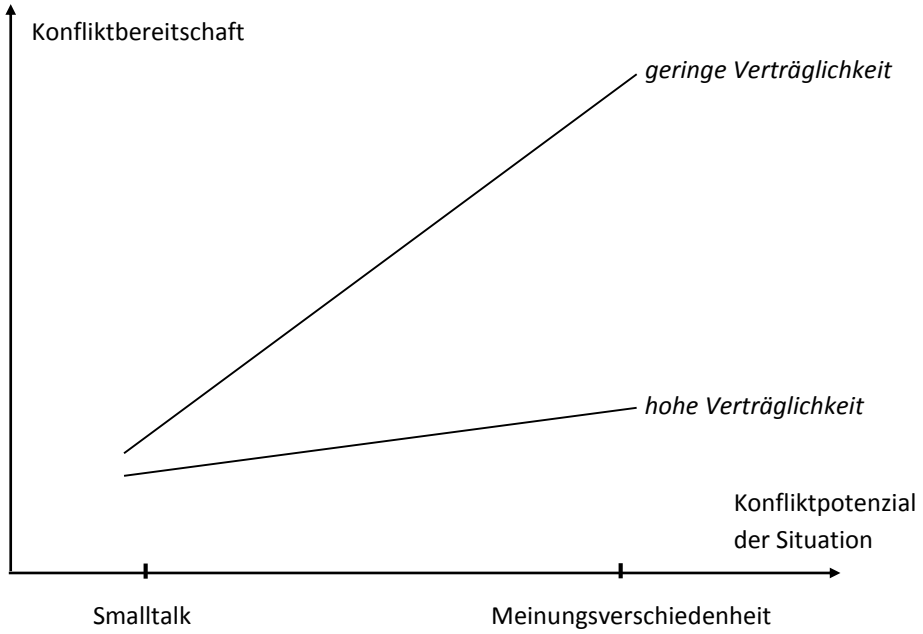

Abbildung 5: Interaktion zwischen der Persönlichkeitsdimension Verträglichkeit und dem Konfliktpotenzial einer Situation

2.2.2 Intraindividuelle Variabilität und die Situations-Verhaltens-Kontingenz

Ausgehend von den im vorherigen Abschnitt (2.2.1) aufgeführten Erkenntnissen untersuchte Fleeson (2001) den Zusammenhang zwischen Traits, States und Verhalten. Fleeson zufolge können bei nahezu jeder Person Verhaltensweisen und States für alle möglichen Ausprägungen eines Traits beobachtet werden. Zum Beispiel kann das Verhalten jeder Person in einigen Situationen als introvertiert, in anderen als extravertiert und in wieder anderen als mittelgradig extravertiert beschrieben werden; gleichermaßen sind die meisten Personen in einigen Situationen sehr ängstlich, in anderen wiederum verspüren sie keine Angst. Diese

Umstände nahm Fleeson zum Anlass, die Dichte-Verteilung von States zu untersuchen. Unter States verstand er die vorrübergehende Art und Weise des Erlebens, des Denkens und des Verhaltens. In einer ESM-Studie wurden Teilnehmer über 13 Tage hinweg mehrmals täglich gebeten, ihr Erleben und Verhalten der jeweils letzten Stunde im Hinblick auf die Big Five einzuschätzen. Im Einklang mit den Ergebnisse von Shoda et al. (1994) und denen von Hartshorne und May (1928, zitiert nach Amelang & Bartussek, 2001; Asendorpf & Neyer, 2012; Mischel, 1968, 2004; Schmitt, 2005) korrelierten dabei die States für beliebige Zeitpunkte kaum miteinander. Hohe Zusammenhänge dagegen zeigten sich für die Mittelwerte und Streuungen der States aus einer Hälfte der Erhebungseinheiten mit denen aus der anderen. Auch die Schiefe und der Exzess der State-Verteilungen waren stabil.

In einer Folgestudie wollte Fleeson (2007) herausfinden, inwiefern sich Personen mit breiter State-Verteilung von denen mit schmaler State-Verteilung unterscheiden. Dazu ging er noch systematischer vor und führte zwei ESM-Studien über zwei bzw. fünf Wochen durch: Mehrmals täglich wurden die Big Five States sowie die Charakteristiken der Situation mit je mehreren Items erfasst. Das erste Ziel von Fleeson war Merkmalsdimensionen („psychoaktive Merkmale“) von Situationen zu identifizieren. Mit einer Faktorenanalyse extrahierte er drei Faktoren: 1. wie freundlich die Stimmung in einer Situation ist, 2. wie anonym eine Situation ist (d. h. wie wenige Personen anwesend sind) und 3. wie strukturiert eine Situation ist. Als Zweites suchte Fleeson nach Kontingenzen zwischen diesen Merkmalsdimensionen und dem Verhalten und erhielt hypothesenkonforme Ergebnisse: Verhalten, das als extravertiert beschrieben werden kann, ging zum Beispiel einher mit der Freundlichkeit Situationen; der Grad der Gewissenhaftigkeit des Verhaltens korrelierte mit der Struktur der Situation. Zu beobachten waren jedoch auch interindividuelle Unterschiede im Ausmaß dieser Kontingenzen. Personen reagieren also unterschiedlich stark auf die Charakteristiken der Situation: Einige Personen sind in freundlichen Situationen extravertierter und in weniger freundlichen Situationen introvertierter, bei anderen Personen sind die Unterschiede zwischen dem Extraversions-State in freundlichen Situationen und dem in unfreundlichen Situationen nur gering. Fleesons drittes Anliegen war, diese interindividuellen Unterschiede zu erklären. Er zeigte, dass die Situations-Verhaltens-Kontingenzen, die sich auch als Reaktivität auf die psychoaktiven Merkmale der Situation interpretieren ließen, mit der Variabilität der States einhergehen: Je breiter die State-Verteilung einer Person war, desto höher war die Situa-

tions-Verhaltens-Kontingenz. Mit anderen Worten wurde das Verhalten von Personen mit hoher Variabilität stärker von den Charakteristiken der Situation bestimmt.

2.2.3 Fazit: Intraindividuelle Variabilität in der Person-Situation-Debatte

Die Beschreibung von Traits als Dichte-Verteilung von States hat weiter zum Abebben der Person-Situation-Debatte beigetragen (Fleeson & Leicht, 2006). Der Mittelwert solcher Verteilungen – die Trait-Ausprägung – ist stabil. Ebenso stabil ist die Breite oder Streuung von State-Verteilungen, die intraindividuelle Variabilität. Mit den Ergebnissen von Fleeson (2007) liegt auch Evidenz dafür vor, dass Variabilität im Verhalten mit Reaktivität auf die Merkmale von Situationen einhergeht. Variabilität lässt sich also als Moderator des Zusammenhangs zwischen situativen Faktoren und Verhalten verstehen: Verhalten von Personen mit hoher Variabilität wird stärker durch die Situation beeinflusst. Das Verhalten von Personen mit niedriger Variabilität hingegen hängt stärker von Traits ab; schließlich ist die Verteilung ihrer States enger und die Abweichung vom Mittelwert, der Trait-Ausprägung, in der Regel kleiner.

Die Zusammenhänge lassen sich mit Blick auf die in Abschnitt 2.2.1 vorgestellten Determinanten von Verhalten wie in Abbildung 6 vereinfacht darstellen: Damit das Verhalten von Personen mit niedriger Variabilität stark durch die Situation beeinflusst wird, muss die Situation sehr „mächtig“ sein. Verhalten von Personen mit hoher Variabilität wird dagegen schon bei schwächeren Situationen stark von situativen Determinanten bestimmt. Umgekehrt verhält es sich beim Einfluss der Persönlichkeit auf das Verhalten.

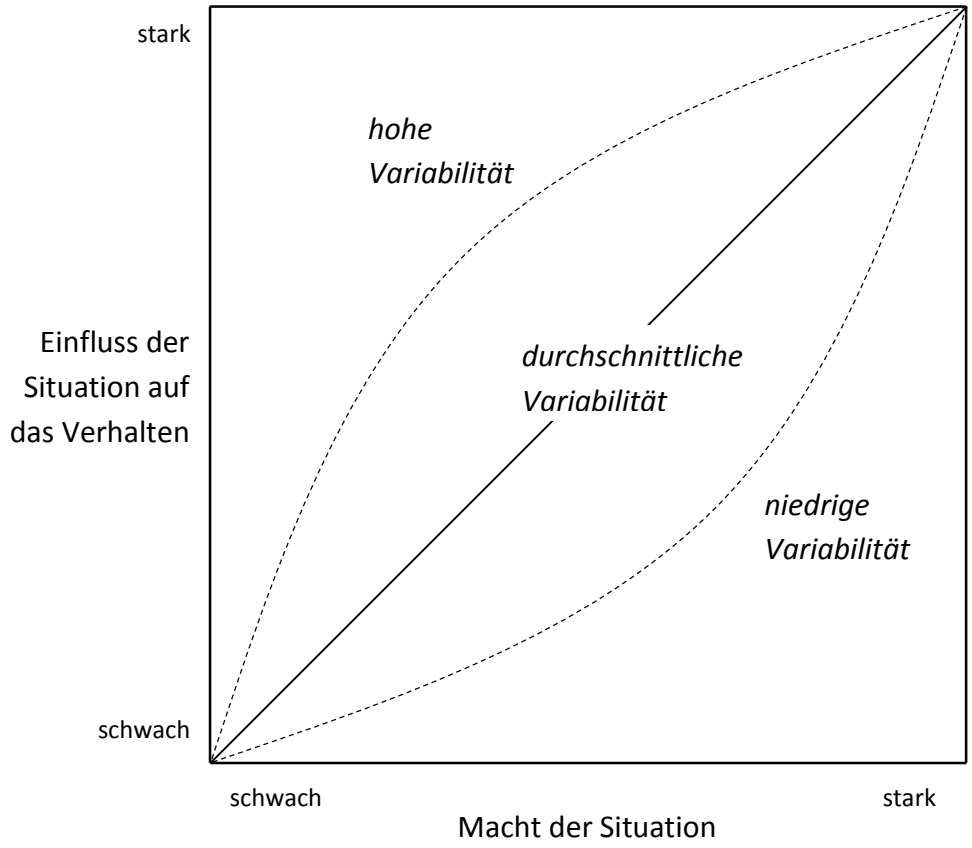

Abbildung 6: Der Einfluss von Variabilität und der Macht der Situation auf den Einfluss der Situation auf das Verhalten

Dieser theoretische Rahmen lässt sich auch auf soziale Rollen übertragen: Personen mit hoher Variabilität verhalten sich eher rollenkonform statt konform mit ihren Persönlichkeitsmerkmalen und Personen mit niedriger Variabilität verhalten sich über verschiedene Rollen hinweg ähnlich. Für die Interaktion von Person und Situation im Beispiel der Konfliktbereitschaft (vgl. Abbildung 5, Seite 24) kann gefolgert werden, dass sich bei Personen mit hoher Variabilität stärkere Interaktionseffekte zeigen als bei Personen mit niedriger Variabilität: Die Konfliktbereitschaft von Personen mit hoher Variabilität hängt stärker von der Situation ab als von Personen mit niedriger Variabilität. Die Zusammenhänge werden in Abbildung 7 dargestellt.

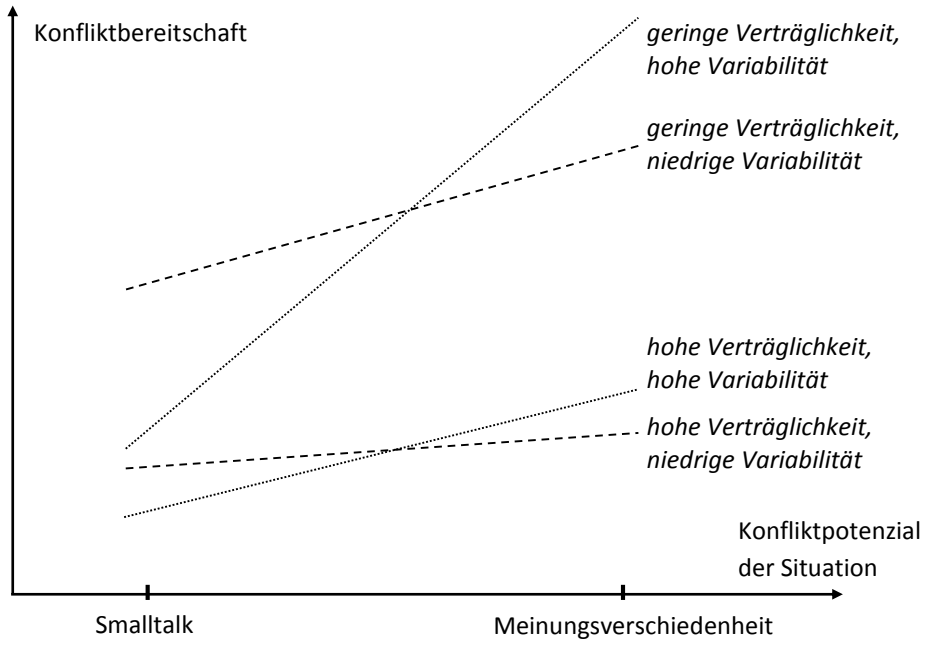

Abbildung 7: Interaktionen zwischen Variabilität, der Persönlichkeitsdimension Verträglichkeit und dem Konfliktpotenzial einer Situation

2.3 Intraindividuelle Variabilität in Persönlichkeitsfragebogen

Bereits im Rahmen der Metatraits-Forschung wurde Variabilität auf Skalen von Persönlichkeitsfragebogen erfasst und untersucht (Abschnitt 2.1.1). Zu Beginn dieses Jahrzehnts wurden drei weitere Studien zu Variabilität in Persönlichkeitsfragebogen veröffentlicht (Biderman & Reddock, 2012; Fleisher et al., 2011; Reddock et al., 2011). Diese widmen sich insbesondere der Moderation der Kriteriumsvalidität von Fragebogen durch Variabilität; die methodischen Probleme bei der Erfassung von Variabilität (Abschnitt 2.1.3) werden im Vergleich zu den Studien zu Metatraits überwunden. In den Studien von Biderman und Reddock (2012) sowie von Reddock et al. (2011) wird aufbauend auf den Ergebnissen von Baird et al. (2006; siehe Abschnitt 2.1.4) Variabilität für mehrere Skalen aggregiert, wohingegen Fleisher et al. (2011) die Variabilität für einzelne Traits erfassen.

Bei der Suche nach Moderatoreffekten kontrollierten Fleisher und Kollegen (2011) die Inter-Item-SD – anders als in den (meisten) Studien zu Metatraits – um den Item-Mittelwert und das Quadrat des (z-standardisierten) Item-Mittelwerts. Dennoch zeigten sich die erwarteten Effekte: Variabilität auf Verträglichkeit moderierte den Zusammenhang zwischen Verträglichkeit und *Team Performance*, Variabilität auf Gewissenhaftigkeit den Zusammenhang zwi-

schen Gewissenhaftigkeit und *Team Performance*: Der Zusammenhang war jeweils unter Personen mit niedriger Variabilität stärker als unter Personen mit hoher Variabilität. Eine Besonderheit bei der Studie von Fleisher et al. war, dass Variabilität mit einem Itemformat erfasst wurde, bei dem die Variabilität pro Item berechnet werden konnte, dem *Frequency-Estimation*-Format. Dieses Format ist in Abbildung 8 beispielhaft abgebildet. Bearbeiter mussten einschätzen, in wie viel Prozent der Gelegenheiten im letzten halben Jahr eine Aussage sie gut, weder gut noch schlecht oder schlecht beschreibt. Damit wurde für jedes Items eine trimodale Verteilung erfasst. Zur Messung der Variabilität, wurde jeweils die Streuung dieser Verteilung berechnet und für alle Items einer Skala aggregiert. Die Reliabilität dieses Aggregats für die erhobenen Skalen war jeweils zufriedenstellend ($\alpha > .80$). Ein globales Maß von Variabilität wurde nicht bestimmt.

„Ich genieße es, im Mittelpunkt zu stehen.“		
Bezogen auf die letzten sechs Monate, beschreibt mich die Aussage in		
_____ %	_____ %	_____%
der Gelegenheiten *eher gut.*	der Gelegenheiten *weder gut noch schlecht.*	der Gelegenheiten *eher schlecht.*

Abbildung 8: *Frequency-Estimation*-Format (Fleisher, Woehr, Edwards & Cullen, 2011)

Anders als Fleisher et al. (2011) verwendeten Reddock et al. (2011) Likert-Items. Sie berichteten ein globales Maß für Variabilität: Sie korrigierten die Inter-Item-SD auf Big-Five-Skalen jeweils um den Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts auf der entsprechenden Skala. Anschließend addierten sie die fünf Indizes zu einem globalen Variabilitätsindex, dessen Reliabilität anhand der Daten auf $\alpha = .81$ geschätzt wurde[8]. Die hohe Konsistenz werteten sie als Beleg dafür, dass Variabilität ein globaler Trait ist. Dieser Trait moderierte erwartungskonform die Reliabilität der Skalen sowie den Zusammenhang zwischen Gewissenhaftigkeit und Semester-Noten an der Universität: Die Reliabilität und der Zusammenhang waren unter Personen mit niedriger Variabilität höher als unter Personen mit hoher Variabilität. Interessanterweise korrelierte Variabilität bei Reddock et al. negativ mit Intelligenz und sagte ebenfalls Noten vorher (je niedriger die Variabilität, desto besser die No-

[8] Reddock, Biderman und Nguyen (2011) geben zwar nicht an, wie die Reliabilität des Variabilitätsindex geschätzt wurde. Dass Cronbachs Alpha verwendet wurde lässt sich jedoch erschließen; u. a. wird in dem Artikel an anderer Stelle Cronbachs Alpha als Reliabilitätsschätzung berichtet.

te). Die Vorhersage von Noten blieb sogar bestehen, wenn sie um den Einfluss der Intelligenz kontrolliert wurde. Eine Erklärung für dieses Phänomen lieferten die Autoren nicht.

Ein Jahr später adressierten Biderman und Reddock (2012) eine Einschränkung der Studie von Reddock et al. (2011): Als Moderator der Vorhersage von Kriterien war stets Variabilität auf dem Fragebogen untersucht worden, mit dem der Prädiktor erfasst worden war. Ein zufälliger Messfehler in einem Fragebogen könnte allerdings bei einigen Personen zu höherer Variabilität im Antwortverhalten und gleichzeitig zu geringerer Reliabilität und somit geringeren Validitätskoeffizienten führen. Um Messfehler als Ursache für die schlechtere Vorhersage ausschließen zu können und die bislang gezeigten Moderatoreffekte eindeutig auf das Merkmal Variabilität zurückzuführen, ließen Biderman und Reddock Studierende drei verschiedene Fragebogen zur Erfassung der Big Five bearbeiten. Mit einem der Fragebogen bestimmten sie die Variabilität. Mit den Skalen der anderen beiden Fragebogen berechneten sie die Reliabilität, die Konstruktvalidität und die Kriteriumsvalidität und untersuchten, ob Variabilität diese Testgütekriterien moderiert. Variabilität wurde sowohl als Summe der Inter-Item-SD der fünf Skalen als auch als Summe der korrigierten Inter-Item-SD ermittelt. Die Ergebnisse im Hinblick auf die Moderation der Reliabilität und Validität waren parallel, weshalb die Autoren die Ergebnisse für die Summe der nicht-korrigierten Inter-Item-SD berichteten. Zur Analyse auf Moderatoreffekte teilten die Autoren ihre studentische Stichprobe in drei Gruppen ein: inkonsistent, mittel konsistent und konsistent Bearbeitende. Zwischen diesen Gruppen unterschieden sich die Reliabilitätskoeffizienten für die meisten Skalen: je höher die Variabilität, desto niedriger war die Reliabilität. In gleicher Weise wurde die konvergente Konstruktvalidität für drei der Big-Five-Skalenpaare moderiert (Extraversion, Neurotizismus und Verträglichkeit): Sie war unter Personen mit niedriger Variabilität höher als unter Personen mit hoher Variabilität. Für die Skalen zu Offenheit für Erfahrungen und Gewissenhaftigkeit wurde der Effekt auf die Konstruktvalidität nicht gefunden. Die Kriteriumsvalidität der beiden infrage stehenden Fragebogen wurde als Vorhersage von Studiennoten durch Gewissenhaftigkeit operationalisiert. Die höchsten Validitätskoeffizienten zeigten sich für Personen mit mittlerer Variabilität, unter den Personen mit hoher Variabilität waren sie am niedrigsten. Während die Befunde zur Reliabilität und Konstruktvalidität konform mit bisherigen Forschungsergebnissen waren, fiel die Moderation der Kriteriumsvalidität nicht hypothesenkonform aus, wofür Biderman und Reddock keine Erklärung anbieten. Offen

bleibt auch, warum die Stichprobe in drei Gruppen (à 68 bzw. 69 Personen) eingeteilt wurde. Auch diskutieren Biderman und Reddock nicht den möglichen Schluss auf den Zusammenhang von Variabilität, Reliabilität und Kriteriumsvalidität, der sich aus ihren Ergebnissen ziehen lässt: Die Abhängigkeit der Kriteriumsvalidität von der Variabilität wird nicht durch die Reliabilität mediiert. Denn für die Dimension Gewissenhaftigkeit zeigte sich ein Moderatoreffekt bei der Kriteriumsvalidität, nicht aber bei der Reliabilität.

Zwar diskutierten Biderman und Reddock (2012) die Befunde zur Reliabilität und Validität nicht hinreichend, führten allerdings zwei Einschränkungen an: Zum einen bestand die Möglichkeit, dass Unterschiede zwischen der Verarbeitung von positiv und von negativ gepolten Items bestanden und dass Variabilität zum Teil auf diese Unterschiede zurückzuführen ist. Diese Hypothese konnten Biderman und Reddock widerlegen: Die berichteten Effekte zeigten sich sowohl für die negativ als auch für die positiv gepolten Items separat. Zum anderen vermuteten die Autoren, dass ERS einen Einfluss auf die Ergebnisse hatte. ERS korrelierte zu $r = .42$ mit Variabilität, hatte aber keinen Einfluss auf die Reliabilität. Die Konstruktvalidität wurde in erwartungswidriger Richtung moderiert: Unter extrem Antwortenden fiel sie höher aus als unter nicht extrem Antwortenden. Für die Kriteriumsvalidität waren die Ergebnisse parallel zu denen für Variabilität. Eine Erklärung für diese Ergebnisse liefern Biderman und Reddock (2012) nicht. Fraglich ist, ob die Befunde zu ERS stabil sind; die Autoren fordern daher:

> A more detailed investigation of the joint relationships of inconsistency and extreme response style to reliability and validity is also called for. (S. 651)

Die Zusammenhänge von Variabilität und ERS werden in dieser Arbeit empirisch untersucht (Kapitel 5 und 6). Im nächsten Kapitel (3) werden Forschungsbefunde zu ERS vorstellt.

3 Extreme Response Style

Beim Beantworten von Fragebogen reagieren Personen nicht nur im Sinne des zu erfassenden Merkmals auf den jeweiligen Reiz, das Item. Das Antwortverhalten wird zusätzlich von Antwortstilen beeinflusst, die sachlogisch nicht mit dem erfassten Merkmal verknüpft sind (Baumgartner & Steenkamp, 2001; Cronbach, 1946; Van Vaerenbergh & Thomas, 2013). Ein Antwortstil, der in den letzten Jahren in der Forschung viel Aufmerksamkeit erfahren hat, ist ERS (Weijters et al., 2010b). Unter ERS wird die Tendenz verstanden, extrem, d. h. an den Endpunkten einer (Likert-)Skala, zu antworten (Berg & Collier, 1953; Greenleaf, 1992b; Hamilton, 1968; Van Vaerenbergh & Thomas, 2013). Untersucht wurde ERS hauptsächlich in Einstellungsfragebogen (z. B. Baumgartner & Steenkamp, 2001; Kieruj & Moors, 2013; Meisenberg & Williams, 2008; Weijters et al., 2010b, 2010c); aber auch zu ERS in Persönlichkeitsfragbogen liegen Studien vor (Austin, Deary & Egan, 2006; Iwawaki & Zax, 1969; M. Johnson, 2013; Wetzel, Böhnke, Carstensen, Ziegler & Ostendorf, 2013a; Wetzel et al., 2013b). Da die Übergänge zwischen beiden Fragebogentypen fließend sind und mit beiden Dispositionseigenschaften erfasst werden, lassen sich die Forschungsergebnisse übertragen (Schmitt, 1992; Sherman & Fazio, 1983).

Wie relevant ERS für die empirische Sozialforschung ist, lässt sich an der großen Zahl an Publikationen zum Thema ablesen – und am *Common Sense*, dass ERS eine Bedrohung für die Validität von Fragebogen-Messungen ist (Baumgartner & Steenkamp, 2001; Cronbach, 1946; De Beuckelaer, Weijters & Rutten, 2010; Naemi et al., 2009; Van Vaerenbergh & Thomas, 2013; Weijters et al., 2010b; Wetzel et al., 2013b). Allgemeingültige Aussagen darüber, wie die Validität genau beeinträchtigt wird, lassen sich trotz der zahlreichen Befunde der letzten 60 Jahre nicht treffen. Auch die Ursachen bzw. die Entstehung von ERS sind weitgehend ungeklärt. In den folgenden Abschnitten werden die einschlägigen Forschungsergebnisse dargestellt und ein Überblick über die Problemfelder gegeben. Es folgt eine Übersicht über die Forschung zur Erfassung und Beschreibung (Abschnitt 3.1) sowie zu den Ursachen und Korrelaten (Abschnitt 3.2) von ERS. Abschließend werden Studien zur Beeinträchtigung der Validität von Fragebogen durch ERS referiert und diskutiert (Abschnitt 3.3).

3.1 Die Beschreibung und Erfassung von Extreme Response Style

Viele Forschergruppen sehen ERS als stabile Verhaltenstendenz bzw. als Personenmerkmal an (Berg & Collier, 1953; Bolt & Newton, 2011; De Beuckelaer et al., 2010; Greenleaf, 1992b; Merrens, 1970; Naemi et al., 2009; Weijters et al., 2010b, 2010c). Voraussetzungen für diese Einordnung sind valide Messungen von ERS sowie die Generalisierbarkeit und die Stabilität dieser Messungen. Im Folgenden werden Methoden zur Erfassung von ERS vorgestellt (Abschnitt 3.1.1) und Evidenz für die Generalisierbarkeit und die Stabilität von ERS aufgeführt (Abschnitt 3.1.2). Wenige Forschungsarbeiten zeigen auch einen Zusammenhang zwischen der Inter-Item-SD in Fragebogen und ERS, die entsprechenden Studien werden am Ende dieses Unterkapitels skizziert (Abschnitt 3.1.3).

3.1.1 Methoden zur Erfassung von Extreme Response Style

Zur Erfassung von ERS werden mehrere Ansätze beschrieben, die sich im Wesentlichen in zwei Klassen einteilen lassen (Kieruj & Moors, 2013): Methoden, bei denen ERS durch Abzählen der Extremwerte bestimmt wird, und statistische Schätzmethoden.

Abzähl-Methoden zur Bestimmung von ERS

Unter den Abzähl-Methoden dominerte lange Zeit das Erfassen von ERS als Anteil von Extremantworten an allen Antworten im Fragebogen (Berg & Collier, 1953; Borgatta & Glass, 1961; Crandall, 1973, 1982; G. Marín, Gamba & B. V. Marín, 1992; Meisenberg & Williams, 2008; Merrens, 1970). Diese Methode erlaubt jedoch nur dann valide Messungen, wenn ein Fragebogen ausreichend lang ist und die Iteminhalte heterogen sind; andernfalls spiegelt der Anteil extremer Antworten auch einen gemeinsamen Inhalt der Skalen wider (Bolt & Newton, 2011; Greenleaf, 1992b; Van Vaerenbergh & Thomas, 2013). Schließlich können Extremantworten nicht nur ein Indikator von ERS, sondern auch von hohen Ausprägungen auf den erfassten Traits sein. Ausschließen lässt sich die Konfundierung der Extremwerthäufigkeit im gesamten Fragebogen mit Trait-Ausprägungen nur schwer, weil sich auch in mehrdimensionalen, heterogenen Fragenbogen, wie z. B. Big-Five-Fragebogen, unter Umständen ein globaler Methoden- (Biderman, Nguyen, Cunningham & Ghorbani, 2011) oder Meta-Faktor (van der Linden, te Nijenhuis & Bakker, 2010) findet.

Eine Möglichkeit, den Einfluss der Trait-Ausprägungen auf die ERS-Messung zu verringern, schlägt Greenleaf (1992b) vor. Er verwendet eine Skala zur Erfassung von ERS, deren 16 Likert-kodierte (1 bis 6) Items nicht miteinander korrelieren, d. h. diese Items haben keinen gemeinsamen Inhalt und messen weder dasselbe Konstrukt noch denselben Methodenfaktor. ERS wird als Anzahl der extremen Antworten auf diese 16 Items operationalisiert, also als Summe der 16 Items in dichotomem Format (Extremantwort: 1, andere Antwort: 0)[9]. Anhand stochastischer Modelle arbeitet Greenleaf heraus, dass ERS reliabler gemessen wird, je mehr Items in diese Skala eingehen. Greenleaf zufolge wird ERS mit dieser Methode auch reliabler gemessen, wenn der Anteil an Extremantworten bei den betreffenden Items – wie bei allen Skalen, die aus dichotomen Items bestehen (Bühner, 2011; Schmidt-Atzert & Amelang, 2012) – nicht zu niedrig ist.

Ganz ähnlich wie mit der Greenleaf-Skala wird ERS mit der RIRS-Methode (*Representative Indicators for Response Styles*) von Weijters (2006) erfasst: Aus verschiedenen Fragebogeninventaren wird eine repräsentative bzw. zufällige Stichprobe von mindestens 15 Items gewählt, die bei der jeweiligen Studie *zusätzlich* erhoben werden müssen und deren Extremwerthäufigkeit dann ERS indiziert. De Beuckelaer et al. (2010) vergleichen diese Methode mit der Häufigkeit von Extremwerten bei bedeutsamen Skalen, d. h. mit der Abzähl-Methode bei einem Fragebogen, der relevante Merkmale misst, und stellen konvergente Validität fest: Die Häufigkeit der Extremantworten in diesem Fragebogen korreliert hoch mit zwei verschiedene RIRS-Skalen à 15 Items. ERS lässt sich also prinzipiell unabhängig von der Methode messen.

Statistische Schätzmethoden zur Bestimmung von ERS

Über Abzähl-Methoden hinaus wurde ERS insbesondere in den letzten 10 Jahren auch als latente Variable in SEM oder Modellen der Item-Response-Theorie (IRT) ermittelt. Laut der meisten Studien in diesem Bereich liegen die Vorteile dieser Methoden in der simultanen Messung von ERS und der relevanten Merkmale auch ohne zusätzliche Items. Für die Konzeption von ERS in SEM sei hier beispielhaft der Ansatz von Weijters et al. (2010b) genannt:

[9] Die Summe der 16 dichotomisierten Items von Greenleaf (1992b) wird von Naemi, Beal und Payne (2009) als „Greenleaf-Skala“ bezeichnet; Naemi et al. bilden eine solche Skala auch mit anderen Items und bezeichnen diese als „eigene Greenleaf-Skala“. In dieser Arbeit wird der Begriff „Greenleaf-Skala“ ebenfalls verwendet. Sofern nicht anders ausgewiesen, ist damit allgemein die Summe der Extremantworten auf eine Auswahl von 16 Items gemeint, die – im Likert-Format – nicht bzw. nur gering miteinander korrelieren.

Weijters et al. ließen Untersuchungsteilnehmer einen Fragebogen von 112 Items bearbeiten und teilten die Items in fünf Blöcke (à 22 bis 23 Items). Für diese Item-Blöcke ermittelten sie die Häufigkeiten extremer Antworten und überprüften für mehrere SEM, wie gut die Daten zum Modell passten. Den besten Fit erzielte ein Modell mit einem tau-äquivalenten ERS-Faktor und autoregressiven Effekten (vgl. Abbildung 9): In diesem Modell werden die Extremwerthäufigkeiten der fünf Item-Blöcke (y_1 bis y_5) sowohl von einem gemeinsamen Faktor als auch von der Extremwerthäufigkeit im jeweils vorangehenden Item-Block bestimmt (autoregressive Effekte). Der Einfluss des ERS-Faktors auf jede der Extremwerthäufigkeiten ist gleich groß, d. h. alle Ladungen sind gleich ($\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5$). Die autoregressiven Effekte sind ebenfalls gleich groß ($\beta_1 = \beta_2 = \beta_3 = \beta_4$), allerdings deutlich kleiner als die Ladungen auf dem Faktor. Die Extremwerthäufigkeit wird also stärker von der zugrunde liegenden Eigenschaft bestimmt als von der Extremwerthäufigkeit des jeweils vorangehenden Item-Blocks. Dieses Modell passte nicht nur gut zu den Daten von Weijters et al., sondern auch zu denen von Hui und Triandis (1985), die Weijters et al. mit ihrem SEM ebenfalls reanalysierten.

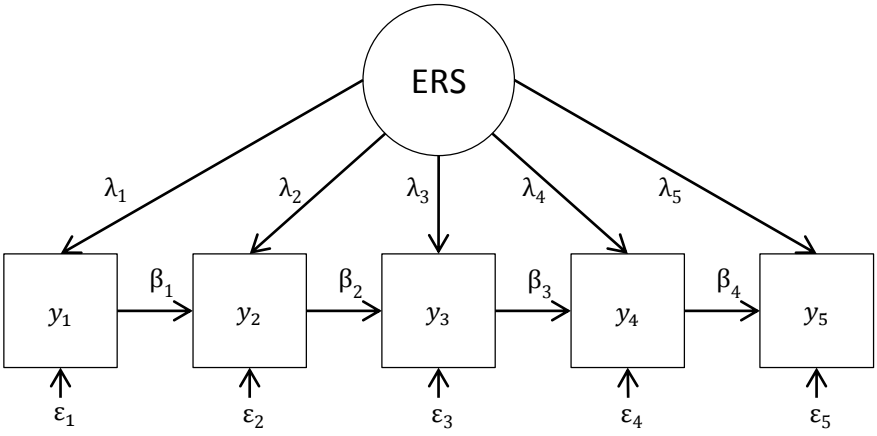

Abbildung 9: Strukturgleichungsmodell zur Erfassung von ERS als tau-äquivalenter Faktor von Extremwerthäufigkeiten mit autoregressiven Effekten (β_1 bis β_4) (Weijters, Geuens & Schillewaert, 2010b)

Als Beispiele für IRT-Modelle zur Messung von ERS seien die Ansätze von Bolt und Newton (2011) sowie Wetzel et al., (2013b) aufgeführt: Bolt und Newton zeigen anhand der Ergebnisse einer Simulation, dass sich ERS als kontinuierliche Variable in einem mehrdimensionalen ordinalen Rasch-Modell darstellen lässt[10]. ERS wird dabei als latente Dimension betrach-

[10] Eine ausführliche Beschreibung ordinaler Rasch-Modelle finden interessierte Leser bei Andrich (1978), bei Masters (1982) und im Lehrbuch von Rost (2004).

tet, zu der das Antwortverhalten ebenso wie zur inhaltlichen Dimension (das zu erfassende Merkmal) in einem stochastischen Zusammenhang steht. Entsprechend geht ERS auch als latente Dimension in die Wahrscheinlichkeits-Gleichung für die Antwortkategorien ein:

$$P(U_j = k|\theta_{\mathrm{m}}, \theta_{\mathrm{ERS}}) = \frac{\exp(a_{jkm}\theta_{\mathrm{m}} + a_{jkERS}\theta_{\mathrm{ERS}} + c_{jk})}{\sum_{h=1}^{k} \exp(a_{jhm}\theta_{\mathrm{m}} + a_{jhERS}\theta_{\mathrm{ERS}} + c_{jh})}$$

Die Gleichung beschreibt für eine Person mit der Ausprägung θ_{m} auf dem Merkmal und der Ausprägung θ_{ERS} auf der ERS-Dimension die Wahrscheinlichkeit, die Antwortkategorie k zu wählen. Der Parameter a_{jkm} bezieht sich auf die Position der Antwortkategorie in der Likert-Skala, (für das Beispiel von fünf Kategorien, $1 \leq k \leq 5$, mit gleichem Abstand z. B. -2; -1; 0; 1; 2). Der Parameter a_{jkERS} kann Bolt und Newton (2011) zufolge zwei Ausprägungen annehmen: eine für die Extrempunkte ($k = 1$ und $k = 5$) und eine für die übrigen Kategorien ($k = 2 \leq k \leq 4$). Der Parameter θ_{ERS} wirkt sich also ausschließlich auf die Wahrscheinlichkeit aus, extrem oder nicht extrem zu antworten.

Im Gegensatz zu Bolt und Newton (2011) beschreiben Wetzel et al. (2013b) ERS als Gruppenvariable, nicht als Dimension: Personen können entweder zur Gruppe bzw. Klasse der extrem Antwortenden oder zur Gruppe der nicht extrem Antwortenden gehören. Im ordinalen Rasch-Modell bedeutet dies Wetzel et al. zufolge eine horizontale Verschiebung der itemcharakteristischen Kurven (Schwellenfunktionen) für das Überschreiten der ersten und das Überschreiten der vorletzten Antwortkategorie. Der Sachverhalt wird in Abbildung 10 für ein vier-stufiges Likert-Format (1 bis 4) veranschaulicht. In der Abbildung oben sind die Schwellenfunktionen für die Klasse extrem Antwortender aufgeführt, unten für die Klasse nicht extrem Antwortender. Bei extrem Antwortenden ist die Schwelle zur zweiten Antwortkategorie höher (d. h. weiter rechts) als bei nicht extrem Antwortenden, was bedeutet, dass sie erst bei einem höheren Itemparameter die „2“ (und nicht mehr die „1“) ankreuzen. Die Schwelle zur vierten Antwortkategorie liegt bei extrem Antwortenden niedriger (weiter links); das bedeutet, das Ankreuzen der „4“ ist wahrscheinlicher als bei nicht extrem Antwortenden.

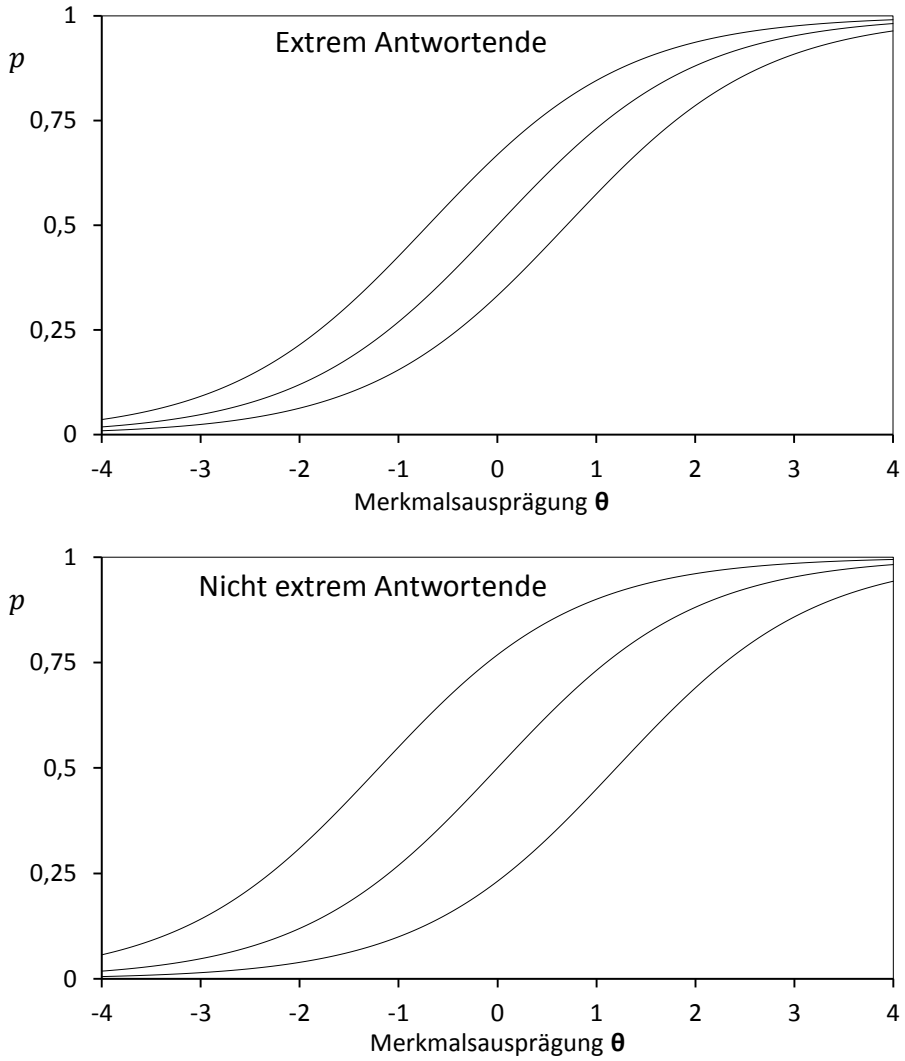

Abbildung 10: Operationalisierung von ERS als latente Klasse im ordinalen Rasch-Modell (Wetzel, Carstensen & Böhnke, 2013b)

p: Wahrscheinlichkeit, dass Schwelle überschritten wird. Dargestellt sind die Schwellen von der ersten zur zweiten, von der zweiten zur dritten und der dritten zur vierten Antwortkategorie einer vier-stufigen Likert-Skala.

Zusammenfassend lässt sich für statistische Methoden zur Schätzung von ERS festhalten, dass wie bei Abzählmethoden ausschließlich das Ankreuzverhalten der Extremkategorien betrachtet wird. Derzeit bestehen mehrere Ansätze – unklar ist, mit welcher Methode ERS am besten erfasst werden kann. Diese Unklarheit wird verstärkt durch die Implikationen der einzelnen Methoden; so kann ERS nicht latente Dimension und latente Klasse zugleich sein. Ein weiterer Nachteil statistischer Modellierung von ERS stellen der Analyse-Aufwand und das dafür nötige Know-how dar (Van Vaerenbergh & Thomas, 2013). Derzeit ist nicht abzusehen, inwieweit solche Ansätze den Abzählmethoden tatsächlich überlegen sind.

3.1.2 Extreme Response Style als stabiles Personenmerkmal

Frühe Evidenz dafür, dass ERS eine stabile Verhaltenstendenz ist, die sich generalisieren lässt, lieferte Merrens (1970): Versuchspersonen in seiner Studie beurteilten auf semantischen Differenzialen mit einer sieben-stufigen Likert-Skala, wie angenehm (vs. unangenehm) und wie spannend (vs. entspannend) 10 visuelle und 10 auditive Stimuli auf sie wirkten. Die Stimuli waren weitgehend inhaltslos und ambivalent in Bezug auf die beiden Beurteilungsdimensionen. ERS, als Häufigkeit der Endpunktwahl, konnte also nicht auf die Bewertung der Stimuli zurückgeführt werden. Die Konvergenz zwischen ERS bei visuellen und auditiven Stimuli ($r = .68$) ließ darauf schließen, dass ERS konsistent und unabhängig von den zu beurteilenden Stimuli auftritt. Die Konvergenz mit den Ergebnissen eines Retests eine Woche später ($.60 \leq r \leq .75$) verdeutlichte die Stabilität von ERS. Die Stabilität wurde bereits von Berg und Collier (1953) festgestellt und wie die Befunde zur Generalisierbarkeit mehrfach repliziert (z. B. Bachman & O'Malley, 1984; Wetzel et al., 2013b). Weijters et al. (2010c) stellten fest, dass ERS über ein Jahr hinweg stabil bleibt: Im Mittel waren 65 % der Varianz zeitunabhängig. Folglich lässt sich die Aussage treffen, dass ERS stabil und personengebunden ist.

ERS lässt sich nicht nur über Instrumente und Zeitpunkte generalisieren, sondern auch über Methoden, wie De Beuckelaer et al. (2010) für die RIRS und die Abzähl-Methode festgestellt haben (siehe Abschnitt 3.1.1). Auch Naemi et al. (2009) berichteten, dass die ERS-Skala von Greenleaf (1992b; siehe Abschnitt 3.1.1) sowie eine weitere Greenleaf-Skala moderat bis hoch mit der Extremwerthäufigkeit korrelierten ($.45 \leq r \leq .70$) und dass sich für diese drei ERS-Indikatoren ein Faktor extrahieren ließ, auf dem alle drei hoch luden ($a > .80$). In der Studie von Kieruj und Moors (2013) zeigte sich Konvergenz zwischen einer Greenleaf-Skala mit 18 Items und einem latenten Faktor für ERS, der mit anderen Items zu einem anderen Zeitpunkt[11] gemessen wurde: Zwar lagen die Korrelationen bei $.37 \leq r \leq .49$, die Autoren werteten dies aufgrund der zeitversetzten Erhebung dennoch als hohe Überstimmung.

[11] Aus dem Artikel von Kieruj und Moors (2013) geht nicht hervor, wie lang das Intervall zwischen den Erhebungen war. Hervor geht, dass die Studie auf Panel-Erhebungen basiert und die beiden Maße mit Items aus unterschiedlichen Erhebungswellen berechnet wurden.

3.1.3 Extreme Response Style und die Inter-Item-Standardabweichung

In der Literatur zu Antwortstilen wird – ungeachtet der Befunde zu Metatraits und intraindividueller Variabilität (u. a. Baird et al., 2006; Biderman & Reddock, 2012; Britt, 1993; Dwight et al., 2002) – häufig auch die Inter-Item-SD als Antworttendenz beschrieben, die mit ERS zusammenhängt (Baumgartner & Steenkamp, 2001; Greenleaf, 1992a). Greenleaf (1992a) stellt ohne Begründung und ohne Belege anzuführen, fest:

> Standard deviation is sometimes compared with extreme response style, the tendency to mark extreme scale intervals; though the two are typically highly correlated, they are not identical. (S. 176)

Übereinstimmend mit Greenleafs (1992a) Feststellung berichteten Baumgartner und Steenkamp (2001), dass die Inter-Item-SD eng mit ERS verknüpft ist. Aufgrund einer Korrelation von $r = .92$ zwischen ERS und der Inter-Item-SD über alle Items ihres mehrdimensionalen Fragebogens hinweg aggregierten Baumgartner und Steenkamp die Inter-Item-SD und ERS sogar zu einem Index[12]. Eine theoretische Begründung lieferten sie allerdings nicht. Auch ein Bezug zur Metatraits-Forschung findet sich in ihrem Artikel nicht.

3.2 Ursachen und Korrelate von Extreme Response Style

Trotz der Vielzahl an Forschungsarbeiten liegen derzeit keine zufriedenstellenden Erklärungen für das Zustandekommen von ERS vor. Zwar wird ERS mit der Beschaffenheit von Fragebogen und Items (Abschnitt 3.2.1), mit demografischen Merkmalen (Abschnitt 3.2.2) und mit verschiedenen Persönlichkeitseigenschaften (Abschnitt 3.2.3) in Verbindung gebracht, die Befundlage ist jedoch uneinheitlich. Die Ansätze werden im Folgenden vorgestellt.

3.2.1 Stimuli als Ursachen von Extreme Response Style

Als mit dem Stimulus verknüpfte Quellen von ERS wurden das Skalenformat, der Modus der Datenerhebung, die Sprache und das infrage stehende Merkmal oder Thema untersucht (vgl. Van Vaerenbergh & Thomas, 2013). Insgesamt wurden jedoch nur wenige oder wenig konsistente Befunde berichtet. Hinsichtlich des Modus der Datenerhebung stellten Weijters,

[12] Es sei darauf hingewiesen, dass die Inter-Item-SD hier – im Gegensatz zu den Studien zu Metatraits und Variabilität – für die Items *verschiedener* Skalen eines Einstellungsfragebogens berechnet wurde. Daher ging auch die intraindividuelle Streuung der Mittelwerte der verschiedenen Skalen in die Inter-Item-SD ein.

Schillewaert und Geuens (2008) fest, dass Online-Erhebungen weniger ERS hervorrufen als Papier-Bleistift-Erhebungen. Gibbons, Zellner und Rudek (1999) zufolge zeigen Personen mehr ERS, wenn der Fragebogen nicht in ihrer Muttersprache präsentiert wird und wenn das Thema für sie persönlich wichtig ist.

Etwas umfassender fällt die Forschung zum Einfluss des Skalenformats auf ERS aus. Ein Faktor ist die Benennung der Antwortkategorien (Weijters, Cabooter & Schillewaert, 2010a): Weniger ERS tritt bei Likert-Skalen auf, bei denen sämtliche Kategorien benannt sind, im Vergleich zu Likert-Skalen, bei denen nur Endpunkte benannt sind. Was die Länge und Anzahl der Stufen der Likert-Skala angeht, berichteten Weijters et al., dass Personen häufiger extrem antworten, wenn die zu bearbeitende Likert-Skala keine mittlere Antwortkategorie aufweist und weniger Antwortkategorien umfasst. Den Einfluss der Länge der Likert-Skala auf ERS untersuchten auch Kieruj und Moors (2013): Anders als Weijters et al. prüften Sie nicht die Häufigkeit extremer Antworten, sondern ob sich für unterschiedliche Skalenlängen gleichermaßen ein ERS-Faktor zeigt. Dies wurde für einen Bereich von 5 bis 11 Likert-Stufen bestätigt, was darauf hindeutet, dass ERS bzw. interindividuelle Unterschiede bezogen auf ERS unabhängig von der Skalenlänge auftreten.

Weitere Studien gehen der Fragestellung nach, ob der Antwortprozess sich auf die Häufigkeit extremer Antworten auswirkt. Albaum, Roster, Yu und Rogers (2006) fragten Personen zuerst nach der Richtung ihrer Antwort (z. B. „*How effective do you believe …?*“ mit den Antworten „*Effective*“, „*Ineffective*“, „*No opinion*“) und dann nach der Intensität („*Very*“ vs. „*Somewhat*“). Bei diesen Zwei-Stufen-Items beobachteten Albaum et al. häufiger Extremantworten als bei Likert-Items, die nur eine Verarbeitungsstufe umfassen. Arce-Ferrer (2006) versuchte, die unterschiedlichen Verarbeitungsstufen grafisch zu operationalisieren, und konnte die Ergebnisse von Albaum et al. nicht replizieren. Personen antworteten bei einem einstufigen Prozess und bei einem vermeintlich zweistufigen Prozess gleich häufig extrem. Das Itemformat von Arce-Ferrer ist in Abbildung 11 abgebildet: Ein-Prozess-Items enthielten eine gerade Linie, auf der der Grad der Zustimmung bzw. Ablehnung eingezeichnet werden sollte. In Zwei-Prozess-Items war diese Linie unterbrochen; die Annahme Arce-Ferrers war, dass Personen sich zunächst für einen der Teilstriche und dann für die Ausprägung auf diesem entscheiden. Als extreme Antworten wurden Markierungen gewertet, die maximal einen Zentimeter vom Ende eines Pols entfernt waren. Womöglich geht der Nullbefund darauf

zurück, dass die Operationalisierung von zwei mentalen Prozessen bei der Beantwortung der Items nicht gelungen ist und der Beantwortung jeweils nur ein Prozess zugrunde lag.

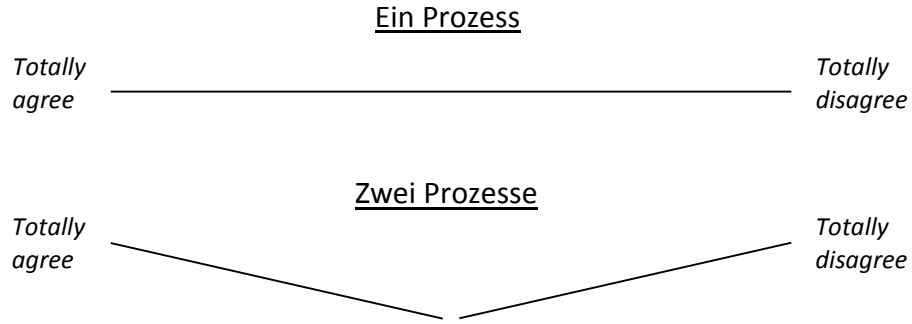

Abbildung 11: Itemformate mit einem und zwei implizierten Verarbeitungsprozessen (Arce-Ferrer, 2006)

3.2.2 Korrelationen mit demografischen Merkmalen

Unterschiede bezüglich der Häufigkeit extremer Antworten zwischen Geschlechtern, im Hinblick auf das Alter, im Hinblick auf das Bildungsniveau und den sozioökonomischen Status (SÖS) sowie im Hinblick auf den kulturellen Hintergrund wurden häufig im Rahmen der Marketing- und Einstellungsforschung erforscht.

Über Alterseffekte lassen sich keine allgemeinen Aussagen treffen: Greenleaf (1992b), Weijters et al. (2010c) sowie Kieruj und Moors (2013) berichteten mehr ERS unter älteren als unter jüngeren Personen, wohingegen die jüngeren Studienteilnehmer von Austin et al. (2006) häufiger extrem antworteten als die älteren. Keine Alterseffekte fanden Eid und Rauber (2000). Möglicherweise ist der Zusammenhang zwischen Alter und ERS auch nicht linear: In der Studie von Light, Zax und Gardiner (1965) zeigten Kinder mehr ERS, je jünger sie waren. Das und Dutta (1969) zufolge antworten Personen nach der Adoleszenz seltener extrem als zuvor und geben im mittleren Erwachsenenalter wieder vermehrt extreme Antworten.

Hinsichtlich der Geschlechtsunterschiede lassen sich in der Literatur zwei Befunde ausmachen. Einige Forschergruppen fanden, dass Männer und Frauen gleichermaßen extrem antworten (Bachman & O'Malley, 1984; Greenleaf, 1992b; Kieruj & Moors, 2013; G. Marín et al. 1992; Naemi et al., 2009). Den Ergebnissen anderer Studien zufolge geben Frauen häufiger Extremantworten als Männer (Austin et al., 2006; Eid & Rauber, 2000; Weijters et al., 2010c).

Laut der Studie von Crandall (1973) zeigt sich der Geschlechtsunterschied nur am positiven Pol, d. h. Frauen stimmen den Aussagen in Fragebogen häufiger sehr stark zu.

Ähnlich inkonsistent ist die Befundlage bezüglich des Ausbildungsniveaus: In einigen Studien zeigten sich keine Effekte (Bachman & O'Malley, 1984; Kieruj & Moors, 2013), in anderen antworteten Teilnehmer umso häufiger extrem, je niedriger ihr Bildungsniveau war (Greenleaf, 1992b; Weijters et al., 2010c). Parallel verhält es sich beim SÖS: Bei Bachman und O'Malley (1984) sowie bei Kieruj und Moors (2013) zufolge wird kein Zusammenhang zwischen ERS und SÖS berichtet. Bei Greenleaf (1992b) tritt der Effekt dagegen – zumindest hinsichtlich des Bildungsniveaus – auf: je höher der Bildungsgrad, desto weniger ERS.

Dass sich im Antwortverhalten kulturelle Unterschiede zeigen, demonstrierten Mõttus et al. (2012), deren Studienteilnehmer aus 20 verschiedenen Ländern die Gewissenhaftigkeit von Personen in Kurzgeschichten beurteilten. Die mittlere Häufigkeit von Extremantworten unterschied sich zum Teil deutlich von Land zu Land: Während etwa in Hongkong 49 % der Antworten extrem ausfielen, waren es in Burkina Faso 71 %. Kulturelle Unterschiede innerhalb eines Landes werden ebenfalls berichtet. Für Daten von US-Stichproben berichteten Bachman und O'Malley (1984), dass Afroamerikaner häufiger extrem antworten als Personen europäischer Abstammung. In der Studien von G. Marín et al. (1992) gaben Personen mit lateinamerikanischen Wurzeln mehr Extremantworten als Personen mit europäischen Wurzeln. Diesen Effekt fanden Hui und Triandis (1989) nur für eine 5-stufige Likert-Skala, bei einer 10-stufigen antworteten Personen beider Bevölkerungsgruppen gleich häufig extrem.

Insgesamt zeigt sich ein uneinheitliches Bild. Die Effekte sind klein und vermutlich stark abhängig vom Fragebogen, von den gemessenen Merkmalen und von den untersuchten Stichproben. Die meisten Ergebnisse lassen sich nicht ohne weiteres replizieren.

3.2.3 Extreme Response Style und Persönlichkeitsmerkmale

ERS wurde in den vergangenen 60 Jahren mit mehreren Persönlichkeitsmerkmalen in Verbindung gebracht. Dabei wurden ebenfalls in der Regel kleine Effekte berichtet, und die Befunde waren nicht immer konsistent. Die meisten Studien sind dem Ziel kaum näher gekommen, Ursachen interindividueller Unterschiede in der Tendenz, extrem zu antworten, zu finden und ERS im nomologischen Netz zu verankern. Ein Grund dafür könnte sein, dass die

Zusammenhänge zwischen ERS und Persönlichkeitsmerkmalen bisher vorwiegend in Studien untersucht wurden, in denen auch anderen Fragestellungen nachgegangen wurde, und daher nicht systematisch genug erforscht wurden.

Mehrfach ist ERS mit Ängstlichkeit oder Neurotizismus und entsprechend mit schlechter Anpassung in Verbindung gebracht worden. So fanden Berg und Collier (1953) sowie Lewis und Taylor (1955) einen Zusammenhang zwischen Ängstlichkeit und der Häufigkeit von Extremantworten bei der Beurteilung abstrakter visueller Stimuli. R. P. Norman (1969) replizierte die Ergebnisse für ERS auch bei der Beurteilung von bedeutsamem Bildmaterial. Auch Iwawaki und Zax (1969) berichteten, dass neurotische Personen häufiger extrem antworten als nicht-neurotische. Dagegen fanden Borgatta und Glass (1961) diesen Zusammenhang nicht.

Auch Extraversion wird mit ERS in Verbindung gebracht: So berichtete Crandall (1982), dass ERS mit sozialem Interesse zusammenhängt. Konform mit diesem Befund stellen Austin et al. (2006) sowie Kieruj und Moors (2013) eine Korrelation zwischen Extraversion und ERS fest. Die Zusammenhänge waren jedoch durchweg klein und lagen im Bereich von $r \approx .20$.

Über Korrelate im Bereich von Neurotizismus und Extraversion hinaus wurde Intelligenz als eine Determinante von ERS vorgeschlagen. Dies liegt in Anbetracht der Zusammenhänge zwischen ERS und dem Ausbildungsniveau nahe: Intelligenz könnte die Urteilsfähigkeit bzw. die Fähigkeit, in der Urteilsbildung zu differenzieren, beeinflussen; Intelligente differenzieren ihre Urteile möglichweise präziser als nicht Intelligente und wählen deshalb seltener extreme Antworten. Tatsächlich zeigte sich ein solcher Zusammenhang in einigen Studien (Light et al., 1965; Wilkinson, 1970). Bei Zuckerman und Norton (1961) und bei Naemi et al. (2009) tritt der Effekt hingegen nicht auf.

Festgestellt wurden auch Zusammenhänge zwischen ERS und Suggestibilität (Das & Dutta, 1969), Rigidität (Brengelmann, 1960) sowie Gewissenhaftigkeit (Austin et al., 2006). Diese Befunde gehen jedoch auf einzelne Studien zurück, andere Autoren fanden z. B. die Beziehung zu Gewissenhaftigkeit nicht (Borgatta & Glass, 1961). Naemi et al. (2009) merkten an, dass ERS zwar häufig mit Persönlichkeitsmerkmalen in Verbindung gebracht wurde, dass aber bislang keine theoretische Erklärung präsentiert wurde. Entsprechend stellten sie drei Persönlichkeitseigenschaften vor, die inhaltlich mit ERS verknüpft sein könnten: Ambiguitätsintoleranz (Tendenz, Unsicherheit als unangenehm wahrzunehmen), vereinfachtes Denken

und Entscheidungsfreude. Extreme Antworten könnten laut Naemi et. nämlich Folge von Präferenzen für sichere, einfache und starke Entscheidungen sein. Die Ergebnisse ihrer Studie stützten diese Theorie: ERS korrelierte mit den drei Eigenschaften. Die Zusammenhänge waren jedoch gering ($.25 \leq r \leq .29$); für Ambiguitätsintoleranz und vereinfachtes Denken waren sie größer, je schneller Personen den Fragebogen bearbeiteten.

Fazit: Der Verhaltensstil ERS lässt sich derzeit weder anhand von Persönlichkeitsmerkmalen erklären noch im nomologischen Netz repräsentieren. Bis auf die Ergebnisse von Naemi et al. (2009) entspricht dies dem Stand von 1968, als Hamilton in einem Review folgerte:

> Many authors have offered explanatory hypotheses to account for ERS findings, and a few of the more common proposals are briefly presented here. The term "theoretical" in its present usage is an exaggeration of the actual state of affairs. Most of the following accounts are derived from speculations expressed in the "Discussion" sections of articles reviewed above. (S. 199)

Als Beispiel dafür, dass die Folgerung Hamiltons auch heute noch Bestand hat, dient die Studie von Kieruj und Moors (2013), in der vorrangig der Einfluss der Skalenlänge auf ERS untersucht wurde: Die Autoren erfassten sieben Persönlichkeitseigenschaften, die möglicherweise mit ERS in Verbindung stehen (Extraversion, Verträglichkeit, Indifferenz, wie stark Personen ihre Ansichten vertreten, soziale Fähigkeiten, Schwarz-weiß-Denken sowie Intellekt). Ihre Hypothesen leiteten Kieruj und Moors aus wenigen Forschungsbefunden ab, u. a. aus denen von Lewis und Taylor (1955) oder Austin et al. (2006), die ihrerseits kaum theoretische Erklärungen für die Zusammenhänge angeboten hatten. Gleichzeitig referierten Kieruj und Moors den Zusammenhang zwischen ERS und Gewissenhaftigkeit (Austin et al., 2006), ignorierten diesen aber beim Studiendesign. Lediglich die Orientierung an den Befunden von Naemi et al. (2009) kann als theoriegeleitet angesehen werden.

3.3 Extreme Response Style und die Validität von Fragebogen

Die Feststellung von Cronbach (1946), dass ERS die logische Validität von Fragebogen beeinträchtigt, gilt als allgemein anerkannt (Baumgartner & Steenkamp, 2001; De Beuckelaer et al., 2010; Naemi et al., 2009; Van Vaerenbergh & Thomas, 2013; Weijters et al., 2010b; Wetzel et al., 2013b). Das Ausmaß und die Art und Weise dieser Beeinträchtigung sind jedoch weitgehend unbekannt. Die meisten Studien zum Thema ERS sind auf die Erfassung, die Ur-

sachen und die demografischen Korrelate von ERS fokussiert. Das geringe Forschungsinteresse spiegelt sich auch im Ergebnis einer Abfrage in der wissenschaftlichen Datenbank PsycINFO wider: Die Suchbegriffe „Extreme Response Style“ und „Validity“ werden gemeinsam in den Zusammenfassungen von nur 11 Artikeln gefunden, von denen die meisten die Validität der Messung von ERS behandeln[13]. Auch Van Vaerenbergh und Thomas (2013) gehen in einem systematischen Review zu Antworttendenzen in der Umfrageforschung nur wenig auf die Beziehung zwischen Antworttendenzen und der Validität ein.

Eine der wenigen Studien, die Van Vaerenbergh und Thomas anführen und die sich mit dem Einfluss von Antwortstilen auf die Validität befasst, führten Baumgartner und Steenkamp (2001) durch. Diese Autoren waren der Auffassung, dass Antwortstile (auch ERS) linear in die Antworten auf Fragebogen-Items eingehen und dass die Zusammenhänge zwischen Fragebogenskalen durch Antwortstile sowohl über- als auch unterschätzt werden können. Diese Annahme stützten sie auf eine Datenanalyse mit einer Multi-Level-Regressionsanalyse, die sich auf mehrere Antworttendenzen gleichzeitig bezog. Mit hohem ERS geht Baumgartner und Steenkamp zufolge eher eine Erhöhung der Itemantworten einher, wenn der Gruppen-Mittelwert oberhalb der Mitte der Likert-Skala liegt, und niedrigere Itemantworten, wenn der Gruppen-Mittelwert unterhalb der Mitte der Likert-Skala liegt. Inhaltlich bedeutet dies, dass Personen mit hohem ERS, deren Ausprägung auf einer Likert-Skala oberhalb der Mitte liegt, eher die höchste Kategorie ankreuzen als Personen mit niedrigem ERS. Personen, deren Ausprägung auf der Likert-Skala unterhalb der Mitte liegt, kreuzen eher die niedrigste Kategorie an, wenn sie eine hohe ERS-Ausprägung aufweisen. ERS beeinflusst die Antwort auf ein Likert-Item laut Baumgartner und Steenkamp also als *Bias* (siehe Abbildung 12).

[13] Bei einer Erweiterung der Suche mit den Suchbegriffen „Response Style“ und „Validity“ erhält man bei PsycINFO 122 Treffer. Die meisten der so gefundenen Artikel beziehen sich auf positive oder negative Selbstdarstellung. Andere Artikel haben einen starken Fokus auf Methoden zur Erfassung von Antwortstilen, in wiederum anderen werden Validitätsskalen von verschiedenen Fragebogen verglichen. Aussagekräftige Ergebnisse über den Zusammenhang zwischen ERS und der Validität von Fragebogen finden sich nicht.

	Ablehnung	Zustimmung
niedrige ERS-Ausprägung	1 2 3 4 5 ☐ ☒ ☐ ☐ ☐	1 2 3 4 5 ☐ ☐ ☐ ☒ ☐
hohe ERS-Ausprägung	1 2 3 4 5 ☒ ☐ ☐ ☐ ☐	1 2 3 4 5 ☐ ☐ ☐ ☐ ☒

Abbildung 12: ERS als Bias-Komponente bei der Beantwortung von Likert-Items (nach Baumgartner & Steenkamp, 2001; eigene Darstellung)

Für den Bereich psychologischer Messungen befassten sich McGrath et al. (2010) in einem Review mit der Frage, ob Indikatoren für Antworttendenzen (Bias-Indikatoren) valide sind und in der angewandten Diagnostik eingesetzt werden sollen. McGrath et al. zufolge wäre dies der Fall, wenn die Bias-Indikatoren die Kriteriumsvalidität entweder als Moderator oder als Suppressor beeinflussen. Forschungsergebnisse, die eindeutig für den Einsatz von Bias-Indikatoren, wie z. B. ERS-Maßen, sprechen, sind den Autoren zufolge jedoch rar: Es lassen sich keine Aussagen über die Auswirkungen von Antworttendenzen auf die Validität treffen.

Zu ERS fanden McGrath et al. (2010) in der Literatur nur wenige Befunde. In einer Studie, die sie zitierten, untersuchten McCrae, Stone, Fagan und Costa (1998) den Einfluss von ERS auf die Übereinstimmung zwischen Selbst- und Fremdbeschreibung für die Big Five und fanden keinen Zusammenhang. Dagegen fanden Arce-Ferrer und Ketterer (2003) in einem Fragebogen zur Selbstwirksamkeit bei Karriere-Entscheidungen Unterschiede zwischen Personen, die häufig extrem antworteten, und solchen, die nur wenige extreme Antworten in Fragebogen gaben: Unter häufig extrem Antwortenden klärten fünf theoriegeleitet extrahierte Faktoren weniger Varianz auf als unter nicht extrem Antwortenden. Für nicht extrem Antwortende ließen sich die Faktoren zudem deutlich besser interpretieren als für extrem Antwortende. Hohe ERS-Ausprägung geht also möglicherweise mit schlechterer Konstruktvalidität und somit schlechterer Interpretierbarkeit der gemessenen Eigenschaften einher. Die Ergebnisse von Arce-Ferrer und Ketterer sind ein erstes Indiz dafür, dass ERS ein Moderator der Validität sein kann. Sie sind jedoch weder repliziert noch für andere Arten bzw. Indikatoren der Validität berichtet worden. Es mangelt auch an einer Erklärung der Befunde.

4 Fazit und Implikationen für die empirischen Studien

In diesem Kapitel werden die Forschungsbefunde zu Variabilität und zu ERS kritisch bewertet. In Abschnitt 4.1 wird eine kurze Zusammenfassung der Forschung zur Erfassung und Struktur von Variabilität gegeben und die offenen Fragen werden in Hypothesen überführt. In Abschnitt 4.2 wird die Forschung zur Beschreibung und Erklärung von ERS zusammengefasst und kritisch bewertet. Anknüpfend an diese Bewertung werden eine Theorie und eine Hypothese über die Entstehung von ERS aufgestellt. In Abschnitt 4.3 werden Forschungsergebnisse zum Einfluss von Variabilität auf die Reliabilität und die Validität von Persönlichkeitsfragebogen sowie auf die Stabilität von Persönlichkeitseigenschaften aufgegriffen und bewertet. Daran anschließend werden auf der Basis der Theorien zu Variabilität Hypothesen formuliert. In Abschnitt 4.4 folgt schließlich ein Ausblick auf die im Rahmen dieser Arbeit präsentierten empirischen Studien.

4.1 Die Erfassung und Struktur von intraindividueller Variabilität

Zwar ist Variabilität schon weitreichend untersucht und auch die Probleme bei der Erfassung von Variabilität scheinen überwunden, dennoch geben einige Aspekte Anlass zu weiterer Forschung: In den Forschungsarbeiten der letzten Jahren wird davon ausgegangen, dass Variabilität ein stabiler, globaler Trait ist (Baird et al., 2006; Biderman & Reddock, 2012; Reddock et al., 2011). Insbesondere Baird et al. (2006) zeigen, dass Variabilität für verschiedene Traits konvergiert, und schließen daher auf ein unidimensionales Konstrukt. Dies steht den Implikationen der Ergebnisse von Paunonen (1988) gegenüber: Paunonen erhebt Variabilität explizit für verschiedene Traits und berichtet, dass Personen weniger variabel in der jeweiligen Trait-Ausprägung sind, je weiter ihre Trait-Ausprägung vom Gruppen-Mittelwert entfernt ist (Abschnitt 2.1.3). Dieser Zusammenhang liegt Paunonen zufolge auf der inhaltlichen bzw. Konstrukt-Ebene. Wenn Paunonens Messung von Variabilität valide ist, dann ist Variabilität – konform mit der Metatraits-Theorie (Abschnitt 2.1.1) – für unabhängige Eigenschaftsdimensionen unterschiedlich. Schließlich konvergieren die Abweichungen vom Gruppen-Mittelwert für verschiedene unabhängige (d. h. unkorrelierte) Traits nicht.

Baird et al. (2006) etablieren als Maß für Variabilität dagegen die Summe der korrigierten Inter-Item-SD für Items, die für verschiedene Rollen erhoben wurden. Dabei wird die Inter-

Item-SD um den jeweiligen Mittelwert und dessen Abweichung vom Gruppen-Mittelwert (operationalisiert als Quadrat des z-Werts) korrigiert. Aufgrund dieser Korrektur ist die Variabilität auf einem Trait von diesem unabhängig. Dass die Korrektur um den Mittelwert die Messung von Variabilität um einen Einfluss der Mess-Methode bereinigt, belegen Baird et al. anhand ihrer Daten: Je schiefer die Verteilung der Item-Mittelwerte ist, desto stärker hängt die Inter-Item-SD mit dem Item-Mittelwert zusammen. Dies stützt die Annahme von Baird et al. (2006), dass die Zusammenhänge zwischen den Inter-Item-SD und den Item-Mittelwerten auf methodische Artefakte zurückgehen (vgl. Abschnitt 2.1.3). Dass die Korrektur der Inter-Item-SD um das Quadrat des z-standardisierten Mittelwerts die Validität der Messung von Variabilität erhöht, begründen Baird et al. nur theoretisch: Bei Personen mit hohem oder niedrigem Item-Mittelwert kann die Inter-Item-SD nur niedrig sein und bei Personen mit mittlerem Item-Mittelwert kann die Inter-Item-SD hoch oder niedrig ausfallen (vgl. Abschnitt 2.1.3). Aufgrund dieses Umstandes – einer Restriktion der Messung der Variabilität auf einem Trait – hängen die Inter-Item-SD laut Baird et al. mit der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert zusammen. Allerdings zeigte sich dieser Zusammenhang bei Paunonen (1988) auch, obwohl diese methodische Restriktion nicht bestand; denn die Variabilität eines Traits wurde von Paunonen explizit und somit methodisch unabhängig von den Trait-Maßen erfasst.

Fraglich bleibt also, ob der Zusammenhang zwischen der Inter-Item-SD und der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert wie von Baird et al. (2006) angenommen tatsächlich auf die Beschaffenheit der Methode zurückzuführen ist oder wie von Paunonen gefolgert auf die Merkmalsausprägung (=Konstrukt). Im ersten Fall (Zusammenhang geht auf Methode zurück) führt die Korrektur der Inter-Item-SD um die Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert zu valideren Messungen von Variabilität. Es wäre dann davon auszugehen, dass die selbstberichtete Variabilität auf einem Trait (explizites Maß, vgl. Paunonen, 1988) ein anderes Konstrukt ist als die anhand der Streuung auf einer Skala erschlossene Variabilität. Im zweiten Fall (Zusammenhang geht auf Konstrukt zurück) würde die Korrekturmethode von Baird et al. die Validität der Messung von Variabilität auf einem Trait mindern. Welcher der beiden Fälle zutrifft, soll im Rahmen der vorliegenden Arbeit geklärt werden. Da die von Baird et al. beschriebene methodische Restriktion nicht von der Hand zu weisen ist und für die korrigierte Inter-Item-SD von Skalen der gleiche Moderatoref-

fekt auf Zusammenhänge zwischen Traits und Kriterien berichtet wird (Fleisher et al., 2011; vgl. Abschnitt 2.3) wie für die nicht korrigierte Inter-Item-SD von Skalen (Baumeister & Tice, 1988; vgl. Abschnitt 2.1.1), wird erwartet, dass der erste Fall zutrifft, d. h. die Korrekturmethode von Baird et al. ist methodisch begründet.

In dieser Arbeit soll belegt werden, dass die Korrektur der Inter-Item-SD um den Einfluss der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert methodisch begründet ist. Belegt werden soll auch – zusätzlich zum Beleg von Baird et al. (2006) – dass die Korrektur der Inter-Item-SD um den Item-Mittelwert methodisch begründet ist. Zusammengefasst lauten die Hypothesen hinsichtlich der Erfassung von Variabilität:

H1A: Durch die Korrektur der Inter-Item-SD um den Item-Mittelwert wird die Validität der Messungen von Variabilität verbessert.

H1B: Durch die Korrektur der Inter-Item-SD um die Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert (d. h. um das Quadrat des z-standardisierten Mittelwerts) wird die Validität der Messung von Variabilität verbessert.

Diese beiden Hypothesen lassen sich auf zwei Wegen testen:

(1) Geprüft werden kann zum einen, ob die Korrekturen vom Inhalt der Skalen abhängen: Die Korrektur würde jeweils dann zu *weniger validen* Messungen von Variabilität führen, wenn sich die gemeinsame Varianz zwischen Inter-Item-SD und dem Item-Mittelwert bzw. der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert sachlogisch auf das mit der Skala erfasste Merkmal bezieht. Wäre dies der Fall, so würden die Inter-Item-SD zweier Skalen jeweils ähnlich hoch mit dem Item-Mittelwert bzw. der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert korrelieren, wenn die Skalen ähnliche Dimensionen messen. Sie würden jeweils unterschiedlich hoch mit dem Item-Mittelwert bzw. der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert korrelieren, wenn die Skalen unterschiedliche Dimensionen messen.

(2) Zum anderen kann geprüft werden, ob die Korrekturen von der Methode abhängen: Die Korrektur würde zu *valideren* Messungen von Variabilität führen, wenn sich die gemeinsame Varianz zwischen Inter-Item-SD und dem Item-Mittelwert bzw. der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert auf die Messung (bzw. Methode) zurückführen lässt. Wäre dies der Fall, wäre der Zusammenhang zwischen Inter-Item-SD und

dem Item-Mittelwert bzw. der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert nicht vom mit der Skala gemessenen Merkmal abhängig, sondern von der Lage des Gruppen-Mittelwerts relativ zur Mitte der Likert-Skala: Der Zusammenhang zwischen der Inter-Item-SD und dem Item-Mittelwert fiele größer aus, wenn der Abstand zwischen Gruppen-Mittelwert einer Skala und der Mitte der Likert-Skala größer ist; der Zusammenhang zwischen Inter-Item-SD und der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert wäre in diesem Fall kleiner.

Rückschlüsse auf die Struktur von Variabilität sind eng mit Rückschlüssen auf die richtige Methode der Erfassung von Variabilität verknüpft. Wenn sich die von Baird et al. (2006) vorgeschlagene Korrektur nicht auf die Messmethode zurückführen lässt, treffen die Annahmen der Metatraits-Theorie zu: Variabilität wäre als multidimensionales Konstrukt zu betrachten. Ist die Korrektur dagegen – wie hier erwartet wird – methodisch begründet, dann lässt sich keine Aussage für die Konstrukt-Struktur von Variabilität treffen. Zwar folgern Baird et al., Biderman und Reddock (2012) sowie Reddock et al. (2011), dass Variabilität ein eindimensionaler Trait ist. Ihre Folgerung stützen sie allerdings lediglich auf die hohe interne Konsistenz der Summe der korrigierten Inter-Item-SD der Skalen zur Messungen der Big Five ($.60 \leq \alpha \leq .95$) sowie auf die hohen Korrelationen zwischen diesen korrigierten Inter-Item-SD. Für die Eindimensionalität fehlt ein klarer Beleg. Da in den referierten Studien Variabilität stets für die Big Five, d. h. für voneinander unabhängige Merkmalsdimensionen, erfasst wurde, ließ sich eine logische Verknüpfung wie „zu ähnlichen Traits gehören auch ähnliche Metatraits“ bislang weder widerlegen noch nachweisen. Aufgrund der relativ hohen Zusammenhänge zwischen den korrigierten Inter-Item-SD für verschiedene Dimensionen, aufgrund der Annahme, dass diese Messungen auch einen signifikanten Messfehler enthalten und aufgrund der gemeinsamen Konzeption von Variabilität als Ausmaß, in dem Verhalten von der Situation determiniert wird (Abschnitt 2.2.3), soll in dieser Arbeit ein Nachweis für die Eindimensionalität erbracht werden. Folgende Hypothese wird formuliert:

H1C: Variabilität ist ein eindimensionaler globaler Trait, der sich – anders als von der Metatraits-Theorie impliziert – nicht auf einzelne Traits bezieht.

Wenn diese Hypothese zutrifft, korrelieren die Messungen von Variabilität auf verschiedenen Skalen unabhängig davon, ob die mit den Skalen erfassten Merkmale miteinander korre-

lieren. Trifft die Hypothese nicht zu, dann korrelieren Maße von Variabilität höher für Skalen, bei denen die erfassten Traits stärker zusammenhängen.

4.2 Zur Erklärung von Extreme Response Style

ERS wurde bislang stets als Antworttendenz beschrieben, also als Merkmal, das spezifisch auf das Antwortverhalten in Fragebogen bezogen ist (Berg & Collier, 1953; Bolt & Newton, 2011; De Beuckelaer et al., 2010; Greenleaf, 1992b; Merrens, 1970; Naemi et al., 2009; Weijters et al., 2010b, 2010c). Dass dieses Merkmal zeitstabil und über verschiedene Fragebogen und Methoden der Erfassung generalisierbar ist, gilt als gesichert (Bachman & O'Malley, 1984; De Beuckelaer et al., 2010; Merrens, 1970; Naemi et al., 2009; Weijters et al., 2010c; Wetzel et al., 2013b). Im Einklang mit der Konzeption von ERS als stabile, personengebundene Antworttendenz steht die Vorstellung, dass ERS in Fragebogen ein Bias ist, der die Validität mindert (Baumgartner & Steenkamp, 2001; Cronbach, 1946; De Beuckelaer et al., 2010; Naemi et al., 2009; Van Vaerenbergh & Thomas, 2013; Weijters et al., 2010b; Wetzel et al., 2013b). Empirisch wurde dies kaum untersucht; so begründen Van Vaerenbergh und Thomas (2013) in einem umfassenden Review zu Antworttendenzen die Relevanz von Antwortstilen mit einer Beeinträchtigung der Validität und beziehen sich dabei lediglich auf die Studie von Baumgartner und Steenkamp (2001) für den Bereich der Marketingforschung (vgl. Abschnitt 3.3). Baumgartner und Steenkamp verstehen ERS als Bias, der in die Itemantworten eingeht und sich daher auf die Zusammenhänge zwischen Variablen auswirkt. Für den Bereich der Persönlichkeitsdiagnostik hängt die Validität laut McCrae et al. (1998) nicht von ERS ab, laut Arce-Ferrer und Ketterer (2003) hat ERS möglicherweise einen Einfluss auf die Konstruktvalidität von Persönlichkeitsfragebogen. In Anbetracht der spärlichen Befundlage lässt sich kein übergreifendes Urteil über eine Beeinträchtigung der Validität durch ERS treffen. Dies hängt nicht zuletzt damit zusammen, dass ERS bisher nicht ausreichend erklärt wurde.

In den meisten Studien wird ERS wie ein Trait konzeptualisiert, für den Korrelate im Bereich der Persönlichkeit gesucht werden. Dabei handelt es sich bei ERS um eine Reaktion auf eine Messung, die – den Definitionen der in Kapitel 3 berichteten Studien gemäß – *nur* bei Messungen auftritt. Als Analogie kann hier das Ankreuzverhalten in einem Fragebogen zur Gewissenhaftigkeit gesehen werden: Nicht das Ankreuzverhalten selbst ist ein Trait, sondern Gewissenhaftigkeit, also eine Eigenschaft, die sich methodeninvariant erfassen lässt. Dieser

Beleg fehlt für ERS: Gemäß Definition zeigt sich ERS nicht in anderen Situationen als dem Bearbeiten von Fragebogen. Zwar wird ERS immer wieder als stabil und generalisierbar beschrieben (vgl. Abschnitt 3.1.2), es mangelt jedoch an einer Beschreibung dieses Konstrukts. Dieser Mangel zeigt sich auch bei Ansätzen, ERS als latente Variable in IRT-Modellen zu erfassen (Bolt & Newton, 2011; Wetzel et al., 2013b). Diesen Ansätzen folgend ist ERS eine latente Variable, die sich ausschließlich auf die Wahrscheinlichkeit der Wahl von Endpunkten (vs. Nicht-Endpunkten) einer Likert-Skala auswirkt, also als Bias (vgl. Abschnitt 3.1.1). Die Annahmen, die dafür nötig sind, werden jedoch weder explizit genannt noch in irgendeiner Form überprüft. Auch die Forschung zu personengebundenen Determinanten von ERS ist bislang wenig ergiebig; denn es wurde kaum mit Hilfe nützlicher und fundierter Theorien geforscht (vgl. Abschnitt 3.2.3). In wenigen Studien wurden Annahmen über die Entstehung von ERS hinreichend begründet (z. B. bei Naemi et al., 2009); allerdings sind hypothesenkonforme Befunde rar und die damit verbundenen Effekte klein, so dass sich festhalten lässt: Der größte Teil der interindividuellen Varianz von ERS kann derzeit nicht erklärt werden.

Ein Ziel der vorliegenden Arbeit ist es, einen Beitrag zur Erklärung von ERS zu leisten: Eine Ursache von ERS könnte Variabilität sein. So merkt Greenleaf (1992a) an, dass beide Konzepte zusammenhängen, und Baumgartner und Steenkamp (2001) berichten eine Korrelation von $r = .92$. Inhaltlich wurde der Zusammenhang bislang nicht erläutert. Dies liegt vermutlich daran, dass die Forschungszweige zu ERS und Variabilität sehr unterschiedlich sind und die Forschung jeweils auf anderen Ebenen stattfindet: Studien zu ERS sind größtenteils der Markt- und der Einstellungsforschung zuzuordnen. Dort werden vorwiegend Korrelationen mit demografischen Merkmalen und Ergebnisse zur Abhängigkeit von ERS von Stimuli berichtet. ERS gilt dabei stets als Antwort-Bias; wird die Inter-Item-SD auch untersucht, so wird sie ebenfalls als Bias gesehen. Auch in der psychologischen Literatur wird Forschung zu ERS berichtet; diese konzentriert sich mittlerweile eher auf Messmodelle, insbesondere im Bereich der IRT. Vom Forschungszweig zu Variabilität ausgehend befassten sich Forscher lange mit inhaltlichen Theorien; der Fokus lag viel stärker auf der Beschreibung und Erklärung von Variabilität und ihrer Effekte. Ausgehend von dieser Forschung haben Biderman und Reddock (2012) den Zusammenhang zwischen Variabilität und ERS festgestellt. Dieser fällt jedoch gegenüber der von Baumgartner und Steenkamp berichteten Korrelation mit $r = .42$ sehr niedrig aus. Dafür lassen sich zwei Gründe ausmachen: Erstens messen Biderman und

Reddock Variabilität als Summe der (nicht korrigierten) Inter-Item-SD der Skalen für die Big Five. Dieser Index ist konfundiert mit der Ausprägung auf diesen Dimensionen (vgl. Abschnitte 2.1.3 und 4.1), was sich im günstigsten Fall reliabilitätsmindernd auswirkt. Zweitens verwenden die Autoren als Maß für ERS die Summe der Extremantworten über den gesamten Fragebogen hinweg. Wenn sich dabei für Extremwerthäufigkeiten der fünf Skalen unterschiedliche Gruppen-Mittelwerte und Standardabweichungen zeigen, gehen sie jeweils unterschiedlich stark in die Gesamtsumme der Extremantworten ein. Dies ist insofern problematisch, als dass Personen mit extremer Ausprägung auf einer Skala auch eher extreme Antworten geben (vgl. Abschnitt 3.1.1). Für einen Big-Five-Fragebogen bedeutet dies, dass die Summe der Extremwerthäufigkeiten u. U. auch die Ausprägung einer oder mehrerer Skalen widerspiegeln kann, selbst wenn die Skalen voneinander unabhängig sind.

Die Tendenz zu extremen Antworten in einem Fragebogen wird also – wie auch die Kritik an Abzähl-Methoden zur Erfassung von ERS deutlich macht (Abschnitt 3.1.1) – auch durch die Ausprägung auf einer Skala bestimmt. Neben der Skalenausprägung hat wahrscheinlich auch Variabilität einen Einfluss auf ERS. Dies ist insofern plausibel, als dass Variabilität die Streuung von Verhalten (auch Antwortverhalten) um den durch die Eigenschaft vorhergesagten Wert darstellt. Je größer die Streuung, desto wahrscheinlicher sind (bei Konstanthalten des Item-Mittelwerts) extreme Antworten. Zu dem gleichen Schluss führt auch der Ansatz von Fleeson (2001, 2007; vgl. Abschnitt 2.2.2): Bei jeder Person lassen sich Traits als Dichte-Verteilungen von States beschreiben; diese Verteilungen haben einen konstanten Mittelwert und eine konstante Streuung. Bei breiten Streuungen liegen mehr extreme Ausprägungen der States vor, die sich in Fragebogen in häufigeren extremen Antworten zeigen. Die Einflussfaktoren auf extremes Antworten auf einer Skala werden in Abbildung 13 in Abhängigkeit der Dichte-Verteilung für einen Trait veranschaulicht. Dargestellt sind vier verschiedene Verteilungen: Oben sind zwei Antwortmuster von Personen mit mittlerer Ausprägung abgebildet, unten zwei Antwortmuster von Personen mit extremerer (hier: hoher) Ausprägung. Die beiden Antwortmuster links zeigen eine niedrige Variabilität, die beiden Antwortmuster rechts eine hohe Variabilität.

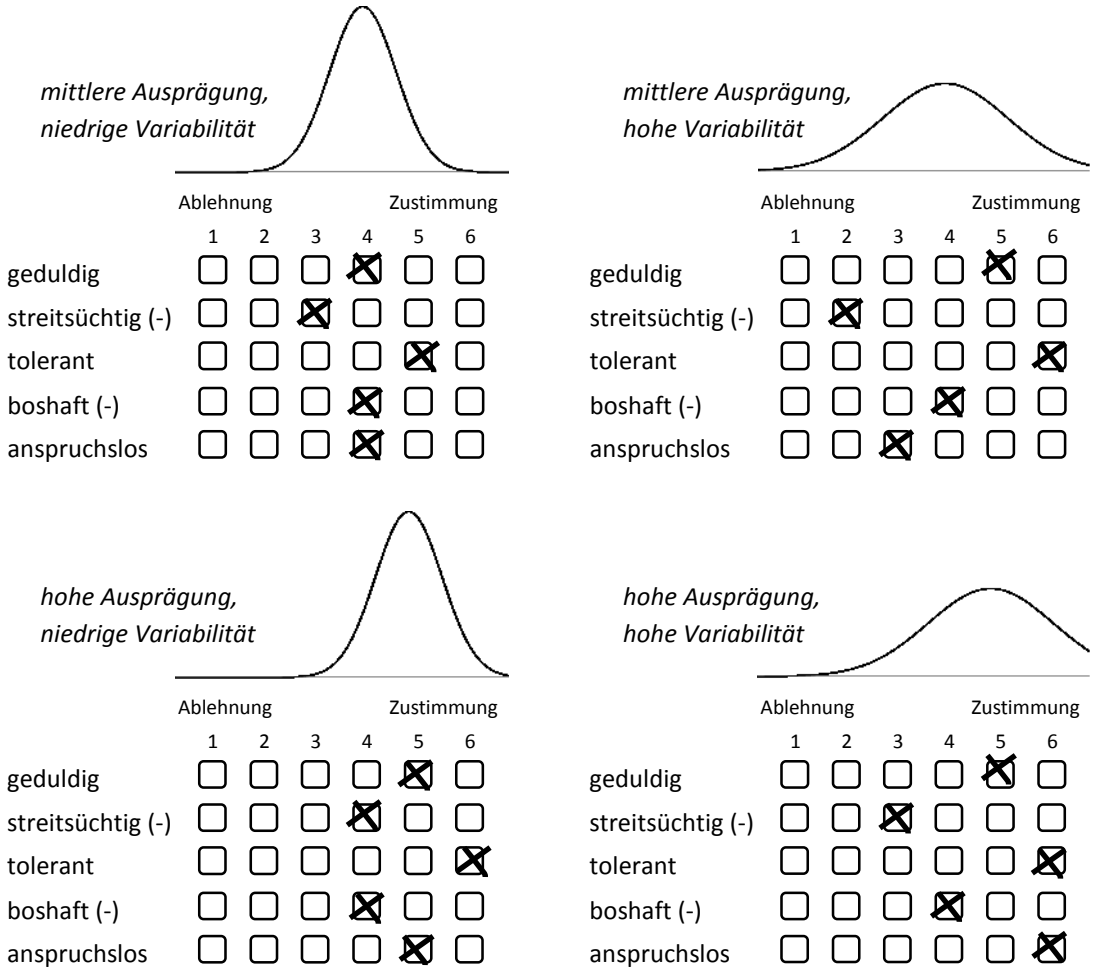

Abbildung 13: Antwortmuster auf fünf Items in Abhängigkeit der Ausprägung und der Streuung der Dichte-Verteilung des zugrunde liegenden Traits

Ersichtlich ist, dass extreme Antworten mit Variabilität und mit hoher Ausprägung auf der Skala einhergehen. Wenn ERS unabhängig von der Trait-Ausprägung erfasst werden kann, z. B. für viele verschiedene unabhängige Dimensionen, sollte lediglich die Variabilität einen Einfluss auf ERS haben. Aufgrund der Repräsentation von Antwortverhalten als Dichte-Verteilung von States für einen Trait und angesichts der hohen Korrelation zwischen Variabilität und ERS (Baumgartner & Steenkamp, 2001) wird Folgendes erwartet:

H2: ERS ist Indikator von intraindividueller Variabilität.

Diese Annahme und die Veranschaulichung in Abbildung 13 erklären auch die Befunde von Weijters et al. (2010a): Je mehr Stufen eine Likert-Skala hat, desto kleiner ist der Bereich, den eine Likert-Kategorie in der Dichte-Verteilung einnimmt, und desto seltener antworten Personen extrem. Auch der Befund von Kieruj und Moors (2013), dass ERS unabhängig von

der Länge der Likert-Skala über verschiedene Skalen hinweg konsistent ist, lässt sich einordnen: Extremen Antworten liegt jeweils Variabilität zugrunde.

4.3 Die Effekte in Persönlichkeitsfragebogen

Variabilität lässt sich laut der Befunde von Fleeson (2001, 2007; Abschnitt 2.2.2) nicht nur als Breite der Verteilung von States für einen Trait beschreiben, sondern auch als Ausmaß, in dem Verhalten von Situationen bestimmt wird (vgl. Abschnitt 2.2.3). Niedrige Variabilität oder Konsistenz lässt sich entsprechend als Übereinstimmung von Verhalten bzw. States mit Persönlichkeitseigenschaften beschreiben. Wenn Variabilität sich auf die Zusammenhänge zwischen Traits und Verhalten auswirkt, liegt der Schluss nahe, dass auch der Zusammenhang zwischen Personenmerkmalen und *Life-Outcomes* bzw. Kriterien von Variabilität moderiert wird. Schließlich sind diese – wie Verhalten – sowohl von Persönlichkeitsmerkmalen (z. B. Barrick & Mount, 1991; Judge & Bono, 2001; Schmidt & Hunter, 1998) als auch von Situationen, von den Lebensumständen bzw. vom Kontext (z. B. Haney et al., 1973; Osherow, 1988; Rosenthal & Jacobson, 1968) abhängig. Entsprechend finden Reddock et al. (2011) auch Moderatoreffekte bei der Vorhersage von akademischem Erfolg: Unter Personen, die einen Fragebogen konsistent bearbeiten, lässt sich dieser besser anhand von Gewissenhaftigkeit vorhersagen als unter Personen, die einen Fragebogen variabel bearbeiten. Biderman und Reddock (2012) greifen diese Befunde auf; sie erwarten, dass Variabilität nicht nur Einfluss auf die Kriteriumsvalidität, sondern auch auf die Konstruktvalidität hat. Die Kriteriumsvalidität wird in ihrer Studie allerdings nicht hypothesenkonform moderiert (vgl. Abschnitt 2.3): Zwar zeigt sich unter Personen mit hoher Variabilität der geringste Zusammenhang zwischen Gewissenhaftigkeit und akademischer Leistung und unter Personen mit niedriger Variabilität ist der Zusammenhang höher, der höchste Zusammenhang zeigt sich allerdings für Personen mit mittlerer Variabilität. Eine Erklärung für dieses Ergebnis liefern Biderman und Reddock nicht. Hinsichtlich der Konstruktvalidität entsprechen die Befunde den Erwartungen der Autoren: Je geringer die Variabilität von Personen ist, desto höher ist die Konstruktvalidität. Als weiteres Ergebnis wird berichtet, dass Variabilität einen Einfluss auf die Reliabilität hat; die Effekte haben die gleichen Vorzeichen und sind ähnlich ausgeprägt wie die zur Konstruktvalidität.

Diese Befunde sind zwar – wie Biderman und Reddock (2012) anmerken – praxisrelevant, die Herleitung der Hypothesen und die Diskussion greifen allerdings zu kurz: Variabilität kann als Eigenschaft gesehen werden, die die Kontingenz zwischen Persönlichkeitseigenschaften und Verhalten bzw. Kriterien moderiert (Abschnitt 2.2.3 und Abschnitt 2.3); dies impliziert nicht, dass Variabilität die Beziehung zwischen zwei Traits beeinflusst. Ausgehend von der Konzeption von Traits als Verteilungen von States (Fleeson, 2001; vgl. Abschnitt 2.2.2) spricht wenig dafür, dass die Zusammenhänge der Mittelwerte dieser Verteilungen von Variabilität moderiert werden. Biderman und Reddocks (2012) Ergebnis hinsichtlich der Konstruktvalidität lässt sich vermutlich auf den Einfluss von Variabilität auf die Reliabilität zurückführen. Dass nämlich die Reliabilität von Variabilität beeinflusst wird, ist wiederum plausibel: Die Reliabilität lässt sich als der Varianzanteil einer Messung definieren, der durch das erfasste Konstrukt erklärt wird, also als Zusammenhang zwischen Konstrukt und Messwert bzw. Antwortverhalten (Schmidt-Atzert & Amelang, 2012). Tatsächlich berichten Biderman und Reddock (2012) auch, dass die Reliabilität der meisten der erhobenen Skalen von der Variabilität abhängt. In Fällen, in denen Variabilität keinen Einfluss auf die Reliabilität hat, hängt auch die Konstruktvalidität nicht von Variabilität ab. Es wird daher Folgendes angenommen:

H3A: Variabilität hat einen Einfluss auf die Reliabilität von Persönlichkeitsfragebogen; bei Personen mit niedriger Variabilität ist die Reliabilität höher als bei Personen mit hoher Variabilität.

H3B: Der Zusammenhang zwischen zwei Konstrukten hängt *nicht* von der Variabilität ab.

Hinsichtlich des Einflusses von Variabilität auf die Reliabilität werden kleine Effekte erwartet: Die Items eines Persönlichkeitsfragebogens entsprechen nämlich schwachen und v. a. standardisierten Situationen, mit denen die Ausprägung eines Trait erfasst werden soll. Möglicherweise wird der Trait (gemessen als Mittelwert der Items) mit einer Skala bereits so genau geschätzt, dass der Messfehler bei Personen mit hoher Variabilität sich nur geringfügig vom Messfehler bei Personen mit niedriger Variabilität unterscheidet. Diese Annahme passt auch zu den Ergebnissen von Biderman und Reddock (2012), schließlich wird in deren Studie die Reliabilität einiger Skalen nicht von Variabilität beeinflusst. Hinsichtlich der Konstruktvalidität gilt: Nur der Zusammenhang zwischen den Eigenschaften und der jeweiligen Messung hängt von Variabilität ab, nicht der Zusammenhang zwischen zwei Traits. Mit anderen Wor-

ten ist die Abhängigkeit der Reliabilität von Variabilität eine notwendige Bedingung dafür, dass die Konstruktvalidität einer Skala von Variabilität abhängt.

Baird et al. (2006) stellen die Hypothese auf, dass Variabilität mit Veränderungen von Persönlichkeitseigenschaften über die Zeit einhergeht; auch Baumeister (1991) trifft diese Vorhersage für Metatraits. Allerdings zeigt sich bei Baumeister kein hinreichender Beleg in den Daten; Baird et al. finden nur in Teilen Unterstützung für die Hypothese. Möglicherweise geht dieser Effekt – sofern er sich überhaupt zeigt – auf den Einfluss von Variabilität auf die Reliabilität zurück. Denn wenn Persönlichkeitseigenschaften als Verteilung ihrer States konzeptualisiert werden (Abschnitt 2.2.2), lässt sich parallel zur Konstruktvalidität nicht schließen, dass Variabilität mit Veränderungen der Mittelwerte dieser Verteilungen über die Zeit einhergeht. Als Hypothese soll geprüft werden:

H3C: Die Stabilität von Persönlichkeitseigenschaften hängt *nicht* von der Variabilität ab.

Als Operationalisierung dieser Hypothese soll der Einfluss von Variabilität auf die Retestreliabilität von Eigenschaftsmessungen untersucht werden. Diese hängt von der Stabilität der Eigenschaften ab und von der Reliabilität der Messungen. Trifft Hypothese 3C zu, ist ein Einfluss der Variabilität auf die Reliabilität eine notwendige Bedingung dafür, dass die Retestreliabilität von Variabilität abhängt. Wenn sich ein Einfluss von Variabilität auf die Retestreliabilität zeigte und andere Reliabilitätsschätzungen nicht von Variabilität abhingen, wäre Hypothese 3C zu verwerfen: Variabilität hätte in diesem Fall einen Einfluss auf die Stabilität der Persönlichkeitseigenschaften.

Die Erwartungen hinsichtlich der Kriteriumsvalidität stehen im Einklang mit den Ergebnissen von Reddock et al. (2011), mit den Befunden der Metatraits-Forschung (Abschnitt 2.1.1) sowie mit der theoretischen Konzeption von Variabilität (Abschnitt 2.2.3). Verhalten von Personen mit hoher Variabilität wird stärker durch den Kontext bzw. die Situation beeinflusst als Verhalten von Personen mit niedriger Variabilität, welches stärker von Persönlichkeitseigenschaften beeinflusst wird. Dies sollte sich auch auf Kriterien auswirken. Erwartet wird:

H3D: Der Zusammenhang von Persönlichkeitseigenschaften und Kriterien hängt von der Variabilität ab; er ist bei Personen mit niedriger Variabilität stärker als bei Personen mit hoher Variabilität.

Hypothese 3D kann geprüft werden, indem der Einfluss von Variabilität auf die Kriteriumsvalidität betrachtet wird. Die Kriteriumsvalidität hängt vom Zusammenhang zwischen den gemessenen Eigenschaften und dem Kriterium sowie von der Reliabilität der Messungen ab. Die Hypothese gilt dann als bestätigt, wenn die Kriteriumsvalidität stärker von der Variabilität abhängt als die Reliabilität.

Zusammengefasst wird für die Eigenschaftsdiagnostik vorhergesagt, dass Variabilität den Zusammenhang von Eigenschaften und Verhalten sowie von Eigenschaften und Kriterien moderiert. Die Zusammenhänge zwischen Eigenschaften und die Stabilität von Eigenschaften werden nicht von Variabilität moderiert. Da vermutet wird, dass ERS ein Indikator von Variabilität ist (Abschnitt 4.2), ist davon auszugehen, dass sich für ERS die gleichen Effekte auf die Reliabilität und Validität von Eigenschaftsmessungen sowie auf die Stabilität von und Zusammenhänge zwischen Persönlichkeitseigenschaften zeigen wie für Variabilität.

4.4 Ausblick auf die empirischen Studien

Die in den Abschnitten 4.1 bis 4.3 vorgestellten Hypothesen werden im Rahmen von zwei empirischen Studien geprüft. Erstmals werden hier Variabilität und ERS gemeinsam, ihr Zusammenhang sowie ihr Einfluss auf die Messung von Persönlichkeitseigenschaften systematisch untersucht. Die Hypothesen werden dabei größtenteils mit unterschiedlichen Methoden überprüft. Da beide Studien online durchgeführt werden, wird im Folgenden zunächst kurz in das Thema Online-Test eingeführt (Abschnitt 4.4.1), daran anschließend wird ein Ausblick auf Studie 1 (Abschnitt 4.4.2) und Studie 2 (Abschnitt 4.4.3) gegeben.

4.4.1 Einführung in das Thema „Online-Studien“

Das Interesse an Online-Assessments in der Arbeits- und Organisationspsychologie steigt seit mehr als 10 Jahren an (cut-e, 2013; Lefever, Dal & Matthíasdottóttir, 2007; Lievens & Harris, 2003; A. M. Ryan & Ployhart, 2014) und mittlerweile setzen vier von fünf Unternehmen auf irgendeine Form von Online-Assessment (cut-e, 2013); folgerichtig sollten Erkenntnisse zu diagnostischen Verfahren und zu den erfassten Konstrukten auch online gewonnen werden. Die Vorteile dieses Durchführungsmodus liegen auch in geringeren Kosten der Datengewinnung (Fricker & Schonlau, 2002; Tuten, Urban & Bosnjak, 2002) bei vergleichbarer Qualität

der Daten (Truell, 2003), vergleichbaren Studienergebnissen (Huang, 2006) und vergleichbarer psychometrischer Qualität der eingesetzten Verfahren (Bartram & Brown, 2004; Chuah, Drasgow & Roberts, 2006; Coyne, Warszta, Beadle & Sheehan, 2005; Hertel, Naumann, Konradt & Batinic, 2002; Meade, Michels & Lautenschlager, 2007; Stanton, 1998). Weitere Vorteile sind die Möglichkeit, leicht große Stichproben zu rekrutieren (Lefever et al., 2007; Wilhelm & McKnight, 2002), und die Unabhängigkeit der Durchführung von Zeit und Ort (Tuten et al., 2002). Als möglicher Nachteil im Bereich Persönlichkeitsforschung können allenfalls die im Vergleich zu Papier-Bleistift-Testungen niedrigeren Ausprägungen auf der Dimension Extraversion und die höheren Ausprägungen auf der Dimension Offenheit für Erfahrungen angesehen werden (Marcus, Machilek & Schütz, 2006); diese Unterschiede sind jedoch klein und wirken sich – unabhängig davon, ob sie auf die Unterschiede in der Methode oder auf Unterschiede zwischen Stichproben zurückgehen – lediglich auf Normen aus, nicht auf die Reliabilität und die Validität. Schließlich handelt es sich lediglich um eine Verschiebung der Messwerte und keine Verschiebung innerhalb der Rangreihen; somit bleiben multivariate Verteilungen unbeeinflusst.

4.4.2 Studie 1: Einsatz eines Dimensions- und eines Facetten-Fragebogens

In Studie 1 werden zwei Fragebogen eingesetzt, davon einer zu einem späteren Zeitpunkt ein zweites Mal, sowie mehrere Kriterien erhoben. Mit einem der Fragebogen werden breite, weitgehend unabhängige Persönlichkeitsdimensionen gemessen, mit dem anderen Fragebogen Persönlichkeitsfacetten. Eine Besonderheit an der Studie ist, dass Variabilität und ERS mit einem Dimensions-Fragebogen und gleichzeitig – erstmals – mit einem Facetten-Fragebogen erfasst werden. Relevant sind Facetten-Fragebogen, weil sie in der Praxis oftmals bessere Vorhersagen von Kriterien ermöglichen als Dimensions-Fragebogen (Beermann, 2011; Dudley et al., 2006; Paunonen & Ashton, 2001; Vinchur et al., 1998). Für die Untersuchung von Variabilität und ERS hat der Einsatz dieser zwei Fragebogen-Typen mehrere Vorteile: Erstens lassen sich mit dem Facetten-Fragebogen die Hypothesen zur Erfassung und zur Struktur von Variabilität prüfen. Schließlich werden – anders als beim Einsatz von Big-Five-Fragebogen (wie bei Biderman & Reddock, 2012; Reddock et al., 2011) – miteinander verknüpfte Eigenschaften und somit korrelierte Skalen erfasst. Zweitens können Variabilität und ERS mit beiden Fragebogen gemessen und diese Messungen verglichen werden.

Schließlich kann drittens der Einfluss von Variabilität und ERS, gemessen auf dem einen Fragebogen, auf die Reliabilität, die Stabilität und die Validität des anderen Fragebogens untersucht werden. In Kapitel 5 wird Studie 1 ausführlich beschrieben und die inhaltlichen formulierten Hypothesen werden statistisch spezifiziert und getestet.

4.4.3 Studie 2: Vergleich von Auswahl- und Nicht-Auswahl-Daten

Studie 2 liegt ein Anwendungsfall zugrunde: Zur Personalauswahl wurde ein kurzer Facetten-Fragebogen eingesetzt. Die Daten werden hier analysiert und mit Daten aus einem Nicht-Auswahl-Kontext verglichen. Sowohl der Einsatz eines Facetten-Fragebogens als auch der Kontext, eine Auswahlsituation, stellen die Besonderheiten von Studie 2 dar. Bislang liegen keine Ergebnisse zu Variabilität und zu ERS vor, die auf im Personalauswahlkontext erhobenen Daten basieren. Praxisrelevant sind diese Ergebnisse insbesondere angesichts der potenziellen Moderation des Zusammenhangs zwischen Tests und Kriterien (Biderman & Reddock, 2012; Reddock et al., 2011; vgl. Abschnitt 2.3 und 4.2). Denn als Moderatoren könnten Variabilität oder ERS Hinweise auf die Relevanz von Persönlichkeitsmaßen für das Arbeitsverhalten geben bzw. darauf, inwieweit das diagnostische Urteil auf den Persönlichkeitsfragebogen gestützt werden kann. Hauptgütekriterien, deren Abhängigkeit von Variabilität und ERS in Studie 2 untersucht wird, sind die Split-Half-Reliabilität und die Kriteriumsvalidität.

Da davon auszugehen ist, dass die Antworten von Bewerbern sich von denen von Untersuchungsteilnehmern im Nicht-Auswahl-Kontext unterscheiden, werden die Forschungsbefunde zu Persönlichkeitsfragebogen in der Personalauswahl im Folgenden erläutert. Anschließend werden die Folgerungen, die sich daraus für Studie 2 ergeben, aufgeführt.

Persönlichkeitsfragebogen in der Personalauswahl

Es gilt als gesichert, dass mit Likert-Items erhobene Persönlichkeitsmaße prinzipiell verfälschbar sind (Alliger & Dwight, 2000; Ellingson, Sackett & Hough, 1999; B. A. Martin, Bowen & Hunt, 2002) und die Mittelwerte bei Bewerbern höher (bei Neurotizismus: niedriger) ausfallen als bei Angestellten (Bott et al., 2007; Kanning & Holling, 2001; Tsaousis & Nikolaou, 2001). In einigen Studien werden sowohl für Persönlichkeitsfragebogen (Bott et al., 2007; Tsaousis & Nikolaou, 2001) als auch für Maße von emotionaler Intelligenz (Lievens, Klehe & Libbrecht, 2011) auch geringere Varianzen unter Bewerbern als unter Nicht-Bewerbern be-

richtet. Das Setting hat darüber hinaus einen Einfluss auf die multivariaten Verteilungen der Skalen. Zwar zeigen sich bei den Skaleninterkorrelationen in einigen Fällen keine oder nur geringe Unterschiede zwischen Auswahl und Nicht-Auswahl (Lievens et al., 2011), in anderen Fällen fallen die Skaleninterkorrelationen im Auswahlkontext jedoch größer (= stärkere Zusammenhänge) aus (Bott et al., 2007; Schmit & A. M. Ryan, 1993; Tsaousis & Nikolaou, 2001). Entsprechend klärt der erste Faktor einer Faktorenanalyse von Persönlichkeitsskalen bei der Auswahl mehr Varianz auf als im Nicht-Auswahl-Setting (Collins & Gleaves, 1998; Kanning & Holling, 2001; vgl. Marcus, 2003). Trotz dieser Beeinträchtigung der Konstruktvalidität konnte für Auswahlsettings gute Kriteriumsvalidität von Persönlichkeitsfragebogen nachgewiesen werden (Ones & Viswesvaran, 1998), weshalb Marcus (2003) zufolge Verfälschungen von Persönlichkeitsfragebogen in der Praxis ignoriert werden können. Anderer Auffassung ist Kersting (2004), der von zur Auswahl eingesetzten konstruktorientierten Verfahren Konstruktvalidität einfordert. Diese sei notwendig für einen hypothesengeleiteten und anforderungsbezogenen Einsatz. In anderen Worten ist ein Analogieschluss, d. h. ein Schluss von einer Eigenschaft, die anhand von Indikatoren erschlossen wird, auf ein Kriterium, unzulässig, wenn die infrage stehende Eigenschaft überhaupt nicht erschlossen wird. Mittlerweile liegen neue Erkenntnisse dafür vor, dass die Skalen eines Big-Five-Fragebogens bei der Auswahl die Big Five zwar messen, aber gleichzeitig auch gemeinsam einen sechsten Faktor erfassen (Klehe et al., 2012; Schmit & A. M. Ryan, 1993), den Klehe et al. wie auch Schmit und A. M. Ryan als *Ideal Employee Factor* (IEF) bezeichnen. Dieser korreliert mit beruflicher Leistung, was die hohe Kriteriumsvalidität trotz mangelnder Konstruktvalidität bei der Auswahl erklärt. Klehe et al. zufolge verschwindet dieser Zusammenhang jedoch, wenn die Fähigkeit, Bewertungskriterien zu identifizieren (*Ability to Identify Criteria*, ATIC), konstant gehalten wird. Dass bedeutet, ATIC hat sowohl auf den IEF als auch auf berufliche Leistung einen Einfluss und dieser Einfluss ist verantwortlich für die Korrelation des IEF mit beruflicher Leistung.

Folgerungen für Studie 2

Die Besonderheiten der Daten, die mit Persönlichkeitsfragebogen im Personalauswahlkontext gewonnen werden, – höhere Mittelwerte, möglicherweise geringere Streuungen und ein stärkerer gemeinsamer Faktor als im Nicht-Auswahl-Kontext – erlauben, Rückschlüsse aus einem Vergleich zwischen Auswahl- und Nicht-Auswahl-Setting zu ziehen: und zwar hin-

sichtlich der Messung und Struktur von Variabilität sowie hinsichtlich des Zusammenhangs zwischen ERS und Variabilität. Die Korrektur von Variabilität soll auf die Beschaffenheit von Messungen zurückgeführt werden: Da sich die univariaten Verteilungen von Persönlichkeitsskalen bei der Auswahl von denen im Nicht-Auswahl-Setting unterscheiden, kann geprüft werden, ob die Korrekturmethode von Baird et al. (2006; vgl. Abschnitte 2.1.3 und 4.1) die Validität der Messung von Variabilität erhöht. Angesichts des vermutlich stärkeren ersten gemeinsamen Faktors der Skalen im Auswahl- gegenüber dem Nicht-Auswahl-Setting wird erwartet, dass ERS bei der Auswahl höher mit diesem gemeinsamen Faktor korreliert. Extremantworten sollten im Auswahlsetting aufgrund der Verschiebung der univariaten Verteilungen häufiger auftreten als im Nicht-Auswahl-Setting. Eine detaillierte Beschreibung von Studie 2 und eine Überführung der inhaltlichen in statistische Hypothesen sowie deren Überprüfung finden sich in Kapitel 6.

5 Studie 1

Studie 1 basiert auf zwei Untersuchungen zum ITB Personality Structure Assessment (ITB-PESA; Beermann, 2011, 2013), einem berufsbezogenen Persönlichkeitsfragebogen auf Facetten-Ebene. Mit den Untersuchungen wurden verschiedene Zwecke verfolgt, u. a. sollten die Retestreliabilität, die Konstrukt- und Kriteriumsvalidität sowie die Akzeptanz des ITB-PESA überprüft werden. Ergebnisse zur Validität und Akzeptanz wurden bereits an anderen Stellen veröffentlicht (Beermann & Heilmann, 2014; Beermann et al., 2013). In der vorliegenden Arbeit werden die Daten genutzt, um Variabilität und ERS im ITB-PESA sowie in einem – ebenfalls in der ersten Untersuchung eingesetzten – Dimensions-Fragenbogen, der deutschsprachigen Version der revidierten Fassung des HEXACO-Persönlichkeitsinventars (englische Originalversion: Lee & Ashton, 2004), zu untersuchen. Die Methode wird in Abschnitt 5.1 berichtet, in Abschnitt 5.2 werden die statistischen Hypothesen vorgestellt und geprüft, und in Abschnitt 5.3 werden die Befunde zusammengefasst.

5.1 Methode

Ein Überblick über die Durchführung der beiden Untersuchungen findet sich in Abschnitt 5.1.1, die Stichprobe wird in Abschnitt 5.1.2 vorgestellt und die Instrumente und Messungen werden in Abschnitt 5.1.3 erläutert. Ergänzungen zur Stichprobe werden in Anhang A aufgeführt, Ergänzungen zu den Messungen in Anhang B.

5.1.1 Stichprobenakquise und Durchführung der Untersuchungen

Zur Teilnahme an beiden Untersuchungen wurden Studierende des Bachelor-Studiengangs Psychologie der Fernuniversität Hagen rekrutiert. Die Akquise für die erste Untersuchung erfolgte über eine Mitteilung auf der Website des dortigen Instituts für Psychologie. Als Anreiz erhielten die Studierenden nach Bearbeitung Feedback über ihr Abschneiden im ITB-PESA sowie Versuchspersonenstunden[14]. Zur Teilnahme an der zweiten Untersuchung, dem Retest, wurden Studierende eingeladen, die an der ersten Untersuchung teilgenommen und dort ihre E-Mail-Adresse angegeben hatten. Die Einladungen zur Teilnahme an der Retest-

[14] Das Ansammeln einer bestimmten Zahl an Versuchspersonenstunden durch die Teilnahme an Studien ist an der Fernuniversität obligatorischer Bestandteil des Bachelor-Studiengangs Psychologie.

Untersuchung wurden vier bis fünf Monate nach der ersten Untersuchung an diese E-Mail-Adressen versendet und, nachdem die Einladungen versendet wurden, nahmen die Studierenden innerhalb eines Monats an der Retest-Untersuchung teil. Als Anreiz zur Teilnahme wurden wieder Versuchspersonenstunden vergeben, Feedback bereitgestellt und unter Personen, die die Verfahren vollständig bearbeiteten, Amazon-Gutscheine verlost. Die Durchführung beider Untersuchungen erfolgte online mit Hilfe der frei zugänglichen webbasierten Software testMaker (Hartweg, Milbradt, Zimmerhofer & Hornke, 2009). Teilnehmer konnten während der Erhebungszeiträume zu jeder Zeit per Link auf die Untersuchung zugreifen.

Die erste Untersuchung war in vier Teile gegliedert: Im ersten Teil wurde über die Untersuchung und den Datenschutz informiert, und Teilnehmer gaben demografische Daten sowie einen Code zur anonymisierten Zuordnung ihres Datensatzes zu ihrem Datensatz beim Retest an. Im zweiten Teil bearbeiteten die Teilnehmer zwei Persönlichkeitsfragebogen: das ITB-PESA sowie die revidierte Fassung des HEXACO-Persönlichkeitsinventars (HEXACO-PI-R) von Lee und Ashton (2004) in der deutschsprachigen 100-Item-Version[15]. Die Items der beiden Persönlichkeitsfragebogen wurden mit einer identischen sechs-stufigen Likert-Skala dargeboten. Die Kategorien der Likert-Skala wurden mit den Ziffern 1 bis 6 versehen, die Endpunkt wurden zusätzlich verbal beschrieben: 1 stand für „trifft überhaupt nicht zu“, 6 für „trifft voll zu“. Die Instruktionen werden in Anhang B.1 aufgeführt. Die eingesetzte Version des ITB-PESA umfasst 284 Items, von denen 200 für die in dieser Arbeit berichteten Skalen verwendet werden. Die Items wurden in randomisierter Reihenfolge dargeboten, und zwar so, dass zwischen zwei Items einer Skala mindestens ein Item einer anderen Skala lag. Die 100 Items des HEXACO-PI-R wurden einzeln an zufällig gewählten Positionen zwischen den Items des ITB-PESA präsentiert, innerhalb dieser 100 Items wurde die ursprüngliche Reihenfolge beibehalten. Die 384 Items wurden in 24 Blöcken à 16 Items pro (Web-)Seite präsentiert. Im dritten Teil der ersten Untersuchung wurden folgende Kriterien erhoben: (i) Schulnoten sowie Noten eines ggf. zuletzt abgeschlossenen Studiums und Zwischennoten des aktuellen Studiums, (ii) allgemeine Arbeitszufriedenheit und allgemeine Arbeitsleistung,

[15] Für diese Version wurden keine Kennwerte veröffentlicht. Ashton zufolge wurde die psychometrische Qualität jedoch bereits anhand eines großen Datensatzes (N=1122) belegt (M. Ashton, persönliche Kommunikation am 12.08.2013). Dieser Datensatz entstammt laut Ashton den Forschungsarbeiten von Hilbig und Zettler (2009), Hilbig, Zettler und Heydasch (2012), Hilbig, Zettler, Moshagen und Heydasch (2012), Zettler, Friedrich und Hilbig (2011), Zettler und Hilbig (2010), Zettler, Hilbig und Haubrich (2011) sowie Zettler, Hilbig und Heydasch (2013). In der vorliegenden Arbeit bezieht sich das Akronym HEXACO-PI-R stets auf die deutschsprachige 100-Item-Version.

(iii) aufgabenbezogene Arbeitsleistung und kontextbezogene Arbeitsleistung sowie (iv) kontraproduktives Arbeitsverhalten. Schließlich erhielten Teilnehmer im vierten Teil Feedback zum ITB-PESA.

Die zweite Untersuchung lässt sich in fünf Teile gliedern: Zunächst wurden Teilnehmer über die Untersuchung und den Datenschutz aufgeklärt, und der Code zur anonymisierten Zuordnung der Datensätze zu denen der ersten Untersuchung wurde erhoben. Im zweiten Teil wurde ein Teil der Aufgabengruppe „Diagramme und Tabellen“ der Demoversion des Tests für Masterstudiengänge in Wirtschafts- und Sozialwissenschaften (TM-WISO; ITB Consulting GmbH, 2012) bearbeitet[16] und anschließend ein Fragebogen zur Akzeptanz dieser Aufgabengruppe. Im dritten Teil war ein numerischer Intelligenztest (Teil „Umgang mit Zahlen“ der Intelligenz-Basis-Faktoren; Ibrahimović, Bulheller, Horn, Gitter & Institut für Test- und Begabungsforschung GmbH, 2006) zu bearbeiten und im Anschluss daran wiederum ein Fragebogen zur Akzeptanz des Tests. Der vierte Teil bestand aus dem ITB-PESA und einem Fragebogen zur Akzeptanz. Das ITB-PESA wurde mit derselben sechs-stufigen Likert-Skala erhoben wie in der ersten Untersuchung. Die eingesetzte Fragebogenversion umfasste 300 Items, von denen 200 mit den relevanten Items aus der ersten Untersuchung übereinstimmten. Die Items wurden randomisiert präsentiert, zwischen zwei Items einer Skala lag mindestens ein Item einer anderen Skala. Die 300 Items wurden in 20 Blöcken à 15 Items pro Seite präsentiert. Die Akzeptanz wurde jeweils mit einem Fragebogen aus der Akzept!-Fragebogen-Reihe (Kersting, n.d.) erfasst. Im fünften Teil konnten Teilnehmer die Studie bewerten, kommentieren und Feedback zu den Fähigkeitstests und zum Persönlichkeitsfragebogen erhalten.

In der vorliegenden Arbeit werden aus der ersten Untersuchung die Daten zu den Persönlichkeitsfragebogen sowie die Angaben zu den Kriterien „Arbeitszufriedenheit“ und „Note im Hochschulabschluss“ berichtet. Aus der zweiten Untersuchung sind die Daten zum ITB-PESA relevant. Als Kriterium wurde Arbeitszufriedenheit gewählt, da sie ein subjektives Kriterium darstellt, das nicht leistungsbezogen und somit wenig verfälschungsanfällig ist. Vorhergesagt wird es mit der Skala „Erfolgszuversicht“ des ITB-PESA, da diese Selbstwirksamkeit im Beruf erfasst (Beermann, 2011) und Selbstwirksamkeit Arbeitszufriedenheit gut vorhersagt (Judge & Bono, 2001). Die Note im Hochschulabschluss wurde als weiteres Kriterium gewählt, da sie

[16] Der Vollständigkeit halber sei darauf verwiesen, dass zur Untersuchung einer hier nicht relevanten Fragestellung zwischen zwei Testversionen der Aufgabengruppe „Diagramme und Tabellen“ variiert wurde.

als echtes Leistungsmaß angesehen werden kann, das valide mittels Selbstauskünften erfasst werden kann (Greiff, 2006). Vorhergesagt wird es mit der Skala „Leistungsstreben und Erfolgsmotivation“ des ITB-PESA, da diese Leistungsmotivation erfasst und Leistungsmotivation als ein für Persönlichkeitseigenschaften vergleichsweise guter Prädiktor akademischer Leistung gilt (Robbins et al., 2004; Schmidt-Atzert, 2005).

5.1.2 Beschreibung der Stichprobe

In der ersten Untersuchung wurde die Bearbeitung der Persönlichkeitsfragebogen 632 Mal begonnen und bei 417 der 632 Datensätze (66.0 %) vollständig abgeschlossen. Bei einem Datenscreening und einer anschließenden Datenbereinigung (J. A. Johnson, 2005) wurde ein doppelter Fall identifiziert, von dem der später erfasste gelöscht wurde. Zwei weitere Fälle wiesen sehr kurze Bearbeitungszeit auf (weniger als 4 Sekunden pro Item) und bei neun Fällen traten lange Folgen gleicher Antworten in den Persönlichkeitsfragebogen (selbe Antworten auf mehr als 10 Likert-Items in Folge) auf; auch diese 11 Datensätze wurden eliminiert, so dass die Daten von 405 der 632 Fälle (64.1 %) in die Analysen eingingen. Die demografischen Merkmale der 405 Personen werden in Tabelle 2 zusammengefasst (Untersuchungsgruppe 1A): Die meisten der 334 Frauen (82.5 %) und 71 Männer (17.5 %) waren im Alter von 20 bis 49 Jahren, ein Großteil der Personen hatte mehrere Jahre Berufserfahrung. Einen akademischen Abschluss hatten 152 Personen (37.5 %) erreicht.

Tabelle 2: Demografische Merkmale der Untersuchungsgruppe 1A

Alter	Häufigkeit	Höchster Bildungsabschluss	Häufigkeit	Berufserfahrung	Häufigkeit
unter 20 Jahren	3 (0.7 %)	Fachhochschulreife	21 (5.2 %)	Keine	40 (9.9 %)
20 bis 24 Jahre	70 (17.3 %)	Abitur	155 (38.3 %)	unter 1 Jahr	26 (28.9 %)
25 bis 29 Jahre	65 (16.0 %)	Berufsausbildung	72 (17.8 %)	1 bis 5 Jahre	91 (22.5 %)
30 bis 39 Jahre	145 (35.8 %)	Bachelor	20 (4.9 %)	6 bis 10 Jahre	86 (21.2 %)
40 bis 49 Jahre	97 (24.0 %)	Diplom (FH)	45 (11.1 %)	11 bis 15 Jahre	69 (17.0 %)
über 49 Jahre	24 (5.9 %)	Diplom / Master	80 (19.8 %)	16 bis 20 Jahre	35 (8.6 %)
		Promotion	7 (1.7 %)	21 bis 30 Jahre	42 (10.4 %)
				über 30 Jahre	9 (2.2 %)
keine Angabe	1 (0.2 %)	*keine Angabe*	5 (1.2 %)	*keine Angabe*	7 (1.7 %)

Von den 405 Personen beantworteten 394 Personen (97.3 %) alle Items zur Erfassung des Kriteriums Arbeitszufriedenheit (Untersuchungsgruppe 1B). Die Abschlussnote eines früheren Hochschulstudiums berichteten 144 der 405 Personen (35.6 %, Untersuchungsgruppe 1C). Von den 144 Personen haben 45 (31.3 %) ihren höchsten erreichten Studienabschluss im Studienfeld Wirtschaftswissenschaften, 19 in den sogenannten MINT[17]-Fächern (13.2 %), 15 in Sprachwissenschaften (10.4 %) und je 14 (9.7 %) in Sozialwissenschaften und Pädagogik gemacht. Ein Studium in einem anderen Fach schlossen 8 Personen (5.6 %) ab, 29 (20.1 %) machten keine Angabe. Die demografischen Merkmale der Personen in den Untersuchungsgruppen 1B und 1C werden in Anhang A (Tabelle A - 1) aufgeführt.

Die Bearbeitung des Retests wurde 329 Mal begonnen. In 157 Fällen (47.7 %) wurde der Persönlichkeitsfragebogen vollständig bearbeitet. Darunter wurden zwei Bearbeitungen wegen zu langer Folgen gleicher Antworten und fünf wegen zu kurzer Bearbeitungszeiten ausgeschlossen. Zwanzig weitere Fälle waren doppelt (gleicher Code zur Zuordnung der Daten mit denen der ersten Untersuchung): Davon wurde bei 9 der 10 Paare der weniger plausible Fall eliminiert (alle Items in einem Leistungstest falsch, deutlich kürzere Bearbeitungszeit, unvollständiger Datensatz, kein Feedback angesehen), von dem anderen Datensatz-Paar wurden aufgrund mangelnder Unterscheidbarkeit beide Fälle von den Analysen ausgeschlossen. Von den verbleibenden 139 Fällen lagen zu 93 (66.9 %) Angaben vor, anhand derer die Ergebnisse mit denen der ersten Untersuchung verglichen werden konnten. In dieser Arbeit werden aus der Retest-Erhebung die Daten dieser 93 Fälle berichtet (Stichprobe 1D). Die demografischen Merkmale werden in Anhang A (Tabelle A - 1) aufgeführt.

5.1.3 Instrumente und Messungen

In den folgenden Abschnitten werden die für diese Arbeit relevanten Instrumente und Messungen ausführlich beschrieben. Berichtet werden zunächst die Skalen, der Messbereich und die Gütekriterien des HEXACO-PI-R und des ITB-PESA, daran anschließend die Items zur Erfassung der Kriterien und Kennwerte zu deren psychometrischer Qualität. Abschließend werden die Methoden zur Bestimmung von Variabilität und von ERS erläutert.

[17] Mathematik, Informatik, Naturwissenschaften, Technik

HEXACO-PI-R (Lee & Ashton, 2004), deutschsprachige Fassung der 100-Item-Version

Mit dem HEXACO-PI-R werden mit je 16 Items die sechs Dimensionen des HEXACO-Modells – Ehrlichkeit-Bescheidenheit (H: *Honesty-Humility*), Emotionalität (E: *Emotionality*), Extraversion (X: *Extraversion*), Verträglichkeit versus Ärger (A: *Agreeableness vs. Anger*), Gewissenhaftigkeit (C: *Conscientiousness*) und Offenheit für Erfahrungen (O: *Openness to Experience*) – gemessen[18]. Zudem lassen sich mit je vier Items je vier Facetten pro Dimensionen erfassen, sowie eine weitere Facette, die mit mehreren der HEXACO-Dimensionen assoziiert ist, „Altruismus“ (siehe auch Anhang B.2, Tabelle B.2 - 1, erste Spalte; Lee & Ashton, 2009). Wie von Beermann und Heilmann (2014) skizziert, zeigen viele Veröffentlichungen seit der Jahrtausendwende, dass das HEXACO-Modell einen breiteren Messbereich hat als das Fünf-Faktoren-Modell (FFM) (u. a. Ashton, Lee, Perugini et al., 2004; M. K. Johnson, Rowatt & Petrini, 2011; Saucier, 2009). Ebenso wie das FFM geht das HEXACO-Modell auf einen lexikalischen Ansatz zurück. Anders als beim FFM werden allgemeine Eigenschaftsbeschreibungen zu sechs statt fünf Faktoren gruppiert (Ashton, Lee & Goldberg, 2004; Lee & Ashton, 2004). Dabei stimmen drei Dimensionen (Extraversion, Gewissenhaftigkeit, Offenheit für Erfahrungen) des HEXACO-Modells weitgehend mit denen des FFM überein, die anderen beiden FFM-Dimensionen finden sich nicht deckungsgleich im HEXACO-Modell wieder: Gegenüber FFM-Neurotizismus umfasst Emotionalität die Facette Sentimentalität (FFM: Verträglichkeit) und gegenüber dem FFM gehört die Facette Ärger (FFM: Neurotizismus) im HEXACO-Modell zu Verträglichkeit (bzw. ihrem Gegenpol). Ferner enthält das HEXACO-Modell die Dimension Ehrlichkeit-Bescheidenheit, die im FFM nicht beschrieben ist. Neben dem größeren Eigenschaftsbereich bieten die HEXACO-Dimensionen auch bessere Vorhersagen von Kriterien als die Dimensionen des FFM (Ashton, Lee, Perugini et al., 2004; Saucier, 2009; M. K. Johnson et al., 2011). Diese Überlegenheit ist vermutlich historisch und methodologisch bedingt (vgl. Beermann & Heilmann, 2014, S. 70): Für die Forschungsarbeiten, auf Basis derer das FFM formuliert wurde, wurden lediglich 342 Eigenschaftswörter empirisch untersucht (Cattell, 1943). Die Big Five wurden daraufhin als Ergebnisse von Faktorenanalysen von 20 (W. T. Norman, 1963), 22 (Fiske, 1949) oder 35 Variablen (Tupes & Christal, 1961) extrahiert. Das HEXACO-Modell basiert dagegen auf den Ergebnissen einer Faktorenanalyse mit 1710 Eigenschaftswörtern (Ashton, Lee & Goldberg, 2004).

[18] Die in dieser Arbeit aufgeführten Übersetzungen der HEXACO-Dimensionen und der dazugehörigen Facetten stammen von Moshagen, Hilbig & Zettler (2014).

Die Gütekennwerte des HEXACO-PI-R wurden anhand der Daten von Untersuchungsgruppe 1A ($N = 405$) bestimmt und zum Teil bereits von Beermann und Heilmann (2014) berichtet; für die Analyse in Studie 1 werden die Faktor-Skalen verwendet. Deren Konstruktvalidität belegt eine Faktorenanalyse der Facetten, bei der sich das erwartete Sechs-Faktoren-Muster (siehe Anhang B.2, Tabelle B.2 - 1) zeigt. Ferner sind die Schätzungen der Reliabilität der Faktor-Skalen zufriedenstellend ($.74 < \alpha < .84$ bzw. $.74 < r_{tt} < .89$) und die Korrelationen zwischen diesen Skalen niedrig. In Anhang B.3 werden die Skalen einschließlich des Items mit der höchsten Trennschärfe und der Reliabilitätsschätzungen (Tabelle B.3 - 1) sowie die Skaleninterkorrelationen (Tabelle B.3 - 2) aufgeführt.

ITB-PESA (Beermann, 2011, 2013; Beermann & Heilmann, 2014)

Das ITB-PESA ist eine Testbatterie zur Erfassung kompetenzorientierter Persönlichkeitseigenschaften. Erfasst werden 23 Eigenschaftsfacetten, von denen sich die meisten psychometrisch einer der sechs HEXACO-Dimensionen zuordnen lassen. Augenscheinlich lassen sich die Eigenschaften in von Unternehmen häufig genutzte Kompetenzmodelle (Höft & Obermann, 2010; Lievens & Thornton, 2005; Obermann, 2009, S.86) integrieren. Die 23 Facetten können folglich auch als Facetten der HEXACO-Dimensionen aufgefasst werden, die an Kompetenzen orientiert sind und daher stärkeren Bezug zum Berufsleben und höhere Vorhersagekraft für berufliche Leistung aufweisen als mit einem allgemeinen Persönlichkeitsfragebogen gemessene Facetten (Beermann & Heilmann, 2014). Im Gegensatz zu vielen anderen berufsbezogenen Persönlichkeitsfragebogen ist der Messbereich des ITB-PESA breit – mit jeder HEXACO-Dimension korreliert mindestens eine ITB-PESA-Facette hoch (Beermann & Heilmann, 2014). Gleichzeitig werden berufsrelevante Eigenschaftsfacetten differenziert erfasst, insbesondere im Bereich der Gewissenhaftigkeit und der Extraversion.

Die 23 Skalen des ITB-PESA werden mit 212 der in der ersten Untersuchung eingesetzten 284 Items gebildet (Beermann & Heilmann, 2014). Hier werden jedoch nur 22 Skalen berichtet, die mit 200 der Items gebildet werden. Nicht im Rahmen der vorliegenden Arbeit berücksichtigt werden 3 der 212 Items, die nicht in der Retest-Untersuchung eingesetzt wurden. Unberücksichtigt bleibt auch die Skala „Integrität“ (im Kompetenzbereich *Integrität und Verlässlichkeit*), da vier der elf zu dieser Skala zugeordneten Items auch für andere Skalen gewertet werden. Diese vier Items werden hier den Skalen „Ehrlichkeit“ bzw. „Regelbe-

wusstsein“ zugeordnet. Neben den drei nicht im Retest eingesetzten Items und den sieben nicht berichteten Items der Skala „Integrität“ werden zwei weitere Items hier nicht aufgeführt, die ebenfalls jeweils zwei Skalen zugeordnet werden[19]. In Tabelle 3 findet sich eine Übersicht über die 22 Skalen des ITB-PESA, denen die in dieser Arbeit berichteten Ergebnisse zugrunde liegen, und ihre inhaltliche Nähe zu den HEXACO-Dimensionen. Aufgeführt sind auch die jeweiligen Itemzahlen und Reliabilitätsschätzungen. In Anhang B.4 werden zur Illustration der Skalen die Items mit der jeweils höchsten Trennschärfe aufgeführt (Tabelle B.4 - 1). In Anhang B.5 werden weitere Skalenstatistiken und die Skaleninterkorrelationen des ITB-PESA angeführt (Tabelle B.5 - 1), in Anhang B.6 das Ergebnis einer Faktorenanalyse der ITB-PESA-Skalen (Tabelle B.6 - 1).

Die Reliabilität der meisten Skalen ist zufriedenstellend. Die konvergenten und die diskriminanten Validitäten zum HEXACO-PI-R sind ebenfalls gut (Beermann & Heilmann, 2014). Die Skaleninterkorrelationen sind zwar mitunter hoch (Anhang B.5, Tabelle B.5 - 1), sie lassen sich jedoch gut erklären und für die Praxis nutzen (Beermann & Heilmann, 2014). Auch die Kriteriumsvalidität ist auf den ersten Blick positiv zu bewerten: So werden Selbstberichte allgemeiner Arbeitszufriedenheit und allgemeiner Arbeitsleistung (Beermann, 2011, 2013) sowie von aufgaben- und kontextbezogener Arbeitsleistung (Beermann & Heilmann, 2014) von jeweils für diese Kriterien theoretisch relevanten Skalen vorhergesagt. Lediglich Zusammenhänge mit objektiven Maßen von Berufserfolg liegen noch nicht vor.

[19] Bei Beermann und Heilmann (2014) werden die Skalen *einschließlich* der hier unberücksichtigten Items berechnet. Erwähnt wird jedoch, dass Skalen, die ein gemeinsames Item enthalten, im Ernstfalleinsatz nicht zusammen eingesetzt werden sollten.

Tabelle 3: Überblick über die Skalen des ITB-PESA, ihre Einordnung in ein Kompetenzmodell, ihre Korrelationen mit den HEXACO-Dimensionen sowie die Skalenstatistiken

	Korrelation zu		Skalenstatistiken				
Kompetenzbereich und Skala	r ≥ .50	.50 > r ≥ .30	$n\ (neg.)$	α	r_{tt}	$\overline{r_{it}}$	r_{TT}
Soziale Kompetenz							
Kontaktfreude	X		8 (4)	.85	.89	.61	.91
Kommunikationsvermögen	X	O	8 (1)	.69	.51	.40	.76
Geselligkeit	X		10 (7)	.77	.79	.44	.85
Einfühlungsvermögen	E		8 (4)	.71	.74	.42	.67
Konsensorientierung		A	10 (8)	.68	.70	.35	.72
Aufgeschlossenheit und Neugier		X, O	8 (0)	.63	.63	.33	.69
Führungskompetenz							
Leadership	X	C	10 (1)	.80	.80	.48	.82
Steuerungsvermögen		X, C	10 (0)	.89	.91	.63	.64
Führungswille und Machtmotivation	X	A(-)	10 (4)	.87	.91	.61	.82
Souveränität		E(-), X, A	9 (9)	.70	.68	.38	.72
Unternehmerische Kompetenz							
Ganzheitlich-strategische Denkweise	O	X	8 (4)	.68	.62	.38	.75
Kundenorientierung	X		8 (1)	.69	.65	.39	.59
Mut und Risikobereitschaft		E(-), X	9 (4)	.78	.73	.48	.74
Eigeninitiative	X	C, O	9 (1)	.78	.80	.47	.70
Markt- und Wettbewerbsorientierung		X	9 (3)	.82	.87	.53	.72
Ergebnisorientierung							
Arbeitsdisziplin	C	X	11 (7)	.88	.89	.59	.82
Ausdauer und Belastbarkeit	C	E(-), X	10 (4)	.81	.83	.51	.71
Sorgfalt	C		9 (1)	.80	.83	.50	.85
Erfolgszuversicht	X	E(-)	9 (4)	.86	.89	.59	.77
Leistungsstreben und Erfolgsmotivation		C	9 (2)	.77	.74	.46	.71
Integrität & Verlässlichkeit							
Ehrlichkeit	HH	A, C	10 (9)	.72	.75	.39	.64
Regelbewusstsein	C		8 (6)	.72	.69	.42	.61

n: Itemzahl, neg: Zahl negativ gepolter Items, α: Cronbachs Alpha, r_{tt}: Split-Half-Reliabilität (odd-even, Spearman-Brown-korrigiert), $\overline{r_{it}}$: mittlere Trennschärfe Items der Skala (Part-Whole-korrigiert, berechnet mit Fishers Z-Transformation, Fisher, 1918), erhoben an Untersuchungsgruppe 1A, $N = 405$

r_{TT}: Retestreliabilität, erhoben an Untersuchungsgruppe 1D, $N = 93$

H: Ehrlichkeit-Bescheidenheit, E: Emotionalität, X: Extraversion, A: Verträglichkeit versus Ärger, C: Gewissenhaftigkeit, O: Offenheit für Erfahrungen

Kriterien

Das Kriterium „Arbeitszufriedenheit“ wurde mittels Selbstauskünften erfasst. Es wurden dieselben drei Likert-Items verwendet wie bei Beermann (2011), die Antwortkategorien dieser Items waren wie folgt ausschließlich verbal umschrieben:

(I) „Wie hoch würden Sie Ihre Lebenszufriedenheit bewerten?“
(6) „sehr hoch“, (5) „hoch“, (4) „eher hoch“, (3) „eher niedrig“, (2) „niedrig“, (1) „sehr niedrig“

(II) „Wie zufrieden sind Sie mit Ihrer beruflichen Situation allgemein?“
(6) „sehr zufrieden“, (5) „zufrieden“, (4) „eher zufrieden“, (3) „eher unzufrieden“, (2) „unzufrieden“, (1) „sehr unzufrieden“

(III) „Wie zufrieden sind Sie mit Ihren beruflichen Aufgaben?“
(6) „sehr zufrieden“, (5) „zufrieden“, (4) „eher zufrieden“, (3) „eher unzufrieden“, (2) „unzufrieden“, (1) „sehr unzufrieden“

Zwar bezieht sich das erste der drei Items auf die Lebenszufriedenheit, da jedoch Lebens- und Arbeitszufriedenheit sehr hoch korrelieren ($r = .94$; Judge, Heller & Mount, 2002) und die Skala bei Beermann (2011) mit $\alpha = .73$ eine hohe interne Konsistenz hatte, lässt sich die Erfassung von Arbeitszufriedenheit durch diese drei Items rechtfertigen. In der ersten Untersuchung (Untersuchungsgruppe 1B) lag der Mittelwert bei $M = 12.23$ ($SD = 2.93$) im Wertebereich von 3 bis 18. Die interne Konsistenz lag bei $\alpha = .76$. Ein Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilung (K-S-Test) war zwar signifikant ($Z = 2.054$, $p < .001$) und die Verteilung von Arbeitszufriedenheit war leicht schief (siehe Histogramm, Anhang B.7, Abbildung B.7 - 1), dies kann jedoch vermutlich auf die Häufung der Ausprägung $x = 15$ (Modalwert) zurückgeführt werden. Der Median lag bei $Med = 12$ und damit in der Nähe des Mittelwerts. Vier weitere Gründe sprachen dafür, zur Berechnung der Ergebnisse die nicht transformierten Werte zu verwenden: Erstens waren die Werte logarithmiert auch nicht normalverteilt; zweitens wurde bei den Likert-Items Intervallskalenniveau angenommen (u. a. bei der Reliabilitätsschätzung und der Berechnung des Mittelwerts); drittens berichtete Beermann (2011) mit parametrischen Verfahren verlässliche Ergebnisse für die

Summe der drei eingesetzten Items; und viertens sollte sich die schiefe Verteilung nicht validitätssteigernd auswirken (Havlicek & Peterson, 1977).

Das zweite Kriterium war die Abschlussnote im höchsten abgeschlossenen Hochschulstudium. Gefragt wurde:

> „Falls Sie bereits ein Studium abgeschlossen haben, welche Durchschnittsnote haben Sie erreicht (zwischen 1.0 und 4.0)? (Falls Sie mehrere Abschlüsse haben, geben Sie bitte die Note des höchsten erreichten Abschlusses an.)“

Das Antwortformat war frei. Der Mittelwert der Abschlussnote lag studienfeldübergreifend bei $M = 1.99$ ($SD = 0.56$; Untersuchungsgruppe 1C). Zwischen den Studienfeldern bestanden keine Unterschiede (univariate ANOVA: $F_{(4;107)} = 1.906$, $p = .115$, $\eta^2 = .070$). Ebenso wie das Kriterium „Arbeitszufriedenheit“ waren auch die Noten nicht normalverteilt (K-S-Test: $Z = 1.407$, $p = .038$); im Histogramm (Anhang B.7, Abbildung B.7 - 2) sind besonders viele Werte mit $x = 2.0$ zu erkennen, was unter Umständen darauf zurückzuführen ist, dass ein Teil der Untersuchungsteilnehmer ihre Note in ganzen Zahlen eingegeben ($N = 13$) und somit Nachkommastellen vernachlässigt hat. Dieses Phänomen tritt jedoch nicht bei $x = 1.0$ oder $x = 3.0$ auf. Daher, und weil Mittelwert, Median und Modus ($Mod = Med = 2.0$) nahe beieinander liegen, weil Intervallskalenniveau angenommen wird und weil die Nicht-Normalverteilung die Validität nicht steigert (Havlicek & Peterson, 1977), wurden die Werte wie bei „Arbeitszufriedenheit“ nicht transformiert. Beide Kriterien korrelierten nicht signifikant miteinander ($r_{(140)} = -.165$, $p = .067$; 140 Personen machten Angaben zu beiden Kriterien).

Intraindividuelle Variabilität

Orientiert an Baird et al. (2006), Biderman und Reddock (2012) sowie Reddock et al. (2011) wird zur Berechnung der Variabilität zunächst jeweils die Inter-Item-SD der Skalen mittels Regressionen um den Einfluss des Mittelwerts und um den des Quadrats des (z-standardisierten) Mittelwerts korrigiert. Anschließend wird abweichend von früheren Forschungsarbeiten nicht die Summe der korrigierten SD berechnet, sondern für jeden Fragebogen ein Faktor. Gegenüber der Summe geht in den Faktor ausschließlich die gemeinsame Varianz der korrigierten Inter-Item-SD ein und die Gruppen-Varianz der Variabilität einer Skala spielt keine Rolle, da der Faktor auf der Korrelationsmatrix beruht. Bei der Summe da-

gegen hätte die korrigierte Inter-Item-SD nämlich mehr Gewicht, wenn ihre Gruppen-Varianz größer ist. Schließlich gehen in die Varianz einer Summe die Kovarianzen ein.

Als Extraktionsmethode wird den Empfehlungen von Costello und Osborne (2005), Reise, Waller und Comrey (2000) sowie Russell (2002) folgend eine Hauptachsenanalyse (*Principal Axis Factor Analysis*, PAF) verwendet: Verglichen mit der PAF hätte eine PCA den Nachteil, dass sie Ladungen und somit Kommunalitäten und Eigenwerte überschätzt. Aus theoretischer Sicht werden Hauptachsen anders als Hauptkomponenten zudem als Ursachen der Ausprägungen auf den zu faktorisierenden Variablen angenommen (Bühner, 2011; Fabrigar, Wegener, MacCallum & Strahan, 1999; Reise et al., 2000).

Extreme Response Style

ERS wird parallel zu Variabilität ebenfalls faktorenanalytisch bestimmt: Berechnet wird jeweils die Anzahl extremer Antworten („1“ oder „6“) pro Skala und dann wird aus diesen Summen für jeden Fragebogen ein ERS-Faktor mittels PAF extrahiert. Zur Überprüfung der Validität der Faktoren als Operationalisierung von ERS werden die relativen Häufigkeiten der Extremantworten pro Fragebogen sowie drei Greenleaf-Skalen (nach Greenleaf, 1992b; vgl. Abschnitt 3.1.1) gebildet. Die Greenleaf-Skalen werden für das ITB-PESA, für das HEXACO-PI-R sowie einmal für Items, die eingesetzt wurden, aber weder für die Skalen des ITB-PESA noch für die des HEXACO-PI-R gewertet werden, berechnet. Sie werden jeweils mit einem Teil der Items aus der jeweiligen Gruppe gebildet. Die Extremwerthäufigkeit der ausgewählten Items ist jeweils der Score für die Greenleaf-Skala. Die Itemauswahl ist orientiert an den Vorgaben von Greenleaf (1992b): Die Items sollten Likert-skaliert nicht miteinander korrelieren und möglichst häufig extrem beantwortet werden. In der vorliegenden Studie wird zur Itemauswahl jeweils eine PCA mit anschließender Varimax-Rotation und dem Kaiser-Guttmann-Kriterium (Eigenwert > 1; siehe z. B. Bortz & Schuster, 2010) gerechnet. Verwendet werden sollen jeweils Items, die am stärksten auf ihrer Hauptkomponente laden (mit höchstem Betrag); diese Items korrelieren nämlich eher gering miteinander. Zudem sollte der Gruppen-Mittelwert der zu verwendenden Items Likert-skaliert nicht zu extrem sein ($2.5 < M < 4.5$ im Wertebereich von 1 bis 6). Von den Items, die diese beiden Kriterien erfüllen, sollen wie von Greenleaf (1992b) empfohlen die 16 Items mit der größten Extremwerthäufigkeit gewählt werden.

5.2 Ergebnisse

Als erstes wird im Folgenden beschrieben, wie Variabilität und ERS im ITB-PESA sowie im HEXACO-PI-R erfasst wurden, und die Analysen für die jeweiligen Messungen werden berichtet (Abschnitte 5.2.1 und 5.2.2). Im dritten Abschnitt (5.2.3) dieses Unterkapitels werden die Analysen der Zusammenhänge zwischen Variabilität und ERS vorgestellt. Schließlich werden in den letzten beiden Abschnitten Analysen zu den Fragen präsentiert, ob Variabilität und ERS die Split-Half-Reliabilität und die Retestreliabilität (Abschnitt 5.2.4) sowie die Konstrukt- und die Kriteriumsvalidität (Abschnitt 5.2.5) der Persönlichkeitsfragebogen moderieren. Sofern nicht anders angegeben, beziehen sich die Ergebnisse auf die Untersuchungsgruppe 1A. Zur Datenanalyse wurde die Software SPSS 19.0.0 verwendet; konfirmatorische Faktorenanalysen bzw. SEM wurden mit SPSS Amos 19.0.0 (Arbuckle, 2010) gerechnet. Wurde eine Hypothese für einen Fragebogen mit einem statistischen Verfahren für mehrere ausgewählte Variablensets parallel getestet, erfolgte eine Korrektur des Alpha-Fehlerniveaus. Für die vorliegende Arbeit wurde dafür eine neue Korrekturmethode entwickelt, die in Anhang C beschrieben wird. Da diese Methode erstmals eingeführt wird, werden zum Vergleich auch die Ergebnisse nach der Bonferroni-Holm-Korrektur (Holm, 1979) berichtet.

5.2.1 Die Erfassung und Struktur von intraindividueller Variabilität

Zunächst wurden für jede Skala des ITB-PESA bzw. des HEXACO-PI-R die Inter-Item-SD berechnet und mittels Regression um den Mittelwert und das Quadrat des jeweiligen (z-standardisierten) Mittelwerts bereinigt. Dadurch wurden zwischen 1.9 % und 28.0 % der Varianz der Inter-Item-SD für die Skalen des ITB-PESA aufgeklärt und eliminiert, beim HEXACO-PI-R waren es zwischen 3.5 % und 32.5 %. Die einzelnen Werte sowie die deskriptiven Statistiken der Inter-Item-SD sind in Anhang D.1 und D.2 aufgeführt (Tabelle D.1 - 1 und Tabelle D.2 - 1). Die Residuen, die korrigierten Inter-Item-SD, wurden als Indikator für die Variabilität auf den Skalen verwendet. Infolge der Korrektur sollte Variabilität somit unabhängig von der Lage der Mittelwerte auf den Likert-Skalen indiziert werden.

Als Prüfung, ob die Korrektur der Inter-Item-SD um den Item-Mittelwert die Validität der Messung von Variabilität erhöht, diente ein Vergleich der Korrelation der Skalen und der Inter-Item-SD mit den Korrelationen innerhalb der Skalen. Die Korrektur um den Mittelwert

und das Quadrat des jeweiligen (z-standardisierten) Mittelwerts wäre nämlich dann im Hinblick auf die Validität der Variabilitätsmessung unangemessen, wenn bedeutsame Varianz aus den Inter-Item-SD eliminiert würde. Ein bedeutsamer Anteil an Varianz der Inter-Item-SD ist sachlogisch mit dem jeweils erfassten Merkmal verknüpft; das bedeutet, wenn zwei Skalen ein ähnliches Konstrukt messen, dann muss die Inter-Item-SD dieser Skalen auch in ähnlicher Weise korrigiert werden. Ist dies nicht der Fall, wird mit der oben genannten Korrektur inhaltlich irrelevante Varianz aus den Inter-Item-SD entfernt. Die korrigierten Anteile ließen sich dann auf methodische Artefakte zurückführen. Die Hypothese wurde wie folgt operationalisiert und getestet:

(i) Zunächst wurde für jede Skala die Korrelation zwischen Mittelwert und Inter-Item-SD berechnet. Diese Korrelationen geben Auskunft darüber, wie stark die Inter-Item-SD bei der Messung von Variabilität um den Einfluss des Item-Mittelwerts „bereinigt“ wird. Die resultierenden Werten wurden Z-transformiert (Fisher, 1918) und waren somit verhältnisskaliert (Bortz & Schuster, 2010).

(ii) Für jedes Paar an Skalen wurden nun erst die Differenz der unter (i) ermittelten Z-transformierten Korrelationen und dann der Betrag dieser Differenz berechnet. Dieser Betrag zeigt an, wie ähnlich die Inter-Item-SD von zwei Skalen korrigiert werden: Niedrige Werte zeigen, dass die Inter-Item-SD von zwei Skalen in ähnlichem Ausmaß um den Mittelwert korrigiert werden, hohe Werte zeigen, dass beide Skalen in unterschiedlichem Ausmaß korrigiert werden.

(iii) Als nächstes wurde der Zusammenhang zwischen dem unter (ii) errechneten Betrag für ein Skalenpaar und der Korrelation zwischen den Item-Mittelwerten dieses Skalenpaars untersucht. Inhaltlich bedeutet das, es wurde getestet, ob Inter-Item-SD ähnlich stark um den Einfluss des Item-Mittelwerts korrigiert werden (niedrige Differenz nach Schritt ii), wenn sie ähnliche Konstrukte messen (hohe Korrelation zwischen den Item-Mittelwerten). Wenn sich ein deutlicher negativer Zusammenhang zeigt, dann verringert die Korrektur der Inter-Item-SD um den Item-Mittelwert die Validität der Messung von Variabilität; denn es wäre davon auszugehen, dass die Inter-Item-SD jeweils mit dem Inhalt der Skala verknüpft ist. Ist die Korrektur angemessen, zeigt sich dieser Zusammenhang nicht.

(iv) Spearmans Rangkorrelationskoeffizienten für diesen Zusammenhang war niedrig und nicht signifikant ($rho_{(231)} = -.104$, $p = .116$). Zwar beträgt die Teststärke bei einem kleinen Effekt ($r = -.100$, einseitig) nur $1 - \beta = .452$ (berechnet mit GPower 3.19; Faul, Erdfelder, Lang & Buchner, 2007); wäre die Korrektur der Inter-Item-SD jedoch validitätsmindernd, sollten sich – insbesondere aufgrund der großen Bandbreite der Skaleninterkorrelationen des ITB-PESA (vgl. Anhang B.5, Tabelle B.5 - 1) – größere Zusammenhänge zeigen. Die Teststärke für einen mittleren Effekt ($r = -.300$, einseitig) betrug $1 - \beta = .999$.

Ob die Korrektur der Inter-Item-SD um den Einfluss des Quadrats des (z-standardisierten) Mittelwerts angemessen ist, wurde nach gleichem Muster überprüft: Bei gleicher Teststärke zeigte sich ein nicht signifikantes Ergebnis ($rho_{(231)} = .042$, $p = .527$). Auch die Korrektur um den Einfluss des Quadrats des (z-standardisierten) Mittelwerts ist also – im Sinne der Validität der Messung von Variabilität – angemessen. Die Inter-Item-SD wird also – wie intendiert – um methodische Artefakte bzw. um Einflüsse der Messmethode korrigiert und diese Korrektur ist unabhängig von den mit den Skalen erfassten Merkmalen.

Als globale Maße für Variabilität wurden im nächsten Schritt Faktoren der korrigierten Inter-Item-SD für das ITB-PESA und für das HEXACO-PI-R berechnet. Ob die Voraussetzungen für die PAF gegeben sind, wurde jeweils mit dem Kaiser-Mayer-Olkin-Koeffizienten (KMO), mit dem Bartlett-Test auf Sphärizität sowie mit K-S-Tests auf Ablehnung der Normalverteilungsannahme untersucht. Da beim ITB-PESA der KMO-Wert hoch und der Bartlett Test signifikant ausfielen, wurde trotz einiger signifikanter K-S-Tests angenommen, dass die Daten für die PAF geeignet sind. Beim HEXACO-PI-R deuten alle Kennwerte darauf hin, dass sich die Daten gut für eine PAF eignen. In Tabelle 4 werden die KMO-Koeffizienten und die Ergebnisse der Bartlett-Tests aufgeführt und die K-S-Tests zusammengefasst. Die einzelnen K-S-Tests werden in Anhang D.1 und D.2 (Tabelle D.1 - 1 und Tabelle D.2 - 1) berichtet.

Tabelle 4: Prüfung der Voraussetzungen für eine Faktorenanalyse der korrigierten Inter-Item-SD der Skalen des ITB-PESA und der Skalen des HEXACO-PI-R

		Bartlett-Test auf Sphärizität			signifikante
Faktor der korrigierten Inter-Item-SD	KMO	χ^2	df	p	K-S-Tests
für die Skalen des ITB-PESA	.96**	3290.35	231	<.001	9 von 22
für die Skalen des HEXACO-PI-R	.88*	854.56	15	<.001	0 von 6

Untersuchungsgruppe 1A, $N = 405$

KMO: Kaiser-Mayer-Olkin-Koeffizient, * gute Eignung, ** sehr gute Eignung (vgl. Bühner, 2011)

Bartlett-Test: χ^2: Teststatistik, df: Freiheitsgrade, p: Signifikanzniveau; $p < .05$ deutet auf gute Eignung der Daten für eine Faktorenanalyse hin.

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis)

Ob den korrigierten Inter-Item-SD jeweils tatsächlich genau ein Faktor zugrunde liegt, wurde mittels *Minimum-Average-Partial*-Test (MAP-Test, Velicer, 1976) und Scree-Test (Cattell, 1966) kontrolliert. Beide Verfahren legten erwartungsgemäß die Extraktion jeweils nur eines Faktors nahe[20]. Dieser klärt jeweils einen großen Varianzanteil auf. Die Ladungsmuster sind ebenfalls hypothesenkonform: In beiden PAF (für das ITB-PESA und für das HEXACO-PI-R) laden alle korrigierten Inter-Item-SD hoch auf dem Faktor[21]. Die Ergebnisse werden in Tabelle 5 zusammengefasst. Aufgeführt werden die Ergebnisse von MAP- und Scree-Test, der Anteil durch den Faktor aufgeklärter Varianz sowie das Minimum und Maximum der Ladungen der korrigierten Inter-Item-SD auf dem jeweiligen Faktor. Die Ladungen werden im Einzelnen in Anhang D.1 und D.2 (Tabelle D.1 - 1 und Tabelle D.2 - 1) aufgeführt; die Scree-Plots sind in Anhang D.3 (Abbildung D.3 - 1 und Abbildung D.3 - 2) abgebildet.

[20] Eine alternative Methode wäre die Parallelanalyse nach Horn (1965). Mit dieser Analyse wird laut Beauducel (2001) die Faktorenzahl jedoch unterschätzt, wenn der erste Faktor – wie für den vorliegenden Fall erwartet – einen hohen Eigenwert hat.

[21] Bei einer PAF der *nicht* korrigierten Inter-Item-SD zeigt sich ein ähnliches Bild. Allerdings sind der Anteil durch den Faktor aufgeklärter Varianz sowie die Ladungen jeweils kleiner, da die Inter-Item-SD für Variabilität irrelevante Varianzanteile enthält.

Tabelle 5: Ergebnisse der Faktorenanalyse der korrigierten Inter-Item-SD der Skalen von ITB-PESA und HEXACO-PI-R sowie Konsistenzwerte für die Faktoren

	Anzahl der Faktoren		aufgekl. Var	Ladungen		Konsistenz	
Faktor der korrigierten Inter-Item-SD	MAP-Test	Scree-Test	(1. Fakt.)	Min	Max	α	ω
für die Skalen des ITB-PESA	1	1	36.0 %	.49	.68	.93	.92
für die Skalen des HEXACO-PI-R	1	1	48.5 %	.66	.72	.85	.85

Untersuchungsgruppe 1A, $N = 405$

aufgekl. Var (1. Fakt.): durch den (ersten) Faktor aufgeklärte Varianz, Min: niedrigste Ladung, Max: höchste Ladung; alle Ladungen waren positiv.

α: Cronbachs Alpha (hier – da es sich um Faktoren handelt – für standardisierte Werte berechnet), ω: Omega

In den beiden rechten Spalten von Tabelle 5 wird die interne Konsistenz, operationalisiert durch Cronbachs Alpha und McDonalds Omega (McDonald, 1978, 1999), berichtet. Da jeweils der Faktor (nicht die Summe) als Maß für die Variabilität dient, wurde Alpha mit den standardisierten Variablen berechnet. Schließlich basiert der Faktor der korrigierten Inter-Item-SD auf Korrelationen, im Gegensatz zur Summe, der Kovarianzen zugrunde liegen. Omega ist Revelle und Zinbarg (2009) sowie Stone et al. (2013) zufolge ein besserer Schätzer der internen Konsistenz eines Faktors als Alpha. Den Empfehlungen von Revelle und Zinbarg folgend werden hier beide Werte berichtet. Omega wurde im Rahmen der PAF ermittelt. Zur Berechnung wird ein Quotient gebildet: Im Zähler werden die Ladungen a_i auf dem Faktor summiert und die Summe quadriert. Im Nenner wird zu diesem Quadrat die Summe der quadrierten Ladungen der Variablen auf den ihnen spezifischen (Fehler-)Faktoren hinzuaddiert:

$$\omega = \frac{(\sum a_i)^2}{(\sum a_i)^2 + \sum {\delta_i}^2}$$

a_i sind dabei die Ladungen des Items i auf dem Faktor. δ_i ist die Ladung des Items auf dem ihm spezifischen (Fehler-)Faktor; ${\delta_i}^2$ ist entsprechend die Spezifität des Items und kann wie folgt berechnet werden: ${\delta_i}^2 = 1 - {a_i}^2$.

Den Ergebnissen der Reliabilitätsanalysen zufolge wurde Variabilität für beide Fragebogen reliabel erfasst. Der K-S-Test war jeweils nicht signifikant, die Normalverteilungsannahme wurde für die Variabilitäts-Faktoren beider Fragebogen beibehalten (ITB-PESA: $Z = 0.814$, $p = .515$; HEXACO-PI-R: $Z = 0.743$, $p = .639$). Die beiden Faktoren korrelierten hoch mit-

einander ($r_{(405)} = .879,\ p < .001$). Die Retestreliabilität des Variabilitäts-Faktors im ITB-PESA war für das Vier-Monats-Intervall hoch ($r_{(93)} = .758,\ p < .001$, Untersuchungsgruppe 1D). Sogar der Variabilitäts-Faktor im HEXACO-PI-R bei der ersten Erhebung korrelierte hoch mit dem Variabilitäts-Faktor im ITB-PESA beim Retest ($r_{(93)} = .670,\ p < .001$, Untersuchungsgruppe 1D). Die hohen Zusammenhänge zeigen, dass mit den Variabilitäts-Faktoren in beiden Instrumenten dasselbe Konstrukt erfasst wird. Variabilität kann reliabel, stabil und unabhängig vom verwendeten Fragebogen erfasst werden.

Der hohe Anteil an Varianz, den der erste Faktor jeweils aufklärt, und die schmale Bandbreite der Ladungen (siehe Tabelle 5, fünfte und sechste Spalte) lassen vermuten, dass Variabilität nicht als Metatraits (vgl. Baumeister & Tice, 1988; Britt, 1993; Dwight et al., 2002), die jeweils auf einen bestimmten Trait bezogen sind, verstanden werden kann. Eher handelt es sich wie von Baird et al. (2006), Biderman und Reddock (2012) sowie Reddock et al. (2011) berichtet um einen globalen Trait. Dennoch ist festzustellen, dass sich mindestens die Hälfte der Varianz einer korrigierten Inter-Item-SD einer Skala nicht durch den Faktor erklären lässt. Unklar ist, ob dieser Varianzanteil bedeutsam ist, d. h. ob er zu dem jeweiligen Trait gehört, oder nicht. Gehört er zum Trait, ließe sich die Metatraits-Theorie halten, Metatraits wäre Facetten der Variabilität. Wenn die eigene Varianz der korrigierten Inter-Item-SD einer Skala nicht zum jeweils erfassten Trait gehört, wäre sie auf die Skala, d. h. auf das Messinstrument, oder auf einen Messfehler zurückzuführen und theoretisch irrelevant. Als empirische Prüfung diente ein Vergleich der Korrelationen innerhalb der Item-Mittelwerte der Skalen des ITB-PESA mit den Korrelationen innerhalb der korrigierten Inter-Item-SD dieser Skalen[22]. Der Annahme von Metatraits folgend müssten die korrigierten Inter-Item-SD stärker für Traits konvergieren, die ihrerseits stärker miteinander zusammenhängen: Im Extrembeispiel wird mit zwei Skalen, die das gleiche Merkmal messen (als Item-Mittelwert), auch derselbe Metatrait erfasst (als korrigierte Inter-Item-SD); im anderen Extrembeispiel korrelieren die Metatraits zweier Skalen, die verschiedene Merkmale erfassen, nicht über die durch den Variabilitäts-Faktor aufgeklärte Varianz hinaus.

[22] Da mit dem HEXACO-PI-R sechs voneinander unabhängige Faktoren erfasst werden und die Facettenskalen mit nur vier Items vermutlich keine reliable Bestimmung der Variabilität erlauben, wurde diese Berechnung nur für das ITB-PESA durchgeführt.

Ob nun Variabilität mit den erfassten Merkmalen verknüpft ist und somit mehrere Variabilitäts-Facetten bestehen oder ob Variabilität ein eindimensionales Konstrukt ohne bedeutsame Varianz über die Dimension hinaus ist, wurde mit zwei verschiedenen Methoden geprüft:

(1) Erstens wurde der Zusammenhang zwischen den Korrelationen[23] innerhalb der Item-Mittelwerte und den Korrelationen innerhalb der entsprechenden korrigierten Inter-Item-SD mit Spearmans Rangkorrelationskoeffizient berechnet. Mit anderen Worten wurde berechnet, ob die korrigierten Inter-Item-SD von zwei assoziierten Skalen höher zusammenhängen als die korrigierten Inter-Item-SD von nicht assoziierten Skalen. Dieser Zusammenhang fiel gering aus ($rho_{(231)} = .028$, $p = .676$); die Teststärke für einen mittleren Effekte war hoch ($1 - \beta = .999$). Die Korrelationen zwischen den Item-Mittelwerten der Skalen des ITB-PESA sind in Anhang B.5 (Tabelle B.5 - 1) aufgeführt, die Korrelationen zwischen den korrigierten Inter-Item-SD in Anhang D.4 (Tabelle D.4 - 1).

(2) Zweitens wurde eine konfirmatorische Faktorenanalyse mit Maximum-Likelihood-Schätzung für ein Modell gerechnet, bei dem jede korrigierte Inter-Item-SD ein Indikator des gemeinsamen Faktors war und zudem einen eigenen latenten Faktor (eigene und Fehlervarianz, hier mit Fehlerfaktor bezeichnet) aufwies. Die eigenen Faktoren waren voneinander unabhängig. Ein Chiquadrat-Test (${\chi_{(209)}}^2 = 409.17$, $p < .001$) und das Ergebnis des Bollen-Stine Bootstrap-Verfahrens ($p = .001$) waren zwar signifikant, die Stichprobe jedoch groß, und die Fit-Indizes deuteten auf einen guten bis akzeptablen Modellfit mit nur geringen Modellfehlspezifikationen hin ($CFI = .94$; $SRMR = .043$; $RMSEA = .049$, $CI\ 90\%: .042 - .056$)[24]. Bei der Inspektion der Modifikationsindizes fiel auf, dass sich weder die Pfade mit den Modifikationsindizes $M.I > 4$ (Voreinstellung von SPSS Amos, vgl. Bühner, 2011, S. 454) noch die Pfade mit den fünf höchsten Modifikationsindizes überwiegend auf Paare von Fehlerfaktoren beziehen, deren Skalen (Item-Mittelwerte) hoch miteinander korrelieren. Die Modifikationsindizes sind in Anhang D.5 (Tabelle D.5 - 1) aufgeführt. Die bedeutsame Varianz der korrigierten Inter-Item-SD geht also ausschließlich mit der Varianz des Faktors einher. Varianz der Inter-Item-SD einer

[23] Hier wurde der Betrag als Indikator der Ähnlichkeit zweier Merkmalsdimensionen verwendet. Schließlich ist im dimensionalen Eigenschaftsmodel der Metatrait einer Eigenschaft gleich dem Metatrait ihres Gegenpols. Zum Beispiel haben die Pole der vierten HEXACO-Dimension, Verträglichkeit und Ärger denselben Metatrait.

[24] Bühner (2011) empfiehlt bei – wie vorliegend – nicht multivariat normalverteilten Variablen, den p-Wert mittels Bollen-Stine Bootstrap zu berechnen. Als Fit-Indizes werden in dieser Arbeit die von Beauducel und Wittmann (2005) sowie Schweizer (2010) empfohlenen berichtet. Diese werden nach den Cut-Offs von Hu und Bentler (1999) sowie Schermelleh-Engel, Moosbrugger und Müller (2003) bewertet.

Skala, die nicht durch diesen Faktor aufgeklärt wird, ist nicht mit dem durch die Skala erfassten Merkmal verknüpft. Entsprechend wird mit der (korrigierten) Inter-Item-SD einer Skala kein Metatrait erfasst – Variabilität ist ein eindimensionaler universeller Trait.

5.2.2 Die Erfassung von Extreme Response Style

Im ITB-PESA wurden im Mittel 22.6 % ($M = 0.226$, $SD = 0.151$), im HEXACO-PI-R 29.9 % der Items ($M = 0.299$, $SD = 0.146$) extrem beantwortet. Für jede Skala wurde die Extremwerthäufigkeit ihrer Items summiert, und als globales Maß für ERS wurde für die beiden Fragebogen jeweils der Faktor dieser Summen (d. h. der Extremwerthäufigkeiten pro Skala) mittels PAF bestimmt. Die Voraussetzungen dafür waren bis auf die Annahme der Normalverteilung gegeben; der KMO-Koeffizient, die Ergebnisse des Bartlett-Tests auf Sphärizität und eine Zusammenfassung der K-S-Tests werden in Tabelle 6 berichtet. Die einzelnen K-S-Tests und die deskriptiven Statistiken der Extremwerthäufigkeiten der Skalen sind in Anhang D.6 (Tabelle D.6 - 1) und in Anhang D.7 (Tabelle D.7 - 1) aufgeführt.

Tabelle 6: Prüfung der Voraussetzungen für eine Faktorenanalyse der Extremwerthäufigkeiten auf den Skalen des ITB-PESA und des HEXACO-PI-R

		Bartlett-Test auf Sphärizität			signifikante
ERS-Faktor	KMO	χ^2	df	p	K-S-Tests
für die Skalen des ITB-PESA	.96**	4862.99	231	<.001	22 von 22
für die Skalen des HEXACO-PI-R	.86*	786.89	15	<.001	6 von 6

Untersuchungsgruppe 1A, $N = 405$

KMO: Kaiser-Mayer-Olkin-Koeffizient, * gute Eignung, ** sehr gute Eignung (vgl. Bühner, 2011)

Bartlett-Test: χ^2: Teststatistik, df: Freiheitsgrade, p: Signifikanzniveau; p < .05 deutet auf gute Eignung der Daten für eine Faktorenanalyse hin.

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis)

MAP-Test und Scree-Test zufolge liegt den Extremwerthäufigkeiten der Skalen von ITB-PESA und HEXACO-PI-R jeweils ein Faktor zugrunde. Bei der PAF klärt der erste Faktor jeweils einen relativ großen Anteil der Varianz auf (> 45 %); die Scree-Plots sind in Anhang D.8 (Abbildung D.8 - 1 und Abbildung D.8 - 2) aufgeführt. Die Ladungen der Extremwerthäufigkeiten auf den einzelnen Skalen sind hoch und haben eine geringe Bandbreite. Die internen Konsistenzen für die ERS-Faktoren sind ebenfalls hoch. Die Ergebnisse werden in Tabelle 7 angeführt. Die einzelnen Ladungen werden in Anhang D.6 (Tabelle D.6 - 1) und in Anhang D.7

(Tabelle D.7 - 1) berichtet. Mit K-S-Tests wurde die Verteilung der ERS-Faktoren überprüft: Für den Faktor im ITB-PESA wurde die Normalverteilungsannahme verworfen ($Z = 1.799$, $p = .003$), für den Faktor im HEXACO-PI-R ließ sie sich beibehalten ($Z = 1.160$, $p = .136$).

Tabelle 7: Ergebnisse der Faktorenanalyse der Extremwerthäufigkeiten auf den Skalen von ITB-PESA und von HEXACO-PI-R sowie Konsistenzwerte für die Faktoren

	Anzahl der Faktoren		aufgekl. Var	Ladungen		Konsistenz	
ERS-Faktor	MAP-Test	Scree-Test	(1. Fakt.)	Min	Max	α	ω
der Skalen des ITB-PESA	1	1	45.4 %	.56	.82	.95	.95
der Skalen des HEXACO-PI-R	1	1	45.6 %	.62	.73	.83	.83

Untersuchungsgruppe 1A, $N = 405$

aufgekl. Var (1. Fakt.): durch den (ersten) Faktor aufgeklärte Varianz, Min: niedrigste Ladung, Max: höchste Ladung; alle Ladungen waren positiv.

α: Cronbachs Alpha (hier – da es sich um Faktoren handelt – für standardisierte Werte berechnet), ω: Omega

Für die drei Greenleaf-Skalen wurden jeweils 16 Items ausgewählt. Likert-skaliert und in Richtung der jeweiligen Skala gepolt hingen diese Items etwas stärker zusammen als die der Greenleaf-Skalen bei Greenleaf (1992b) und Naemi et al. (2009) ($-.172 \leq r_{(405)} \leq .390$ im ITB-PESA; $-.217 \leq r_{(405)} \leq .256$ im HEXACO-PI-R; $-.189 \leq r_{(405)} \leq .213$ unter den zusätzlichen Items). Die Korrelationen waren jedoch gleichmäßig um $r = .00$ verteilt, der mittlere Zusammenhang war klein und die interne Konsistenz für die Summe war ebenfalls niedrig. Zur Berechnung des jeweiligen Greenleaf-Scores wurden die Items umkodiert: Die Endpunkte („1“ und „6“) wurden mit 1 gewertet, die übrigen Werte („2“ bis „5“) mit 0. Von den je 16 Items wurde im Mittel zwischen einem Sechstel und einem Viertel extrem beantwortet. ERS-kodiert zeigten sich überwiegend positive Zusammenhänge zwischen den Items (ITB-PESA: $-.028 \leq r_{(405)} \leq .335$, HEXACO-PI-R: $-.090 \leq r_{(405)} \leq .239$, bei zusätzlichen Items: $-.067 \leq r_{(405)} \leq .276$). Die interne Konsistenz lag jeweils im mittleren bis hohen Bereich. Eine Übersicht findet sich in Tabelle 8.

Tabelle 8: Statistiken zu den Greenleaf-Skalen, links für die Likert-Kodierung, rechts für die ERS-Kodierung

	Likert-Kodierung (1 bis 6)		ERS-Kodierung (Endpunkte: 1, „2“ bis „5“: 0)				
Greenleaf-Skala	α	$\overline{r_{ii}}$	M	SD	h_{rel}	α	$\overline{r_{ii}}$
mit Items des ITB-PESA	.20	.01	3.55	2.69	22.2 %	.67	.11
mit Items des HEXACO-PI-R	.21	.02	3.88	2.55	24.2 %	.60	.09
mit zusätzlichen Items	.39	.04	2.97	2.50	18.6 %	.66	.12

Untersuchungsgruppe 1A, $N = 405$

α: Cronbachs Alpha, $\overline{r_{ii}}$: mittlere Korrelation zwischen den Items (berechnet mit Fishers Z-Transformation, Fisher, 1918), M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, h_{rel}: relative Häufigkeit von Extremantworten

Zur Bestimmung der konvergenten Konstruktvalidität wurden die Zusammenhänge zwischen den verschiedenen ERS-Maßen überprüft. Wie Tabelle 9 zeigt, sind die Korrelationen durchweg sehr hoch. Mit allen Maßen wird also dasselbe Konstrukt – die Tendenz, extrem zu antworten – erfasst. Es spielt auch keine Rolle, für welchen Fragebogen die jeweiligen Maße erhoben werden. Für die Greenleaf-Skalen war die Konvergenz etwas niedriger, vermutlich aufgrund der etwas geringeren Reliabilität. Für den ERS-Faktor des ITB-PESA wurde die Retestreliabilität für das Vier-Monats-Intervall berechnet; diese betrug $r_{(93)} = .731$ ($p < .001$, Untersuchungsgruppe 1D). Damit liegt ein weiterer Beleg für die Reliabilität des ERS-Faktors im ITB-PESA vor und weitere Evidenz für die Stabilität von ERS.

Tabelle 9: Korrelationen zwischen den ERS-Maßen

	Korrelation zu					
Skala	2	3	4	5	6	7
1. ERS-Faktor ITB-PESA	.891	.998	.882	.826	.710	.746
2. ERS-Faktor HEXACO-PI-R		.895	.993	.770	.811	.724
3. ERS-Häufigkeit ITB-PESA			.887	.830	.713	.747
4. ERS-Häufigkeit HEXACO-PI-R				.762	.790	.722
5. Greenleaf-Skala ITB-PESA					.632	.650
6. Greenleaf-Skala HEXACO-PI-R						.576
7. zusätzliche Greenleaf-Skala						

Untersuchungsgruppe 1A, $N = 405$; für alle Korrelationen gilt p < .001.

5.2.3 Intraindividuelle Variabilität und Extreme Response Style

Die Hypothese, dass die Faktoren von ERS und Variabilität denselben breiten Trait indizieren, wurde zunächst anhand von Korrelationen überprüft. Tabelle 10 zeigt die Zusammenhänge: Die Variabilitäts-Faktoren beider Instrumente korrelieren hoch mit den ERS-Faktoren. Die Korrelationen liegen nur unwesentlich unterhalb der in den Abschnitten 5.2.1 und 5.2.2 berichteten Schätzungen für die Reliabilität. Die Variabilitäts-Faktoren konvergieren auch Fragebogen-übergreifend mit den ERS-Faktoren.

Tabelle 10: Korrelationen zwischen den Variabilitäts- und den ERS-Faktoren

	Korrelation zu ERS-Faktor	
Variabilitäts-Faktor	des ITB-PESA	des HEXACO-PI-R
des ITB-PESA	.860	.852
des HEXACO-PI-R	.800	.904

Untersuchungsgruppe 1A, $N = 405$; für alle Korrelationen gilt $p < .001$.

In einem zweiten Schritt wurde die Übereinstimmung von Variabilität und ERS mit SEM getestet. Als manifeste Variablen dienten die korrigierten Inter-Item-SD sowie die Extremwerthäufigkeiten der Skalen. Für beide Fragebogen wurden jeweils zwei Modelle aufgestellt. Für das ITB-PESA wurden die Modelle mit je 44 manifesten Variablen (d. h. den korrigierten Inter-Item-SD sowie den Extremwerthäufigkeiten der 22 Skalen) berechnet, für das HEXACO-PI-R mit jeweils 12. Beide Modelle sind in Abbildung 14 (Seite 89) exemplarisch für das HEXACO-PI-R abgebildet[25]. In Modell 1 lag den manifesten Variablen je ein Faktor zugrunde, in Modell 2 je zwei Faktoren, davon einer als Ursache der korrigierten Inter-Item-SD und einer als Ursache der Extremwerthäufigkeiten. In beiden Modellen waren die Fehlerfaktoren der korrigierten Inter-Item-SD und der Extremwerthäufigkeiten für eine Skala korreliert (z. B. der Fehlerfaktor der korrigierten Inter-Item-SD der Skala „Ehrlichkeit-Bescheidenheit“ mit dem Fehlerfaktor der Extremwerthäufigkeit derselben Skala), weitere Korrelationen zwischen Fehlerfaktoren wurden nicht angenommen.

Ergänzt sei, dass weitere Modelle aufgestellt wurden. In diesen war der Variabilitäts-Faktor als Ursache des ERS-Faktors definiert. Diese Modelle sind jedoch äquivalent zu Modell 2. Das

[25] Aufgrund der Größe des Modells, und da dies keinen Mehrwert für den Leser bedeuten würde, wird auf eine Darstellung der Modelle für das ITB-PESA verzichtet.

bedeutet, diese Modelle haben denselben Fit wie die hier berichteten Modelle des Typs *Modell 2*. SEM eignen sich also nicht zur Überprüfung eines kausalen Zusammenhangs.

Die Ergebnisse zur Maximum-Likelihood-Schätzung für die vier Modelle sind in Tabelle 11 (Seite 90) aufgeführt. Beim ITB-PESA weisen beide Modelle geringe Fehlspezifikationen auf. Für Modell 2 (zwei Faktoren) zeigt sich ein besserer Fit als für Modell 1 (ein Faktor). Allerdings korrelieren Variabilitäts- und ERS-Faktoren in diesem Modell zu $r = .903$. Kein Unterschied zwischen Modell 1 und Modell 2 zeigt sich beim HEXACO-PI-R: Beide Modelle weisen nach den Kriterien von Hu und Bentler (1999) sowie Schermelleh-Engel, Moosbrugger und Müller (2003) einen guten Fit auf. Da Variabilitäts- und ERS-Faktor in Modell 2 zu $r = .993$ korrelieren, liegt es nahe, dem Prinzip der Sparsamkeit folgend von der Ein-Faktoren-Lösung auszugehen und Variabilität und ERS als Indikatoren eines Traits zu beschreiben.

Zur weiteren Prüfung wurden in einem dritten Schritt SEM fragebogenübergreifend berechnet. Zum einen wurden Modell 1 und Modell 2 für die 22 korrigierten Inter-Item-SD der Skalen des ITB-PESA und die sechs Extremwerthäufigkeiten der Skalen des HEXACO-PI-R berechnet. In Modell 1 waren alle 28 manifesten Variablen dem einen Faktor zugeordnet, in Modell 2 waren die korrigierten Inter-Item-SD einem Faktor und die Extremwerthäufigkeiten dem anderen Faktor zugeordnet. Die beiden Faktoren waren korreliert. Zusammenhänge zwischen Fehlerfaktoren wurden nicht angenommen. Zum anderen wurden Modell 1 und Modell 2 nach gleichem Muster für die sechs korrigierten Inter-Item-SD der HEXACO-PI-R-Skalen und für die 22 Extremwerthäufigkeiten der ITB-PESA-Skalen berechnet. Die Ergebnisse sind in Tabelle 12 (Seite 90) aufgeführt. Mit den korrigieren Inter-Item-SD der Skalen des ITB-PESA und den Extremwerthäufigkeiten der Skalen des HEXACO-PI-R weisen beide Modelle einen akzeptablen Fit auf; die Daten passen etwas besser zu Modell 2 als zu Modell 1. Allerdings korrelieren in Modell 2 die beiden Faktoren zu $r = .976$, was eher für das Ein-Faktoren-Modell (Modell 1) spricht. Gehen die korrigierten Inter-Item-SD der Skalen des HEXACO-PI-R und die Extremwerthäufigkeiten der Skalen des ITB-PESA in die Berechnung ein, so ist der Fit beider Modell noch akzeptabel; die Daten passen jedoch deutlich besser zu Modell 2 als zu Modell 1. Die beiden Faktoren in Modell 2 korrelieren zu $r = .891$, so dass eher davon ausgegangen werden kann, dass diesen Variablen zwei latente Merkmale zugrunde liegen. Eine Interpretation dieser und der zuvor beschriebenen Ergebnisse (vgl. Tabelle 11, Seite 90) findet sich in der nach dem Ergebnisteil folgenden Diskussion (Abschnitt 5.3).

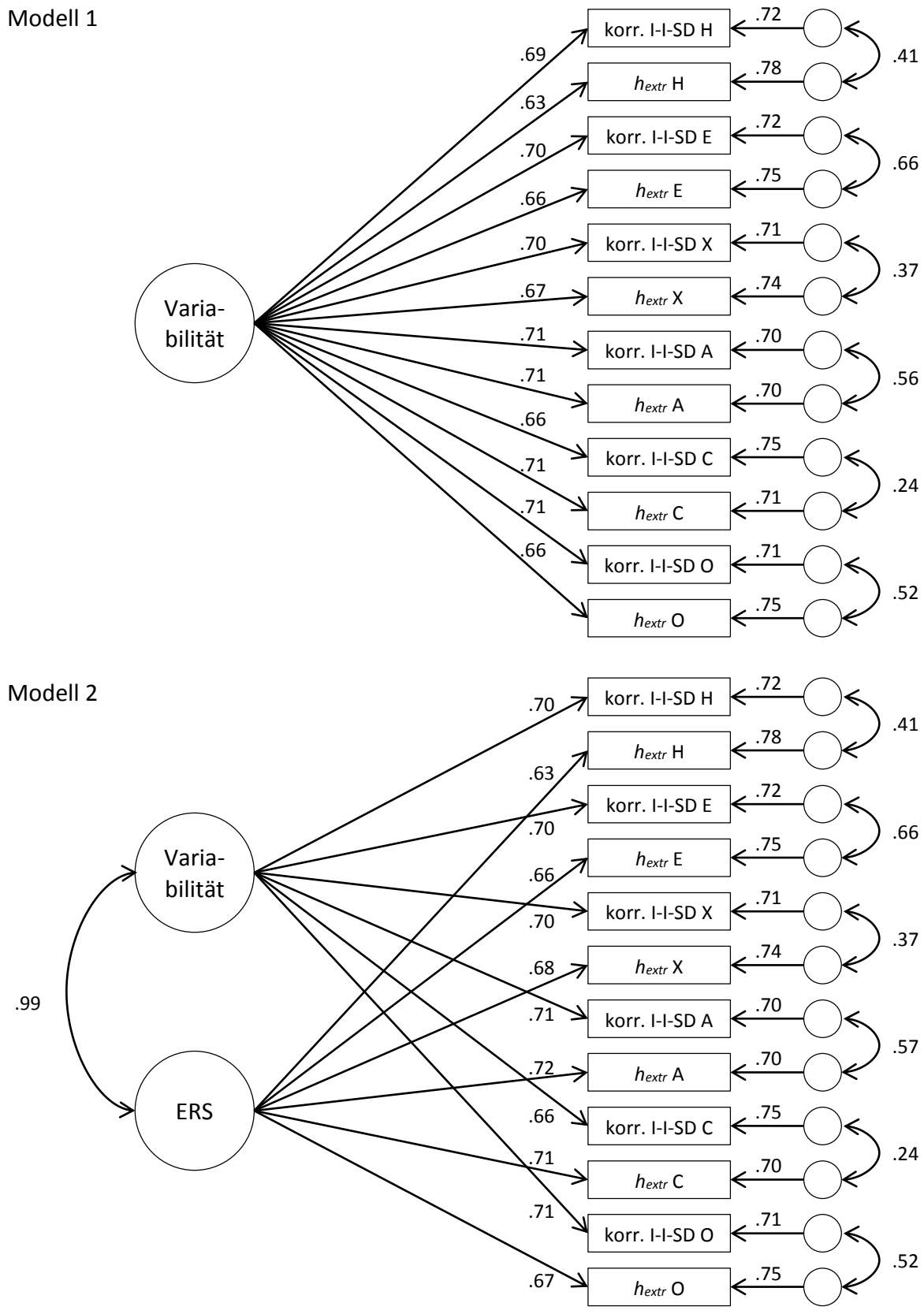

Abbildung 14: Strukturgleichungsmodelle für das HEXACO-PI-R einschließlich standardisierter Regressionsgewichte und Korrelationen

korr. I-I-SD: korrigierte Inter-Item-Standardabweichung, h_{extr}: Extremwerthäufigkeit, jeweils berechnet für die Skalen zur Messung von Ehrlichkeit-Bescheidenheit (H), Emotionalität (E), Extraversion (X), Verträglichkeit versus Ärger (A), Gewissenhaftigkeit (C) sowie Offenheit für Erfahrungen (O)

Tabelle 11: Analyse des Zusammenhangs von Variabilität und ERS mittels Strukturgleichungsmodellen für das ITB-PESA und für das HEXACO-PI-R

	Modelltest					Fit-Indizes			Modell 1 vs. 2		
Analyse für	χ^2	df	p	p_{BSB}	AIC	*CFI*	*SRMR*	*RMSEA (CI 90)*	$\Delta\chi^2$	Δdf	p
das ITB-PESA											
Modell 1	2438.75	880	<.001	.001	2658.75	.85	.049	.066 (.063-.069)			
Modell 2	2044.11	879	<.001	.001	**2266.11**	.89	.049	.057 (.054-.061)	394.64	1	<.001
das HEXACO-PI-R											
Modell 1	95.87	48	<.001	.002	**155.87**	.98	.028	.050 (.035-.064)			
Modell 2	94.93	47	<.001	.001	156.93	.98	.028	.050 (.036-.065)	0.94	1	.331

Untersuchungsgruppe 1A, $N = 405$

Modelltest: χ^2: Teststatistik, df: Freiheitsgrade, p: Signifikanzniveau; p < .05 deutet auf Modellfehlspezifikationen hin.

p_{BSB}: p-Wert für den Bollen-Stine-Bootstrap, AIC: *Akaike Information Criterion*, *CFI*: *comparative Fit Index*, *SRMR*: *standardized Root Mean Square Residual*, *RMSEA (CI 90)*: *Root Mean Square Error of Approximation* (90-Prozent-Konfidenzintervall)

Modell 1 vs. 2: $\Delta\chi^2$: Differenz der Chiquadrat-Werte zwischen Modell 1 und Modell 2, Δdf: Differenz der Freiheitsgrade, p: Signifikanzniveau; p < .05 deutet auf einen Unterschied zwischen den Modellen hin.

Tabelle 12: Analyse des fragebogenübergreifenden Zusammenhangs von Variabilität und ERS mittels Strukturgleichungsmodellen

	Modelltest					Fit-Indizes			Modell 1 vs. 2		
Analyse für	χ^2	df	p	p_{BSB}	*AIC*	*CFI*	*SRMR*	*RMSEA (CI 90)*	$\Delta\chi^2$	Δdf	p
Variabilität: ITB-PESA, ERS: HEXACO-PI-R											
Modell 1	649.20	350	<.001	.001	761.20	.93	.041	.046 (.040-.051)			
Modell 2	644.22	349	<.001	.001	**758.22**	.94	.041	.046 (.040-.051)	4.98	1	.026
Variabilität: HEXACO-PI-R, ERS: ITB-PESA											
Modell 1	1048.30	350	<.001	.001	1160.30	.89	.048	.070 (.065-.075)			
Modell 2	958.68	349	<.001	.001	**1077.68**	.90	.046	.066 (.061-.071)	89.62	1	<.001

Untersuchungsgruppe 1A, $N = 405$

Modelltest: χ^2: Teststatistik, df: Freiheitsgrade, p: Signifikanzniveau; p < .05 deutet auf Modellfehlspezifikationen hin.

p_{BSB}: p-Wert für den Bollen-Stine-Bootstrap, AIC: *Akaike Information Criterion*, *CFI*: *comparative Fit Index*, *SRMR*: *standardized Root Mean Square Residual*, *RMSEA (CI 90)*: *Root Mean Square Error of Approximation* (90-Prozent-Konfidenzintervall)

Modell 1 vs. 2: $\Delta\chi^2$: Differenz der Chiquadrat-Werte zwischen Modell 1 und Modell 2, Δdf: Differenz der Freiheitsgrade, p: Signifikanzniveau; p < .05 deutet auf einen Unterschied zwischen den Modellen hin.

5.2.4 Der Einfluss von Variabilität und ERS auf die Split-Half-Reliabilität und auf die Retestreliabilität

Die Reliabilität und die Stabilität wurden anhand der Split-Half-Reliabilität und anhand der Retestreliabilität operationalisiert. Inwieweit diese Gütekennwerte von Variabilität und ERS moderiert werden, wurde mit zwei verschiedenen Analyseverfahren überprüft: Zum einen wurden individuelle Profilübereinstimmungen und ihr Zusammenhang zu Variabilität und ERS berechnet, zum anderen wurden moderierte multiple Regressionen durchgeführt.

Moderation der Split-Half-Reliabilität und der Retestreliabilität als Profilübereinstimmung

Als Profilübereinstimmung im Sinne der Split-Half-Reliabilität wurden die Skalen des ITB-PESA und des HEXACO-PI-R jeweils mit der Hälfte ihrer Items (an ungeraden Positionen) gebildet und die Übereinstimmung des resultierenden Profils mit dem der anderen Skalenhälften (Items an geraden Positionen) berechnet. Für die Retestreliabilität wurde die Profilübereinstimmung des ITB-PESA bei der ersten Erhebung mit dem Profil beim Retest gebildet. Als Maß der Übereinstimmung diente aufgrund seiner günstigen Eigenschaften der *Index of Profile Agreement* (I_{pa})[26]. I_{pa} ist intervallskaliert und bei mehr als zwei Paaren an Skalen(hälften) im Profil normalverteilt (McCrae, 1993, 2008; McCrae et al., 1998). I_{pa} wird wie folgt mit z-standardisierten Skalenhälften bzw. Prä- und Post-Messungen berechnet:

$$I_{pa} = \frac{k + \sum M^2 - \sum d^2}{\sqrt{10k}}$$

k ist dabei die Anzahl der Skalenpaare, $\sum M^2$ die Summe der quadrierten Mittelwerte der Skalenpaare und $\sum d^2$ das Quadrat der euklidischen Distanz zwischen den beiden Profilen. Hohe Werte stehen bei diesem Index für hohe Übereinstimmung, niedrige Werte für niedrige Übereinstimmung. Der besonderer Vorteil, den I_{pa} bietet, ist die höhere Gewichtung von Abweichungen in der Nähe des Gruppen-Mittelwertes. Beispielsweise würde sich – was in-

[26] Andere Maße waren weniger geeignet. Die Korrelation zwischen einem Profil und dem anderen (vgl. Jacksons individuelle Reliabilität, Jackson, 1976; zitiert nach J. A. Johnson, 2005) vergleicht lediglich die Form, nicht die Lage beider Profile (McCrae, 2008). Zudem hängt sie von der Streuung innerhalb des Profils ab; wenn nämlich die Streuung der Skalen-Mittelwerte für eine Person klein ist, so mindern bereits kleine Abweichungen zwischen den Profilhälften die Übereinstimmung stark verglichen mit dem Profil einer Person, deren Eigenschaften mal sehr hoch und mal sehr niedrig ausgeprägt sind. Gleiches trifft auf die „double-entry Intraclass Correlation" (ICC_{DE}) zu. Dabei wird jedes Datenpaar zweimal, davon einmal mit vertauschten Elementen, aufgelistet und eine Korrelation für die Datenpaare berechnet. Eine Erläuterung des ICC_{DE} findet sich ebenso wie eine Beschreibung und Diskussion weiterer Maße der Profilübereinstimmung bei McCrae (2008).

haltlich sinnvoll sein kann – die Abweichung zwischen den T-Werten 40 und 60 stärker negativ auf I_{pa} auswirken als die Abweichung zwischen 60 und 80 (vgl. McCrae, 2008; McCrae et al., 1998). Erwartet wurde, dass I_{pa} negativ mit Variabilität und ERS zusammenhängt: Bei hoher Variabilität oder hohem ERS stimmen die Profile der ersten Skalenhälften mit denen der anderen Skalenhälften bzw. die Profile von Prä- und Postmessung schlechter überein.

Die Zusammenhänge sind in Tabelle 13 aufgeführt: Nur die Profilübereinstimmung der Skalenhälften im HEXACO-PI-R hängt signifikant mit Variabilität, gemessen mit dem ITB-PESA oder mit dem HEXACO-PI-R, zusammen: Je variabler eine Person die Skalen ITB-PESA oder des HEXACO-PI-R beantwortet, desto weniger stimmt beim HEXACO-PI-R das Profil aus den ersten Skalenhälften (Mittelwerte der Items an jeweils ungeraden Positionen innerhalb der Skala) mit dem Profil der anderen Skalenhälften (Mittelwerte der Items an jeweils geraden Positionen) überein. Die Split-Half-Reliabilität im HEXACO-PI-R ist also unter Personen mit niedriger Variabilität größer als unter Personen mit hoher Variabilität. Mit den ERS-Faktoren korreliert die Profilübereinstimmung der Skalenhälften im HEXACO-PI-R nicht. Die Profilübereinstimmung der Skalenhälften im ITB-PESA korreliert weder mit einem der Variabilitäts- noch mit einem der ERS-Faktoren negativ. Sie korreliert sogar positiv mit dem ERS-Faktor des ITB-PESA. Hinsichtlich der Retestreliabilität im ITB-PESA zeigen sich keine Effekte. Die Profilübereinstimmung von Prä- und Postmessung korreliert mit keinem der Variabilitäts- bzw. ERS-Faktoren.

Tabelle 13: Korrelationen der Profilübereinstimmungen für die Profile der Skalenhälften und für die Profile von Prä- und Postmessung mit den Variabilitäts- und ERS-Faktoren von ITB-PESA und HEXACO-PI-R

	Korrelation zu I_{pa} für die Skalenhälften (Untersuchungsgruppe 1A)		Korrelation zu I_{pa} zwischen Prä- und Postmessung (Untersuchungsgruppe 1D) für das ITB-PESA
	des ITB-PESA	des HEXACO-PI-R	
ITB-PESA			
Variabilitäts-Faktor	-.016 (.372)	**-.097** (.027)	-.036 (.367)
ERS-Faktor	.188 (>.999)	-.077 (.061)	.015 (.443)
HEXACO-PI-R			
Variabilitäts-Faktor	.073 (.930)	**-.177** (<.001)	.115 (.137)
ERS-Faktor	.121 (.992)	-.030 (.277)	.118 (.131)

Untersuchungsgruppe 1A, $N = 405$

I_{pa}: *Index of Profile Agreement* (Index der Profilübereinstimmung)

In Klammern werden die Signifikanzniveaus aufgeführt (einseitige Testungen). Signifikante Korrelationen sind fett gedruckt.

Moderierte multiple Regressionen für die Split-Half-Reliabilität

Der Einfluss von Variabilität und ERS auf die Split-Half-Reliabilität wurde auch einzeln für jede Skala geprüft. Dafür wurden moderierte multiple Regressionen verwendet, mit denen für jede Skala der Mittelwert der Items an geraden Positionen mit dem Mittelwert der Items an ungeraden Positionen vorhergesagt wurde. Zunächst wurden Skalenhälften sowie Variabilitäts- und ERS-Faktoren z-standardisiert. Im ersten Schritt wurden dann als Prädiktoren jeweils die erste Skalenhälfte sowie der jeweilige Moderator (Variabilität oder ERS) eingegeben. Der Moderator wurde auf dem jeweils anderen Fragebogen erfasst, d. h. bei einer Skala des ITB-PESA wurde der Variabilitäts- bzw. ERS-Faktor im HEXACO-PI-R erfasst und umgekehrt. Damit sollte Effekten, die auf Messfehler zurückgehen, vorgebeugt werden (vgl. Biderman & Reddock, 2012). Im zweiten Schritt wurde dann das Produkt aus dem jeweiligen Prädiktor und dem Moderator hinzugefügt. Das bedeutet beispielsweise für die Moderation der Split-Half-Reliabilität der ITB-PESA-Skala „Kontaktfreude“ durch Variabilität: Als abhängige Variable wurde die (z-standardisierte) Summe der Items an *geraden* Positionen der Skala verwendet. Als Prädiktoren wurden zunächst die (z-standardisierte) Summe der Items an *ungeraden* Positionen der Skala sowie der (z-standardisierte) Variabilitäts-Faktor des HEXACO-PI-R eingegeben. Im zweiten Schritt wurde das Produkt dieser beiden (z-standardisierten) Variablen hinzugefügt. Das Regressionsgewicht dieses Produkts steht in der Regressionsgleichung für die Moderation der Vorhersage einer Skalenhälfte von „Kontaktfreude“ auf Basis der anderen – in anderen Worten für den Einfluss von Variabilität im HEXACO-PI-R auf die Split-Half-Reliabilität der ITB-PESA-Skala „Kontaktfreude“.

Die Ergebnisse der Regressionsanalysen für die Split-Half-Reliabilität sind in Tabelle 14 (ITB-PESA) und in Tabelle 15 (HEXACO-PI-R) aufgeführt. Zur besseren Übersichtlichkeit werden nicht sämtliche Werte der Regressionen berichtet, sondern jeweils das standardisierte Regressionsgewicht (Beta) für die Moderation, das Ergebnis des entsprechenden einseitigen T-Tests sowie die Effektstärke (Änderung von R^2). Mit den T-Tests wurde einseitig getestet, weil erwartet wurde, dass Variabilität und ERS in eine bestimmte Richtung moderieren: Angenommen wurde eine höhere Split-Half-Reliabilität bei niedriger Variabilität bzw. niedrigem ERS, statistisch entspricht dies einem negativen Beta-Gewicht.

Bei der Moderation der Split-Half-Reliabilität durch Variabilität lag das Signifikanzniveau für zwei Skalen des ITB-PESA bei $p \leq .05$, alpha-adjustiert war nur ein Regressionsgewicht signifikant verschieden von Null. ERS moderierte alpha-adjustiert die Split-Half-Reliabilität einer Skala signifikant, insgesamt fielen zwei p-Werte unter $p = .05$[27].

Tabelle 14: Moderation der Split-Half-Reliabilität der Skalen des ITB-PESA durch Variabilität und durch ERS, gemessen mit dem HEXACO-PI-R

	Moderator Variabilität				Moderator ERS			
Moderation der Split-Halt-Reliabilität der Skala		T-Test				T-Test		
	β	T	p	ΔR^2	β	T	p	ΔR^2
Soziale Kompetenz								
Kontaktfreude	-.056	-1.773	**.039**	.003[a]	-.036	-1.168	.122	.001[a]
Kommunikationsvermögen	.071	1.507	.934	.005[b]	.072	1.547	.938	.005[b]
Geselligkeit	-.015	-0.374	.355	.000[a]	-.032	-0.819	.207	.001[a]
Einfühlungsvermögen	.093	2.271	.988	.008[b]	.107	2.608	.995	.011[b]
Konsensorientierung	.064	1.488	.931	.004[b]	.086	2.005	.977	.007[b]
Aufgeschlossenheit und Neugier	.036	0.815	.792	.001[b]	.037	0.841	.799	.001[b]
Führungskompetenz								
Leadership	.017	0.442	.670	.000[b]	.037	0.990	.838	.001[b]
Steuerungsvermögen	.046	1.571	.941	.002[b]	.060	2.109	.982	.003[b]
Führungswille und Machtmotivation	-.017	-0.562	.287	.000[a]	-.004	0.893	.447	.000[a]
Souveränität	.018	0.400	.655	.000[b]	.028	0.638	.738	.001[b]
Unternehmerische Kompetenz								
Ganzheitlich-strategische Denkweise	-.135	-3.007	**.001**[c]	.017[a]	-.091	-2.001	**.023**	.008[a]
Kundenorientierung	-.052	-1.212	.113	.003[a]	-.053	-1.225	.111	.003[a]
Mut und Risikobereitschaft	.024	0.575	.717	.001[b]	.021	0.489	.687	.000[b]
Eigeninitiative	.027	0.712	.761	.001[b]	.008	0.203	.580	.000[b]
Markt- und Wettbewerbsorientierung	.056	1.717	.956	.003[b]	.059	1.804	.964	.003[b]
Ergebnisorientierung								
Arbeitsdisziplin	.044	1.443	.925	.002[b]	.077	2.523	.994	.006[b]
Ausdauer und Belastbarkeit	-.011	-0.307	.380	.000[a]	-.021	-0.572	.284	.000[a]
Sorgfalt	.018	0.508	.694	.000[b]	.041	1.147	.874	.002[b]
Erfolgszuversicht	.020	0.673	.749	.000[b]	-.005	-0.152	.440	.000[a]
Leistungsstreben und Erfolgsmotivation	-.064	-1.571	.059	.004[a]	-.098	-2.446	**.007**[c]	.009[a]
Integrität & Verlässlichkeit								
Ehrlichkeit	.052	1.270	.897	.003[b]	.043	1.079	.858	.002[b]
Regelbewusstsein	.023	0.510	.694	.000[b]	.052	1.181	.881	.002[b]

Untersuchungsgruppe 1A, $N = 405$

β: standardisiertes Regressionsgewicht für den Moderatoreffekt, T: Teststatistik des Signifikanztests für β, p: Signifikanzniveau (einseitig), ΔR^2: Effektstärke (Änderung von R^2)

[a] Effekt in erwarteter Richtung, [b] Effekt in nicht erwarteter Richtung, p (einseitig) < .05 fett gedruckt, [c] nach Alpha-Adjustierung signifikant

[27] Die hier berichtete Alpha-Korrektur wird in Anhang C vorgestellt. Bonferroni-Holm-korrigiert (Holm, 1979) war lediglich einer der Effekte signifikant, und zwar der Einfluss von Variabilität auf die Split-Half-Reliabilität der Skala „Ganzheitlich-strategische Denkweise“.

Beim HEXACO-PI-R lag das Signifikanzniveau für einen Moderatoreffekte von Variabilität (gemessen mit dem ITB-PESA) und einen Moderatoreffekt von ERS (gemessen mit dem ITB-PESA) unter $p = .05$. Alpha-adjustiert war lediglich der Moderatoreffekt von Variabilität signifikant[28]. Bei den meisten Skalen zeigte sich wie beim ITB-PESA kein Moderatoreffekt.

Tabelle 15: Moderation der Split-Half-Reliabilität der Skalen des HEXACO-PI-R durch Variabilität und durch ERS, gemessen mit dem ITB-PESA

	Moderator Variabilität				Moderator ERS			
Moderation der Split-Halt-Reliabilität der Skala		T-Test				T-Test		
	β	T	p	ΔR^2	β	T	p	ΔR^2
Ehrlichkeit-Bescheidenheit	-.126	-2.917	**.002**[c]	.015[a]	-.100	-2.326	**.010**	.010[a]
Emotionalität	.002	0.046	.518	.000[b]	.007	0.144	.557	.000[b]
Extraversion	.021	0.594	.723	.000[b]	.012	0.332	.630	.000[b]
Verträglichkeit versus Ärger	.076	-1.968	.987	.005[b]	.060	1.571	.941	.003[b]
Gewissenhaftigkeit	.008	0.204	.581	.000[b]	-.021	-0.525	.300	.000[a]
Offenheit für Erfahrungen	-.068	-1.546	.062	.005[a]	-.054	-1.247	.107	.003[a]

Untersuchungsgruppe 1A, $N = 405$

β: standardisiertes Regressionsgewicht für den Moderatoreffekt, T: Teststatistik des Signifikanztests für β, p: Signifikanzniveau (einseitig), ΔR^2: Effektstärke (Änderung von R^2)

[a] Effekt in erwarteter Richtung, [b] Effekt in nicht erwarteter Richtung, p (einseitig) < .05 fett gedruckt, [c] nach Alpha-Adjustierung signifikant

Moderierte multiple Regressionen für die Retestreliabilität

Die Analysen für die Retestreliabilität waren parallel zu denen für die Split-Half-Reliabilität: Mittels moderierter multipler Regressionen wurde der Retest-Wert anhand des Ergebnisses bei der ersten Erhebung vorhergesagt. Erwartet wurden auch hier negative Beta-Gewichte für die Moderation. In Tabelle 16 werden die Ergebnisse (Untersuchungsgruppe 1D) aufgeführt: Variabilität bzw. ERS (gemessen mit dem HEXACO-PI-R) moderierten (auch alpha-adjustiert) die Retestreliabilität von fünf bzw. drei Skalen[29]. Bei Skalen, deren Retestreliabilität durch ERS moderiert wurde, wurde diese auch durch Variabilität moderiert.

[28] Dieser Effekt ist auch nach der Bonferroni-Holm-Korrektur (Holm, 1979) signifikant.
[29] Bonferroni-Holm-korrigiert (Holm, 1979) ist jeweils einer der Effekte signifikant.

Tabelle 16: Moderation der Retestreliabilität der Skalen des ITB-PESA durch Variabilität und durch ERS, gemessen mit dem HEXACO-PI-R

Moderation der Retestreliabilität der Skala	Moderator Variabilität				Moderator ERS			
		T-Test				T-Test		
	β	T	p	ΔR^2	β	T	p	ΔR^2
Soziale Kompetenz								
Kontaktfreude	. 002	0.036	.514	.000[b]	.036	0.705	.758	.001[b]
Kommunikationsvermögen	-.158	-2.218	**.015**[c]	.021[a]	-.164	-2.197	**.015**[c]	.021[a]
Geselligkeit	-.019	-0.302	.382	.000[a]	.040	0.621	.732	.001[b]
Einfühlungsvermögen	-.034	-0.376	.354	.001[a]	.019	0.202	.580	.000[b]
Konsensorientierung	.141	1.919	.971	.019[b]	.186	2.505	.993	.032[b]
Aufgeschlossenheit und Neugier	-.053	-0.608	.273	.002[a]	-.019	-0.209	.418	.000[a]
Führungskompetenz								
Leadership	-.213	-3.455	**<.001**[c]	.038[a]	-.227	-3.628	**<.001**[c]	.041[a]
Steuerungsvermögen	-.176	-1.963	**.026**[c]	.024[a]	-.143	-1.559	.061	.016[a]
Führungswille und Machtmotivation	-.021	-0.334	.370	.000[a]	.014	0.232	.591	.000[b]
Souveränität	.014	0.191	.575	.000[b]	.068	0.935	.824	.004[b]
Unternehmerische Kompetenz								
Ganzheitlich-strategische Denkweise	.089	1.259	.894	.007[b]	.090	1.294	.900	.008[b]
Kundenorientierung	-.247	-2.810	**.003**[c]	.053[a]	-.229	-2.516	**.007**[c]	.043[a]
Mut und Risikobereitschaft	-.009	-0.119	.453	.000[a]	.042	0.582	.719	.002[b]
Eigeninitiative	-.040	-0.468	.321	.001[a]	.001	0.008	.503	.000[b]
Markt- und Wettbewerbsorientierung	.102	1.348	.909	.010[b]	.103	1.357	.911	.010[b]
Ergebnisorientierung								
Arbeitsdisziplin	-.095	-1.559	.062	.009[a]	-.041	-0.663	.205	.002[a]
Ausdauer und Belastbarkeit	.055	0.725	.764	.003[b]	.098	1.297	.901	.009[b]
Sorgfalt	.065	1.013	.843	.003[b]	.062	1.007	.841	.003[b]
Erfolgszuversicht	-.130	-1.843	**.034**[c]	.014[a]	-.024	-0.336	.369	.001[a]
Leistungsstreben und Erfolgsmotivation	-.018	-0.237	.407	.000[a]	-.066	-0.866	.195	.004[a]
Integrität & Verlässlichkeit								
Ehrlichkeit	-.051	-0.593	.277	.002[a]	-.018	-0.211	.417	.000[a]
Regelbewusstsein	-.045	-0.513	.305	.002[a]	.054	0.617	.730	.003[b]

Untersuchungsgruppe 1D, $N = 93$

β: standardisiertes Regressionsgewicht für den Moderatoreffekt, T: Teststatistik des Signifikanztests für β, p: Signifikanzniveau (einseitig), ΔR^2: Effektstärke (Änderung von R^2)

[a] Effekt in erwarteter Richtung, [b] Effekt in nicht erwarteter Richtung, p (einseitig) < .05 fett gedruckt, [c] nach Alpha-Adjustierung signifikant

In Fällen, in denen sich ein Moderatoreffekt für Variabilität und ERS zeigt, wurde geprüft, ob ERS (Variabilität) für diesen Fragebogen den Zusammenhang über die Variabilität (ERS) hinaus moderiert. Dazu wurde die Regression neu berechnet: Im ersten Schritt wurden – jeweils z-standardisiert – das Ergebnis der Prä-Messung, der Variabilitäts-Faktor (ERS-Faktor) des Fragebogens, das Produkt aus dem Ergebnis der Prä-Messung und dem Variabilitäts-Faktor

(ERS-Faktor) sowie der ERS-Faktor (Variabilitäts-Faktor) eingegeben. Im zweiten Schritt wurde dann das Produkt aus dem Ergebnis der Prä-Messung und dem ERS-Faktor (Variabilitäts-Faktor) hinzugefügt. Führte der zweite Schritt zu einer Verbesserung der Vorhersage, so moderierte der ERS-Faktor (Variabilitäts-Faktor) die Retestreliabilität über den Variabilitäts-Faktor (ERS-Faktor) hinaus.

In keinem der drei Fälle moderierte ERS die Retestreliabilität über Variabilität hinaus („Kommunikationsvermögen“: $\beta = -.118,\ T = -0.411,\ p = .341,\ \Delta R^2 = .001$; „Leadership“: $\beta = -.249, T = -1.116, p = .134, \Delta R^2 = .004$; „Kundenorientierung“: $\beta = .299,\ T = 0.787,\ p = .783,\ \Delta R^2 = .004$). Ebenso moderierte Variabilität nicht über ERS hinaus („Kommunikationsvermögen“: $\beta = -.034,\ T = -0.119,\ p = .483,\ \Delta R^2 = .000$; „Leadership“: $\beta = .032,\ T = 0.148,\ p = .558,\ \Delta R^2 = .000$; „Kundenorientierung“: $\beta = -.547,\ T = -1.457, p = .074, \Delta R^2 = .014$).

5.2.5 Der Einfluss von Variabilität und ERS auf die Konstruktvalidität und auf die Kriteriumsvalidität

Im Folgenden wird beschrieben, wie die Moderation der Konstruktvalidität und Kriteriumsvalidität geprüft wurde. Dargestellt sind jeweils Operationalisierungen und Analysen.

Moderation der Konstruktvalidität

Die Konstruktvalidität wurde als Zusammenhang zwischen den Faktor-Skalen des HEXACO-PI-R und jeweils dazu passenden Skalen des ITB-PESA operationalisiert: Von den ITB-PESA-Skalen wurde aus mehreren Markier-Skalen (Beermann & Heilmann, 2014) für jede HEXACO-Dimensionen jeweils die ausgewählt, die am höchsten mit der Faktor-Skala des HEXACO-PI-R korreliert. Da die HEXACO-Dimensionen auf einer globaleren Abstraktionsebene angesiedelt sind, wurden moderierte multiple Regressionen für die Vorhersage der Skalen des ITB-PESA auf Basis der HEXACO-Dimensionen gerechnet: Die HEXACO-Dimension Ehrlichkeit-Bescheidenheit war Prädiktor für die ITB-PESA-Skala „Ehrlichkeit“ ($r_{(405)} = .518, p < .001$), die Dimension Emotionalität für die Skala „Einfühlungsvermögen“ ($r_{(405)} = .641,\ p < .001$), Extraversion für „Kontaktfreude“ ($r_{(405)} = .723, p < .001$), Verträglichkeit versus Ärger für „Konsensorientierung“ ($r_{(405)} = .480, p < .001$), Gewissenhaftigkeit für „Regelbewusstsein“ ($r_{(405)} = .640, p < .001$) und Offenheit für Erfahrungen für „Ganzheitlich-strategische Denk-

weise“ ($r_{(405)} = .655$, $p < .001$). Die Moderatoren, Variabilität bzw. ERS, wurden mit dem ITB-PESA erfasst, da der Prädiktor jeweils mit dem HEXACO-PI-R erfasst wurde.

Die Regressionsanalysen wurden wie bei den Analysen auf Moderation der Split-Half-Reliabilität und der Retestreliabilität (Abschnitt 5.2.4) durchgeführt. Im ersten Schritt wurden jeweils der Prädiktor und der Moderator eingegeben, im zweiten Schritt das Produkt aus Prädiktor mal Moderator (jeweils z-standardisiert). Auch hier werden aus Gründen der Übersichtlichkeit nur die standardisierten Regressionsgewichte für den Moderatoreffekt (Produkt aus Prädiktor mal Moderator) sowie entsprechende einseitige T-Tests, Signifikanzniveaus und Effektstärken präsentiert. Die Ergebnisse sind in Tabelle 17 aufgeführt: Für keines der untersuchten Skalenpaare zeigte sich ein Moderatoreffekt – weder Variabilität noch ERS moderierte die Konstruktvalidität signifikant.

Tabelle 17: Moderation der konvergenten Konstruktvalidität der Skalen des HEXACO-PI-R und der jeweils passenden Markier-Skala des ITB-PESA (Beermann & Heilmann, 2014) durch Variabilität und durch ERS

	Moderator Variabilität				Moderator ERS			
Moderation der konvergenten Konstruktvalidität im Bereich		T-Test				T-Test		
	β	T	p	ΔR^2	β	T	p	ΔR^2
Ehrlichkeit-Bescheidenheit	.008	0.190	.575	.000[b]	-.012	-0.284	.388	.000[a]
Emotionalität	.035	0.887	.812	.001[b]	.035	0.858	.804	.001[b]
Extraversion	.041	1.179	.880	.002[b]	.111	3.149	.999	.012[b]
Verträglichkeit versus Ärger	-.021	-0.458	.324	.000[a]	-.000	-0.003	.499	.000[a]
Gewissenhaftigkeit	.025	0.655	.743	.001[b]	.099	2.483	.993	.009[b]
Offenheit für Erfahrungen	-.029	-0.776	.219	.001[a]	-.026	-0.723	.235	.001[a]

Untersuchungsgruppe 1D, $N = 93$

Prädiktor ist jeweils die Faktor-Skala im HEXACO-PI-R, abhängige Variable die im Text aufgeführte zugehörige Markier-Skala für die jeweilige HEXACO-Dimension. Variabilität und ERS wurden mit dem ITB-PESA gemessen.

β: standardisiertes Regressionsgewicht für den Moderatoreffekt, T: Teststatistik des Signifikanztests für β, p: Signifikanzniveau (einseitig), ΔR^2: Effektstärke (Änderung von R^2)

[a] Effekt in erwarteter Richtung, [b] Effekt in nicht erwarteter Richtung

Moderation der Kriteriumsvalidität

Bei der Untersuchung auf Moderation der Kriteriumsvalidität wurden wie in Abschnitt 5.1.1 beschrieben die Zusammenhänge zwischen der ITB-PESA-Skala „Erfolgszuversicht“ und Arbeitszufriedenheit sowie zwischen der ITB-PESA-Skala „Leistungsstreben und Erfolgszuversicht“ und der Note im Hochschulabschluss betrachtet. Die Moderatoren, Variabilität und ERS, wurden auf dem HEXACO-PI-R erfasst, so dass eine Abhängigkeit zwischen Prädiktoren

und Moderatoren ausgeschlossen werden konnte. Erwartet wurden jeweils besserer Vorhersagen für Personen mit niedriger Variabilität und niedrigem ERS.

Die moderierten multiplen Regressionen für Arbeitszufriedenheit sind in Tabelle 18 aufgeführt (Untersuchungsgruppe 1B). In der oberen Hälfte findet sich die Regression für den Moderator Variabilität, in der unteren für den Moderator ERS. Die Ergebnisse zeigen, dass „Erfolgszuversicht“ Arbeitszufriedenheit vorhersagt und dass diese Vorhersage signifikant von Variabilität sowie von ERS moderiert wird. In Abbildung 15 und Abbildung 16 werden die Moderatoreffekte illustriert: Je höher die Variabilität, desto schlechter wird Arbeitszufriedenheit durch die Skala „Erfolgszuversicht“ vorhergesagt. Die Vorhersage ist auch schlechter, je stärker die Tendenz zu extremen Antworten war.

Tabelle 18: Moderierte multiple Regressionen zur Vorhersage von Arbeitszufriedenheit mit dem Prädiktor „Erfolgszuversicht“ aus dem ITB-PESA und dem Moderator Variabilität bzw. ERS aus dem HEXACO-PI-R

Regressionen zur Vorhersage von Arbeitszufriedenheit				T-Test			
	B	s_E	β	T	p	R^2	ΔR^2
Schritt 1						.200	
Erfolgszuversicht	1.295	0.133	.444	9.728	**<.001**[a]		
Variabilität im HEXACO-PI-R	0.065	0.136	.022	0.479	.633[b]		
Schritt 2						.207	.007
Erfolgszuversicht (Ez)	1.338	0.135	.459	9.925	**<.001**[a]		
Variabilität im HEXACO-PI-R (V)	0.043	0.136	.015	0.319	.750[b]		
Ez x V	-0.251	0.138	-.084	-1.821	**.035**[a]		
Schritt 1						.201	
Erfolgszuversicht	1.278	0.136	.438	9.391	**<.001**[a]		
ERS im HEXACO-PI-R	0.102	0.138	.035	0.743	.458[b]		
Schritt 2						.220	.019
Erfolgszuversicht (Ez)	1.359	0.137	.466	9.905	**<.001**[a]		
ERS im HEXACO-PI-R (ERS)	0.097	0.136	.033	0.711	.478[b]		
Ez x ERS	-0.394	0.128	-.140	-3.069	**.001**[a]		

Untersuchungsgruppe 1B, $N = 394$

Die Prädiktoren wurden vor der Analyse z-standardisiert, das Produkt aus Prädiktor und Moderator wurde aus den z-standardisierten Werten berechnet.

B: Regressionsgewicht, s_E: Standardfehler des Regressionsgewichts, β: standardisiertes Regressionsgewicht; Signifikanztests für β (T-Test): Teststatistik T, Signifikanzniveau p (einseitig); R^2: Effektstärke (Determinationskoeffizient), ΔR^2: Änderung des Determinationskoeffizienten

p < .05 fett gedruckt, [a] einseitige Testung (für Effekt in erwarteter Richtung), [b] zweiseitige Testung

Da beide Moderatoren hoch korrelierten, wurde analog zu den Analysen auf Moderation der Retestreliabilität getestet, inwieweit ein Moderator über den jeweils anderen hinaus moderiert. In einer moderierten multiplen Regression wurden im ersten Schritt „Erfolgszuversicht“, die beiden Moderatoren und das Produkt des einen Moderators mit dem Prädiktor eingegeben. Im zweiten Schritt wurde das Produkt des anderen Moderators mit dem Prädiktor hinzugefügt und der zusätzliche Moderatoreffekt bestimmt. Während ERS über Variabilität hinaus moderierte ($\beta = -.287$, $T = -3.035$, $p = .001$, $\Delta R^2 = .018$), zeigte sich im umgekehrten Fall kein Effekt – Variabilität moderierte nicht über ERS hinaus ($\beta = .165$, $T = 1.749$, $p = .958$, $\Delta R^2 = .006$).

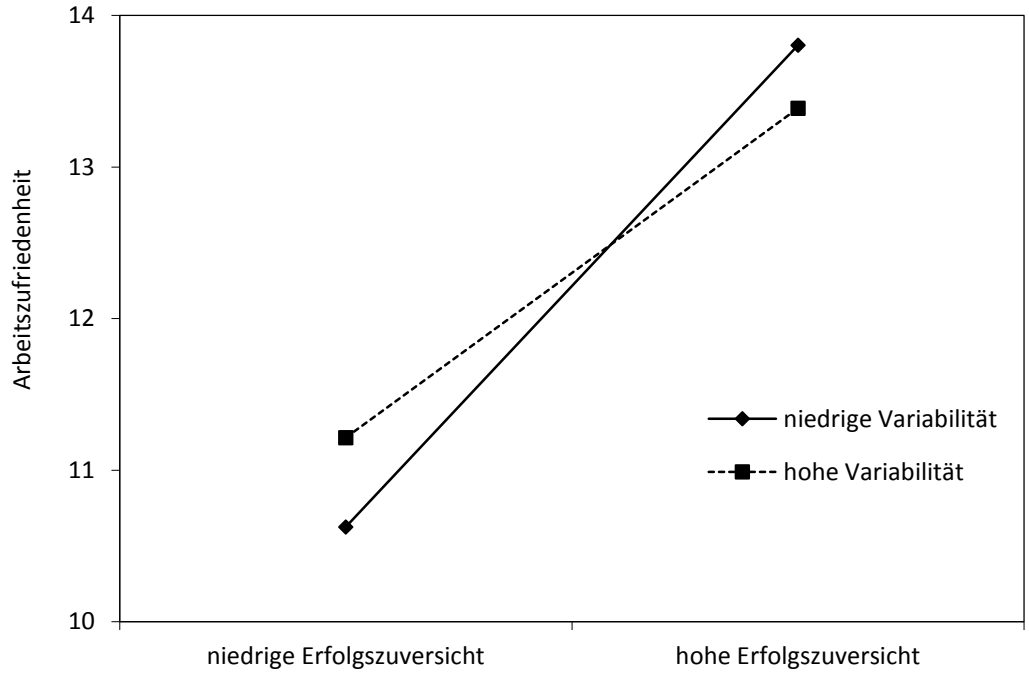

Abbildung 15: Moderation der Vorhersage von Arbeitszufriedenheit durch „Erfolgszuversicht“ (ITB-PESA) durch den Moderator Variabilität

Untersuchungsgruppe 1B, $N = 394$

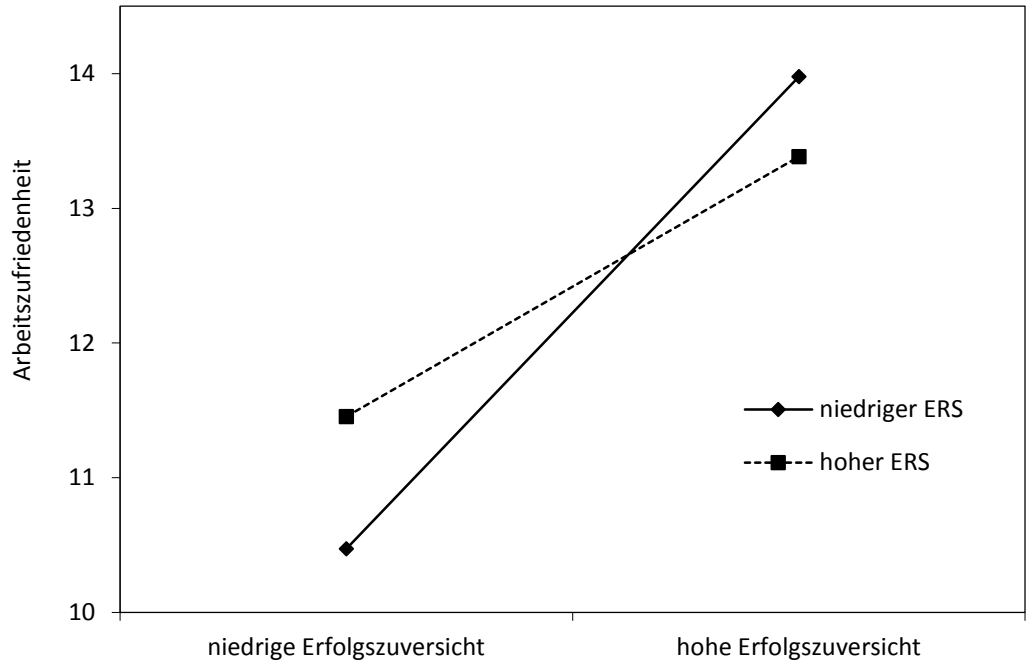

Abbildung 16: Moderation der Vorhersage von Arbeitszufriedenheit durch „Erfolgszuversicht“ (ITB-PESA) durch den Moderator ERS

Untersuchungsgruppe 1B, $N = 394$

Für die Vorhersage der Note im Hochschulabschluss wurden die gleichen Analysen mit Untersuchungsgruppe 1C durchgeführt. Die Ergebnisse der moderierten multiplen Regressionen werden in Tabelle 19 berichtet sowie in Abbildung 17 und in Abbildung 18 veranschaulicht. Deutlich wird, dass die Note im Hochschulabschluss durch die Skala „Leistungsstreben und Erfolgsmotivation“ vorhergesagt wird. Die Vorhersage wird von Variabilität moderiert, ERS moderiert marginal signifikant. Entsprechend moderiert Variabilität den Zusammenhang auch über ERS hinaus ($\beta = .364$, $T = 1.946$, $p = .027$, $\Delta R^2 = .025$), ERS moderiert nicht über Variabilität hinaus ($\beta = -.192$, $T = -1.028$, $p = .847$, $\Delta R^2 = .007$).

Tabelle 19: Moderierte multiple Regressionen zur Vorhersage der Note im Hochschulabschluss mit dem Prädiktor „Leistungsstreben und Erfolgsmotivation“ aus dem ITB-PESA und dem Moderator Variabilität bzw. ERS aus dem HEXACO-PI-R

Regressionen zur Vorhersage der Note im Hochschulabschluss				T-Test			
	B	s_E	β	T	p	R^2	ΔR^2
Schritt 1						.044	
Leistungsstreben und Erfolgsmotivation	-0.109	0.049	-.184	-2.221	**.014**[a]		
Variabilität im HEXACO-PI-R	0.068	0.047	.121	1.457	.147[b]		
Schritt 2						.078	.034
Leistungsstreben und Erfolgsmotivation (LE)	-0.132	0.049	-.222	-2.665	**.004**[a]		
Variabilität im HEXACO-PI-R (V)	0.057	0.047	.100	1.215	.226[b]		
LE x V	0.127	0.056	.191	2.282	**.012**[a]		
Schritt 1						.038	
Leistungsstreben und Erfolgsmotivation	-0.111	0.050	-.187	-2.232	**.014**[a]		
ERS im HEXACO-PI-R	0.053	0.049	.076	1.087	.279[b]		
Schritt 2						.055	.018
Leistungsstreben und Erfolgsmotivation (LE)	-0.126	0.050	-.212	-2.505	**.007**[a]		
ERS im HEXACO-PI-R (ERS)	0.044	0.049	.076	0.902	.368[b]		
LE x ERS	0.088	0.054	.137	1.624	.053[a]		

Untersuchungsgruppe 1C, $N = 144$

Die Prädiktoren wurden vor der Analyse z-standardisiert, das Produkt aus Prädiktor und Moderator wurde aus den z-standardisierten Werten berechnet.

B: Regressionsgewicht, s_E: Standardfehler des Regressionsgewichts, β: standardisiertes Regressionsgewicht; Signifikanztests für β (T-Test): Teststatistik T, Signifikanzniveau p (einseitig); R^2: Effektstärke (Determinationskoeffizient), ΔR^2: Änderung des Determinationskoeffizienten

p < .05 fett gedruckt, [a] einseitige Testung (für Effekt in erwarteter Richtung), [b] zweiseitige Testung

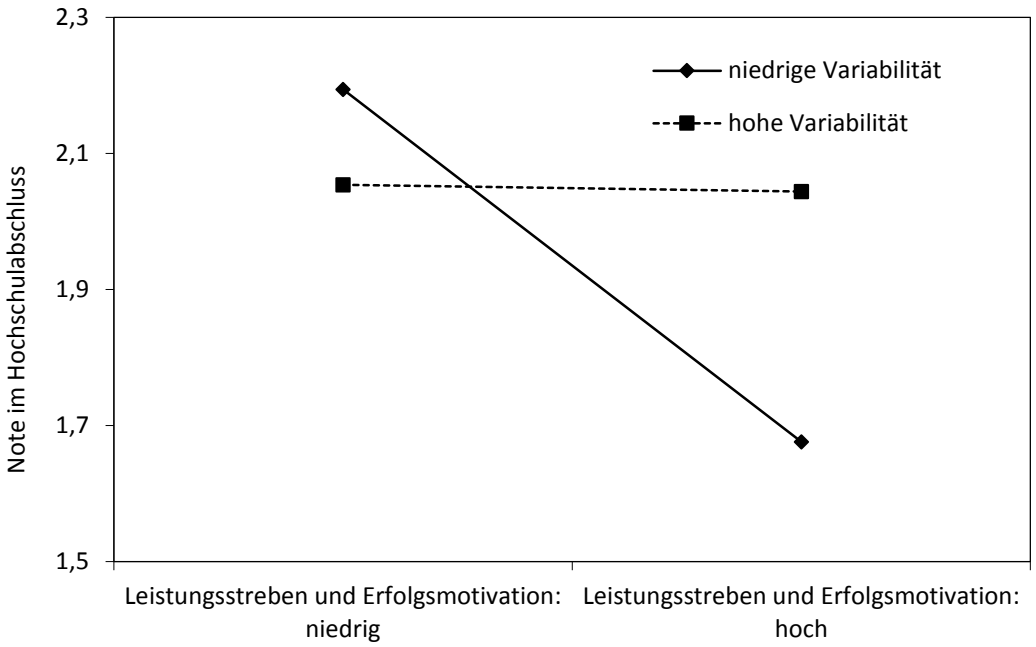

Abbildung 17: Moderation der Vorhersage der Note im Hochschulabschluss durch „Leistungsstreben und Erfolgsmotivation“ (ITB-PESA) durch den Moderator Variabilität

Untersuchungsgruppe 1C, $N = 144$

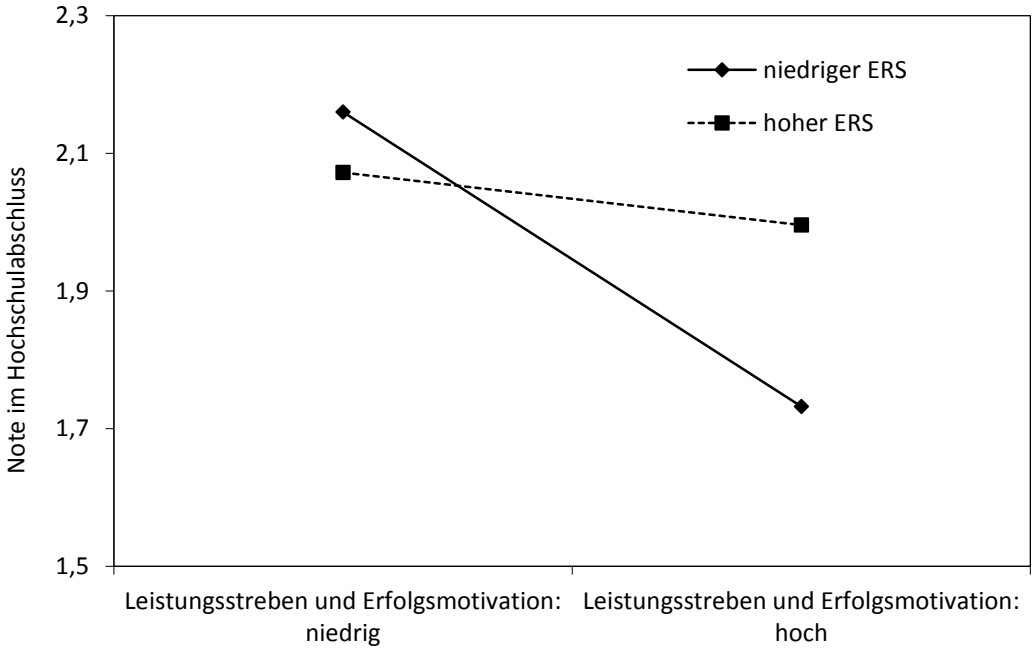

Abbildung 18: Moderation der Vorhersage der Note im Hochschulabschluss durch „Leistungsstreben und Erfolgsmotivation“ (ITB-PESA) durch den Moderator ERS

Untersuchungsgruppe 1C, $N = 144$

5.3 Diskussion

Basierend auf einer Erhebungen eines Facetten-Fragebogens, des ITB-PESA, sowie eines Dimensions-Fragebogens, des HEXACO-PI-R, und der Kriterien *Arbeitszufriedenheit* und *Note im Hochschulabschluss* sowie einer Retest-Erhebung des ITB-PESA wurden Hypothesen in drei Bereichen getestet: Erstens wurden die Erfassung und Struktur von Variabilität untersucht, zweitens der Zusammenhang zwischen Variabilität und ERS und drittens der Einfluss von Variabilität und ERS auf die Reliabilität und Validität der Persönlichkeitsfragebogen.

Als Variabilitätsindizes wurden die Inter-Item-SD der Skalen des ITB-PESA und des HEXACO-PI-R berechnet und um die Einflüsse der Item-Mittelwerte und der Abweichung dieser vom jeweiligen Gruppen-Mittelwert korrigiert. Da das ITB-PESA miteinander in Beziehung stehende Eigenschaften misst, konnte überprüft werden, ob bei der Korrektur inhaltlich relevante Varianz aus den Inter-Item-SD eliminiert wird oder inhaltlich irrelevante und methodisch begründete. Die Analyse ergab, dass die Korrelationen zwischen den Inter-Item-SD und den Item-Mittelwerten bzw. den Abweichungen der Item-Mittelwerte vom Gruppen-Mittelwert auf methodische Restriktionen und auf die Beschaffenheit der Messungen zurückzuführen sind. Entsprechend bilden die korrigierten Inter-Item-SD Variabilität valider ab als die *nicht* korrigierten Inter-Item-SD. Als globales Maß für Variabilität wurde für beide Fragebogen ein Faktor der korrigierten Inter-Item-SD gebildet. Dieser Faktor ist jeweils reliabel und klärt einen großen Teil der Varianz der korrigierten Inter-Item-SD auf. Die Faktoren für beide Fragebogen korrelierten so hoch, dass davon auszugehen ist, dass dasselbe Merkmal erfasst wurde. Für den Variabilitäts-Faktor im ITB-PESA konnte eine hohe Retestreliabilität nachgewiesen werden. Untersucht wurde auch, ob die korrigierten Inter-Item-SD neben der gemeinsamen Varianz weitere bedeutsame Varianz aufweisen, d. h. ob sich inhaltlich relevante Facetten beobachten lassen. Mit einer Analyse der Korrelationen zwischen den korrigierten Inter-Item-SD sowie mit SEM konnte kein Hinweis auf eine Facettenstruktur gefunden werden: Varianz der korrigierten Inter-Item-SD, die sich nicht dem gemeinsamen Faktor zuordnen lässt, hängt nicht vom mit der Skala gemessenen Merkmal ab. Variabilität ist also eine globale und eindimensionale Eigenschaft, die sich reliabel und valide mit einem Faktor der korrigierten Inter-Item-SD erfassen lässt.

ERS wurde ebenfalls global erfasst, und zwar jeweils als Faktor der Extremwerthäufigkeiten der Skalen eines Fragebogens. Dieser Faktor klärte in beiden Fragebogen jeweils einen substanziellen Anteil der Varianz der Extremwerthäufigkeiten auf und hatte eine hohe Konsistenz. Für den ERS-Faktor im ITB-PESA wurde auch eine hohe Retestreliabilität festgestellt. Eine hohe Korrelation zwischen den Faktoren für ITB-PESA und HEXACO-PI-R sowie hohe Korrelation mit drei Greenleaf-Skalen und mit den Extremwerthäufigkeiten im gesamten ITB-PESA und im gesamten HEXACO-PI-R deuteten darauf hin, dass ERS ebenfalls valide erfasst wurde. Der Zusammenhang zwischen ERS und Variabilität wurde zunächst anhand von Korrelationen exploriert: Die Variabilitäts- und die ERS-Faktoren korrelierten fragebogenunabhängig sehr hoch, so dass vermutet wurde, beiden Faktoren liege dieselbe Eigenschaft zugrunde. Dies wurde mittels SEM getestet: Für das ITB-PESA hat ein Modell, das zwei Faktoren, einen für Variabilität und einen für ERS, beinhaltet, einen besseren Fit als ein Modell mit einem einzigen Faktor, der den korrigierten Inter-Item-SD und den Extremwerthäufigkeiten aller Skalen zugrunde liegt. Allerdings korrelieren die beiden Faktoren im Zwei-Faktoren-Modell sehr hoch miteinander. Für das HEXACO-PI-R hatte das Ein-Faktoren-Modell einen mindestens ebenso guten Fit wie das Zwei-Faktoren-Modell, so dass von einem Faktor ausgegangen werden muss. Eine mögliche Erklärung für diese unterschiedlichen Ergebnisse sind die Skaleninterkorrelationen der beiden Fragebogen: Diese sind beim ITB-PESA moderat bis hoch (mit einer hohen Bandbreite; siehe Anhang B.5, Tabelle B.5 - 1), während die Skalen des HEXACO-PI-R weitgehend unkorreliert sind (siehe Anhang B.3, Tabelle B.3 - 2). Im ITB-PESA haben die Skalen entsprechend einen gemeinsamen Messbereich – es ließe sich ein Faktor extrahieren, auf dem ein großer Teil der Skalen lädt. Weil Personen mit hoher oder niedriger Ausprägung auf den Skalen auch häufiger extrem antworten als Personen mit mittlerer Ausprägung und die Gruppen-Mittelwerte der Skalen im ITB-PESA in der Regel oberhalb der Mitte der Likert Skala liegen, hängt extremes Antworten wahrscheinlich nicht nur von Variabilität, sondern auch von dem gemeinsamen Merkmalsbereich der Skalen des ITB-PESA, d. h. vom gemeinsamen Faktor der ITB-PESA-Skalen, ab. Dieser Faktor umfasst vorwiegend die Merkmalsbereiche Extraversion und Gewissenhaftigkeit, denn die Skalen des ITB-PESA zielen überwiegend auf diese (für das Berufsleben relevanten) Eigenschaften ab.

Dass die unterschiedlichen Modellfits dadurch zustande kommen, dass der ERS-Faktor im ITB-PESA nicht ausschließlich von Variabilität abhängt, wurde in weiteren SEM offensichtlich:

Ein Ein-Faktoren-Modell, dessen Faktor die korrigierten Inter-Item-SD der Skalen des ITB-PESA und die Extremwerthäufigkeiten der Skalen des HEXACO-PI-R indiziert, zeigte einen ähnlich guten Fit wie ein Zwei-Faktoren-Modell mit einem Variabilitäts- und einem ERS-Faktor. Wenn jedoch Variabilität von den korrigierten Inter-Item-SD *der Skalen des HEXACO-PI-R* und ERS von den Extremwerthäufigkeiten *der Skalen des ITB-PESA* repräsentiert wurde, war das Zwei-Faktoren-Modell dem Ein-Faktoren-Modell deutlich überlegen.

Die Vermutung, dass ERS auch von einem gewichtigen bzw. dominanten Faktor der Skalen eines Fragebogens abhängt, wird in Studie 2 überprüft: Dort werden ein Auswahl- und ein Nicht-Auswahl-Datensatz desselben Fragebogens miteinander verglichen; da der erste Faktor in einem Fragebogen bei der Auswahl üblicherweise stärker ausfällt als in einem Nicht-Auswahl-Setting (siehe Abschnitt 4.4.3; Collins & Gleaves, 1998; Kanning & Holling, 2001; vgl. Marcus, 2003), sollte ERS bei der Auswahl auch stärker mit dem gemeinsamen Faktor des Fragebogens korrelieren.

Hinsichtlich der Effekte von Variabilität und ERS auf die Reliabilität entsprechen die Ergebnisse nur zum Teil den Erwartungen: Die Split-Half-Reliabilität für das Profil des HEXACO-PI-R hängt von Variabilität ab, jedoch nicht die für das Profil des ITB-PESA; im HEXACO-PI-R ist die Übereinstimmung zwischen dem Profil, gemessen mit der einen Hälfte der Items, mit dem mit der anderen Hälfte der Items gemessenen Profil größer, je niedriger die Variabilität ist. ERS hat weder einen Effekt auf die Profil-Reliabilität des ITB-PESA noch auf die Profil-Reliabilität des HEXACO-PI-R. Auf Skalenebene zeigen sich nur vereinzelt erwartungskonforme Effekte: Die Split-Half-Reliabilität je einer Skala des ITB-PESA hängt von Variabilität bzw. ERS ab. Im HEXACO-PI-R wird die Konvergenz von Skalenhälften (=Split-Half-Übereinstimmung) nur bei einer Skala von Variabilität moderiert. ERS hat keinen Einfluss auf die Split-Half-Reliabilität der Skalen des HEXACO-PI-R. Für die Retestreliabilität zeigen sich ähnliche Ergebnisse: Die Übereinstimmung der Profile von Prä- und Post-Messung im ITB-PESA hängt weder mit Variabilität noch mit ERS zusammen, und die Retestreliabilität eines nur kleinen Teils der Skalen hängt signifikant von Variabilität (5 der 22 Skalen) bzw. ERS (3 der 22 Skalen) ab. Bemerkenswert ist, dass die Retestreliabilität in Fällen, in denen sie von ERS abhängt, auch von Variabilität abhängt. Eine weitere Analyse zeigte, dass die Effekte sogar deckungsgleich sind. Das bedeutet der Effekt von ERS auf die Retestreliabilität lässt sich auf den Effekt der Variabilität auf die Retestreliabilität zurückführen. Umgekehrt gilt dies nur für drei der

Skalen, deren Retestreliabilität von Variabilität moderiert wird. Vermutlich ist die Eigenschaft „Variabilität“ verantwortlich für die Moderatoreffekte und wahrscheinlich wird diese Eigenschaft besser mit dem Variabilität-Faktor erfasst als mit dem ERS-Faktor. Zusammenfassend lässt sich festhalten, dass die Split-Half-Reliabilität und die Retestreliabilität nicht oder nur sehr gering von Variabilität (und ERS) abhängen.

Ein Einfluss von Variabilität auf die Konstruktvalidität zeigte sich nicht: Der Zusammenhang der Skalen des HEXACO-PI-R mit jeweils einer HEXACO-Markier-Skala aus dem ITB-PESA wurde weder von Variabilität noch von ERS moderiert. Dagegen zeigten sich bei der Kriteriumsvalidität klare Effekte: Variabilität moderiert den Zusammenhang zwischen der ITB-PESA-Skala „Erfolgszuversicht“ und der Arbeitszufriedenheit sowie den Zusammenhang zwischen der ITB-PESA-Skala „Leistungsstreben und Erfolgsmotivation“ und der Note im Hochschulabschluss. Unter Personen mit niedrigerer Variabilität waren jeweils stärkere Zusammenhänge zu beobachten als unter Personen mit hoher Variabilität. Für ERS zeigte sich der Effekt nur im ersten Fall: Je höher ERS ausgeprägt ist, desto schlechter kann Arbeitszufriedenheit anhand der Skala „Erfolgszuversicht“ vorhergesagt werden. Der Einfluss von ERS auf die Vorhersage der Note im Hochschulabschluss durch die Skala „Leistungsstreben und Erfolgsmotivation“ war knapp nicht signifikant. Dennoch lässt sich für den Einfluss von Variabilität und ERS auf die Validität insgesamt ein – im Sinne der Hypothesen – positives Fazit ziehen: Variabilität und ERS haben keinen Einfluss auf die Konstruktvalidität, aber auf die Vorhersage von Kriterien durch Persönlichkeitsmaße. Das Ergebnis ist nicht auf Unterschiede in der Messgenauigkeit zwischen Personen mit hoher Variabilität und Personen mit niedriger Variabilität zurückzuführen.

6 Studie 2

Grundlage für die zweite Studie war der Ernstfalleinsatz einer Vertriebsversion des ITB-PESA zur Personalauswahl bei einem Versicherungsunternehmen. Mit dieser Version wurden nur einige der Eigenschaften gemessen, die mit der in Studie 1 eingesetzten Version erfasst werden. Da alle Items der Vertriebsversion auch in der ersten Untersuchung von Studie 1 enthalten waren, können die entsprechenden Daten von Studie 1 verwendet und die Ergebnisse für den Ernstfalleinsatz (für Studie 2 neu gewonnene Daten) mit denen aus dem Nicht-Auswahl-Setting (Daten aus der ersten Untersuchung in Studie 1) verglichen werden. In Abschnitt 6.1 werden die Methoden vorgestellt und die Gütekennwerte der verwendeten Version berichtet, in Abschnitt 6.2 werden die Analysen im Hinblick auf die Hypothesen beschrieben und die entsprechenden Ergebnisse berichtet und in Abschnitt 6.3 werden die Befunde zusammengefasst.

6.1 Methode

Der Einsatz des ITB-PESA zur Personalauswahl fand an verschiedenen Orten in den Räumen des Versicherungsunternehmens statt. Gleichzeitig bearbeiteten zwischen einer und acht Personen zunächst einen Fragebogen zum sozialen Umfeld, der speziell für diesen Anwendungsfall entwickelt wurde, und anschließend die 84 Items umfassende, vertriebsspezifische Version des ITB-PESA. Die Bearbeitung erfolgte online im Testsystem *iona* (ITB Consulting GmbH, 2011). Im Fragebogen zum sozialen Umfeld wurde unter anderem nach der Anzahl zur letzten Geburtstagsfeier eingeladener Gäste gefragt, die Antwort dient in dieser Studie als Kriterium für die Skala „Kontaktfreude“ des ITB-PESA. Im Folgenden werden Informationen zur Stichprobe aufgeführt (Abschnitt 6.1.1) sowie Instrumente und Messungen (Abschnitt 6.1.2) näher beschrieben. Ergänzungen zu den Messungen finden sich in Anhang E.

Die Ergebnisse für den Ernstfalleinsatz werden verglichen mit den Ergebnissen für eine identische Fragebogenversion im Nicht-Auswahl-Setting. Dafür werden die Daten aus der ersten in Studie 1 berichteten Untersuchung verwendet (vgl. Abschnitt 5.1). Sie beziehen sich auf Untersuchungsgruppe 1A (vgl. Abschnitt 5.1.2).

6.1.1 Beschreibung der Stichprobe

Die Teilnehmer haben sich bei dem Versicherungsunternehmen für eine Stelle als „Tippgeber“ beworben. Aufgabe von Tippgebern war das Herstellen von Kontakten zwischen möglichen Versicherungsnehmern und Versicherungsvermittlern. Tippgeber wurden als freie Mitarbeiter angestellt und erhielten für einen bestimmten Zeitraum ein fixes Gehalt; abhängig von der Zahl und dem Volumen der abgeschlossenen Versicherungsverträge erhielten sie darüber hinaus einen kleinen Anteil an variabler Vergütung. Eine Weiterbeschäftigung über den Zeitraum hinaus, in dem das Fixgehalt bezahlt wurde, sowie die Konditionen dieser Weiterbeschäftigung waren ebenfalls abhängig vom Erfolg im Anfangszeitraum. Die Bewerberakquise erfolgte vorwiegend über Zeitungsannoncen des Versicherungsunternehmens. Mit der Bearbeitung des ITB-PESA haben 401 Personen begonnen, 367 davon bearbeiteten das ITB-PESA vollständig (Untersuchungsgruppe 2A), darunter 203 Männer (55.3 %) und 164 Frauen (44.7 %). Zum Zeitpunkt der Bearbeitung waren 23 Personen unter 20 Jahren (6.3 %), 94 zwischen 20 und 24 Jahren (25.6 %), 60 zwischen 25 und 29 Jahren (16.3 %), 70 zwischen 30 und 39 Jahren (19.1 %), 53 zwischen 40 und 49 Jahren (14.4 %) und 67 älter als 49 Jahre (18.3 %). Weitere Angaben zu demografischen Merkmalen waren aufgrund des Settings (Personalauswahl bei einem Auftraggeber der ITB Consulting) nicht zugänglich.

Angaben zum Kriterium „Anzahl zur letzten Geburtstagsfeier eingeladener Gäste“ lagen für 343 Personen vor (Untersuchungsgruppe 2B). Von den 187 Männern (54.5 %) und 156 Frauen (45.5 %) waren 22 jünger als 20 Jahre (6.4 %), 86 zwischen 20 und 24 Jahren (25.1 %), 58 zwischen 25 und 29 Jahren (16.9 %), 66 zwischen 30 und 39 Jahren (19.2 %), 48 zwischen 40 und 49 Jahren (14.0 %) und 63 älter als 49 Jahre (18.4 %). Die Messung des Kriteriums wird im folgenden Abschnitt (6.1.2) beschrieben.

6.1.2 Instrumente und Messungen

Im Folgenden werden die Erfassung des Kriteriums beschrieben, die eingesetzte Version des ITB-PESA vorstellt und die Analysen zur Messung von Variabilität und ERS skizziert.

Messung und Analyse des Kriteriums

Der Fragebogen zum sozialen Umfeld enthielt neun Fragen zu Hobbys, Bekanntenkreis, Familie und Alltag der Bearbeitenden. Berichtet wird ausschließlich das für die vorliegende Arbeit verwendete Kriterium, die Anzahl zur letzten Geburtstagsfeier eingeladener Gäste, die mit der Skala „Kontaktfreude“ des ITB-PESA vorhergesagt werden sollte. Erfasst wurde das Kriterium mit der zweiten Frage im Fragebogen zum sozialen Umfeld: „Wie viele Personen haben Sie zu Ihrer letzten Geburtstagsfeier eingeladen?“. Das Antwortformat war frei. Gewertet wurden alle Antworten, die als Zahl oder Zahlenspanne definiert waren. Bei Zahlenspannen (z. B. „25-30“) wurde der Mittelwert (im Beispiel 27.5) verwendet. Adverbien wie „ca.“ oder „etwa“ wurden ignoriert. Im Mittel luden die Teilnehmer 17.24 Personen zu ihrer letzten Geburtstagsfeier ein, die Standardabweichung war größer als der Mittelwert ($SD = 19.12$), der Median lag deutlich unter dem Mittelwert ($Med = 12.00$) und die Normalverteilungsannahme musste verworfen werden (K-S-Test: $Z = 4.041$, $p < .001$). Wie diese drei Werte vermuten lassen, war die Anzahl zur letzten Geburtstagsfeier eingeladener Gäste schief verteilt. Diese Vermutung wird durch das Histogramm (Anhang E.1, Abbildung E.1 - 1) bestätigt, also wurde der Logarithmus der Werte (zur Basis e) überprüft. Die logarithmierten Werte konnten für 24 Personen der Untersuchungsgruppe 2B nicht berechnet werden, da sie die Antwort „0“ gaben. Unter den übrigen Personen war die Verteilung nicht schief (Histogramm in Anhang E.1, Abbildung E.1 - 2; $M = 2.57$, $SD = 0.83$, $Med = 2.71$), eine Normalverteilung konnte jedoch nicht angenommen werden (K-S-Test: $Z = 1.424$, $p = .035$). Zwar bieten die logarithmierten Werte den Vorteil einer gleichmäßigeren Verteilung, dennoch wurden die nicht transformierten Daten für die Analysen verwendet; denn die logarithmierten Werte waren auch nicht normalverteilt und basierten auf weniger Datensätzen; ein Weglassen der Personen, die „0“ geantwortet hatten, wäre nicht gerechtfertigt gewesen (der Modalwert lag bei 20). Neben der Verteilung wurde auch die Abhängigkeit des Kriteriums vom Alter der Personen bestimmt und davon, ob es sich um einen besonderen Geburtstag handelt (18. Geburtstag oder runder Geburtstag). Das Alter hatte keinen signifikanten Einfluss auf die Anzahl der zur letzten Geburtstagsfeier eingeladenen Gäste ($r_{(343)} = -.090$, $p = .097$, zweiseitig) und es wurden nicht mehr Gäste eingeladen, wenn es sich um einen besonderen Geburtstag handelte ($T_{(43.28)} = 1.375$, $p = .088$, einseitig, $d = 0.397$). Allerdings war die Anzahl zur letzten Geburtstagsfeier eingeladener Gäste bei

besonderen Geburtstagen variabler als bei nicht besonderen ($F_{(2;341)} = 12.449$, $p < .001$). In einer weiteren Analyse wurde die Anzahl zur letzten Geburtstagsfeier eingeladener Gäste mittels Regressionen um den Einfluss von besonderen Geburtstagen bereinigt; da sich mit den bereinigten Werten bei den relevanten Analysen kein Unterschied gegenüber den nicht bereinigten Werten zeigte, werden in der vorliegenden Arbeit die Ergebnisse nur für die unbereinigten Werte berichtet.

Die vertriebsspezifische Version des ITB-PESA und ihre psychometrischen Eigenschaften

Die Vertriebsversion des ITB-PESA umfasste 84 Items, mit denen acht Eigenschaftsfacetten gemessen wurden. Alle 84 Items wurden auch in Studie 1 eingesetzt, so dass Messungen derselben Eigenschaften mit denselben Items im Nicht-Auswahl-Setting (Daten von Studie 1) berichtet werden können. Da es sich bei der Vertriebsversion des ITB-PESA um eine frühere und kundenspezifische Version des ITB-PESA handelt, sind die Skalen dieser Version etwas anders zusammengesetzt als die in Studie 1 berichteten. Das heißt, Skalen in der Vertriebsversion des ITB-PESA umfassen zum Teil andere Items als Skalen mit gleichem Titel, die in Studie 1 berichtet werden. Für den vorliegenden Vergleich von Auswahl und Nicht-Auswahl-Situation wurden die Eigenschaften mit einem identischen Itemsatz erfasst, und zwar mit den Items und der Zusammenstellung, die im hier geschilderten Ernstfalleinsatz verwendet wurden (Item-Skalen-Zusammenstellung der Vertriebsversion). Eine Übersicht über die Skalen und je eine Veranschaulichung durch das Item mit höchster Trennschärfe finden sich in Anhang E.2 (Tabelle E.2 - 1 und Tabelle E.2 - 2).

In Tabelle 20 werden die Itemzahlen und die Skalenstatistiken aufgeführt. Die Skalenstatistiken werden berichtet für die Daten im Ernstfalleinsatz (Auswahl, Untersuchungsgruppe 2A) sowie – für dieselbe Item-Skalen-Zusammenstellung – für die Daten der ersten Untersuchung von Studie 1 (Nicht-Auswahl; Untersuchungsgruppe 1A). Der Tabelle ist zu entnehmen, dass der Mittelwert der meisten Skalen im Auswahlkontext signifikant höher und die Standardabweichung für einige der Skalen signifikant niedriger ist als für das Nicht-Auswahl-Setting aus Studie 1. Die Normalverteilungsannahme muss in beiden Settings bei je vier der acht Skalen abgelehnt werden; die K-S-Tests werden in Anhang E.3 (Tabelle E.3 - 1 und Tabelle E.3 - 2) aufgeführt.

Tabelle 20: Itemzahlen und Skalenstatistiken zur berichteten Version des ITB-PESA bei der Auswahl und im Nicht-Auswahl-Kontext

	Itemzahl			Skalenstatistiken			
				Auswahl		Nicht-Auswahl	
Skala	ges.	neg.	*S1*	M	SD	M	SD
Kontaktfreude	9	4	*8*	4.53**	0.68**	3.48	0.99
Kommunikationsvermögen	10	2	*6*	4.80**	0.58*	4.24	0.67
Geselligkeit	11	7	*8*	4.92**	0.55**	4.25	0.73
Einfühlungsvermögen	10	3	*5*	4.09	0.60	4.10	0.65
Erfolgszuversicht	14	6	*10*	4.74**	0.62**	3.99	0.79
Aufgeschlossenheit und Neugier	10	-	*7*	4.96**	0.58	4.77	0.54
Eigeninitiative	11	2	*6*	4.95**	0.55	4.60	0.55
Leistungsstreben und Erfolgsmotivation	9	1	*8*	4.59**	0.74	4.10	0.74

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

ges: Itemzahl, neg: Zahl negativ gepolter Items, *S1*: Zahl der mit gleichnamiger Skala in Studie 1 gemeinsamen Items, M: Gruppen-Mittelwert der Skalenmittelwerte, SD: Gruppen-Standardabweichung der Skalenmittelwerte

Signifikante Unterschiede der Mittelwerte (mittels T-Tests geprüft) und Standardabweichungen (Levene-Tests der Varianzgleichheit) zwischen Auswahl- und Nicht-Auswahl-Setting werden mit * ($p < .01$) und ** ($p < .001$) indiziert (jeweils zweiseitige Testungen).

Tabelle 21 gibt einen Überblick über die Reliabilitäts- und Konsistenzschätzungen für die in Studie 2 berichteten Skalen. Überprüft wurde auch, ob sich Cronbachs Alpha zwischen der Auswahl- und Nicht-Auswahl-Situation unterscheidet: Dies ist bei vier der acht Skalen der Fall, Alpha fällt bei der Auswahl jeweils niedriger aus. Dies ist möglicherweise auch auf die bei der Auswahl niedrigeren Standardabweichungen dieser Skalen zurückzuführen. Die Korrelationen zwischen den Skalen sind in Anhang E.3 aufgeführt, einmal für Studie 2 (Tabelle E.3 - 1) und einmal für die Nicht-Auswahl-Stichprobe der ersten Untersuchung aus Studie 1 (Tabelle E.3 - 2): Von den 28 Korrelationen unterscheiden sich 17 zwischen beiden Settings nicht signifikant, die anderen 11 sind im Auswahlsetting größer als im Nicht-Auswahl-Setting. Die Skala „Leistungsstreben und Erfolgsmotivation“ korreliert im Auswahlsetting mit jeder der anderen Skalen höher als im Nicht-Auswahl-Setting. Insgesamt scheint die Konstruktvalidität des ITB-PESA bei der Auswahl teilweise beeinträchtigt; die Effekte sind jedoch klein und die meisten der Skalen hängen erwartungskonform zusammen. Die Zusammenhänge liegen überwiegend im mittleren Bereich ($.30 \leq r \leq .60$).

Tabelle 21: Reliabilitäts- und Konsistenzschätzungen zur berichteten Version des ITB-PESA bei der Auswahl und im Nicht-Auswahl-Kontext

	Auswahl			Nicht-Auswahl		
Skala	α	r_{tt}	$\overline{r_{it}}$	α	r_{tt}	$\overline{r_{it}}$
Kontaktfreude	.62**	.65	.33	.84	.85	.57
Kommunikationsvermögen	.67	.69	.36	.70	.73	.38
Geselligkeit	.64**	.63	.31	.76	.72	.41
Einfühlungsvermögen	.55*	.54	.25	.65	.64	.33
Erfolgszuversicht	.79**	.79	.41	.87	.87	.54
Aufgeschlossenheit und Neugier	.66	.69	.33	.63	.70	.30
Eigeninitiative	.68	.72	.36	.68	.64	.36
Leistungsstreben und Erfolgsmotivation	.76	.82	.46	.74	.79	.43

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

α: Cronbachs Alpha, r_{tt}: Split-Half-Reliabilität (odd-even, Spearman-Brown-korrigiert), $\overline{r_{it}}$: mittlere Trennschärfe der Items der Skala (Part-Whole-korrigiert, berechnet mit Fishers Z-Transformation, Fisher, 1918)

Die internen Konsistenzen einer Skala für Auswahl und Nicht-Auswahl-Setting wurden mit der Software Alpha Test (Lautenschlager & Meade, 2008) verglichen: * steht für $p < .05$, ** für $p < .001$.

Zur Erfassung von intraindividueller Variabilität und ERS

Variabilität und ERS wurden wie in Studie 1 berechnet: Für Variabilität werden die Inter-Item-SD der Skalen berechnet und um den Item-Mittelwert und das Quadrat des z-standardisierten Item-Mittelwerts korrigiert. Aus den korrigierten Werten wird mittels PAF ein Faktor als globales Maß von Variabilität extrahiert. Zur Erfassung von ERS wird auf Basis der Anzahl an Extremantworten für jede Skala ebenfalls ein Faktor mittels PAF bestimmt. Außerdem werden zur Validierung dieses Faktors wie in Studie 1 die Summe aller Extremantworten sowie Greenleaf-Skalen für Auswahl- und Nicht-Auswahl-Daten gebildet.

6.2 Ergebnisse

Die Analysen waren weitgehend parallel zu denen in Studie 1: Für den Auswahl- und den Nicht-Auswahl-Kontext wird zunächst die Erfassung von Variabilität und ERS im ITB-PESA beschrieben (Abschnitte 6.2.1 und 6.2.2). Im dritten Abschnitt werden die Zusammenhänge von Variabilität und ERS (6.2.3) und im vierten Abschnitt der Einfluss von Variabilität und ERS auf die Split-Half-Reliabilität der Skalen des ITB-PESA und auf die Kriteriumsvalidität der Skala „Kontaktfreude“ (6.2.4) berichtet. Sofern nicht anders ausgewiesen beziehen sich die Ergebnisse für die Auswahl auf Untersuchungsgruppe 2A. Die Ergebnisse für den Nicht-

Auswahl-Kontext basieren ausschließlich auf den Daten der Untersuchungsgruppe 1A. Zur Datenanalyse wurden SPSS 19.0.0 sowie SPSS Amos 19.0.0 (Arbuckle, 2010) verwendet. Zur Korrektur des Alpha-Fehlerniveaus diente die im Rahmen der vorliegenden Arbeit entwickelte und in Anhang C vorgestellte Methode.

6.2.1 Die Erfassung von intraindividueller Variabilität

Die Varianz der Inter-Item-SD der acht Skalen wurde im Auswahlkontext zu 6.0 % bis 35.4 % durch den Item-Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts aufgeklärt. Im Nicht-Auswahl-Kontext lagen die Werte zwischen 10.2 % und 21.3 %. Im Einzelnen werden die Werte und die deskriptiven Statistiken für die Inter-Item-SD in Anhang F.1 aufgeführt (Tabelle F.1 - 1 und Tabelle F.1 - 2). Im nächsten Schritt wurde geprüft, ob die Korrektur der Inter-Item-SD im Auswahlkontext stärker ausfällt als im Nicht-Auswahl-Kontext. Dafür wurden für jede der acht Skalen die multiplen Korrelationen von Item-Mittelwert und dem Quadrat des (z-standardisierten) Mittelwerts mit der Inter-Item-SD zwischen Auswahl- und Nicht-Auswahl-Setting verglichen. Die Zusammenhänge und die entsprechenden Teststatistiken finden sich in Tabelle 22: Deskriptiv fallen sieben der acht multiplen Korrelationen im Auswahlkontext höher aus als im Nicht-Auswahl-Setting, fünf der Vergleiche werden (auch alpha-korrigiert) signifikant[30]. Die Inter-Item-SD wird also bei den meisten Skalen im Auswahlkontext stärker korrigiert als im Nicht-Auswahl-Kontext.

Tabelle 22: Multiple Korrelationen des Item-Mittelwerts und des Quadrats des z-standardisierten Mittelwerts mit der Inter-Item-SD

	R		Test auf Unterschied	
Skala	Auswahl	Nicht-Auswahl	Fishers Z	p
Kontaktfreude	.429	.396	0.552	.290
Kommunikationsvermögen	.506	.432	1.302	.096
Geselligkeit	.532	.432	1.801	**.036**[a]
Einfühlungsvermögen	.244	.370	-1.920	.973
Erfolgszuversicht	.489	.319	2.822	**.002**[a]
Aufgeschlossenheit und Neugier	.595	.462	2.565	**.005**[a]
Eigeninitiative	.517	.355	2.786	**.003**[a]
Leistungsstreben und Erfolgsmotivation	.508	.394	1.985	**.024**[a]

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

R: multiple Korrelation, p: Signifikanzniveau zu Fishers Z-Test, p (einseitig) < .05 fett gedruckt, [a] nach Alpha-Adjustierung signifikant

[30] Nach der Bonferroni-Holm-Korrektur (Holm, 1979) sind drei der Unterschiede signifikant.

Im nächsten Schritt wurde die Vermutung, dass diese Unterschiede auf die Korrektur der Inter-Item-SD durch den Item-Mittelwert zurückzuführen sind, überprüft: Verglichen wurden jeweils die Korrelationen zwischen Item-Mittelwert und Inter-Item-SD. Die Korrelation sollte bei der Auswahl niedriger (=stärker negativ, vgl. Abschnitte 2.1.3 und 4.1) sein als bei der Nicht-Auswahl, da die Gruppen-Mittelwerte bei der Auswahl in der Regel höher ausfallen als in Nicht-Auswahl-Settings und die Abstände der Gruppen-Mittelwerte von der Mitte der Likert-Skala bei der Auswahl entsprechend größer sind (vgl. Abschnitte 4.4.3 und 6.1.2). Tabelle 23 zeigt die Korrelationen der Inter-Item-SD mit dem Item-Mittelwert. Für fünf Skalen unterscheiden diese sich im Auswahlkontext signifikant von der Nicht-Auswahl-Situation[31]. Die Inter-Item-SD dieser Skalen hängt bei der Auswahl stärker vom Mittelwert ab als in Nicht-Auswahl-Settings. Bei diesen fünf Skalen unterscheidet sich der Gruppen-Mittelwert zwischen Auswahl- und Nicht-Auswahl-Setting auch deutlich (vgl. Tabelle 20, Seite 113).

Tabelle 23: Korrelationen der Inter-Item-SD mit dem Item-Mittelwert

	Korrelation		Test auf Unterschied	
Skala	Auswahl	Nicht-Auswahl	Fishers Z	p
Kontaktfreude	-.325	.129	-6.447	**<.001**[a]
Kommunikationsvermögen	-.418	-.355	-1.034	.151
Geselligkeit	-.483	-.355	-2.153	**.016**[a]
Einfühlungsvermögen	-.200	-.304	1.531	.937
Erfolgszuversicht	-.434	-.255	-2.822	**.002**[a]
Aufgeschlossenheit und Neugier	-.451	-.446	-0.076	.470
Eigeninitiative	-.453	-.316	-2.228	**.013**[a]
Leistungsstreben und Erfolgsmotivation	-.435	-.317	-1.902	**.029**[a]

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

p: Signifikanzniveau zu Fishers Z-Test, p (einseitig) < .05 fett gedruckt, [a] nach Alpha-Adjustierung signifikant

Für das Quadrat des z-standardisierten Mittelwerts wurde bei der Auswahl ein tendenziell geringerer Zusammenhang mit der Inter-Item-SD erwartet als im Nicht-Auswahl-Setting. Schließlich sollte der Zusammenhang dann am stärksten sein, wenn der Gruppen-Mittelwert nahe der Mitte der Likert-Skala liegt (vgl. Abschnitt 2.1.3). Die Ergebnisse in Tabelle 24 zeigen, dass die Inter-Item-SD wie zu erwarten in beiden Kontexten in der Regel negativ mit dem Quadrat des z-standardisierten Mittelwerts korreliert. Das bedeutet, mit hoher Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert geht jeweils eine niedrige Inter-Item-

[31] Dies gilt auch alpha-adjustiert (zur Alpha-Adjustierung: siehe Anhang C). Nach der konservativeren Bonferroni-Holm-Korrektur (Holm, 1979) wäre nur einer der Tests signifikant.

SD einher. Bei einer Skala ist der Zusammenhang im Auswahlkontext signifikant geringer als im Nicht-Auswahl-Kontext[32], bei den anderen Skalen lässt sich dieser Unterschied nicht ausmachen. Der Befund, dass die Inter-Item-SD bei der Auswahl meist stärker korrigiert wird als im Nicht-Auswahl-Kontext, lässt sich für die meisten Skalen auf Unterschiede der Korrelation zwischen Inter-Item-SD und Item-Mittelwert zurückführen. Diese Unterschiede gehen auf eine Verschiebung des Mittelwerts im Auswahlkontext gegenüber dem Nicht-Auswahl-Kontext zurück.

Tabelle 24: Korrelationen der Inter-Item-SD mit dem Quadrat des z-standardisierten Mittelwerts

	Korrelation		Test auf Unterschied	
Skala	Auswahl	Nicht-Auswahl	Fishers Z	p
Kontaktfreude	-.155	-.384	3.436	**<.001**[a]
Kommunikationsvermögen	-.134	-.142	0.112	.455
Geselligkeit	-.027	-.130	1.436	.075
Einfühlungsvermögen	-.125	-.175	0.716	.237
Erfolgszuversicht	-.083	-.127	0.616	.269
Aufgeschlossenheit und Neugier	-.177	.020	-2.759	.997
Eigeninitiative	-.042	-.081	0.542	.294
Leistungsstreben und Erfolgsmotivation	-.098	-.191	1.308	.095

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

p: Signifikanzniveau zu Fishers Z-Test, p (einseitig) < .05 fett gedruckt, [a] nach Alpha-Adjustierung signifikant

Für die weiteren Analysen wurden die korrigierten Inter-Item-SD verwendet: Für vier der acht korrigierten Inter-Item-SD lässt sich die Annahme der Normalverteilung im Auswahlsetting aufrecht erhalten, bei den anderen vier muss sie verworfen werden. Im Nicht-Auswahl-Setting wird die Normalverteilungsannahme bei nur einer Skala abgelehnt. Die einzelnen K-S-Tests werden in Anhang F.1 (Tabelle F.1 - 1 und Tabelle F.1 - 2) berichtet. Zur Prüfung, ob die Daten sich für eine PAF eignen, wurden neben den K-S-Tests der KMO-Koeffizient bestimmt und der Bartlett-Test auf Sphärizität gerechnet – die Ergebnisse werden in Tabelle 25 zusammengefasst: Zwar sind nicht alle der korrigierten Inter-Item-SD normalverteilt, da jedoch die anderen Voraussetzungen erfüllt waren, wurde dennoch mittels PAF je ein Faktor ermittelt.

[32] Dieser Unterschied ist auch nach der in Anhang C berichteten Alpha-Korrektur signifikant. Die konservativere Bonferroni-Holm-Korrektur (Holm, 1979) führt zum selben Ergebnis.

Tabelle 25: Prüfung der Voraussetzungen für eine Faktorenanalyse der korrigierten Inter-Item-SD der Skalen der Vertriebsversion des ITB-PESA

		Bartlett-Test			signifikante
Faktor der korrigierten Inter-Item-SD	KMO	χ^2	df	p	K-S-Tests
Auswahl	.90*	975.04	28	<.001	4 von 8
Nicht-Auswahl	.91*	1529.62	28	<.001	1 von 8

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

KMO: Kaiser-Mayer-Olkin-Koeffizient, * sehr gute Eignung (vgl. Bühner, 2011)

Bartlett-Test: χ^2: Teststatistik, df: Freiheitsgrade, p: Signifikanzniveau; $p < .05$ deutet auf gute Eignung der Daten für eine Faktorenanalyse hin.

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis)

MAP-Tests und Scree-Tests legten nahe, dass den korrigierten Inter-Item-SD jeweils genau ein Faktor zugrunde liegt. Die Scree-Plots finden sich in Anhang F.2 (Abbildung F.2 - 1 und Abbildung F.2 - 2). Der erste Faktor der PAF klärt sowohl im Auswahl- als auch im Nicht-Auswahl-Kontext einen großen Teil der Varianz auf, bei der Auswahl fällt dieser Anteil etwas kleiner aus. Die Ladungen der korrigierten Inter-Item-SD auf dem Faktor sind hoch und liegen jeweils innerhalb einer geringen Bandbreite. Die Ergebnisse werden in Tabelle 26 zusammengefasst, die einzelnen Ladungen sind in Anhang F.1 aufgeführt (Tabelle F.1 - 1 und Tabelle F.1 - 2). Tabelle 26 enthält auch Schätzungen für die Konsistenz der Faktoren: Diese fällt in beiden Kontexten hoch aus. Zusammenfassend ist davon auszugehen, dass Variabilität ein globaler Trait ist, der reliabel erfasst wird.

Tabelle 26: Ergebnisse der Faktorenanalyse der korrigierten Inter-Item-SD der Skalen der Vertriebsversion des ITB-PESA sowie Konsistenzwerte für die Faktoren

	Anzahl der Faktoren		aufgekl. Var	Ladungen		Konsistenz	
Faktor der korrigierten Inter-Item-SD	MAP-Test	Scree-Test	(1. Fakt.)	Min	Max	α	ω
Auswahl	1	1	43.2 %	.61	.71	.71	.86
Nicht-Auswahl	1	1	51.2 %	.62	.83	.89	.89

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

aufgekl. Var (1. Fakt.): durch den (ersten) Faktor aufgeklärte Varianz, Min: niedrigste Ladung, Max: höchste Ladung; alle Ladungen waren positiv.

α: Cronbachs Alpha (hier – da es sich um Faktoren handelt – für standardisierte Werte berechnet), ω: Omega

6.2.2 Die Erfassung von Extreme Response Style

Parallel zu Studie 1 wurden jeweils mehrere Maße für ERS bestimmt: die Extremwerthäufigkeit im gesamten Fragebogen, ein Faktor für extremes Antworten sowie je zwei Greenleaf-Skalen. Im Auswahlkontext beantworteten die Bearbeiter im Mittel 38.7 % der Items an den Extrempunkten ($M = 0.387$, $SD = 0.218$), im Nicht-Auswahl-Setting im Mittel 23.7 % und damit signifikant weniger als bei der Auswahl ($M = 0.237$, $SD = 0.155$; $T_{(655)} = -10.904$, $p < .001$, einseitig). Auch die Varianz der Extremwerthäufigkeit war im Auswahlsetting größer als im Nicht-Auswahl-Setting ($F_{(2;770)} = 46.512$, $p < .001$).

Zur Berechnung der ERS-Faktoren wurde jeweils die Extremwerthäufigkeit für die acht Skalen bestimmt. Zwar war keine dieser Extremwerthäufigkeiten normalverteilt, da die anderen Voraussetzungen für eine PAF jedoch erfüllt waren, wurden Faktoren berechnet. KMO-Koeffizienten, die Ergebnisse der Bartlett-Tests auf Sphärizität sowie eine Übersicht über die K-S-Tests finden sich in Tabelle 27. Die einzelnen K-S-Tests werden ebenso wie die deskriptiven Statistiken der Extremwerthäufigkeiten in Anhang F.3 (Tabelle F.3 - 1 und Tabelle F.3 - 2) berichtet.

Tabelle 27: Prüfung der Voraussetzungen für eine Faktorenanalyse der Extremwerthäufigkeiten auf den Skalen der Vertriebsversion des ITB-PESA

		Bartlett-Test			signifikante
ERS-Faktor	KMO	χ^2	df	p	K-S-Tests
Auswahl	.93*	2185.97	28	<.001	8 von 8
Nicht-Auswahl	.91*	1529.62	28	<.001	8 von 8

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

KMO: Kaiser-Mayer-Olkin-Koeffizient, * sehr gute Eignung (vgl. Bühner, 2011)

Bartlett-Test: χ^2: Teststatistik, df: Freiheitsgrade, p: Signifikanzniveau; p < .05 deutet auf gute Eignung der Daten für eine Faktorenanalyse hin.

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis)

In beiden Kontexten, Auswahl und Nicht-Auswahl, indizierten MAP-Test und Scree-Test für die acht Extremwerthäufigkeiten je einen Faktor. Die Scree-Plots sind in Anhang F.4 (Abbildung F.4 - 1 und Abbildung F.4 - 2) aufgeführt. Der Faktor der PAF klärte je mehr als die Hälfte der Varianz der Extremwerthäufigkeiten der Skalen auf, im Auswahlsetting fällt der Anteil höher aus als im Nicht-Auswahl-Setting. Die Ladungen waren für beide Settings hoch

und hatten eine geringe Bandbreite. Die Konsistenzschätzungen lagen in einem hohen Bereich. Einen Überblick über die Ergebnisse von PAF und Konsistenzanalysen gibt Tabelle 28. Die einzelnen Ladungen finden sich in Anhang F.3 (Tabelle F.3 - 1 und Tabelle F.3 - 2).

Tabelle 28: Ergebnisse der Faktorenanalyse der Extremwerthäufigkeiten auf den Skalen der Vertriebsversion des ITB-PESA sowie Konsistenzwerte für die Faktoren

	Anzahl der Faktoren		aufgekl. Var	Ladungen		Konsistenz	
ERS-Faktor	MAP-Test	Scree-Test	(1. Fakt.)	Min	Max	α	ω
Auswahl	1	1	64.6 %	.74	.87	.94	.94
Nicht-Auswahl	1	1	51.2 %	.62	.83	.89	.89

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

aufgekl. Var (1. Fakt.): durch den (ersten) Faktor aufgeklärte Varianz, Min: niedrigste Ladung, Max: höchste Ladung; alle Ladungen waren positiv.

α: Cronbachs Alpha (hier – da es sich um Faktoren handelt – für standardisierte Werte berechnet), ω: Omega

Zur Itemauswahl für Greenleaf-Skalen wurde das Vorgehen aus Studie 1 wiederholt (Abschnitt 5.1.3); hier wurde eine PCA sowohl mit den Auswahldaten als auch mit den Nicht-Auswahl-Daten durchgeführt. Es wurden zwei Itemsätze à 16 Items bestimmt, einer mit den Auswahldaten und einer mit den Nicht-Auswahl-Daten. Mit beiden Itemsätzen wurde sowohl für den Auswahl- als auch für den Nicht-Auswahl-Kontext der Greenleaf-Score berechnet. Mit anderen Worten wurden zwei Greenleaf-Skalen gebildet, die für den Auswahl- und den Nicht-Auswahl-Kontext berichtet werden; bei je einer Skala basiert

- die Itemauswahl auf Daten aus dem Auswahlkontext,
- die Itemauswahl auf Daten aus dem Nicht-Auswahl-Kontext.

Die Korrelationen zwischen den Greenleaf-Items, die *mit den Auswahldaten* bestimmt wurden, lagen *für die Auswahlsituation* mit Likert-Kodierung und in Richtung der jeweiligen ITB-PESA-Skala gepolt nahe Null ($-.225 \leq r \leq .216$). Im Mittel lagen sie leicht über Null, die interne Konsistenz war moderat. ERS-kodiert (Endpunkte der Likert-Skala: 1; übrige Kategorien: 0) korrelierten sämtliche Items positiv miteinander ($.044 \leq r \leq .329$), auch die mittlere Korrelation und die interne Konsistenz waren höher als bei Likert-Kodierung. Die Ergebnisse werden in der oberen Zeile der oberen Hälfte von Tabelle 29 zusammengefasst. In der unteren Zeile der oberen Hälfte sind die *Ergebnisse für den Nicht-Auswahl-Kontext* aufgeführt: *Dieselben Items* (Itemauswahl *mit Auswahldaten*) hingen Likert-kodiert und in Rich-

tung der jeweiligen ITB-PESA-Skala gepolt stärker zusammen als im Auswahlkontext ($-.349 \leq r \leq .536$), im Mittel lag die Inter-Item-Korrelation über Null. Auch die interne Konsistenz für die Likert-Kodierung war relativ hoch. ERS-kodiert lagen die Korrelationen zwischen den Items überwiegend über Null ($-.013 \leq r \leq .292$) und die interne Konsistenz war höher als bei der Likert-Kodierung.

Die Items der Greenleaf-Skala, die *mit den Nicht-Auswahl-Daten* ermittelt wurden, korrelierten *im Auswahlkontext* Likert-skaliert und in die Richtung der jeweiligen ITB-PESA Skala gepolt gering miteinander ($-.215 \leq r \leq .311$). Im Mittel lagen die Korrelationen leicht über Null. ERS-kodiert lagen alle Inter-Item-Korrelationen über Null ($.032 \leq r \leq .357$), auch die mittlere Korrelation lag deutlich über Null. Die interne Konsistenz fiel für die Likert-Kodierung moderat und für die ERS-Kodierung hoch aus. Eine Übersicht findet sich in der oberen Zeile der unteren Hälfte von Tabelle 29. Im *Nicht-Auswahl-Setting* zeigt sich ein ähnliches Bild: Die Korrelationen zwischen den Likert-kodierten und in Richtung der jeweiligen ITB-PESA-Skala gepolten Items verteilten sich um Null ($-.181 \leq r \leq .286$), ERS-kodiert lagen Korrelationen zwischen den Items überwiegend leicht über Null ($-.016 \leq r \leq .276$). Die mittleren Inter-Item-Korrelationen und die interne Konsistenz waren ERS-kodiert höher als Likert-kodiert. Die Ergebnisse finden sich in der untersten Zeile von Tabelle 29.

Tabelle 29: Statistiken zu den Greenleaf-Skalen, links für die Likert-Kodierung, rechts für die ERS-Kodierung

	Likert-Kodierung (1 bis 6)		ERS-Kodierung (Endpunkte: 1, „2“ bis „5“: 0)				
Greenleaf-Skala	α	$\overline{r_{it}}$	M	SD	h_{rel}	α	$\overline{r_{it}}$
mit Itemauswahl bei Auswahl							
Kennwerte für Auswahl	.45	.05	4.77	3.48	29.8 %	.78	.19
Kennwerte für Nicht-Auswahl	.61	.09	3.71	2.76	23.2 %	.68	.12
mit Itemauswahl bei Nicht-Auswahl							
Kennwerte für Auswahl	.43	.06	5.46	3.62	34.1 %	.79	.19
Kennwerte für Nicht-Auswahl	.45	.05	3.63	2.64	22.7 %	.65	.11

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

α: Cronbachs Alpha, $\overline{r_{it}}$: mittlere Korrelation zwischen den Items (berechnet mit Fishers Z-Transformation, Fisher, 1918), M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, h_{rel}: relative Häufigkeit von Extremantworten

ERS lässt sich mit den Greenleaf-Skalen größtenteils reliabel erfassen, die Auswahl von Items aus der relativ kurzen Vertriebsversion des ITB-PESA führt jedoch bei Likert-Skalierung zu

moderaten Inter-Item-Korrelationen und unerwünscht hohen internen Konsistenzen. Da nicht auszuschließen ist, dass auch von den Likert-Items Gemessenes in die ERS-Skalen eingeht, wurden Korrelationen zwischen den Skalen mit Likert- und mit ERS-Kodierung berechnet: Die Greenleaf-Skala, deren Items *mit den Auswahldaten* ausgewählt wurden, korreliert mit der Summe ihrer Items in der ursprünglichen Likert-Kodierung *bei der Auswahl* moderat ($r_{(367)} = .442,\ p < .001$); *im Nicht-Auswahl-Setting* ist diese Korrelation nicht signifikant ($r_{(405)} = .090,\ p < .072$) und signifikant niedriger als im Auswahlsetting ($Z = 5.326,\ p < .001$). Auch für die Greenleaf-Skala, deren Items *mit den Nicht-Auswahl-Daten* ausgewählt wurden, war die Korrelation *im Auswahlkontext* moderat ($r_{(367)} = .498, p < .001$); *im Nicht-Auswahl-Kontext* war die Korrelation zwar auch signifikant ($r_{(405)} = .282,\ p < .001$), aber signifikant niedriger als bei der Auswahl ($Z = 3.554, p < .001$).

Hinsichtlich der Gruppen-Mittelwerte und Standardabweichungen der Greenleaf-Skalen zeigt sich ein ähnliches Bild wie für die Extremwerthäufigkeit über den gesamten Fragebogen hinweg: Beide Greenleaf-Skalen hatten bei der Auswahl einen höheren Mittelwert und eine größere Standardabweichung als im Nicht-Auswahl-Kontext (auf Auswahldaten basierende Greenleaf-Skala: $T_{(770)} = -4.649,\ p < .001;\ F_{(2;770)} = 14.631,\ p < .001$; auf Nicht-Auswahl-Daten basierend: $T_{(770)} = -7.918, p < .001; F_{(2;770)} = 35.869, p < .001$).

Im nächsten Schritt wurde die konvergente Konstruktvalidität der ERS-Maße bestimmt: Sowohl für die Auswahl als auch für die Nicht-Auswahl zeigten sich hohe Zusammenhänge. Mit den verschiedenen Operationalisierungen wurde also jeweils dasselbe Konstrukt gemessen; die Korrelationen lassen sich Tabelle 30 entnehmen. Interessant sind die hohen Zusammenhänge zwischen den beiden Greenleaf-Skalen, die anzeigen, dass die Reliabilität dieser beiden Maße durch die interne Konsistenz unterschätzt wird.

Aufgrund der hohen Konvergenz, und da Variabilität als Faktor operationalisiert wurde, wird bei den in den folgenden Abschnitten berichteten Analysen jeweils der ERS-Faktor berichtet.

Tabelle 30: Korrelationen zwischen den ERS-Maßen

	Auswahl			Nicht-Auswahl		
Skala	2	3	4	2	3	4
1. ERS-Faktor ITB-PESA	.998	.875	.896	.996	.822	.829
2. ERS-Häufigkeit ITB-PESA		.879	.900		.826	.839
3. Greenleaf-Skala (Itemselektion Auswahl)			.864			.816
4. Greenleaf-Skala (Itemselektion Nicht-Auswahl)						

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$
Für alle Korrelationen gilt $p < .001$.

6.2.3 Intraindividuelle Variabilität und Extreme Response Style

Wie in Studie 1 wurde auf zwei Wegen geprüft, ob die korrigierten Inter-Item-SD und die Extremwerthäufigkeiten der Skalen Indikatoren ein und derselben Eigenschaft sind. Zunächst wurden Korrelationen berechnet: Im Auswahlkontext korrelierten die Faktoren für Variabilität und ERS hoch miteinander ($r_{(367)} = .797$, $p < .001$), allerdings signifikant niedriger als im Nicht-Auswahl-Kontext ($r_{(405)} = .848$, $p < .001$; $Z = 2.210$, $p = .027$, zweiseitig). Als zweites wurden SEM aufgestellt. Zwei Modelle wurden definiert, in denen sowohl die Extremwerthäufigkeiten der Skalen als auch die korrigierten Inter-Item-SD als manifeste Variablen dienten. Modell 1 ist in Abbildung 19 dargestellt: Parallel zu Modell 1 aus Studie 1 (Abschnitt 5.2.3) lag den Extremwerthäufigkeiten und den korrigierten Inter-Item-SD ein Faktor zugrunde; Fehlerfaktoren zu manifesten Variablen, die zu einer Skala gehörten – zum Beispiel zur korrigierten Inter-Item-SD und zur Extremwerthäufigkeit der Skala „Kontaktfreude“, waren korreliert. Darüber hinaus waren Fehlerfaktoren unkorreliert.

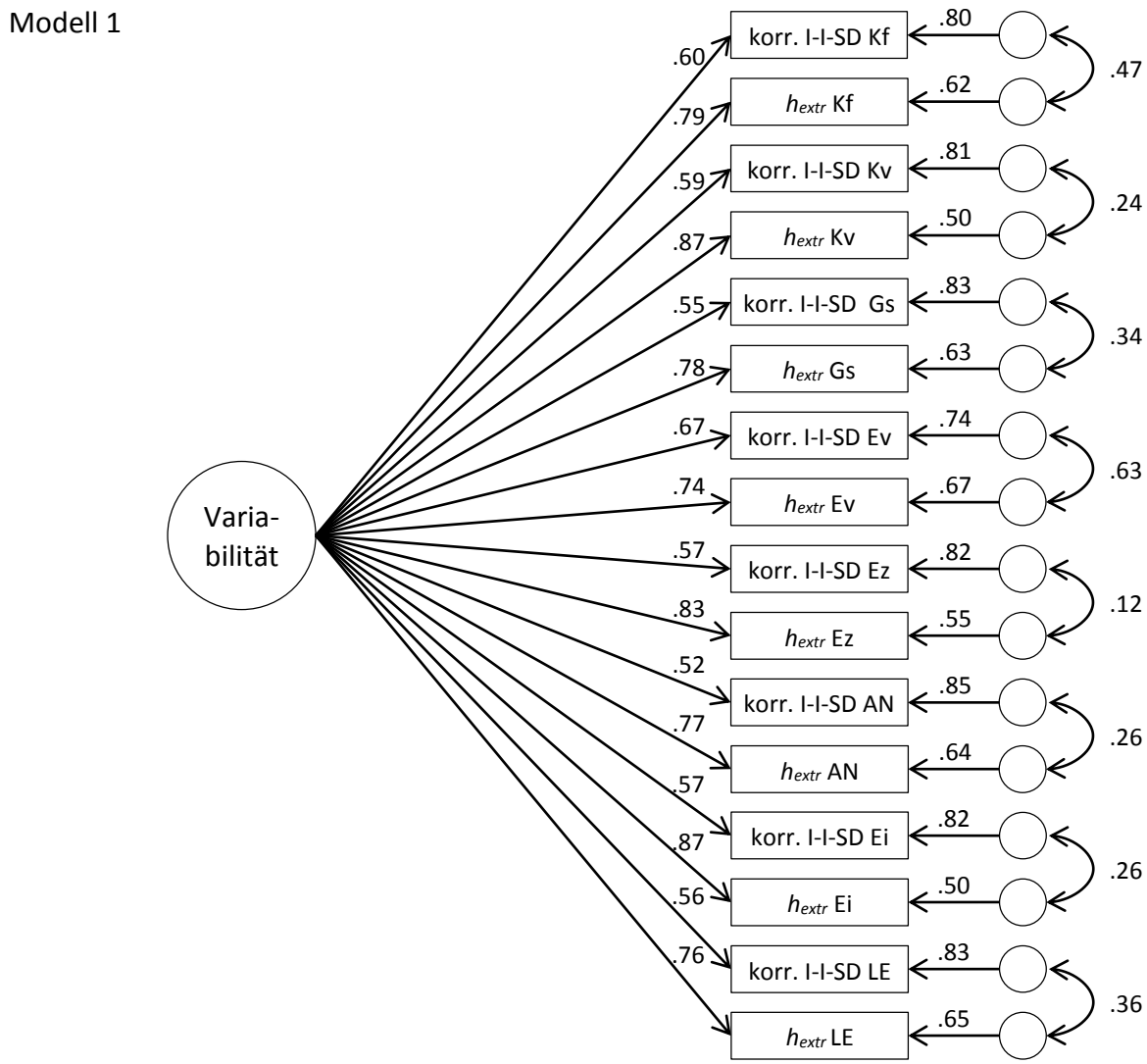

Abbildung 19: Strukturgleichungsmodell 1 mit standardisierten Regressionsgewichten und Korrelationen für den Auswahlkontext

korr. I-I-SD: korrigierte Inter-Item-Standardabweichung, h_{extr}: Extremwerthäufigkeit, jeweils berechnet für die Skalen zur Messung von Kontaktfreude (Kf), Kommunikationsvermögen (Kv), Geselligkeit (Gs), Einfühlungsvermögen (Ev), Erfolgszuversicht (Ez), Aufgeschlossenheit und Neugier (AN), Eigeninitiative (Ei) sowie Leistungsstreben und Erfolgszuversicht (LE)

In Modell 2 wurden die manifesten Variablen zwei Faktoren zugeordnet, einem für Variabilität und einem für ERS. Das Modell ist in Abbildung 20 illustriert. Zusammenhänge zwischen den Fehlerfaktoren waren parallel zu Modell 1; die beiden Globalfaktoren waren korreliert.

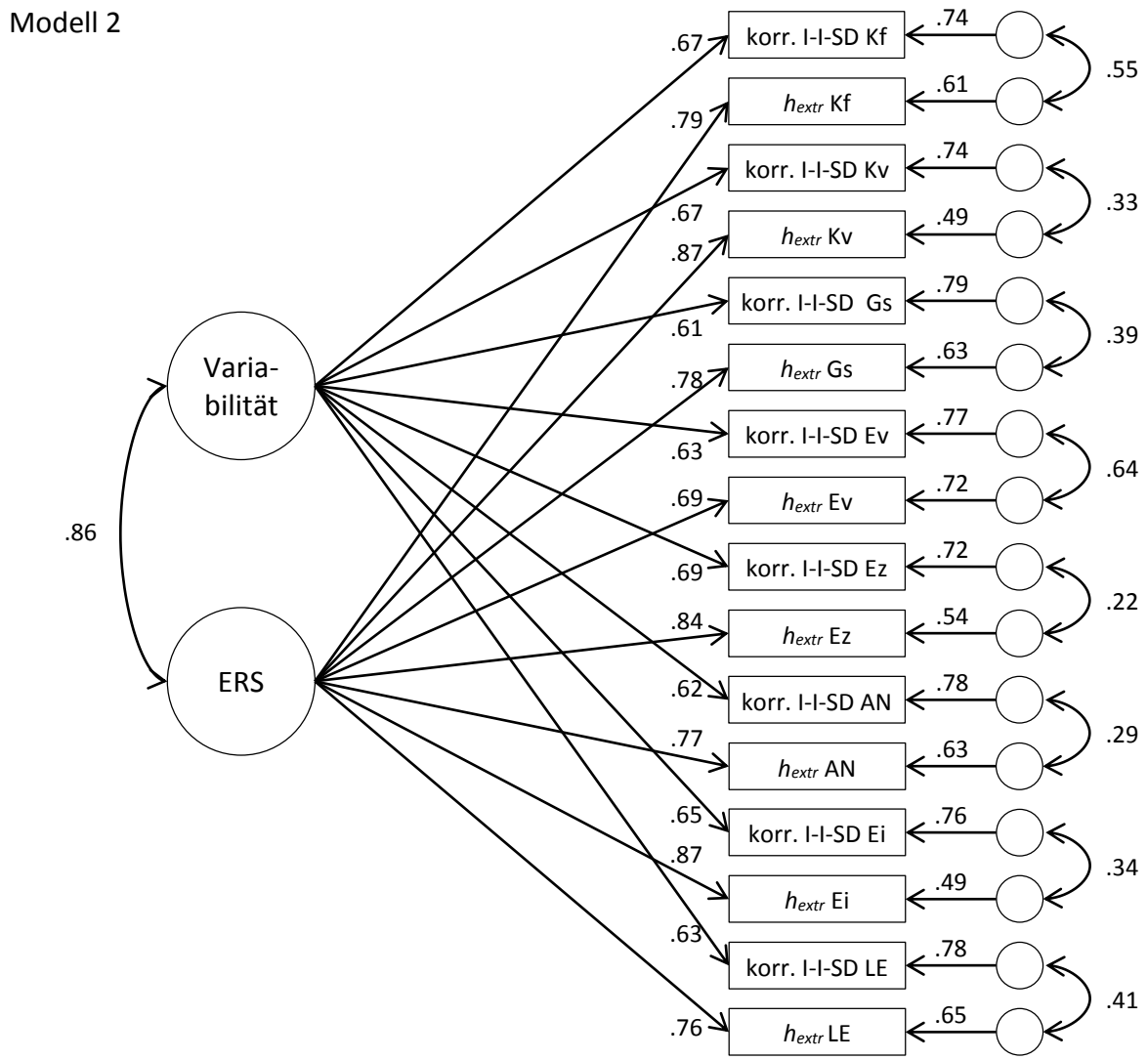

Abbildung 20: Strukturgleichungsmodell 2 mit standardisierten Regressionsgewichten und Korrelationen für den Auswahlkontext

korr. I-I-SD: korrigierte Inter-Item-Standardabweichung, h_{extr}: Extremwerthäufigkeit, jeweils berechnet für die Skalen zur Messung von Kontaktfreude (Kf), Kommunikationsvermögen (Kv), Geselligkeit (Gs), Einfühlungsvermögen (Ev), Erfolgszuversicht (Ez), Aufgeschlossenheit und Neugier (AN), Eigeninitiative (Ei) sowie Leistungsstreben und Erfolgszuversicht (LE)

In Abbildung 19 und Abbildung 20 sind die Regressionsgewichte und Korrelationen exemplarisch für den Auswahlkontext aufgeführt. Für den Nicht-Auswahl-Kontext lagen sie in einem ähnlichen Bereich. Tabelle 31 ist zu entnehmen, dass alle Modelle bis auf Modell 1 im Auswahlkontext einen akzeptablen Fit nach den Empfehlungen von Hu und Bentler (1999) sowie Schermelleh-Engel et al. (2003) erzielten. Beide Modelle wiesen im Nicht-Auswahl-Kontext einen leicht besseren Fit auf als bei der Auswahl. Innerhalb eines Settings wies jeweils Modell 2 (zwei Faktoren) einen besseren Fit auf; dieser Unterschied fiel im Auswahlkontext

deutlicher aus als im Nicht-Auswahl-Kontext. Dennoch korrelierten die Faktoren für Variabilität und ERS im Auswahlkontext zu $r = .864$; im Nicht-Auswahl-Setting lag die Korrelation bei $r = .917$. Der ERS-Faktor teilt also den Großteil seiner Varianz mit Variabilität, er erklärt jedoch auch Varianz auf, die nicht durch Variabilität aufgeklärt wird.

Tabelle 31: Analyse des Zusammenhangs von Variabilität und ERS mittels Strukturgleichungsmodellen

	Modelltest					Fit-Indizes			Modell 1 vs. 2		
Kontext / Modell	χ^2	df	p	p_{BSB}	*AIC*	*CFI*	*SRMR*	*RMSEA (CI 90)*	$\Delta\chi^2$	Δdf	p
Auswahl											
Modell 1	493.66	96	<.001	.001	573.66	.90	.058	.106 (.097-.116)			
Modell 2	303.41	95	<.001	.001	**385.41**	.95	.045	.077 (.068-.087)	190.15	1	<.001
Nicht-Auswahl											
Modell 1	322.95	96	<.001	.001	402.95	.94	.038	.077 (.068-.086)			
Modell 2	228.79	95	<.001	.001	**310.79**	.96	.034	.059 (.049-.069)	94.16	1	<.001

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

Modelltest: χ^2: Teststatistik, df: Freiheitsgrade, p: Signifikanzniveau; p < .05 deutet auf Modellfehlspezifikationen hin.

p_{BSB}: p-Wert für den Bollen-Stine-Bootstrap, AIC: *Akaike Information Criterion*, *CFI*: *comparative Fit Index*, *SRMR*: *standardized Root Mean Square Residual*, *RMSEA (CI 90)*: *Root Mean Square Error of Approximation* (90-Prozent-Konfidenzintervall)

Modell 1 vs. 2: $\Delta\chi^2$: Differenz der Chiquadrat-Werte zwischen Modell 1 und Modell 2, Δdf: Differenz der Freiheitsgrade, p: Signifikanzniveau; p < .05 deutet auf einen Unterschied zwischen den Modellen hin.

Variabilität hängt im Auswahlkontext geringer mit ERS zusammen als im Nicht-Auswahl-Kontext. Und da Extremantworten auch durch hohe Skalenausprägungen bedingt sein können und die Skalen zum Teil im Auswahlsetting höher miteinander korrelieren als im Nicht-Auswahl-Setting (vgl. Abschnitte 4.4.3 und 6.1.2 sowie Anhang E.3, Tabelle E.3 - 1 und Tabelle E.3 - 2), lässt sich vermuten, dass auch der erste den Fragebogenskalen zugrunde liegende Faktor ERS vorhersagt (vgl. Abschnitt 5.3). Dieser klärt womöglich Varianz von ERS auf, die nicht durch Variabilität erklärt wird, und der Anteil dieser Varianz müsste – wie auch die Korrelationen zwischen Likert- und ERS-Kodierung bei den Greenleaf-Skalen bereits andeuten (vgl. Abschnitt 6.2.2) – im Auswahlsetting größer sein als im Nicht-Auswahl-Setting. Zur Prüfung dieser Hypothesen wurden hierarchische Regressionen verwendet. Zunächst wurde mittels PAF jeweils ein Faktor aus den Skalen der Vertriebsversion des ITB-PESA extrahiert. Dieser Faktor klärt im Auswahlkontext mehr Varianz der Skalen auf (43.9 %) als im Nicht-Auswahl-Kontext (35.9 %). Weitere Ergebnisse der Faktorenanalysen werden in Anhang F.5

berichtet. Der gemeinsame Faktor der Skalen korreliert hoch mit ERS und diese Korrelation fällt im Auswahlsetting höher aus ($r_{(367)} = .682,\ p < .001$) als im Nicht-Auswahl-Setting ($r_{(405)} = .512,\ p < .001$; Vergleich: $Z = 3.681,\ p < .001$). Der Zusammenhang zwischen dem Faktor der Skalen und Variabilität ist gering und hängt nicht vom Setting ab (Auswahl: $r_{(367)} = .276,\ p < .001$; Nicht-Auswahl: $r_{(405)} = .199,\ p < .001$; Vergleich: $Z = 1.124$, $p = .261$). Die Regressionen zur Vorhersage von ERS sind in Tabelle 32 aufgeführt: Oben finden sich die Ergebnisse für den Auswahlkontext, unten für den Nicht-Auswahl-Kontext. Im ersten Schritt wurde jeweils Variabilität als Prädiktor eingegeben, im zweiten Schritt der gemeinsame Faktor der Fragebogenskalen. Dieser klärte ERS erwartungsgemäß über Variabilität hinaus auf. Der zusätzliche Anteil aufgeklärter Varianz fällt im Auswahlkontext größer aus als im Nicht-Auswahl-Kontext ($Z = 4.006, p < .001$)[33].

Tabelle 32: Hierarchische Regressionen zur Vorhersage von ERS durch Variabilität und den gemeinsamen Faktor der Skalen der Vertriebsversion des ITB-PESA

				T-Test			
Regressionen zur Vorhersage von ERS	B	s_E	β	T	p	R^2	ΔR^2
Auswahl							
Schritt 1						.635	
Variabilität	0.833	0.033	.797	25.187	**<.001**		
Schritt 2						.866	.231
Variabilität	0.689	0.021	.659	32.996	**<.001**		
Faktor der ITB-PESA-Skalen	0.515	0.021	.500	25.041	**<.001**		
Nicht-Auswahl							
Schritt 1						.719	
Variabilität	0.875	0.027	.848	32.059	**<.001**		
Schritt 2						.842	.123
Variabilität	0.802	0.021	.777	38.351	**<.001**		
Faktor der ITB-PESA-Skalen	0.367	0.021	.358	17.660	**<.001**		

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

B: Regressionsgewicht, s_E: Standardfehler des Regressionsgewichts, β: standardisiertes Regressionsgewicht; Signifikanztests für β (T-Test): Teststatistik T, Signifikanzniveau p (einseitig); R^2: Effektstärke (Determinationskoeffizient), ΔR^2: Änderung des Determinationskoeffizienten

$p < .05$ fett gedruckt

[33] Verglichen wurden die partiellen Korrelationen, d. h. die Korrelationen des gemeinsamen Faktors der ITB-PESA-Skalen mit ERS, bereinigt um den Einfluss von Variabilität auf ERS.

6.2.4 Der Einfluss von Variabilität und ERS auf die Split-Half-Reliabilität und auf die Kriteriumsvalidität

Ob Variabilität und ERS einen Einfluss auf die Split-Half-Reliabilität haben, wurde wie in Studie 1 mit zwei Methoden geprüft. Zunächst wurde der Index of Profile Agreement (I_{pa}) für die Übereinstimmung des Profils (McCrae, 1993, 2008; McCrae et al., 1998; vgl. Abschnitt 5.2.4), gemessen durch die jeweils einen Hälften der Skalen (Mittelwert der Items an ungeraden Positionen), mit dem Profil, gemessen durch die jeweils anderen Hälften der Skalen (Mittelwert der Items an geraden Positionen), berechnet. Die Korrelationen zwischen diesem Index und Variabilität bzw. ERS geben Auskunft über deren Einfluss auf die Split-Half-Reliabilität. Die Koeffizienten finden sich Tabelle 33: Sowohl im Auswahl- als auch im Nicht-Auswahl-Kontext korrelierte die Profilübereinstimmung der Hälften der Skalen hypothesenkonform negativ mit Variabilität. Signifikante Zusammenhänge mit ERS zeigten sich nicht.

Tabelle 33: Korrelationen der Profilübereinstimmungen für die Profile der Skalenhälften mit dem Variabilitäts- und dem ERS-Faktor in der Vertriebsversion des ITB-PESA

	Korrelation zu I_{pa} für die Skalenhälften	
	Auswahl	Nicht-Auswahl
Variabilitäts-Faktor	**-.161** (.002)	**-.144** (.004)
ERS-Faktor	-.009 (.433)	.033 (.748)

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

I_{pa}: *Index of Profile Agreement* (Index der Profilübereinstimmung)

In Klammern werden die Signifikanzniveaus aufgeführt (einseitige Testungen). Signifikante Korrelationen sind fett gedruckt.

Als zweite Prüfmethode dienten moderierte multiple Regressionen. Mit diesen wurde untersucht, ob die Split-Half-Reliabilität der einzelnen Skalen durch Variabilität bzw. ERS moderiert wird (vgl. Studie 1, Abschnitt 5.2.4): Es wurde jeweils die Hälfte einer Skala (Mittelwert der Items an geraden Positionen) mit der anderen Hälfte (Mittelwert der Items an ungeraden Positionen) vorhergesagt; im ersten Schritt der Regressionsanalyse wurde auch der jeweilige Moderator eingegeben, im zweiten Schritt dann das Produkt von Prädiktor und Moderator (jeweils z-standardisiert). Das Regressionsgewicht dieses Produkts indiziert den Moderatoreffekt. Die Ergebnisse *für den Auswahlkontext* werden in Tabelle 34 berichtet: Variabilität moderiert die Split-Half-Reliabilität von keiner der acht Skalen signifikant. Ein Modera-

toreffekt von ERS fällt unter das .05-Signifikanzniveau, alpha-adjustiert zeigt sich jedoch kein signifikanter Effekt[34]. *Für den Nicht-Auswahl-Kontext* zeigt sich – wie die Ergebnisse in Studie 1 erahnen lassen (Abschnitt 5.2.4) – ein ähnliches Bild: Für Variabilität und für ERS fällt je ein Moderatoreffekt unter die .05-Grenze, alpha-adjustiert ist jedoch keiner der Effekte signifikant[35]. Die Ergebnisse sind in Tabelle 35 aufgeführt.

Tabelle 34: Der Einfluss von Variabilität und ERS auf die Split-Half-Reliabilität der Skalen der Vertriebsversion des ITB-PESA im Auswahlkontext

	Moderator Variabilität				Moderator ERS			
Moderation der Split-Halt-Reliabilität der Skala		T-Test				T-Test		
	β	T	p	ΔR^2	β	T	p	ΔR^2
Kontaktfreude	-.071	-1.531	.064	.005[a]	-.096	-2.163	**.016**	.009[a]
Kommunikationsvermögen	-.041	-0.875	.191	.001[a]	.004	0.086	.535	.000[b]
Geselligkeit	-.013	-0.306	.380	.000[a]	.036	0.788	.785	.001[b]
Einfühlungsvermögen	-.024	-0.470	.320	.001[a]	.044	0.906	.817	.002[b]
Erfolgszuversicht	-.035	-0.875	.191	.001[a]	-.013	-0.351	.363	.000[a]
Aufgeschlossenheit und Neugier	-.046	-0.953	.171	.002[a]	.001	0.012	.505	.000[b]
Eigeninitiative	.022	0.489	.687	.000[b]	.031	0.750	.723	.001[b]
Leistungsstreben und Erfolgsmotivation	-.055	-1.378	.085	.003[a]	.009	0.228	.590	.000[b]

Untersuchungsgruppe 2A, $N = 367$

β: standardisiertes Regressionsgewicht für den Moderatoreffekt, T: Teststatistik des Signifikanztests für β, p: Signifikanzniveau (einseitig), ΔR^2: Effektstärke (Änderung von R^2)

[a] Effekt in erwarteter Richtung, [b] Effekt in nicht erwarteter Richtung, p (einseitig) < .05 fett gedruckt

Tabelle 35: Der Einfluss von Variabilität und ERS auf die Split-Half-Reliabilität der Skalen der Vertriebsversion des ITB-PESA im Nicht-Auswahl-Kontext

	Moderator Variabilität				Moderator ERS			
Moderation der Split-Halt-Reliabilität der Skala		T-Test				T-Test		
	β	T	p	ΔR^2	β	T	p	ΔR^2
Kontaktfreude	-.028	-0.796	.213	.001[a]	-.002	-0.046	.432	.000[a]
Kommunikationsvermögen	.024	0.560	.712	.001[b]	.058	1.409	.920	.003[b]
Geselligkeit	-.064	-1.519	.065	.004[a]	-.052	-1.267	.103	.002[a]
Einfühlungsvermögen	-.041	-0.875	.191	.001[a]	.016	0.353	.638	.000[b]
Erfolgszuversicht	-.030	-0.895	.186	.001[a]	-.048	-1.419	.079	.002[a]
Aufgeschlossenheit und Neugier	-.088	-2.032	**.021**	.007[a]	-.067	-1.570	.059	.004[a]
Eigeninitiative	-.057	-1.323	.094	.003[a]	-.074	-1.799	**.036**	.005[a]
Leistungsstreben und Erfolgsmotivation	-.022	-0.556	.290	.000[a]	.032	0.788	.785	.001[b]

Untersuchungsgruppe 1A, $N = 405$

β: standardisiertes Regressionsgewicht für den Moderatoreffekt, T: Teststatistik des Signifikanztests für β, p: Signifikanzniveau (einseitig), ΔR^2: Effektstärke (Änderung von R^2)

[a] Effekt in erwarteter Richtung, [b] Effekt in nicht erwarteter Richtung, p (einseitig) < .05 fett gedruckt

[34] Ausgegangen wird von der in Anhang C vorgestellten Alpha-Adjustierung. Bonferroni-Holm-korrigiert (Holm, 1979) ist ebenfalls keiner der Moderatoreffekte signifikant.

[35] Die Bonferroni-Holm-Korrektur (Holm, 1979) führt auch hier zum selben Ergebnis.

Hinsichtlich der Kriteriumsvalidität im Auswahlkontext wurde geprüft, ob Variabilität und ERS die Vorhersage der Anzahl zur letzten Geburtstagsfeier eingeladener Gäste durch die Skala „Kontaktfreude“ moderieren. Dazu wurden ebenfalls moderierte multiple Regressionen für Untersuchungsgruppe 2B durchgeführt. Die Ergebnisse werden in Tabelle 36 angeführt: Die Skala Kontaktfreude sagt einen substanziellen Anteil der Kriteriumsvarianz vorher, auch Variabilität und ERS verbessern die Vorhersage der Anzahl zur letzten Geburtstagsfeier eingeladener Gäste. Moderatoreffekte zeigen sich hingegen nicht: Weder Variabilität noch ERS haben einen Einfluss darauf, wie gut sich die Anzahl zur letzten Geburtstagsfeier eingeladener Gäste durch die Skala „Kontaktfreude“ vorhersagen lässt.

Tabelle 36: Moderierte multiple Regressionen zur Vorhersage der Anzahl zur letzten Geburtstagsfeier eingeladener Gäste mit dem Prädiktor „Kontaktfreude“ und dem Moderator Variabilität bzw. ERS aus dem ITB-PESA

Regressionen zur Vorhersage der Anzahl zur letzten Geburtstagsfeier eingeladener Gäste	B	s_E	β	T-Test T	T-Test p	R^2	ΔR^2
Schritt 1						.060	
Kontaktfreude	3.940	1.023	.205	3.853	**<.001**[a]		
Variabilität in der ITB-PESA-Vertriebsversion	2.067	1.024	.107	2.018	**.044**[b]		
Schritt 2						.062	.002
Kontaktfreude (Kf)	4.043	1.030	.222	3.925	**<.001**[a]		
Variabilität in der ITB-PESA-Vertriebsversion (V)	1.919	1.039	.100	1.215	.066[b]		
Kf x V	-0.855	0.989	-.046	-0.282	.194[a]		
Schritt 1						.074	
Kontaktfreude	2.695	1.130	.140	2.386	**.009**[a]		
ERS in der ITB-PESA-Vertriebsversion	3.423	1.137	.177	3.010	**.003**[b]		
Schritt 2						.075	.001
Kontaktfreude (Kf)	2.719	1.132	.141	2.403	**.009**[a]		
ERS in der ITB-PESA-Vertriebsversion (ERS)	3.486	1.144	.180	3.047	**.002**[b]		
Kf x ERS	0.088	0.054	-.029	-0.544	.294[a]		

Untersuchungsgruppe 2B, $N = 343$

Die Prädiktoren wurden vor der Analyse z-standardisiert, das Produkt aus Prädiktor und Moderator wurde aus den z-standardisierten Werten berechnet.

B: Regressionsgewicht, s_E: Standardfehler des Regressionsgewichts, β: standardisiertes Regressionsgewicht; Signifikanztests für β (T-Test): Teststatistik T, Signifikanzniveau p (einseitig); R^2: Effektstärke (Determinationskoeffizient), ΔR^2: Änderung des Determinationskoeffizienten

p < .05 fett gedruckt, [a] einseitige Testung (für Effekt in erwarteter Richtung), [b] zweiseitige Testung

6.3 Diskussion

In Studie 2 wurden Daten zum Ernstfalleinsatz eines Facetten-Fragebogens, der Vertriebsversion des ITB-PESA, zur Personalauswahl berichtet. Zusätzlich wurden Vergleichsdaten zum selben Messinstrument aus dem Nicht-Auswahl-Kontext (Daten aus der ersten Untersuchung von Studie 1) herangezogen. Mit diesen zwei Datensätzen wurden Hypothesen in drei Bereichen getestet: hinsichtlich der Erfassung von Variabilität, hinsichtlich der Erklärung von ERS und hinsichtlich des Einflusses von Variabilität und ERS auf die Reliabilität und Validität von Persönlichkeitsfragebogen. Die Erwartungen hinsichtlich der univariaten und der multivariaten Verteilungen der Skalen-Scores des ITB-PESA bei der Auswahl wurden im Wesentlichen erfüllt: Die Mittelwerte waren bei der Auswahl größtenteils höher als im Nicht-Auswahl-Kontext, die Streuungen waren stellenweise geringer und die Skaleninterkorrelationen waren zum Teil höher.

Variabilität wurde wie in Studie 1 als Faktor der korrigierten Inter-Item-SD erfasst. Es zeigte sich, dass die Korrektur der Inter-Item-SD bei der Auswahl stärker ausfällt als im Nicht-Auswahl-Setting. Dies geht auf den höheren Zusammenhang des Item-Mittelwerts mit der Inter-Item-SD im Auswahlsetting zurück: Da die Gruppen-Mittelwerte der Skalen bei der Auswahl deutlich höher liegen als im Nicht-Auswahl-Kontext und entsprechend weiter von der Mitte der Likert-Skala entfernt sind, treten stärkere negative Korrelationen zwischen den Item-Mittelwerten und den Inter-Item-SD auf. Die Korrelation geht also nicht auf den Inhalt der jeweiligen Skala zurück, sondern auf die Verteilung der Messwerte bzw. auf methodische Restriktionen. Für den Variabilitäts-Faktor lässt sich sowohl im Auswahl- als auch im Nicht-Auswahl-Kontext eine hohe interne Konsistenz nachweisen. Die korrigierten Inter-Item-SD laden jeweils hoch auf diesem Faktor und der Faktor klärt einen großen Teil ihrer Varianz auf. Es ist davon auszugehen, dass Variabilität reliabel und valide erfasst wurde.

Bei der Untersuchung von ERS wurde an die Ergebnisse aus Studie 1 angeknüpft. Erwartungskonform zeigte sich, dass Personen bei der Auswahl häufiger extrem antworten als im Nicht-Auswahl-Kontext, dass ERS am stärksten von Variabilität abhängt und dass ERS nicht nur Indikator von Variabilität ist, sondern auch einen den Fragebogenskalen zugrunde liegenden Faktor widerspiegelt. Vier Ergebnisse stützen diese Interpretation: Erstens korrelieren die Faktoren für Variabilität und ERS auch bei der Auswahl sehr hoch, jedoch niedriger

als im Nicht-Auswahl-Setting. Zweitens hatte ein SEM mit zwei Faktoren (Variabilität und ERS) einen besseren Fit als eines mit einem Faktor (nur Variabilität) und dieser Unterschied fällt bei der Auswahl stärker aus. Drittens korreliert ERS bei der Auswahl – d. h. in einem Setting, in dem der erste Faktor der eingesetzten Fragebogenskalen stärker ist – höher mit diesem Faktor als im Nicht-Auswahl-Kontext. Viertens wird ERS im Auswahl- und Nicht-Auswahl-Setting etwa gleich gut von Variabilität und dem gemeinsamen Faktor aufgeklärt, Variabilität sagt ERS im Nicht-Auswahl-Setting besser vorher als im Auswahl-Setting und der gemeinsame Faktor der Skalen sagt ERS besser im Auswahl- als im Nicht-Auswahl-Kontext vorher. Für beide Settings zeigen Regressionen, dass ERS von beiden Einflussfaktoren abhängt; der Teil, den der gemeinsame Faktor der Skalen über Variabilität hinaus vorhersagt, ist im Auswahl-Setting größer.

Auch in Studie 2 wurden die Einflüsse von Variabilität und ERS auf die Split-Half-Reliabilität untersucht. Die Übereinstimmung von Profilen, die jeweils mit der Hälfte der Items gemessen wurden, mit Profilen, die mit der anderen Hälfte der Items gemessenen wurden, hing im Auswahl- und Nicht-Auswahl-Setting von Variabilität, nicht aber von ERS ab. Variabilität moderiert also die Split-Half-Reliabilität des Profils in der Vertriebsversion des ITB-PESA: je höher die Variabilität, desto niedriger die Split-Half-Reliabilität. Dass der Effekt nicht für ERS auftrat, könnte daran liegen, dass der ERS-Faktor Variabilität weniger valider erfasst als der Variabilitäts-Faktor. Auf Ebene der Skalen konnte kein Moderatoreffekt nachgewiesen werden: Weder Variabilität noch ERS moderierten den Zusammenhang zwischen einer Skalenhälfte und der jeweils anderen. Als weiteres Gütekriterium wurde die Kriteriumsvalidität untersucht. Die Skala „Kontaktfreude“ klärt einen Teil der Varianz des Kriteriums „zur letzten Geburtstagsfeier eingeladene Gäste“ auf, diese Korrelation wird weder von Variabilität noch von ERS moderiert – die Ergebnisse zur Kriteriumsvalidität aus Studie 1 wurden bei der Auswahl nicht repliziert.

7 Allgemeine Diskussion

In zwei empirischen Studien wurden Variabilität, ERS sowie deren Einflüsse auf die Reliabilität und die Validität von Persönlichkeitsfragebogen systematisch untersucht. In den folgenden Abschnitten werden die Befunde vor dem Hintergrund der Hypothesen interpretiert und eingeordnet (Abschnitt 7.1), die Beschränkungen aufgeführt und ein Ausblick gegeben (Abschnitt 7.2) sowie ein Fazit gezogen (Abschnitt 7.3).

7.1 Interpretation und Einordnung der Befunde

Im Rahmen der vorliegenden Arbeit wurden Beiträge in drei Bereichen erbracht: zur Erfassung und Struktur von Variabilität (Abschnitt 7.1.1), zur Erklärung von ERS (Abschnitt 7.1.2) sowie zu den Effekten von Variabilität und ERS auf die Reliabilität und Validität von Persönlichkeitsfragebogen und auf die Zusammenhänge zwischen und die Stabilität von Persönlichkeitseigenschaften (Abschnitt 7.1.3). Im Folgenden werden die Befunde jeweils zunächst im Hinblick auf die Hypothesen zusammengefasst und interpretiert. Danach werden jeweils ihre Implikationen und schließlich mögliche Einschränkungen diskutiert.

7.1.1 Die Erfassung und Struktur von intraindividueller Variabilität

Hinsichtlich der Erfassung und der Struktur von Variabilität wurden drei Hypothesen aufgestellt, von denen alle drei als bestätigt angesehen werden können:

H1A: Durch die Korrektur der Inter-Item-SD um den Item-Mittelwert wird die Validität der Messungen von Variabilität verbessert.

H1B: Durch die Korrektur der Inter-Item-SD um die Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert (d. h. um das Quadrat des z-standardisierten Mittelwerts) wird die Validität der Messung von Variabilität verbessert.

H1C: Variabilität ist ein eindimensionaler globaler Trait, der sich – anders als von der Metatraits-Theorie impliziert – nicht auf einzelne Traits bezieht.

Hypothesen 1A und 1B: Die Korrekturen der Inter-Item-SD um den Item-Mittelwert und dessen Abweichung vom Gruppen-Mittelwert erhöhen die Validität der Messung von Variabilität. Wenn der Item-Mittelwert einer Skala und dessen Abweichung vom Gruppen-Mittelwert

mit der Inter-Item-SD zusammenhängen, ist dies Folge methodischer Restriktionen. Dies wurde in beiden Studien deutlich: In Studie 1 zeigte sich, dass mit den Korrekturen keine bedeutsame Varianz der Inter-Item-SD herauspartialisiert wird; denn die Inter-Item-SD ähnlicher Traits hängen in unterschiedlichem Ausmaß vom Item-Mittelwert bzw. von dessen Abweichung vom Gruppen-Mittelwert ab. Würden die Inter-Item-SD um bedeutsame, d. h. konstruktrelevante, Anteile bereinigt werden, so würde die Inter-Item-SD von Skalen, die Ähnliches messen, auch in ähnlichem Ausmaß korrigiert werden. In Studie 2 lagen die Gruppen-Mittelwerte im Auswahlsetting weiter oberhalb der Mitte der Likert-Skala als im Nicht-Auswahl-Setting, und die Inter-Item-SD und die Item-Mittelwerte hingen auch stärker zusammen als im Nicht-Auswahl-Setting. Der Zusammenhang kann also auf die Verteilung der Messwerte relativ zur Likert-Skala zurückgeführt werden. Auf den ersten Blick überraschend erscheint der Befund, dass der Zusammenhang zwischen Inter-Item-SD und der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert bei nur einer der acht Skalen im Auswahlkontext kleiner ist als im Nicht-Auswahl-Kontext. Zu erwarten war nämlich, dass diese Korrelation kleiner wird, je weiter der Gruppen-Mittelwert von der Mitte der Likert-Skala entfernt liegt. Auf den zweiten Blick wirkt dieser Befund jedoch weniger verwunderlich, weil die Gruppen-Mittelwerte der Skalen bereits im Nicht-Auswahl-Setting deutlich oberhalb der Mitte der Likert-Skala liegen. Lediglich der Mittelwert der ITB-PESA-Skala „Kontaktfreude“ ($M = 3.48$) liegt im Nicht-Auswahl-Setting nahe der Mitte der Likert-Skala ($M = 3.50$) und bei dieser Skala ist die Korrelation zwischen der Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert mit der Inter-Item-SD im Nicht-Auswahl-Setting auch stärker (negativ) als im Auswahlsetting.

Hypothese 1C: Variabilität ist nicht bezogen auf den Trait, für dessen Indikatoren sie berechnet wird. Dies wurde in Studie 1 gezeigt: Die (um den Item-Mittelwert und dessen Abweichung von Gruppen-Mittelwert) korrigierten Inter-Item-SD korrelieren nicht höher für Skalen, mit denen ähnliche Merkmale erfasst werden, als für Skalen, mit denen verschiedene bzw. voneinander unabhängige Merkmale erfasst werden. Das bedeutet Variabilität ist nicht mit dem erfassten Merkmal verknüpft, sondern universell.

Fazit: Variabilität ist ein globaler, eindimensionaler Trait und nicht – wie von der Metatraits-Theorie impliziert – Trait-gebunden. Variabilität lässt sich valide erfassen, indem die Inter-Item-SD der Skalen um den Item-Mittelwert und um dessen Abweichung vom Gruppen-

Mittelwert korrigiert werden und aus den korrigierten Werten ein gemeinsamer Index – zum Beispiel ein Faktor – berechnet wird.

Implikationen

Die Abschnitte 2.1.4, 2.2.2 und 2.2.3 deuten an, welchen Stellenwert das Phänomen Variabilität einnehmen kann: Variabilität ist das Bindeglied zwischen personalen und situativen Determinanten von Verhalten (Fleeson, 2004). Bislang bedingten interaktionistische Ansätze das Abebben der Person-Situation-Debatte. Untersucht wurde auch, unter welchen Bedingungen die Situation und unter welchen Bedingungen die Person das Verhalten erklären kann. Die Forschung zu Variabilität zeigt, dass nicht nur die Merkmale der Situation darüber entscheiden, welcher Faktor bei der Erklärung von Verhalten wichtig ist, sondern auch Merkmale der Person: Bei einigen Personen wird das Verhalten eher durch die Situation, bei anderen eher durch eine Persönlichkeitseigenschaft bestimmt. Verglichen mit der Bedeutung dieser Befunde wurden Variabilität und seine Erfassung, insbesondere in Persönlichkeitsfragebogen, bislang nicht ausreichend beschrieben.

Unklar war, wie Variabilität erfasst werden soll und welche Struktur Variabilität aufweist. Zwar lagen bislang plausible und theoretisch brauchbare Empfehlungen dafür vor, wie Variabilität erfasst werden soll (Baird et al., 2006; Reddock et al., 2011); ein Beleg dafür, dass die Inter-Item-SD ein valides Maß darstellt, wenn sie um den Einfluss des Item-Mittelwerts einer Skala und dessen Abweichung vom Gruppen-Mittelwert korrigiert wird, fehlte jedoch. Ebenso fehlte ein Beleg für die Eindimensionalität von Variabilität, der über eine hohe interne Konsistenz von Variabilitätsmaßen für verschiedene Skalen hinausgeht. Beide Lücken werden mit den vorliegenden Ergebnissen geschlossen. Die Annahme der Metatraits-Theorie, dass für jeden Trait ein Metatrait existiert (siehe Abschnitt 2.1.1), lässt sich also nicht halten. Die Ergebnisse bezüglich der Eindimensionalität stehen den Ergebnissen von Eid und Diener (1999) entgegen: Eid und Diener demonstrierten mit SEM, dass Variabilität für Emotionen ein multidimensionales Konstrukt ist. Das Ergebnis ist womöglich darauf zurückzuführen, dass die Autoren die Inter-Item-SD nicht korrigierten (Abschnitt 2.1.4). Schließlich handelt es sich bei Emotionen auch um stabile Dispositionen und somit sollten die Ergebnisse nicht von den in dieser Arbeit berichteten abweichen.

Im Hinblick auf die Studie von Paunonen (1988) lässt sich auf Basis der Ergebnisse der vorliegenden Arbeit schlussfolgern, dass in Persönlichkeitsfragebogen beobachtete Variabilität (implizites Maß, korrigierte Inter-Item SD) nicht mit der subjektiv wahrgenommenen Variabilität (explizites Maß, Selbstbericht) einhergeht. Paunonen stellte fest, dass selbstberichtete Variabilität mit mittlerer Trait-Ausprägung und selbstberichtete Konsistenz mit extremer Trait-Ausprägung einhergeht. Im Gegensatz zur Studie von Paunonen wurde Variabilität hier implizit aus dem Antwortverhalten erschlossen. Entspräche dieses erschlossene Maß dem subjektiven, so würde die Korrektur der Inter-Item-SD um die Abweichung des Item-Mittelwerts vom Gruppen-Mittelwert die Validität des Variabilitätsindex senken und Variabilität wäre multidimensional. Beides wurde widerlegt. Die Einschätzung, wie wichtig eine Persönlichkeitseigenschaft für das eigene Verhalten ist, ist also nicht deckungsgleich mit Variabilität, die Aufschluss über den tatsächlichen Einfluss von Persönlichkeitseigenschaften auf das Verhalten gibt. Die Höhe des Zusammenhangs beider Merkmale lässt sich basierend auf den vorliegenden Ergebnissen nicht abschätzen. Dass die selbsteingeschätzte Variabilität mit extremen Trait-Ausprägungen einhergeht, könnte auf Merkmale der Sprache zurückzuführen sein: Schließlich sind Eigenschaftswörter in der Regel auf einen Pol einer Eigenschaftsdimension bezogen; zum Abstufen müssen meist Adverbien verwendet werden. Im Einklang mit dieser Beobachtung sind die Extremen der Eigenschaftsdimensionen vielleicht salienter als mittlere Ausprägungen, die womöglich schlechter repräsentiert werden. Folglich kommen Personen mit mittlerer Ausprägung auf einer Eigenschaftsdimension (z. B. „etwas kontaktfreudig“ oder „ein wenig schüchtern“) eher selten zu dem Urteil, dass sie sich bezogen auf diese Dimension konsistent verhalten, verglichen mit Personen, die extreme Eigenschaftsausprägungen auf dieser Dimension aufweisen (z. B. „kontaktfreudig“ oder „schüchtern“).

Mögliche Einschränkungen

Trotz der Belege, dass die Korrekturen der Inter-Item-SD bei der Messung von Variabilität die Validität erhöhen, bleiben – wie in der Studie von Baird et al. (2006; Abschnitt 2.1.4) – Zweifel, ob nicht zu viel korrigiert wird: Denn nach der Korrektur sind die Variabilitätsindizes der einzelnen Skalen jeweils unabhängig vom gemessenen Merkmal. Das bedeutet Korrelationen zwischen Persönlichkeitsmerkmalen und Variabilitätsindizes sind allein aufgrund der Messmethode klein. In Fällen, in denen Variabilität für die Skalen eines breiten Fragebogens aggregiert wird, sollten sich Zusammenhänge mit Persönlichkeitsmerkmalen dennoch zeigen:

Im HEXACO-PI-R können fünf der sechs korrigierten Inter-Item-SD mit einer Persönlichkeitsdimension korrelieren. So könnten die korrigierten Inter-Item-SD der fünf anderen Dimensionen mit Extraversion korrelieren, sofern ein Zusammenhang zwischen Extraversion und Variabilität besteht; und wenn der Zusammenhang für fünf der korrigierten Inter-Item-SD auftritt, zeigt er sich wahrscheinlich auch für das Aggregat aller sechs Indizes.

Ein anderes Bild wäre zu erwarten, wenn ein Fragebogen eingesetzt wird, deren Skalen einen ähnlichen Messbereich haben. In dem Fall hängt Variabilität bereits aufgrund der Messmethode *nicht* mit einem der erfassten Merkmale zusammen. Aus mehreren Gründen ist diese Einschränkung jedoch hier zu vernachlässigen: Erstens wurde in Studie 1 gezeigt, dass keine bedeutsame Varianz aus den Inter-Item-SD eliminiert wurde; dies wäre nicht der Fall, wenn Variabilität hoch mit einer der gemessenen Persönlichkeitseigenschaften zusammenhinge. Dann nämlich wäre zumindest die Korrektur der Inter-Item-SD um den Item-Mittelwert der Skalen, die die betreffende Eigenschaft messen, im Sinne der Validität des Variabilitätsindex unangemessen. Zweitens sind nennenswerte Korrelationen zwischen Persönlichkeitseigenschaften und Variabilität aufgrund der Konzeption als Eigenschafts-Verhaltens-Kontingenz (siehe Abschnitt 2.2.3) unplausibel; auch Baird et al. (2006), Biderman und Reddock (2012) sowie Reddock et al. (2011) berichten sehr schwache oder keine Zusammenhänge ($|r| < .20$). Drittens korreliert Variabilität in der Vertriebsversion des ITB-PESA (in Studie 2 berichtete Fragebogenversion, Daten aus der ersten Untersuchung von Studie 1; Untersuchungsgruppe 1A), deren Skalen ein starker erster Faktor zugrunde lag, im Nicht-Auswahl-Kontext $r_{(405)} = .924$ mit Variabilität im gesamten ITB-PESA (in Studie 1 berichtete Fragebogenversion) und $r_{(405)} = .849$ mit Variabilität im HEXACO-PI-R.

Hinsichtlich der Messung von Variabilität ist auch einzuwenden, dass in der vorliegenden Arbeit Unterschiede in der „Schwierigkeit“ (bzw. dem Gruppen-Mittelwert) der Items Einfluss auf die Variabilität haben könnten. Als Beispiel dienen zwei Items, die einen Gruppen-Mittelwert von $M_1 = 3$ und $M_2 = 5$ aufweisen. Eine Person, die beide Items mit „4“ ankreuzt und damit einmal unterhalb und einmal oberhalb des Gruppen-Mittelwerts liegt, hat eine geringere Inter-Item-SD als eine Person, die das erste Item mit „3“ und das zweite Item mit „5“ beantwortet und beide Male relativ zur Referenzgruppe dieselbe Antwort gibt. Dieser Einwand ist sicher theoretisch haltbar und möglicherweise in Fällen relevant, in denen die Itemschwierigkeiten (bzw. Gruppen-Mittelwerte) stark streuen. Wie die Daten von Stu-

die 1 zeigen, haben sie aber keinen Einfluss auf die Ergebnisse: Der Faktor der korrigierten Inter-Item-SD im gesamten ITB-PESA korreliert $r_{(405)} = .969$ mit dem Faktor von korrigierten Inter-Item-SD, bei denen die Unterschiede zwischen den Gruppen-Mittelwerten der Items herausgerechnet wurden. Im HEXACO-PI-R beträgt die Korrelation $r_{(405)} = .924$ (Untersuchungsgruppe 1A).

Der Verdacht, dass Variabilität bei der Auswahl nicht valide erfasst wird, lässt sich ebenfalls ausräumen: Die vorangehende Argumentation macht deutlich, dass Variabilität nicht oder in sehr geringem Ausmaß mit Persönlichkeitsmerkmalen korreliert und unabhängig vom zu erfassenden Trait zu beobachten ist. Selbst wenn die Konstruktvalidität des Persönlichkeitsfragebogens beeinträchtigt ist, werden mit den Likert-Items bedeutsame interindividuelle Differenzen auf einer Persönlichkeitseigenschaft abgebildet, zum Beispiel in Form von sozialer Erwünschtheit (Marcus, 2003) oder in Form des IEF (Klehe et al., 2012). Auch bei diesen Messungen sollte sich Variabilität zeigen und auch bei diesen Messungen sollte die Validität ihrer Messung von der in dieser Arbeit bestätigten Korrekturmethode profitieren.

Ein weiterer denkbarer Kritikpunkt bezieht sich auf die Methode der ersten Untersuchung von Studie 1: Man könnte vermuten, dass die Wiederholung von Iteminhalten – die sich bei großer Itemzahl nicht ausschließen lässt – sich auf die Variabilität zwischen Items auswirkt. Tatsächlich berichten Baird und Lucas (2011), dass die Variabilität für Items, die für mehrere verschiedene Rollen präsentiert werden, höher ausfällt als für Items, die nur für wenige Rollen präsentiert werden. Offenbar interpretieren Personen die wiederholte Vorgabe desselben Items als Aufforderung, neue Information preiszugeben. Dass dies in den Studien dieser Arbeit keine Rolle gespielt hat, lässt sich anhand der Daten ablesen: Zum einen wurde Variabilität im Auswahlkontext mit einem kurzen Fragebogen in ähnlicher Weise mit vergleichbaren Kennwerten erfasst wie in Studie 1 mit längeren Fragebogen. Zum anderen unterscheiden sich die korrigierten Inter-Item-SD von Skalen mit hoher interner Konsistenz (und potenziell höherer Item-Redundanz) nicht von korrigierten Inter-Item-SD von Skalen mit niedrigerer interner Konsistenz; dies betrifft insbesondere ihre Ladungen auf dem gemeinsamen Variabilitäts-Faktor.

Die Störgröße „Reihenfolge-Effekte“ lässt sich ebenfalls ausschließen: In der ersten Untersuchung von Studie 1 wurden die Items von ITB-PESA und HEXACO-PI-R gemeinsam dargebo-

ten. In der Retest-Untersuchung wurden ausschließlich die Items des ITB-PESA, jedoch in einer anderen Reihenfolge, dargeboten – und trotzdem waren die Korrelationen des Variabilitäts-Faktors der Retest-Untersuchung mit den beiden der ersten Untersuchung sehr hoch.

7.1.2 Zur Erklärung von Extreme Response Style

Im Hinblick auf die Ursachen von ERS wurde folgende Hypothese geprüft und bestätigt:

H2: ERS ist Indikator von intraindividueller Variabilität.

ERS wird maßgeblich von Variabilität bestimmt: Personen mit hoher Variabilität geben häufiger extreme Antworten als Personen mit niedriger Variabilität. Dies wurde im Rahmen beider Studien deutlich. In Studie 1 wurde dies mit hohen Korrelationen zwischen den ERS- und den Variabilitäts-Faktoren sowie mit SEM gezeigt. In den SEM wurden jeweils die korrigierten Inter-Item-SD und die Extremwerthäufigkeiten der ITB-PESA- und der HEXACO-PI-R-Skalen durch latente Faktoren erklärt. Unter den Modellen, in die die Extremwerthäufigkeiten der Skalen des ITB-PESA eingingen, passten die Daten besser zu einem Zwei-Faktoren-Modell (mit ERS- und Variabilitäts-Faktor) als zu einem Ein-Faktoren-Modell. In Modellen, in denen die Extremwerthäufigkeiten der Skalen des HEXACO-PI-R repräsentiert werden, zeigte ein Ein-Faktoren-Modell jeweils einen mindestens ebenso guten Fit wie ein Zwei-Faktoren-Modell. Ausgehend von diesem Ergebnis wurde vermutet, dass ERS sich u. U. auch auf einen dominanten Faktor in einem Fragebogen zurückführen lässt – also extreme Antworten auch Indikator von extremen Ausprägungen sind. Denn den Skalen des ITB-PESA liegt ein dominanter Faktor zugrunde, auf dem die meisten Faktoren laden und der zwischen den Dimensionen Gewissenhaftigkeit und Extraversion angesiedelt ist. Ferner liegt der Gruppen-Mittelwert der meisten Skalen des ITB-PESA oberhalb der Mitte der Likert-Skala; es existieren mehr positive als negative Extremantworten. Entsprechend geht extreme Zustimmung auf den Skalen des ITB-PESA in gewissem Ausmaß mit extremer Ausprägung auf den gemessenen Eigenschaften einher. Mit dem HEXACO-PI-R werden dagegen sechs weitgehend unabhängige Dimensionen erfasst, so dass sich kein dominanter Faktor zeigt und extreme Antworten, die auf extreme Ausprägung zurückzuführen sind, „herausgemittelt“ werden.

Diese Erklärungsmöglichkeit wurde in Studie 2 überprüft: In dieser Studie wurde die Vertriebsversion des ITB-PESA eingesetzt, die einen deutlich stärker ausgeprägten ersten Faktor

aufweist, der vorwiegend Extraversion umfasst. Im Auswahlkontext fällt dieser Faktor noch dominanter aus; neben dem intendierten Eigenschaftsbereich bündelt er zusätzlich auch eine Selbstdarstellungstendenz – und zwar das Bestreben, sich als „idealer Mitarbeiter“ zu präsentieren (Klehe et al., 2012). Sowohl im Auswahlkontext als auch im Nicht-Auswahl-Kontext korrelierten Variabilität und ERS sehr hoch miteinander. Der größte Teil der Varianz von ERS wurde durch Variabilität aufgeklärt; bei der Auswahl war er etwas niedriger als im Nicht-Auswahl-Kontext. Wie in Studie 1 wurde geprüft, ob die Daten eher zu einem SEM mit einem oder zu einem SEM mit zwei Faktoren passen: In beiden Settings (Auswahl und Nicht-Auswahl) passten die Daten besser zum Zwei-Faktoren-Modell; der Unterschied zwischen Zwei-Faktoren- und Ein-Faktoren-Modell war bei der Auswahl größer als für das Nicht-Auswahl-Setting. Als nächstes wurde entsprechend des oben geschilderten Erklärungsansatzes je der erste Faktor der eingesetzten Skalen berechnet; dieser klärte im Auswahlsetting deutlich mehr Varianz der Skalen auf als im Nicht-Auswahl-Setting und korrelierte bei der Auswahl höher mit ERS als im Nicht-Auswahl-Setting. Mittels hierarchischer Regressionsanalysen wurde ersichtlich, dass ERS sich nahezu vollständig (~ 85 % der Varianz) von Variabilität und dem gemeinsamen Faktor der Fragebogenskalen erklären lässt. Im ersten Schritt wurde jeweils Variabilität als Prädiktor eingegeben; dieser klärte einen Großteil der Varianz von ERS auf. Im nächsten Schritt wurde der gemeinsame Faktor der Skalen eingegeben, der weitere Varianz von ERS aufklärte. Dieser weitere Varianzanteil war für das Auswahlsetting größer als für das Nicht-Auswahl-Setting.

Implikationen

Bislang war unklar, wie ERS sich erklären lässt und welche Rolle ERS in Fragebogen spielt. Daran hat sich trotz der Bestrebungen in verschiedenen Disziplinen in den letzten 20 Jahren nicht viel geändert. Noch 2006 schloss Weijters in seiner Dissertation zu Antworttendenzen:

> The same response category can have different meanings for different respondents. That is the essence of the response style problem as it has been conceptualized in the current dissertation. Response styles may be the cause that a given level of a latent construct of interest may lead to different levels of observed indicators. (S. 236)

Auf ERS trifft diese Schlussfolgerung nicht zu. ERS ist kein Bias und kein Antwortstil, sondern Indikator der globalen Persönlichkeitseigenschaft Variabilität. Extremes Antworten geht auf die Eigenschaft intraindividuelle Variabilität zurück, die nicht auf Fragebogen begrenzt ist.

Personen, deren Erleben und Verhalten variabler ist, beantworten auch Fragebogen variabler und erzielen somit häufiger Ausprägungen, die vom Item-Mittelwert der Skala abweichen. Damit einhergehend antworten sie auch häufiger extrem. Ihr Item-Mittelwert wird *nicht* verzerrt, möglicherweise nur etwas ungenauer gemessen.

In Fragebogen, in denen sich ein gemeinsamer Faktor extrahieren lässt und in denen die Gruppen-Mittelwerte von der Mitte der Likert-Skala abweichen, ist extremes Antworten auch Ausdruck von extremer Ausprägung auf dem den Fragebogenskalen zugrunde liegenden Faktor. Dabei spielt es keine Rolle, welche Eigenschaftsdimension der Faktor abbildet – die Zusammenhänge sollten universell sein. Das bedeutet, in Fragebogen, deren Skalen vorwiegend Facetten der Extraversion erfassten sind extreme Antworten auch Ausdruck von Extraversion; in Fragebogen, deren Skalen vorwiegend Facetten der Offenheit erfassen, sind extreme Antworten auch Ausdruck von Offenheit; usw. Dies zeigt sich auch, wenn der gemeinsame Faktor Eigenschaften abbildet, die ursprünglich nicht erfasst werden sollen – wie der IEF (Klehe et al., 2012) im Auswahlkontext. Letzteres Phänomen hat vermutlich auch in Studie 2 dazu geführt, dass der den Skalen zugrunde liegende Faktor bei der Auswahl dominanter als im Nicht-Auswahl-Kontext war. Zusammengefasst ist der Einfluss von den (tatsächlich) gemessenen Eigenschaften auf extremes Antworten größer, je dominanter der gemeinsame Faktor der Fragebogenskalen ausfällt und je schiefer die Skalen verteilt sind. Wenn der Mittelwert der Skalen oberhalb der Mitte der Likert-Skala liegt, entspricht der Großteil der extremen Antworten extremer Zustimmung, was eine Korrelation zwischen ERS und dem gemeinsamen Faktor der Skalen nach sich zieht. Die Ergebnisse lassen vermuten, dass ERS nicht – wie viele Forscher annehmen (Baumgartner & Steenkamp, 2001; Bolt & Newton, 2011; Van Vaerenbergh & Thomas, 2013) – Ursache einer Beeinträchtigung der Validität ist, sondern Folge dieser Beeinträchtigung. In anderen Worten führt nicht der Zusammenhang zwischen ERS und den Fragebogenskalen dazu, dass diese bei der Auswahl höher miteinander korrelieren als im Nicht-Auswahl-Setting. Vielmehr führt die höhere Korrelation zwischen den Skalen bei der Auswahl dazu, dass Maße von ERS in stärkerem Ausmaß einen gemeinsamen Messbereich der Skalen widerspiegeln als im Nicht-Auswahl-Kontext. Diese Vermutung wird stark durch die Ergebnisse aus Studie 2 gestützt: Die Korrelation von ERS und dem den Skalen zugrunde liegenden Faktor ist im Auswahlkontext, bei dem die Validität des Fragebogens beeinträchtigt ist und die Skalen stärker auf einem gemeinsa-

men Faktor laden, nämlich höher. Der Einfluss von Variabilität auf extremes Antworten ist bei der Auswahl entsprechend etwas geringer als im Nicht-Auswahl-Setting.

Aufgrund der zuvor geschilderten Konfundierung von ERS und den mit den Fragebogen zu messenden Eigenschaften war das Anliegen der meisten Forschungsansätze stets, ERS unabhängig von den zu messenden Merkmalen zu erfassen. Dies führte dazu, dass viele verschiedene Messmodelle zur Erfassung von ERS aufgestellt wurden, die im Wesentlichen darauf abzielten, die Wahl von Extremwerten isoliert zu betrachten (Bolt & Newton, 2011; Weijters et al., 2010b; Wetzel et al., 2013b). Die Erklärung von ERS durch Variabilität und die Schlussfolgerung, dass Variabilität kein Bias ist, haben ihrerseits Implikationen für die Erfassung von ERS: Erstens sollte stets Variabilität statt ERS erfasst werden. Andere Einflüsse auf ERS sind nicht stabil, sondern abhängig vom Fragebogen und den mit diesem gemessenen Merkmalen. Zweitens ist Variabilität eine kontinuierliche Merkmalsdimension, keine Klasse. Drittens müssen in Messmodelle nicht lediglich die Endpunkte als Indikatoren der Tendenz zu extremen Antworten eingehen, sondern alle Kategorien der Likert-Skala berücksichtigt werden: In SEM sollten die korrigierten Inter-Item-SD von Skalen eingehen. In IRT-Modellen zeigt sich Variabilität konform mit der Konzeption als Eigenschafts-Verhaltens-Kontingenz nicht als latente Klasse und nicht als Merkmalsdimension, die lediglich Einfluss auf die Wahrscheinlichkeit der Extremwertwahl hat.

Vielmehr sollten interindividuelle Unterschiede in der Variabilität und in der Tendenz zu extremen Antworten als Unterschiede in den Steigungen der Itemfunktionen operationalisiert werden. Abbildung 21 veranschaulicht diese Operationalisierung: Variabilität ist eine kontinuierliche Variable, die als Itemdiskriminationsparameter in das ordinale Rasch-Modell eingeht. Exemplarisch sind oben in der Abbildung die Itemfunktionen eines Items für Personen mit hoher Variabilität abgebildet. Für Personen mit hoher Variabilität und niedriger (bzw. hoher) Ausprägung im zu erfassenden Merkmal ist es weniger wahrscheinlich die niedrigste (höchste) Kategorie zu wählen; Personen mit hoher (niedriger) Ausprägung wählen dagegen wahrscheinlicher die niedrigste (höchste) Kategorie. Das Antwortverhalten hängt insgesamt etwas schwächer von der Ausprägung auf dem Merkmal ab als bei Personen mit niedriger Variabilität (Abbildung 21 unten), deren Antwortverhalten stark vom zu erfassenden Merkmal abhängt und die extreme Antwortkategorien nahezu ausschließlich bei extremer Aus-

prägung wählen. Mit geringer Variabilität einher geht entsprechend die Genauigkeit der Messung, mit hoher Variabilität einher geht größere Ungenauigkeit.

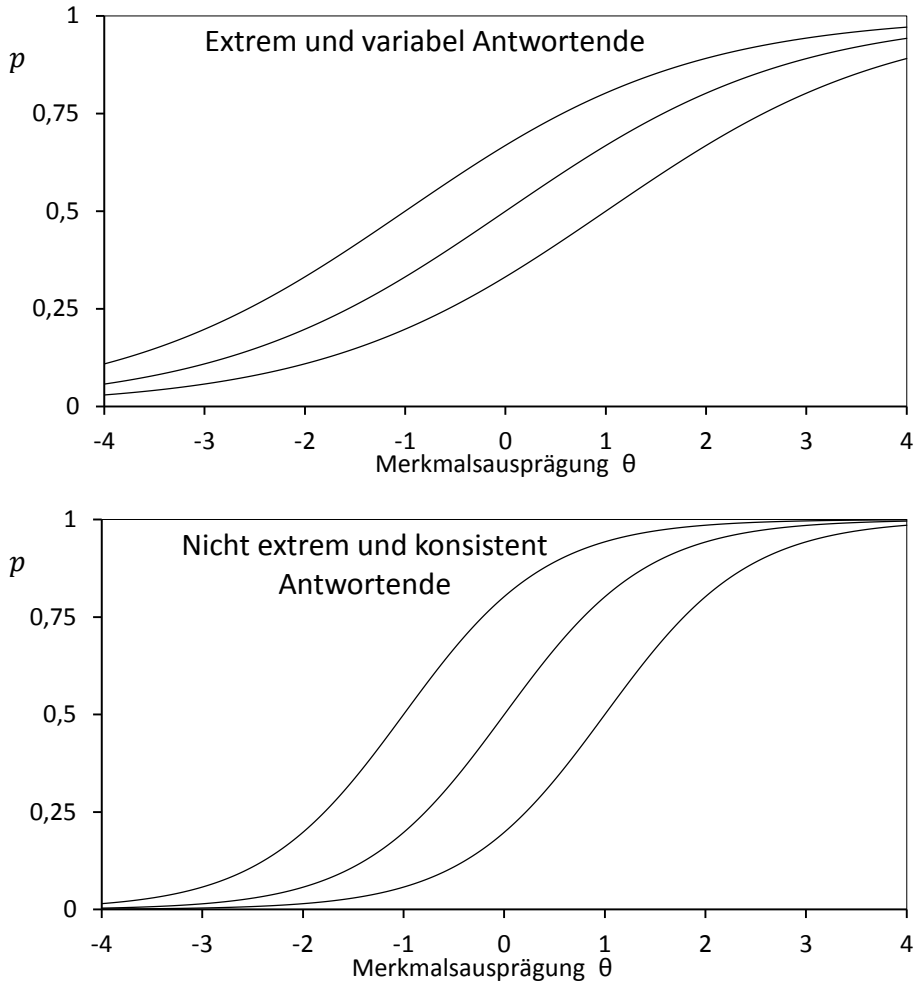

Abbildung 21: Operationalisierung von ERS bzw. Variabilität als Steigungsparameter im ordinalen Rasch-Modell
p: Wahrscheinlichkeit, dass Schwelle überschritten wird. Dargestellt sind die Schwellen von der ersten zur zweiten, von der zweiten zur dritten und der dritten zur vierten Antwortkategorie einer vier-stufigen Likert-Skala.

Mögliche Einschränkungen

Dass Variabilität reliabel und valide gemessen wurde, wurde hinreichend im vorangegangenen Abschnitt (7.1.1) erläutert; allerdings zeigt auch ERS in beiden Studien ähnliche Ergebnisse hinsichtlich der psychometrischen Qualität seiner Messung. Dies wirft die Frage auf: Warum wird hier Variabilität als Eigenschaft vorgestellt, die ERS zugrunde liegt, und nicht ERS als Grundlage von Variabilität, was Biderman und Reddock (2012) als mögliche Einschränkungen der Ergebnisse ihrer Studie einwerfen? Entgegnen lässt sich dieser Kritik sowohl auf theoretischer Ebene als auch auf Ebene der Daten: Variabilität ist eine Eigenschaft,

für die eine kohärente theoretische Grundlage besteht, die sich gut ins nomologische Netz einordnen lässt und die bei der Selbstbeschreibung von States und Traits sowie in Rollen-Fragebogen zu Persönlichkeitseigenschaften erfasst werden kann (Abschnitte 2.1.4 und 2.2). Die Vorhersagen, die über dieses Phänomen getroffen werden, wurden größtenteils bestätigt (Abschnitte 2.2 und 2.3 sowie die Ergebnisse der vorliegenden empirischen Studien), so dass sich folgern lässt: Variabilität ist ein gut untersuchtes und theoretisch fundiertes Konstrukt, das nicht nur in Fragebogen eine Rolle spielt. Von ERS lässt sich dies nicht behaupten. ERS ist lediglich operational definiert, an die Methode (Fragebogen) geknüpft und kann nicht unabhängig von dieser aufgezeigt werden. Wie sich Personen, die extrem antworten, von Personen, die weniger extreme Antworten geben, unterscheiden, konnte bis dato nicht näher beschrieben werden. Auch die Daten zeigen, dass extremes Antworten von Variabilität abhängt, dass aber diese Abhängigkeit nicht über unterschiedliche Kontexte hinweg konstant ist: Bei der Auswahl hängt ERS weniger von Variabilität ab als im Nicht-Auswahl-Setting. Dafür ist der Einfluss eines gemeinsamen Faktors des ITB-PESA bei der Auswahl größer als im Nicht-Auswahl-Kontext. Dies ist darauf zurückzuführen, dass die Skalen bei der Auswahl einen größeren gemeinsamen Messbereich aufweisen und dass sich extreme Antworten, die mit extremen Ausprägungen einhergehen, nicht nivellieren wie in einem Fragebogen, der unabhängige Dimensionen erfasst. Dies zeigt, dass dem Verhalten *extrem Antworten* nicht immer dieselbe Eigenschaft zugrunde liegt. Allerdings spiegeln interindividuelle Unterschiede im extremen Antworten, die nicht von den mit dem Fragebogen zu erfassenden Eigenschaften abhängen, stets interindividuelle Unterschiede im Merkmal Variabilität wider.

Eine weitere Kritik an der hier vertretenen Position knüpft an die Regressionsanalysen in Studie 2, in denen jeweils als erster Prädiktor Variabilität und als zweiter Prädiktor der gemeinsame Faktor der Skalen eingegeben wurde: Wäre nicht denkbar, dass Variabilität einen Teil von ERS aufklärt, der bereits vom gemeinsamen Faktor aufgeklärt wird? Aus rein statistischer Sicht ist dies möglich. Dass der gemeinsame Faktor des Fragebogens auf inhaltlicher Ebene einen Teil von ERS aufklärt, der von Variabilität aufgeklärt wird, ist aus mehreren Gründen fraglich: Erstens wird Variabilität methodisch unabhängig vom gemeinsamen Faktor gemessen (siehe Abschnitt 7.1.1); wenn Variabilität mit dem gemeinsamen Faktor korreliert, dann weil sie mit dem durch den Faktor gemessenen Konstrukt zusammenhängt. Zweitens wird ERS – sofern nicht nur durch hohe Merkmalsausprägung bedingt – als Indikator von

Variabilität verstanden. Die in dieser Arbeit vorgestellten Befunde über verschiedene Fragebogen und Settings hinweg bestätigen dies. Drittens ist der Einfluss des gemeinsamen Faktors der Skalen auf ERS abhängig von Setting und vom Fragebogen. Für die Abhängigkeit von Variabilität von diesem Faktor trifft dies nicht zu.

An den hier durchgeführten empirischen Studien kann auch kritisiert werden, dass ERS hier als Faktor der Extremwerthäufigkeiten von Skalen erfasst wurde und nicht auf einen der Ansätze zurückgegriffen wurde, ERS mittels statistischem Modell zu messen. Neben den Unzulänglichkeiten statischer Messmodelle von ERS, nämlich Theoriearmut, nicht überprüften Annahmen und fraglichem Mehrwert gegenüber konventionellen Ansätzen, sprechen weitere Punkte für die in dieser Arbeit verwendete Methode: Erstens sind die Extremwerthäufigkeiten Indikatoren von Verhalten; als Indizes sind sie nah an den Daten und gut untersucht. Zweitens wird mit dem Faktor zum einen das Gemeinsame der Extremwerthäufigkeiten erfasst, unabhängig von der Streuung der einzelnen Extremwerthäufigkeiten; zum anderen wird ERS parallel zu Variabilität operationalisiert. Drittens werden mit der relativ simplen Methode keine weiteren methodischen Fragen aufgeworfen. Die Methode baut so auf den spärlichen und dünnen Theorien über ERS auf und impliziert keine theoretischen Annahmen über ERS. Viertens sollte das Gemeinsame der Extremwertantworten methodeninvariant sein, so dass aus Gründen der Sparsamkeit eine einfache Methode zu bevorzugen ist.

7.1.3 Die Effekte von Variabilität auf die Gütekriterien von Persönlichkeitsfragebogen

Bezüglich der Effekte in Persönlichkeitsfragebogen wurde in den empirischen Studien der Einfluss von Variabilität und ERS auf die Reliabilität und Validität überprüft. Die entsprechenden Ergebnisse bestätigen die im vorangegangenen Abschnitt ausgeführte Schlussfolgerung: ERS ist Indikator von Variabilität. Daher werden im Folgenden die Befunde im Hinblick auf Variabilität erläutert. Die Hypothesen wurden für Variabilität aufgestellt und auch für ERS überprüft. Im Einzelnen wurde erwartet:

H3A: Variabilität hat einen Einfluss auf die Reliabilität von Persönlichkeitsfragebogen; bei Personen mit niedriger Variabilität ist die Reliabilität höher als bei Personen mit hoher Variabilität.

H3B: Der Zusammenhang zwischen zwei Konstrukten hängt *nicht* von der Variabilität ab.

H3C: Die Stabilität von Persönlichkeitseigenschaften hängt *nicht* von der Variabilität ab.

H3D: Der Zusammenhang zwischen Persönlichkeitseigenschaften und Kriterien hängt von der Variabilität ab; er ist bei Personen mit niedriger Variabilität stärker als bei Personen mit hoher Variabilität.

Hypothese 3A: Die Reliabilität von Persönlichkeitsfragebogen hängt nicht oder nur in geringem Ausmaß von Variabilität ab; die Hypothese wird also *nicht* bestätigt. Variabilität hatte einen geringen Einfluss auf die Split-Half-Reliabilität des Profils einiger der eingesetzten Fragebogen, darunter auch die vertriebsspezifische Version des ITB-PESA bei der Auswahl. Die Reliabilität der Skalen war nur vereinzelt von Variabilität abhängig, die Effekte waren sehr schwach. Wenn der ERS-Faktor einen Einfluss auf die Reliabilität hatte, hatte in der Regel auch der Variabilitätsindex einen Einfluss auf die Reliabilität, der den Einfluss von ERS erklären konnte.

Hypothese 3B: Variabilität hatte keinen Einfluss auf die Konstruktvalidität und somit auf die Zusammenhänge zwischen Skalen zu miteinander in Beziehung stehenden Merkmalen. Die Hypothese wird bestätigt.

Hypothese 3C: Die Stabilität von Persönlichkeitseigenschaften hängt nicht von Variabilität ab. Variabilität hatte keinen Einfluss auf die Stabilität des Profils des ITB-PESA und hing nur in wenigen Fällen mit der Retestreliabilität der Skalen zusammen (Studie 1). Dies wäre Voraussetzung für das Ablehnen der Hypothese gewesen. Somit wird die Hypothese bestätigt.

Hypothese 3D: Der Zusammenhang zwischen Persönlichkeitsmaßen und Kriterien wird in Studie 1 (Nicht-Auswahl-Setting) von Variabilität moderiert. Das „reine“ Variabilitätsmaß moderiert den Zusammenhang in beiden untersuchten Fällen signifikant. Der ERS-Faktor moderiert die Kriteriumsvalidität in einem der beiden Fälle nur marginal. Da die Reliabilität der Persönlichkeitsmaße, anhand derer Hypothese 3D überprüft wurde, nicht von Variabilität beeinflusst wird, lassen sich die Effekte auf die Persönlichkeitseigenschaften zurückführen: Variabilität moderiert nicht nur die Kriteriumsvalidität, sondern den Zusammenhang zwischen Eigenschaften und Arbeitszufriedenheit bzw. Studienerfolg. Dies steht im Einklang mit der Konzeption von Variabilität als Situations- bzw. Eigenschafts-Verhaltens-Kontingenz. Im Auswahlsetting (Studie 2) tritt der Moderatoreffekt zwar nicht auf, dies könnte jedoch

daran liegen, dass das Kriterium nicht gut zum verwendeten Prädiktor passt oder dessen Konstruktvalidität bei der Auswahl leicht beeinträchtigt ist. Da die Effekte für den Nicht-Auswahl-Kontext deutlich sind, gilt Hypothese 3D im Wesentlichen als bestätigt.

Implikationen

Der wichtigste Schluss, der in dieser Arbeit gezogen werden kann, lautet: Variabilität ist Moderator des Zusammenhangs zwischen Eigenschaften und *Life Outcomes*. Arbeitszufriedenheit und Studienerfolg lassen sich bei Personen mit niedriger Variabilität besser anhand von Persönlichkeitseigenschaften vorhersagen als bei Personen mit hoher Variabilität. Während dieser Moderatoreffekt für den Zusammenhang von Eigenschaften und States bzw. Eigenschaften und Verhalten (Eid & Diener, 1999; Fleeson, 2007) gesichert scheint, lagen für Persönlichkeitsfragebogen keine eindeutigen Erkenntnisse vor. Im Gegensatz zur Studie von Biderman und Reddock (2012, siehe Abschnitt 2.3) sind die in Studie 1 gezeigten Effekte linear: je variabler, desto schlechter die Vorhersage. Dieser Befund ist beschränkt auf den Nicht-Auswahl-Kontext.

Die weniger gut theoretisch begründeten Hypothesen, die Forscher über den Einfluss von Variabilität (oder Metatraits) auf die Konstruktvalidität von Persönlichkeitsfragebogen oder die Stabilität von Persönlichkeitseigenschaften aufgestellt haben (z. B. Baird et al., 2006; Biderman & Reddock, 2012; Britt, 1993) können auf Basis der vorliegenden Ergebnisse abgelehnt werden: Variabilität hat jeweils keinen Einfluss. Die Hypothese, dass die Reliabilität von Trait-Messungen von Variabilität abhängt, wurde hier größtenteils abgelehnt. Nichtsdestotrotz legen die Ergebnisse nahe, dass stellenweise schwache Effekte auftreten und dass Persönlichkeitseigenschaften bei Personen mit hoher Variabilität manchmal ungenauer gemessen werden als bei Personen mit niedrigerer Variabilität. Für die IRT wird dies durch die Operationalisierung von Variabilität als Ausprägung des Itemdiskriminationsparameters verdeutlicht (siehe auch Abbildung 21, S. 143). Für die klassische Testtheorie (KTT) hieße das, der Messfehler fällt für Personen mit unterschiedlich ausgeprägter Variabilität unterschiedlich groß aus. Kann daraus geschlossen werden, dass die Annahme der KTT, der Messfehler gehe ausschließlich auf Merkmale des Tests zurück und sei somit bei allen Personen gleich (Schmidt-Atzert & Amelang, 2012), verworfen werden muss? Aus rein theoretischer Perspektive wäre diese Frage zu bejahen; aus praktischer Sicht lautet die Antwort jedoch anders: Für

die hier erhobenen Fragebogen treten die Effekte nicht auf oder sind so klein, dass sich Unterschiede in der Messgenauigkeit nicht bemerkbar machen. In anderen Worten kann der Mittelwert der Verteilung von Indikatoren eines Traits hinreichend genau geschätzt werden. Allerdings muss dies bei Fragebogen mit nur wenigen Items pro Skala oder bei Fragebogen zu heterogeneren Merkmalen nicht zwingend der Fall sein: Möglicherweise moderiert hier Variabilität die Zusammenhänge zwischen den Items einer Skala und somit die Reliabilität stärker als in den vorliegenden Studien.

Die Befunde zur Vorhersage von Kriterien durch Persönlichkeitseigenschaften und die im vorangegangenen Absatz geschilderten Implikationen zur Reliabilität zeigen einmal mehr, dass die Qualität eines Fragebogens in Abhängigkeit von der Zielgruppe variiert: Personen unterscheiden sich darin, inwieweit Prognosen über ihr Verhalten gestellt werden können. Im Falle von Personen mit hoher Variabilität handelt es sich nicht um eine eng umrissene Zielgruppe, für die man a priori die Anwendung eines Persönlichkeitsfragebogens ablehnen könnte. Allerdings kann Variabilität mit Persönlichkeitsfragebogen erfasst werden und als zusätzlicher Indikator Auskunft darüber geben, wie relevant die Ergebnisse für das Verhalten und für das Erreichen bestimmter Kriterien sind. Aufgrund der konsistenteren Befundlage empfiehlt es sich, Variabilität, nicht ERS, zu erfassen und Anwendern von berufsbezogenen Persönlichkeitsfragebogen im Nicht-Auswahl-Kontext (z. B. Berufsberatung, Platzierung, Personalentwicklung) als Relevanz-Indikator anzubieten. Anhand dieses Indikators kann geschlossen werden, inwieweit das eignungsdiagnostische Urteil auf den Persönlichkeitsfragebogen gestützt werden kann bzw. welchen Stellenwert der Persönlichkeitsfragebogen verglichen mit anderen diagnostischen Informationsquellen einnimmt. Bei Personen mit hoher Variabilität wäre es unter Umständen sinnvoller, auf andere als eigenschaftsdiagnostische Verfahren zu setzen.

Neben dem Nutzen in der Eignungsdiagnostik kann die Erfassung von Variabilität eventuell auch einen Nutzen in der Organisationsdiagnostik und -entwicklung bringen. Hohe Variabilität geht mit hoher Reaktivität auf Situationen einher: Entsprechend könnten Veränderungsmaßnahmen in ihrem Ausmaß und ihrer Umsetzungsgeschwindigkeit an die Variabilität der Betroffenen angepasst werden. Die Interpretation von Variabilität als Reaktivität auf die Merkmale von Situationen könnte auch über das Anwendungsfeld der Arbeits- und Organisationspsychologie hinaus nützlich sein. Beispielsweise könnte Variabilität mit höherer Sug-

gestibilität einhergehen; schließlich entspricht diese einer Reaktivität auf äußere Einflüsse. Ist dies der Fall, dann können Urteile über die Variabilität bei klinischen Fällen bei der Suche der Ursachen psychischer Störungen sowie bei deren Behandlung hilfreich sein: Bei Variablen wären die Ursachen möglicherweise eher äußere Einflüsse und bei Personen mit niedriger Variabilität eher personale Faktoren. Auch in der experimentellen Forschung könnte das Berücksichtigen Variabilität – gerade angesichts der Schwierigkeit und des Aufwandes, große Stichproben zu rekrutieren – als Kovariate möglicherweise die Teststärke erhöhen; schließlich sollten sich bei Personen mit hoher Variabilität stärkere von der Situation abhängige Effekte zeigen als bei Personen mit niedriger Variabilität.

Die Ergebnisse von ERS und Variabilität sind weitgehend parallel und stehen im Einklang mit den Erwartungen zu Variabilität. Dies verdeutlicht abermals, dass ERS Indikator von Variabilität und kein Antwortstil oder Bias ist. Skalen-Scores in Persönlichkeitsfragebogen müssen nicht, wie für den Bereich der Einstellungsmessung vorgeschlagen (Baumgartner & Steenkamp, 2001), um ERS korrigiert werden. ERS kann allenfalls als Inkonsistenz-Maß in Persönlichkeitsfragebogen gesehen werden, hier wäre aus den genannten Gründen jedoch ein direktes Maß von Variabilität (z. B. der Faktor der korrigierten Inter-Item-SD) vorzuziehen. In diesem Zusammenhang ist anzumerken, dass Kontrollskalen bereits vielfach entworfen und evaluiert wurden, aber bislang keine zufriedenstellenden Ergebnisse im Sinne der Validitätssteigerung erzielt wurden (McGrath et al., 2010). Dass Kontrollskalen, darunter auch Inkonsistenz-Skalen (Kurtz & Parrish, 2001; Nikolova, Hendry, Douglas, Edens & Lilienfeld, 2012), keine brauchbaren Resultate liefern, liegt möglicherweise auch daran, dass sie üblicherweise operational definiert sind und dass ihnen eine theoretische Fundierung fehlt. Unklar ist überdies, ob die betreffenden Skalen eine Momentaufnahme abbilden oder ein stabiles Personenmerkmal erfassen. Dieser Missstand trifft auf Variabilität als Indikator der Relevanz von Traits für Verhalten bzw. Kriterien nicht zu.

Mögliche Einschränkungen

Die Befunde zum Zusammenhang zwischen Traits und Kriterien und zu den Zusammenhängen zwischen unterschiedlichen Traits sind auf den ersten Blick verwunderlich: Warum wird die Vorhersage eines Kriteriums durch Persönlichkeitseigenschaften moderiert, nicht aber die Vorhersage anderer Eigenschaften? Schließlich können Eigenschaft und Kriterium ähnlich

global und ähnlich stabil sein. Allerdings hängen Persönlichkeitseigenschaften trivialerweise (per Definition) ausschließlich von der Person ab; schließlich sind sie situationsunabhängig.

Betrachtet man dagegen Kriterien und ihre Entstehung, so fällt auf, dass auf diese eine Fülle von Einflussfaktoren einwirken, von denen einige auf die Person zurückgehen und andere auf die Situation. Zum Beispiel hängt die Arbeitszufriedenheit nicht nur von den Persönlichkeitseigenschaften ab, sondern auch von Unterschieden in der Situation (z. B. Arbeitgeber, Kollegen, berufliche Aufgaben). Wie stark die Persönlichkeitseigenschaften und wie stark die Situation sich auf die Arbeitszufriedenheit auswirkt, ist interindividuell unterschiedlich. Diese interindividuellen Unterschiede lassen sich zum Teil von Variabilität aufklären.

7.2 Beschränkungen und Ausblick

Wie im vorangegangenen Abschnitt (7.1) geschildert werden in der vorliegenden Arbeit Beiträge zur Erfassung und Struktur von Variabilität, zur Erklärung von ERS sowie zum Einfluss von Variabilität auf die Reliabilität und Validität von Persönlichkeitsfragebogen und auf die Stabilität von und die Zusammenhänge zwischen Persönlichkeitseigenschaften geleistet. Die Beiträge haben Implikationen für die Forschung, für die Persönlichkeitsdiagnostik, für die Erfassung von Variabilität und ERS sowie für die Praxis. Im Folgenden werden die Grenzen dieser Beiträge aufgezeigt. Daran anschließend werden Vorschläge unterbreitet, die Befunde weiter abzusichern, die Grenzen zu überwinden und die in dieser Arbeit gezogenen Schlüsse auf eine breitere Basis zu stellen. Abschließend werden weitere sich anschließende Forschungsfragen präsentiert und diskutiert.

Bezogen auf vier Aspekte sind die Ergebnisse der vorliegenden empirischen Studien beschränkt: Erstens betreffen die Befunde zu ERS ebenso wie die zu Variabilität Persönlichkeitsfragebogen. Dass sie sich für Einstellungsfragebogen generalisieren lassen, ist – da es sich ebenfalls um Fragebogen, die Dispositionen erfassen, handelt – plausibel, muss allerdings noch belegt werden. In Anbetracht der entsprechenden Belege kann die Erfassung von Variabilität aufgrund ihrer Effekte auf die Vorhersagen von Verhalten oder Kriterien nämlich auch im Bereich der Marketing- und Einstellungsforschung nützlich sein. Zweitens sind die Befunde zu Persönlichkeitsfragebogen begrenzt auf Nicht-Auswahl-Settings. Bei der Auswahl kann Variabilität zwar valide erfasst werden; dafür, dass die Vorhersage von Kriterien auf der

Basis bei der Auswahl gewonnener Persönlichkeitsmaße von Variabilität abhängt, müssen noch Belege erbracht werden. Drittens wurden die Unterschiede zwischen der Auswahl- und der Nicht-Auswahl-Situation in Studie 2 für zwei unterschiedliche Stichproben erfasst. Zwar ist nicht davon auszugehen, dass die Stichproben sich so gravierend in ihren Persönlichkeitseigenschaften unterscheiden, dass die Ergebnisse auf diese Unterschiede zurückzuführen sind. Dennoch sollten Auswahl- und Nicht-Auswahl-Setting künftig auch mittels Within-Subject-Design verglichen werden. Viertens beziehen sich die Befunde zum Einfluss von Variabilität auf den Zusammenhang zwischen Traits und Kriterien in Studie 1 auf ein konkurrentes und ein in der Vergangenheit liegendes Kriterium. Weitere Forschung sollte die berichteten Moderatoreffekte für die Vorhersage zukünftiger Kriterien bestätigen.

Zur Beschreibung von Variabilität liegen mehrere Arbeiten, einschließlich der vorliegenden, vor. Variabilität von Indikatoren für Persönlichkeitseigenschaften zeigt sich zwischen verschiedenen Rollen (Baird et al., 2006), zwischen verschiedenen Zeitpunkten (Baird et al., 2006; Eid & Diener, 1999; Fleeson, 2001) sowie zwischen den Items von Persönlichkeitsskalen (Biderman & Reddock, 2012; Britt, 1993; Dwight et al., 2002; Reddock et al., 2011). Unklar bleibt, auf welcher Hierarchie-Ebene in der Klassifikation von Verhalten und Eigenschaften Variabilität ansetzt: Baird et al. (2006) sowie Eid und Diener (1999) erfassen Variabilität als Streuung von Adjektiv-Items zwischen verschiedenen Zeitpunkten. Baird et al. berichten, dass diese Streuung mit der Streuung von Adjektiv-Items über Rollen hinweg konvergiert. Die Items von Persönlichkeitsfragebogen bilden – je nach zu erfassendem Merkmal – jedoch nicht nur verschiedene Rollen oder Momente ab, sondern auch verschiedene Facetten eines Konstrukts. Zwar ist den Ergebnissen von Studie 1 zu entnehmen, dass Variabilität unabhängig vom Allgemeinheitsgrad der zu erfassenden Traits auftritt, dennoch bleibt offen, inwieweit Variabilität zwischen verschiedenen Situationen, Zeitpunkten und Rollen mit Variabilität zwischen verschiedenen Facetten einer Eigenschaftsdimension einhergeht. Fraglich ist auch, inwieweit die Variabilität zwischen Items mit der Variabilität innerhalb von Verhaltensprofilen korreliert. Dies sollte im Rahmen weiterer Studien untersucht werden: Variabilität sollte auf unterschiedlichen Ebenen erfasst und die Messungen miteinander verglichen werden. Dabei sollte jeweils auch der Einfluss von Variabilität auf die Eigenschafts-Verhaltens-Kontingenz und die Vorhersagekraft von Traits für Kriterien erforscht werden.

Die Interpretation von ERS als Indikator von Variabilität muss weiteren Prüfungen standhalten. Festgestellt wurde hier, dass ERS kein Antwortstil ist; entsprechend sollte sich Variabilität – und damit einhergehend die Tendenz zu extremen Facetten innerhalb einer Eigenschaft – auch bei Fremdbeschreibungen und Verhaltensbeobachtungen zeigen, sofern diese valide sind. Schließlich implizieren variable und extreme Antworten auch Variabilität im Verhalten, die beobachtbar sein sollte. Restzweifel, dass die beiden Maße spezifisch auf die Methode, und zwar Persönlichkeitsfragebogen, bezogen sind, ließen sich auf diese Weise ausräumen.

7.3 Fazit

In der vorliegenden Arbeit wurden Forschungsarbeiten über Variabilität referiert und Variabilität als Link zwischen Persönlichkeitseigenschaften und Situationen als Ursachen von Verhalten und Kriterien beschrieben. Auch die Forschung zu ERS wurde unter die Lupe genommen: Bislang lagen keine zufriedenstellenden Theorien über die Entstehung von ERS und über die Auswirkungen in Persönlichkeitsfragebogen vor. Ebenso fehlte es an einer systematischen Untersuchung der Zusammenhänge zwischen ERS und Variabilität; bisher verlief die Forschung zu beiden Phänomenen ohne nennenswerten Austausch.

Mittels zweier empirischer Studien wurde hier gezeigt, dass Variabilität eine eindimensionale globale Eigenschaft ist, die valide mit einem Aggregat von um Skalenausprägung und -extremität korrigierten Inter-Item-SD erfasst werden kann. Gezeigt wurde auch, dass sich Variabilität in ERS widerspiegelt. Die Auffassung, ERS sei ein Antwortstil bzw. Bias, wurde als Mythos entlarvt: Konsistent extremes Antworten ist Indikator von Variabilität.

Da ERS durch Variabilität erklärt werden kann, zeigen sich auch dieselben Effekte in Persönlichkeitsfragebogen wie für direkte Messungen von Variabilität. Was die Gütekriterien von Persönlichkeitsfragebogen angeht, so brauchen sich Diagnostiker keine Sorgen darüber zu machen, dass diese stark von Variabilität abhängen: Weder die Reliabilität noch die Konstruktvalidität der untersuchten Instrumente hängen bemerkbar von Variabilität ab. Ebenso unabhängig von Variabilität sind die Stabilität von und die Zusammenhänge zwischen Persönlichkeitseigenschaften. Was allerdings – im Nicht-Auswahl-Kontext – von Variabilität abhängt, ist die Vorhersage von Kriterien auf der Basis von Persönlichkeitseigenschaften, und somit auch die Kriteriumsvalidität: Für Personen mit niedriger Variabilität lassen sich bessere

Vorhersagen treffen als für Personen mit hoher Variabilität. Variabilitätsmaße geben nicht Auskunft darüber, ob ein Persönlichkeitsfragebogen reliabel und valide die zu messenden Eigenschaften erfasst, sondern ob die erfassten Persönlichkeitseigenschaften einer Person relevant (d. h. gute Prädiktoren) für bestimmte Kriterien sind. Dieser Moderatoreffekt lässt sich vermutlich in mehreren Anwendungsfeldern der psychologischen Diagnostik, in jedem Fall aber in der Arbeits- und Organisationspsychologie, nutzen.

Nachdem D. J. Bem und Allen 1974 in ihrem Artikel „*On Predicting Some of the People Some of the Time*“ Variabilität als potenziellen Moderator des Zusammenhangs zwischen Persönlichkeitsmaßen und Verhalten eingeführt hatten, betonten D. J. Bem und Funder 1978 im Artikel „*Predicting More of the People More of the Time*“, dass die Interaktion von Person und Situation einen großen Beitrag zur Vorhersage von Verhalten leistet. Das Ausmaß, in dem das Verhalten von der Persönlichkeit bzw. der Situation und der Interaktion abhängt, wird durch interindividuelle Unterschiede in der Eigenschaft Variabilität beziffert. Trotz dieser Erkenntnis bleibt die Ankündigung eines weiteren Artikels von D. J. Bem und Funder (1978) aktuell:

> Our forthcoming monograph, Predicting All of the People All of the Time is, however, still in preparation. (S. 500)

Literaturverzeichnis

Albaum, G., Roster, C., Yu, J. H. & Rogers, R. D. (2007). Simple rating scale formats: Exploring extreme response. *International Journal of Market Research, 49*, 633-650.

Alliger, G. M. & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement, 60*, 59-72.

Alonso-Arbiol, I. & van de Vijver, F. J. (2010). A historical analysis of the European Journal of Psychological Assessment. *European Journal of Psychological Assessment, 26*, 238-247.

Amelang, M. & Bartussek, D. (2001). *Differentielle Psychologie und Persönlichkeitsforschung* (5. Auflage). Stuttgart: Kohlhammer.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

Arbuckle, J. L. (2010). *Amos* (Version 19.0.0) [Computer Program]. Armonk, NY: IBM.

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme-response style: Improving meaning of translated and culturally adapted rating scales. *Educational and Psychological Measurement, 66*, 374-392.

Arce-Ferrer, A. J. & Ketterer, J. J. (2003). The effect of scale tailoring for cross-cultural application on scale reliability and construct validity. *Educational and Psychological Measurement, 63*, 484-501.

Asendorpf, J. B. (2004). *Psychologie der Persönlichkeit* (3., überarbeitete und aktualisierte Auflage). Berlin: Springer.

Asendorpf, J. B. & Neyer, F. J. (2012). *Psychologie der Persönlichkeit* (5., vollständig überarbeitete Auflage). Berlin: Springer.

Ashton, M. C., Lee, K. & Goldberg, L. R. (2004). A hierarchical analysis of 1,710 English personality-descriptive adjectives. *Journal of Personality and Social Psychology, 87*, 707-721.

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., di Blas, L. et al. (2004). A six-factor structure for personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86*, 356-366.

Austin, E. J., Deary, I. J. & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40*, 1235-1245.

Bachman, J. G. & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly, 48*, 491-509.

Baird, B. M., Le, K. & Lucas, R. E. (2006). On the nature of intraindividual personality variability: Reliability, validity, and associations with well-being. *Journal of Personality and Social Psychology*, *90*, 512-527.

Baird, B. M. & Lucas, R. E. (2011). "… and how about now?": Effects of item redundancy on contextualized self-reports of personality. *Journal of Personality, 79*, 1081-1112.

Barrick, M. R. & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Bartram, D. & Brown, A. (2004). Online testing: Mode of administration and the stability of OPQ 32i scores. *International Journal of Selection and Assessment, 12*, 278-284.

Baumeister, R. F. (1991). On the stability of variability: Retest reliability of metatraits. *Personality and Social Psychology Bulletin, 17*, 633-639.

Baumeister, R. F. & Tice, D. M. (1988). Metatraits. *Journal of Personality*, *56*, 571-598.

Baumgartner, H. & Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143-156.

Beauducel, A. (2001). Problems with parallel analysis in data sets with oblique simple structure. *Methods of Psychological Research, 6*, 141-157

Beauducel, A. & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41-75.

Beermann, D. (2011). *Rational-deduktive, anforderungsbezogene und induktive Konstruktion der ITB-Persönlichkeits-Struktur-Analyse*. Universität Düsseldorf: Unveröffentlichte Diplomarbeit [vorgelegt unter dem Geburtsnamen Kusnezow].

Beermann, D. (2013). *Handreichung zum ITB Personality Structure Assessment*. ITB Consulting GmbH, Bonn: Unveröffentlichte Handreichung.

Beermann, D. & Heilmann, K. (2014). Wie passen Kompetenzen und Persönlichkeitseigenschaften zusammen? Ein kompetenzorientierter Ansatz der Persönlichkeitsdiagnostik. *Wirtschaftspsychologie, 12*(1), 66-80.

Beermann, D., Kersting, M., Stegt, S. & Zimmerhofer, A. (2013). Vorteile und Urteile zur Akzeptanz von Persönlichkeitsfragebogen als Instrumente der Personalarbeit. *PersonalQuarterly, 65*(4), 41-45.

Bem, D. J. & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*, 506-520.

Bem, D. J. & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. *Psychological Review, 85*, 485-501.

Bem, S. L. (1975). Sex role adaptability: One consequence of psychological androgyny. *Journal of Personality and Social Psychology, 31*, 634-643.

Berg, I. A. & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement, 13*, 164-169.

Biderman, M. D., Nguyen, N. T., Cunningham, C. J. & Ghorbani, N. (2011). The ubiquity of common method variance: The case of the Big Five. *Journal of Research in Personality, 45*, 417-429.

Biderman, M. D. & Reddock, C. M. (2012). The relationship of scale reliability and validity to respondent inconsistency. *Personality and Individual Differences*, *52*, 647-651.

Block, J. (1961). Ego identity, role variability, and adjustment. *Journal of Consulting Psychology, 25*, 392-397.

Bolt, D. M. & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*, 814-833.

Borgatta, E. F. & Glass, D. C. (1961). Personality concomitants of extreme response set (ERS). *Journal of Social Psychology, 55*, 213-221.

Bortz, J. & Schuster, C. (2010). *Statistik für Human-und Sozialwissenschaftler* (7., vollständig überarbeitete und ergänzte Auflage). Berlin: Springer.

Bott, J. P., O'Connell, M. S., Ramakrishnan, M. & Doverspike, D. (2007). Practical limitations in making decisions regarding the distribution of applicant personality test scores based on incumbent data. *Journal of Business and Psychology, 22*, 123-134.

Bowling, N. A. & Burns, G. N. (2010). A comparison of work-specific and general personality measures as predictors of work and non-work criteria. *Personality and Individual Differences, 49*, 95-101.

Brengelmann, J. C. (1960). Extreme response set, drive level and abnormality in questionnaire rigidity. *The British Journal of Psychiatry, 106*, 171-186.

Britt, T. W. (1993). Metatraits: Evidence relevant to the validity of the construct and its implications. *Journal of Personality and Social Psychology*, *65*, 554-562.

Bühner, M. (2011). *Einführung in die Test-und Fragebogenkonstruktion* (3., aktualisierte Auflage). München: Pearson Deutschland.

Bushman, B. J. (1995). Moderating role of trait aggressiveness in the effects of violent media on aggression. *Journal of Personality and Social Psychology, 69*, 950-960.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology, 38*, 476-506.

Cattell, R. B. (1944). Interpretation of the twelve primary personality factors. *Character and Personality, 13*, 55-90.

Cattell, R. B. (1945). The description of personality: Principles and findings in a factor analysis. *The American Journal of Psychology, 58*, 69-90.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Chaplin, W. F. & Goldberg, L. R. (1984). A failure to replicate the Bem and Allen study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology, 47*, 1074-1090.

Chuah, S. C., Drasgow, F. & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality, 40*, 359-376.

Collins, J. M. & Gleaves, D. H. (1998). Race, job applicants, and the Five-Factor Model of Personality: Implications for Black psychology, industrial/organizational psychology, and the Five-Factor Theory. *Journal of Applied Psychology, 83*, 531-544.

Costello, A. B. & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting most from your analysis. *Practical Assessment & Research Evaluation, 10*, 1-7. Zugriff am 23. Juni 2014 unter http://pareonline.net/getvn.asp?v=10&n=7

Coyne, I., Warszta, T., Beadle, S. & Sheehan, N. (2005). The impact of mode of administration on the equivalence of a test battery: A quasi-experimental design. *International Journal of Selection and Assessment, 13*, 220-224.

Crandall, J. E. (1973). Sex differences in extreme response style: Differences in frequency of use of extreme positive and negative ratings. *Journal of Social Psychology, 89*, 281-293.

Crandall, J. E. (1982). Social interest, extreme response style, and implications for adjustment. *Journal of Research in Personality, 16*, 82-89.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6*, 475-494.

Cucina, J. M. & Vasilopoulos, N. L. (2005). Nonlinear personality–performance relationships and the spurious moderating effects of traitedness. *Journal of Personality, 73*, 227-260.

cut-e (2013). *The cut-e Assessment Barometer 2012/2013*. Hamburg: Cut-e Group.

Das, J. P. & Dutta, T. (1969). Some correlates of extreme response set. *Acta Psychologica, 29*, 85-92.

De Beuckelaer, A., Weijters, B. & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality & Quantity, 44*, 761-775.

Deffenbacher, J. L. (2003). Angry college student drivers: Characteristics and a test of state-trait theory. *Psicologia Conductual, 11*, 163-178.

Donahue, E. M., Robins, R. W., Roberts, B. W. & John, O. P. (1993). The divided self: Concurrent and longitudinal effects of psychological adjustment and social roles on self-concept differentiation. *Journal of Personality and Social Psychology, 64*, 834.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E. & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40-57.

Dwight, S. A., Wolf, P. P. & Golden, J. H. (2002). Metatraits: Enhancing criterion-related validity through the assessment of traitedness. *Journal of Applied Social Psychology*, *32*, 2202-2212.

Eid, M. & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology, 76*, 662-676.

Eid, M. & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*, 20.

Ellingson, J. E., Sackett, P. R. & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155-166.

Eysenck, H. J. (1944). Types of personality: A factorial study of 700 neurotics. *Journal of Mental Science, 90,* 851-861.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Faul, F., Erdfelder, E., Lang, A. G. & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fisher, R. A. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh, 52*, 399-433.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 81*, 329-344.

Fiske, D. W. & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin, 52*, 217-250.

Fisseni, H. J. (2004). *Lehrbuch der psychologischen Diagnostik* (3. Auflage). Göttingen: Hogrefe.

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology, 80*, 1011-1027.

Fleeson, W. (2004). Moving personality beyond the person-situation debate: The challenge and the opportunity of within-person variability. *Current Directions in Psychological Science, 13*, 83-87.

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality, 75*, 825-862.

Fleeson, W. & Leicht, C. (2006). On delineating and integrating the study of variability and stability in personality psychology: Interpersonal trust as illustration. *Journal of Research in Personality, 40*, 5-20.

Fleisher, M. S., Woehr, D. J., Edwards, B. D. & Cullen, K. L. (2011). Assessing within-person personality variability via frequency estimation: More evidence for a new measurement approach. *Journal of Research in Personality, 45*, 535-548.

Fricker, R. D. & Schonlau, M. (2002). Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods, 14*, 347-367.

Gibbons, J. L., Zellner, J. A. & Rudek, D. J. (1999). Effects of language and meaningfulness on the use of extreme response style by Spanish-English bilinguals. *Cross-Cultural Research, 33*, 369-381.

Goffin, R. D., Rothstein, M. G. & Johnston, N. G. (1996). Personality testing and the assessment center: Incremental validity for managerial selection. *Journal of Applied Psychology, 81*, 746-756.

Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research, 29*, 176-188.

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly, 56*, 328-351.

Greiff, S. (2006). *Prädiktoren des Studienerfolgs: Vorhersagekraft, geschlechtsspezifische Validität und Fairness*. Duisburg: WiKu.

Guilford, J. P. & Braly, K. W. (1930). Extroversion and introversion. *Psychological Bulletin, 27*, 96-107.

Guilford, J. P. & Guilford, R. B. (1936). Personality factors S, E, and M and their measurement. *Journal of Personality, 2*, 109-127.

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin, 69*, 192.

Haney, C., Banks, W. & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology, 1*, 69-97.

Hartweg, V., Milbradt, A., Zimmerhofer, A. & Hornke, L. F. (2009). *testMaker – a computer software for web-based assessments* [Online-Testsystem]. Aachen: Rheinisch-Westfälische Technische Hochschule Aachen, Lehrstuhl für Betriebs- und Organisationspsychologie.

Havlicek, L. L. & Peterson, N. L. (1977). Effect of the violation of the assumptions upon significance levels of the Pearson r. *Psychological Bulletin, 84*, 373-377.

Hertel, G., Naumann, S., Konradt, U. & Batinic, B. (2002). Personality assessment via Internet. In B. Batinic, U.-D. Reips & M. Bosnjak (Hrsg.), *Online Social Sciences* (S. 115-133). Seattle: Hogrefe & Huber Publishers.

Hilbig, B. E. & Zettler, I. (2009). Pillars of cooperation: Honesty-Humility, social value orientations, and economic behavior. *Journal of Research in Personality, 43*, 516-519.

Hilbig, B. E., Zettler, I. & Heydasch, T. (2012). Personality, punishment, and public-goods: Strategic shifts towards cooperation as a matter of dispositional Honesty-Humility. *European Journal of Personality, 26*, 245-254.

Hilbig, B. E., Zettler, I., Moshagen, M. & Heydasch, T. (2012). Tracing the path from personality—via cooperativeness—to conservation. *European Journal of Personality, 27*, 319-327.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800-802.

Höft, S. & Obermann, C. (2010). Der Praxiseinsatz von Assessment Centern im deutschsprachigen Raum: Eine zeitliche Verlaufsanalyse basierend auf den Anwenderbefragungen des Arbeitskreises Assessment Center e.V. von 2001 und 2008. *Wirtschaftspsychologie, 12*(2), 5-16.

Holland, B. S. & Copenhaver, M. D. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, *104*, 145-149.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Huang, H.-M. (2006). Do print and web surveys provide the same results? *Computers in Human Behavior, 22*, 334–350.

Hui, C. H. & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly, 49*, 253-260.

Hui, C. H. & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*, 296-309.

Ibrahimović, N., Bulheller, S., Horn, R., Gitter, G. & Institut für Test- und Begabungsforschung GmbH (2006). *IBF. Intelligenz-Basis-Faktoren*. Frankfurt a. M.: Harcourt.

ITB Consulting GmbH (2011). *iona – ITB Online Assessment* [Online-Testsystem]. Bonn: ITB Consulting GmbH. Zugriff am 23. Juni 2014 unter http://www.itb-consulting.de/iona (Demoversion)

ITB Consulting GmbH (2012). *TM-WISO Demotest* [Online-Test]. Bonn: ITB Consulting GmbH. Zugriff am 23. Juni 2014 unter http://www.tm-wiso.de/de/demotest.aspx

Iwawaki, S. & Zax, M. (1969). Personality dimensions and extreme response tendency. *Psychological Reports, 25*, 31-34.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*, 103-129.

Johnson, M. (2013). Patterns of extreme responses to items in self-esteem scales: Does conceptualisation and item content matter? *Personality and Individual Differences, 55*, 622-625.

Johnson, M. K., Rowatt, W. C. & Petrini, L. (2011). A new trait on the market: Honesty–Humility as a unique predictor of job performance ratings. *Personality and Individual Differences, 50*, 857-862.

Judge, T. A. & Bono, J. E. (2001). Relationships of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology, 86*, 80-92.

Judge, T. A., Heller, D. & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*, 530-541.

Kanning, U. P. & Holling, H. (2001). Struktur, Reliabilität und Validität des NEO-FFI in einer Personalauswahlsituation. *Zeitschrift für Differentielle und Diagnostische Psychologie, 22*, 239-247.

Kersting, M. (2004). Zur Bedeutung der Validität und der sozialen Akzeptanz in der Berufseignungsdiagnostik. *Zeitschrift für Personalpsychologie, 3*, 83-86.

Kersting, M. (2005). Zur Relevanz von Persönlichkeitsmerkmalen in der Arbeits- und Organisationspsychologie. In H. Weber & T. Rammsayer (Hrsg.), *Handbuch der Persönlichkeitspsychologie und Differentiellen Psychologie* (S. 535-545). Göttingen: Hogrefe.

Kersting, M. (n.d.). *Akzept! Fragebogen zur Messung der Akzeptanz diagnostischer Verfahren*. Bochum: Martin Kersting. Zugriff am 23. Juni 2014 unter http://kersting-internet.de/testentwicklungen/akzept-fragebogen

Kieruj, N. D. & Moors, G. (2013). Response style behavior: Question format dependent or personal style? *Quality & Quantity, 47*, 193-211.

Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. et al. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance, 25*, 273-302.

König, C. J., Klehe, U.-C., Berchtold, M. & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment, 18*, 17-27.

Kurtz, J. E. & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment, 76*, 315-332.

La Guardia, J. G. & Ryan, R. M. (2007). Why identities fluctuate: Variability in traits as a function of situational variations in autonomy support. *Journal of Personality, 75*, 1205-1228.

Lautenschlager, G. J. & Meade, A. W. (2008). AlphaTest: A Windows program for tests of hypotheses about coefficient Alpha. *Applied Psychological Measurement, 32*, 502-503.

Lee, K. & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research, 39*, 329-358.

Lee, K. & Ashton, M. C. (2009). *The HEXACO Personality Inventory Revised. Scoring Keys for the 100-Item-Version*. Zugriff am 23. Juni 2014 unter http://hexaco.org/ScoringKeys_100.pdf

Lefever, S., Dal, M. & Matthíasdóttir, Á. (2007). Online data collection in academic research: Advantages and limitations. *British Journal of Educational Technology, 38*, 574–582.

Lewis, N. A. & Taylor, J. A. (1955). Anxiety and extreme response preferences. *Educational and Psychological Measurement, 15*, 111-116.

Lievens, F., De Corte, W. & Schollaert, E. (2008). A closer look at the frame-of-reference effect in personality scale scores and validity. *Journal of Applied Psychology, 93*, 268-279.

Lievens, F. & Harris, M. M. (2003). Research on Internet recruiting and testing: Current status and future directions. *International Review of Industrial and Organizational Psychology, 18*, 131-166.

Lievens, F., Klehe, U.-C. & Libbrecht, N. (2011). Applicant versus employee scores on self-report emotional intelligence measures. *Journal of Personnel Psychology, 10*, 89-95.

Lievens, F. & Thornton III, G. C. (2005). Assessment centers: Recent developments in practice and research. In A. Evers, O. Smit-Voskuijl & N. Anderson (Hrsg.), *Handbook of Selection* (S. 243-264). Hoboken, NJ: Blackwell Publishing.

Light, C. S., Zax, M. & Gardiner, D. H. (1965). Relationship of age, sex, and intelligence level to extreme response style. *Journal of Personality and Social Psychology, 2*, 907-909.

Lix, L. M. & Sajobi, T. (2010). Testing multiple outcomes in repeated measures designs. *Psychological Methods, 15*, 268-280.

Marcus, B. (2003). Persönlichkeitstests in der Personalauswahl: Sind „sozial erwünschte" Antworten wirklich nicht wünschenswert? *Zeitschrift für Psychologie, 211*, 138-148.

Marcus, B., Lee, K. & Ashton, M. C. (2007). Personality dimensions explaining relationships between integrity tests and counterproductive behavior: Big Five, or one in addition? *Personnel Psychology, 60*, 1-34.

Marcus, B., Machilek, F. & Schütz, A. (2006). Personality in cyberspace: Personal web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology, 90*, 1014-1031.

Marín, G., Gamba, R. J. & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology, 23*, 498-509.

Martin, B. A., Bowen, C. C. & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*, 247-256.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McCrae, R. R. (1993). Agreement of personality profiles across observers. *Multivariate Behavioral Research, 28*, 25-40.

McCrae, R. R. (2008). A note on some measures of profile agreement. *Journal of Personality Assessment, 90*, 105-109.

McCrae, R. R. & Costa, P. T. Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

McCrae, R. R., Stone, S. V., Fagan, P. J. & Costa, P. T. Jr. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality, 66*, 285-313.

McDonald, R. P. (1978). Generalizability in factorable domains: Domain validity and generalizability. *Educational and Psychological Measurement, 38*, 75-79.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

McGrath, R. E., Mitchell, M., Kim, B. H. & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin, 136*, 450-470.

McReynolds, P., Altrocchi, J. & House, C. (2000). Self-pluralism: Assessment and relations to adjustment, life changes, and age. *Journal of Personality, 68*, 347-381.

Meade, A. W., Michels, L. C. & Lautenschlager, G. J. (2007). Are Internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*, 322-345.

Meisenberg, G. & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences, 44*, 1539-1550.

Merrens, M. (1970). Generality and stability of extreme response style. *Psychological Reports, 27*, 802-802.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371-378.

Milgram, S. (1997). *Obedience to authority*: *An experimental view*. London: Tavistock.

Mischel, W. (1968). Personality and assessment. New York: Wiley.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology, 55*, 1-22.

Mischel, W. & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*, 730-755.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K. & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Moshagen, M., Hilbig, B. E. & Zettler, I. (2014). Faktorenstruktur, psychometrische Eigenschaften und Messinvarianz der deutschsprachigen Version des 60-Item HEXACO Persönlichkeitsinventars. *Diagnostica, 60*, 86-97.

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J. et al. (2012). The effect of response style on self-reported conscientiousness across 20 countries. *Personality and Social Psychology Bulletin, 38*, 1423-1436.

Naemi, B. D., Beal, D. J. & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality, 77*, 261-286.

Nikolova, N. L., Hendry, M. C., Douglas, K. S., Edens, J. F. & Lilienfeld, S. O. (2012). The inconsistency of inconsistency scales: A comparison of two widely used measures. *Behavioral Sciences & the Law, 30*, 16-27.

Norman, R. P. (1969). Extreme response tendency as a function of emotional adjustment and stimulus ambiguity. *Journal of Consulting and Clinical Psychology, 33*, 406-410.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66*, 574-583.

Obermann, C. (2009). *Assessment Center. Entwicklungen, Durchführungen, Trends* (4. Auflage). Wiesbaden: Gabler.

Ones, D. S. & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*, 245-269.

Osherow, N. (1988). Making sense of the nonsensical: An analysis of Jonestown. In E. Aronson (Hrsg.), *Readings about the social animal* (S. 68-86). New York: Freeman.

Paulhus, D. L. & Martin, C. L. (1988). Functional flexibility: A new conception of interpersonal flexibility. *Journal of Personality and Social Psychology, 55*, 88-101.

Paunonen, S. V. (1988). Trait relevance and the differential predictability of behavior. *Journal of Personality, 56*, 599-619.

Paunonen, S. V. & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81*, 524-539.

Paunonen, S. V. & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. *Psychological Review, 92*, 486-511.

Reddock, C. M., Biderman, M. D. & Nguyen, N. T. (2011). The relationship of reliability and validity of personality tests to frame-of-reference instructions and within-person inconsistency. *International Journal of Selection and Assessment, 19*, 119-131.

Reise, S. P., Waller, N. G. & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*, 287-297.

Revelle, W. & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145-154.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288.

Robertson, I. T. & Callinan, M. (1998). Personality and work behaviour. *European Journal of Work and Organizational Psychology, 7*, 321-340.

Rosenthal, R. & Jacobson, L. (1968). Pygmalion in the classroom. *Urban Review, 3*(1), 16-20.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2., überarbeitete und erweiterte Auflage). Bern: Huber.

Rothstein, M. G. & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155-180.

Rushton, J. P., Jackson, D. N. & Paunonen, S. V. (1981). Personality: Nomothetic or idiographic? A response to Kenrick and Stringfield. *Psychological Review, 88*, 582-589.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin, 28*, 1629-1646.

Ryan, A. M. & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65*, 693-717.

Salgado, J. F. (1997). The Five Factor Model of personality and job performance in the European Community. *Journal of Applied Psychology, 82*, 30-43.

Saucier, G. (2009). Recurrent personality dimensions in inclusive lexical studies: Indications for a Big Six structure. *Journal of Personality, 77*, 1577-1614.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23-74. Zugriff am 23. Juni 2014 unter http://www.cob.unt.edu/slides/Paswan../BUSI6280/Y-Muller_Erfurt_2003.pdf

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt-Atzert, L. (2005). Prädiktion von Studienerfolg bei Psychologiestudenten. *Psychologische Rundschau, 56*, 131-133.

Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik* (Lehrbuch mit Online-Materialien). Heidelberg: Springer.

Schmit, M. J. & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*, 966-974.

Schmitt, M. (1992). Interindividuelle Konsistenzunterschiede als Herausforderung für die differentielle Psychologie. *Psychologische Rundschau, 43*, 30-45.

Schmitt, M. (2005). Interaktionistische Ansätze. In H. Weber & T. Rammsayer (Hrsg.), *Handbuch der Persönlichkeitspsychologie und differentiellen Psychologie* (S. 104-115). Göttingen: Hogrefe.

Schuler, H., Hell, B., Trapmann, S., Schaar, H. & Boramir, I. (2007). Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen. *Zeitschrift für Personalpsychologie, 6*, 60-70.

Schweizer, K. (2010). Some guidelines concerning the modeling of traits and abilities in test construction. *European Journal of Psychological Assessment, 26*, 1-2.

Sheldon, K. M., Ryan, R. M., Rawsthorne, L. J. & Ilardi, B. (1997). Trait self and true self: Cross-role variation in the Big-Five personality traits and its relations with psychological authenticity and subjective well-being. *Journal of Personality and Social Psychology, 73*, 1380-1393.

Sherman, S. J. & Fazio, R. H. (1983). Parallals between attitudes and traits as predictors of behavior. *Journal of Personality, 51*, 308-345.

Shoda, Y., Mischel, W. & Wright, J. C. (1993). The role of situational demands and cognitive competencies in behavior organization and personality coherence. *Journal of Personality and Social Psychology, 65*, 1023-1035.

Shoda, Y., Mischel, W. & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology, 67*, 674-687.

Stanton, J. M. (1998). An empirical assessment of data collection using the Internet. *Personnel Psychology, 51*, 709-725.

Stone, L. L., Otten, R., Ringlever, L., Hiemstra, M., Engels, R. C., Vermulst, A. A. et al. (2013). The parent version of the Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment, 29*, 44-50.

Tett, R. P., Jackson D. N. & Rothstein, M. G. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.

Truell, A. D. (2003). Use of Internet tools for survey research. *Information Technology, Learning & Performance Journal, 21*, 31–37.

Tsaousis, I. & Nikolaou, I. E. (2001). The Stability of the Five-Factor model of personality in personnel selection and assessment in Greece. *International Journal of Selection and Assessment, 9*, 290-301.

Tupes, E. C. & Christal, R. E. (1958). *Stability of personality trait rating factors obtained under diverse conditions* [Research Report]. Lackland Air Force Base, Texas: U. S. Air Force.

Tupes, E. C. & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings* [Tech. Rep. No. ASD-TR-61-97]. Lackland Air Force Base, Texas: U. S. Air Force.

Tuten, T. L., Urban, D. J. & Bosnjak, M. (2002). Internet surveys and data quality: A review. In B. Batinic, U.-D. Reips & M. Bosnjak (Hrsg.), *Online Social Sciences* (S. 7-26). Seattle: Hogrefe & Huber Publishers.

van der Linden, D., te Nijenhuis, J. & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*, 315-327.

Van Vaerenbergh, Y. & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research, 25*, 195-217.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327.

Vinchur, A. J., Schippmann, J. S., Switzer, F. S. & Roth, P. L. (1998). A meta-analytic review of predictors of job performance for salespeople. *Journal of Applied Psychology, 83*, 586-597.

Weijters, B. (2006). *Response styles in consumer research*. Universität Gent (Belgien): Unveröffentlichte Dissertation.

Weijters, B., Cabooter, E. & Schillewaert, N. (2010a). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*, 236-247.

Weijters, B., Geuens, M. & Schillewaert, N. (2010b). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement, 34*, 105-121.

Weijters, B., Geuens, M. & Schillewaert, N. (2010c). The stability of individual response styles. *Psychological Methods, 15*, 96-110.

Weijters, B., Schillewaert, N. & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*, 409-422.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M. & Ostendorf, F. (2013a). Do individual response styles matter? Assessing differential item functioning for men and women in the NEO-PI-R. *Journal of Individual Differences, 34*, 69-81.

Wetzel, E., Carstensen, C. H. & Böhnke, J. R. (2013b). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality, 47*, 178-189.

Wilhelm, O. & McKnight, P. E. (2002). Ability and achievement testing on the world wide web. In B. Batinic, U.-D. Reips & M. Bosnjak (Hrsg.), *Online Social Sciences* (S. 151-180). Seattle: Hogrefe & Huber Publishers.

Wilkinson, A. E. (1970). Relationship between measures of intellectual functioning and extreme response style. *The Journal of Social Psychology, 81*, 271-272.

Witt, L. A. (2002). The interactive effects of extraversion and conscientiousness on performance. *Journal of Management, 28*, 835-851.

Witt, L. A., Burke, L. A., Barrick, M. R. & Mount, M. K. (2002). The interactive effects of conscientiousness and agreeableness on job performance. *Journal of Applied Psychology, 87*, 164-169.

Zettler, I., Friedrich, N. & Hilbig, B. E. (2011). Dissecting work commitment: The role of Machiavellianism. *Career Development International, 16*, 20-35.

Zettler, I. & Hilbig, B. E. (2010). Honesty-Humility and a person-situation interaction at work. *European Journal of Personality, 24*, 569-582.

Zettler, I., Hilbig, B. E. & Haubrich, J. (2011). Altruism at the ballots: Predicting political attitudes and behavior. *Journal of Research in Personality, 45*, 130-133.

Zettler, I., Hilbig, B. E. & Heydasch, T. (2013). Two sides of one coin: Honesty-Humility and situational factors mutually shape social dilemma decision making. *Journal of Research in Personality, 47*, 286-295.

Zuckerman, M. & Norton, J. (1961). Response set and content factors in the California F Scale and the Parental Attitude Research Instrument. *The Journal of Social Psychology, 53*, 199-210.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AC	Assessment Center
ATIC	*Ability to Identify Criteria* (Fähigkeit, Bewertungsmaßstäbe zu identifizieren)
ERS	*Extreme Response Style* (Tendenz, extrem zu Antworten)
ESM	*Experience Sampling Methodology* (Methode der Sammlung von „Erfahrungsstichproben“)
FFM	Fünf-Faktoren-Modell
HEXACO	*Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, Openness to Experience* (Ehrlichkeit-Bescheidenheit, Emotionalität, Extraversion, Verträglichkeit, Gewissenhaftigkeit, Offenheit für Erfahrungen)
HEXACO-PI-R	revidierte Fassung des HEXACO-Persönlichkeitsinventars (hier: deutschsprachige 100-Item-Version)
IEF	*Ideal Employee Factor* (Faktor, der die Vorstellung eines idealen Mitarbeiters beschreibt)
Inter-Item-SD	intraindividuelle Standardabweichung innerhalb einer Skala
IRT	Item-Response-Theorie
ITB-PESA	ITB Personality Structure Assessment
KMO	Kaiser-Mayer-Olkin(-Koeffizient)
K-S-Test	Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilung
KTT	klassische Testtheorie
MAP-Test	*Minimum-Average-Partial*-Test (Test auf Minimum der mittleren quadrierten Partialkorrelationen)
PAF	*Principal Axis Factor Analysis* (Hauptachsenanalyse)
PCA	*Principal Component Analysis* (Hauptkomponentenanalyse)
RIRS	*Representative Indicators for Response Styles* (repräsentative Indikatoren von Antwortstilen)
SCD	*Self-Concept-Differentiation* (Ausdifferenziertheit des Selbstkonzepts)
SEM	*Structure Equation Model* (Strukturgleichungsmodell)
SÖS	sozioökonomischer Status
TM-WISO	Test für Masterstudiengänge in Wirtschafts- und Sozialwissenschaften

Anhang A Ergänzungen zu den Untersuchungsgruppen in Studie 1

Tabelle A - 1: Demografische Daten zu den Untersuchungsgruppen in Studie 1

	Häufigkeit in Untersuchungsgruppe		
	1B (N=394)	1C (N=144)	1D (N=93)
Geschlecht			
weiblich	323 (82.0 %)	118 (81.9 %)	74 (79.6 %)
männlich	71 (18.0 %)	26 (18.1 %)	19 (20.4 %)
Alter			
unter 20 Jahren	2 (0.5 %)		
20 bis 24 Jahre	68 (17.3 %)	6 (4.2 %)	17 (18.3 %)
25 bis 29 Jahre	62 (15.7 %)	24 (16.7 %)	14 (15.1 %)
30 bis 39 Jahre	144 (36.5 %)	64 (44.4 %)	29 (31.2 %)
40 bis 49 Jahre	94 (23.9 %)	41 (28.5 %)	28 (30.1 %)
über 49 Jahre	23 (5.8 %)	9 (6.3 %)	5 (5.4 %)
keine Angabe	1 (0.3 %)		
Höchster Bildungsabschluss			
Fachhochschulreife	19 (4.8 %)		3 (3.2 %)
Abitur	151 (38.3 %)		35 (37.6 %)
Berufsausbildung	71 (18.0 %)		16 (17.2 %)
Bachelor	20 (5.1 %)	18 (12.5 %)	5 (5.4 %)
Diplom (FH)	45 (11.4 %)	40 (27.8 %)	15 (16.1 %)
Diplom / Master	76 (19.3 %)	79 (54.9 %)	15 (16.1 %)
Promotion	7 (1.8 %)	7 (4.9 %)	2 (2.2 %)
keine Angabe	5 (1.3 %)		2 (2.2 %)
Berufserfahrung			
Keine	37 (9.4 %)	6 (4.2 %)	7 (7.5 %)
unter 1 Jahr	26 (6.6 %)	9 (6.3 %)	7 (7.5 %)
1 bis 5 Jahre	90 (22.8 %)	31 (21.5 %)	23 (24.7 %)
6 bis 10 Jahre	82 (20.8 %)	39 (27.1 %)	14 (15.1 %)
11 bis 15 Jahre	68 (17.3 %)	30 (20.8 %)	20 (21.5 %)
16 bis 20 Jahre	35 (8.9 %)	9 (6.3 %)	7 (7.5 %)
21 bis 30 Jahre	41 (10.4 %)	16 (11.1 %)	11 (11.8 %)
über 30 Jahre	9 (2.3 %)	3 (2.1 %)	2 (2.2 %)
keine Angabe	6 (1.5 %)	1 (0.7 %)	2 (2.2 %)

Anhang B Ergänzungen zu den Messungen in Studie 1

B.1 Instruktion zu den Persönlichkeitsfragebogen

Instruktion

Die folgenden Fragen und Aussagen beziehen sich auf Eigenschaften und Verhaltensweisen, die insbesondere im Arbeitsleben von Bedeutung sind. Das Profil, das Ihnen am Ende der Bearbeitung angezeigt wird, gibt Ihnen Hinweise, wo Ihre Stärken und wo Ihre Entwicklungsfelder liegen. Damit die Ergebnisse interpretierbar und wertvoll für Sie sind, sollten Sie die Fragen möglichst offen und ehrlich beantworten.

Die Bearbeitung dauert zwischen 50 und 70 Minuten – am unteren Bildrand wird Ihnen Ihr Fortschritt bei der Bearbeitung des Fragebogens angezeigt.

Bitte lesen Sie die Aussagen und beurteilen Sie auf einer Skala von 1 bis 6 spontan, inwieweit diese für Sie zutreffen. „1“ steht für „trifft überhaupt nicht zu“ und „6“ für „trifft voll zu“.

Beispiel:

„Ich sehe gern fern.“

Trifft die Aussage voll zu, dann wählen Sie bitte die „6“.Trifft sie überhaupt nicht zu, dann wählen Sie bitte die „1“. In allen anderen Fällen lässt sich der Grad des Zutreffens mit den Antworten „2“ bis „5“ abstufen.

Einige Aussagen beziehen sich darauf, wie Sie mit Ihren Mitarbeitern umgehen oder mit Ihren Kollegen kommunizieren; andere Aussagen sprechen Ihr Verhalten gegenüber Kunden an. Sollten Sie keine Kunden, Kollegen oder Mitarbeiter haben, stellen Sie sich bitte möglichst plastisch derartige Situationen vor und beantworten die Fragen bitte dennoch.

B.2 Faktorenstruktur des HEXACO-PI-R

Mit den Facetten des HEXACO-PI-R wurde eine Hauptachsenanalyse mit anschließender Oblimin-Rotation ($delta = 0$) durchgeführt. Theoriegeleitet wurden sechs Faktoren extrahiert. Im Folgenden ist die Mustermatrix abgebildet.

Tabelle B.2 - 1: Mustermatrix einer Hauptachsenanalyse der Facetten des HEXACO-PI-R

Facette	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6
Lebhaftigkeit (X)	**.77**	.09	.09	.02	.01	-.02
Soziales Selbstvertrauen (X)	**.66**	.11	.11	-.06	-.04	-.10
Geselligkeit (X)	**.64**	.03	-.09	.20	.01	.15
Soziale Kühnheit (X)	**.61**	-.18	.02	-.04	-.20	.06
Ängstlichkeit (E)	**-.44**	-.24	.14	**.33**	-.01	.08
Geduld (A)	-.03	**.71**	.20	-.22	-.01	.05
Kompromissbereitschaft (A)	.06	**.64**	.03	.07	.14	-.01
Sanftmut (A)	-.06	**.64**	-.09	.16	-.04	-.01
Nachsichtigkeit (A)	.09	**.52**	-.10	-.16	-.08	-.06
Besonnenheit (C)	-.03	.08	**.67**	-.12	.06	-.08
Perfektionismus (C)	-.26	-.06	**.62**	.07	-.12	.05
Organisiertheit (C)	.21	.03	**.60**	.03	.07	.02
Fleiß (C)	**.42**	-.10	**.56**	.00	-.09	-.12
Sentimentalität (E)	.07	.04	.00	**.71**	-.14	-.05
Abhängigkeit (E)	.09	-.05	-.10	**.56**	.04	.02
Altruismus	.08	.29	.08	**.42**	-.18	-.21
Furchtsamkeit (E)	-.14	-.08	.07	**.37**	.10	.01
Unkonventionalität (O)	.03	-.06	-.13	-.01	**-.63**	.08
Sinn für Ästhetik (O)	-.06	.09	.05	.11	**-.61**	-.04
Kreativität (O)	.08	-.01	-.02	.04	**-.59**	.02
Wissbegierde (O)	.00	-.03	.07	-.10	**-.47**	-.03
Aufrichtigkeit (H)	-.03	-.10	.01	-.08	-.07	**-.59**
Materielle Genügsamkeit (H)	-.04	.01	-.10	.04	.04	**-.58**
Selbstbescheidung (H)	-.10	.15	-.05	.10	.03	**-.55**
Fairness (H)	.12	-.03	.17	.00	.04	**-.52**

Untersuchungsgruppe 1A, $N = 405$; Ladungen über $|r| = .3$ sind fett gedruckt.

Korrelationen zwischen Faktoren: $r_{12} = .20$, $r_{15} = .22$, $r_{26} = .28$, für alle anderen Korrelationen gilt $|r| < .20$.

Hinter den Facetten ist jeweils angegeben, zu welcher Dimension die Facette gehört.

H: Ehrlichkeit-Bescheidenheit, E: Emotionalität, X: Extraversion, A: Verträglichkeit versus Ärger, C: Gewissenhaftigkeit, O: Offenheit für Erfahrungen

B.3 Statistiken zu den Faktor-Skalen des HEXACO-PI-R

Tabelle B.3 - 1: Itemzahlen, mittlere Trennschärfen und Reliabilitätsschätzungen der Faktorskalen des HEXACO-PI-R

Skala	Item mit höchster Trennschärfe (Polung) (r_{it})	n_n	$\overline{r_{it}}$	α	r_{tt}
Ehrlichkeit-Bescheidenheit	„Es würde mir viel Freude bereiten, teure Luxusgüter zu besitzen.“ (-) (.53)	10	.42	.81	.86
Emotionalität	„Ich kann mit schwierigen Situationen umgehen, ohne dass ich emotionale Unterstützung von irgendjemandem brauche.“ (-) (.49)	7	.36	.76	.79
Extraversion	„An den meisten Tagen bin ich fröhlich und optimistisch.“ (+) (.66)	7	.50	.86	.89
Verträglichkeit versus Ärger	„Ich werde selten wütend, selbst wenn andere mich ziemlich schlecht behandeln.“ (+) (.56)	8	.41	.81	.79
Gewissenhaftigkeit	„Wenn ich arbeite, habe ich manchmal Schwierigkeiten, weil ich desorganisiert bin.“ (-) (.54)	8	.42	.81	.83
Offenheit für Erfahrungen	„Der Besuch einer Kunstausstellung würde mich langweilen.“ (-) (.52)	8	.34	.74	.74

Untersuchungsgruppe 1A, $N = 405$; n_n: Anzahl negativ gepolter Items, r_{it}: Trennschärfe (Part-Whole-korrigiert), $\overline{r_{it}}$: mittlere Trennschärfe der 16 Items der Skala (berechnet mit Fishers Z-Transformation, Fisher, 1918), α: Cronbachs Alpha, r_{tt}: Split-Half-Reliabilität (odd-even, Spearman-Brown-korrigiert)

Tabelle B.3 - 2: Deskriptive Statistiken, Kolmogorov-Smirnov-Tests auf Ablehnung der Normalverteilung und Skaleninterkorrelationen zu den Faktorskalen des HEXACO-PI-R

			K-S-Test		Korrelation zu				
Skala	M	SD	Z	p	E	X	A	C	O
Ehrlichkeit-Bescheidenheit	4.45	0.76	1.469	.027	-.01	-.01	.21	.17	.04
Emotionalität (E)	3.78	0.66	0.787	.566		-.19	-.23	-.01	.01
Extraversion (X)	4.27	0.77	1.336	.056			.17	.17	.22
Verträglichkeit versus Ärger (A)	3.51	0.66	0.722	.674				.01	-.01
Gewissenhaftigkeit (C)	4.31	0.65	0.946	.333					.09
Offenheit für Erfahrungen (O)	4.41	0.67	1.381	.044					

Untersuchungsgruppe 1A, $N = 405$; M: Gruppen-Mittelwert der Skalenmittelwerte, SD: Gruppen-Standardabweichung der Skalen, K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

B.4 Skalen des ITB-PESA und Item-Beispiele

Tabelle B.4 - 1: Die Skalen des ITB-PESA und das Item mit der jeweils höchsten Trennschärfe

Kompetenzbereich und Skala	Item mit höchster Trennschärfe (Polung) (r_{it})
Soziale Kompetenz	
Kontaktfreude	Es fällt mir leicht, andere anzusprechen. (+) (.74)
Kommunikationsvermögen	Es fällt mir leicht, auch trockene Sachverhalte unterhaltsam zu präsentieren. (+) (.57)
Geselligkeit	Manche halten mich für einen Einzelgänger / eine Einzelgängerin. (-) (.58)
Einfühlungsvermögen	Wenn es anderen Menschen schlecht geht, leide ich mit. (+) (.63)
Konsensorientierung	Ich bin bekannt dafür, dass ich kein Blatt vor den Mund nehme. (-) (.54)
Aufgeschlossenheit und Neugier	Mich faszinieren Menschen, die „anders“ und ungewöhnlich sind. (+) (.40)
Führungskompetenz	
Leadership	Es ist mir schon häufiger gelungen, auch kritische Mitarbeiter / Mitarbeiterinnen für „unsere Sache“ zu begeistern. (+) (.63)
Steuerungsvermögen	Ich spreche mit meinen Mitarbeitern und Mitarbeiterinnen regelmäßig über den Fortschritt ihrer Aufgabenbearbeitung. (+) (.77)
Führungswille und Machtmotivation	Ich fühle mich wohl, wenn ich anderen die Richtung vorgeben soll. (+) (.70)
Souveränität	Ich fühle mich unwohl, wenn ich mich im Job auf andere verlassen muss. (-) (.53)
Unternehmerische Kompetenz	
Ganzheitlich-strategische Denkweise	Wissenschaftliche Themen finde ich spannend. (+) (.52)
Kundenorientierung	Beim Kontakt mit Kunden sind mir sowohl das Ergebnis als auch die Atmosphäre wichtig. (+) (.51)
Mut und Risikobereitschaft	Abenteuerlust kann man mir nun wirklich nicht nachsagen. (-) (.63)
Eigeninitiative	Wenn ich Gegebenheiten für verbesserungswürdig halte, dann packe ich zu und ändere etwas. (+) (.63)
Markt- und Wettbewerbsorientierung	Ich informiere mich regelmäßig darüber, was Wettbewerber meines Unternehmens tun. (+) (.72)
Ergebnisorientierung	
Arbeitsdisziplin	Unangenehme Aufgaben schiebe ich manchmal vor mir her. (-) (.66)
Ausdauer und Belastbarkeit	Ausdauer – auch unter schwierigen Rahmenbedingungen – ist eine meiner Stärken. (+) (.73)
Sorgfalt	Ich nehme die Dinge so genau, dass mich manche als "kleinkariert" bezeichnen. (+) (.62)
Erfolgszuversicht	Auch bei schwierigen Projekten bin ich mir sicher, dass ich sie erfolgreich abschließen werde. (+) (.64)
Leistungsstreben und Erfolgsmotivation	Mit meinen Leistungen bin ich nur zufrieden, wenn ich damit zu den Besten gehöre. (+) (.57)
Integrität & Verlässlichkeit	
Ehrlichkeit	Es ist nicht ungewöhnlich, dass man mitunter lügt, um vor dem Chef / der Chefin besser dazustehen. (-) (.50)
Regelbewusstsein	Es stört mich, wenn andere darauf bestehen, jede einzelne kleine Absprache einzuhalten. (-) (.53)

r_{it}: Trennschärfe (part-whole-korrigiert) ermittelt an Untersuchungsgruppe 1A, $N = 405$

B.5 Skaleninterkorrelationen und –statistiken zum ITB-PESA

Tabelle B.5 - 1: Skaleninterkorrelationen und –statistiken der Skalen des ITB-PESA

			K-S-Test		Korrelation zu																			
Kompetenzbereich und Skala	M	SD	Z	p	Kv	Gs	Ev	Ko	AN	Ls	Sv	FM	So	GD	Ku	MR	Ei	MW	Ad	AB	Sf	Ez	LE	Eh	Rb
Soziale Kompetenz																								
Kontaktfreude	3.45	1.06	1.207	.108	.48	.50	.04	-.27	.31	.41	.34	.41	.39	.32	.31	.44	.38	.38	.25	.33	-.10	.50	.10	.21	-.02
Kommunikationsvermögen (Kv)	4.48	0.69	1.539	.017		.28	.02	-.22	.48	.65	.54	.56	.17	.49	.55	.35	.65	.50	.34	.46	.06	.45	.32	.21	.10
Geselligkeit (Gs)	4.23	0.78	1.565	.015			.29	.00	.31	.29	.32	.23	.50	.16	.29	.22	.22	.19	.13	.19	-.21	.33	-.05	.33	.07
Einfühlungsvermögen (Ev)	3.98	0.76	1.029	.240				.20	.14	.10	.04	-.14	-.10	-.03	.05	-.12	-.01	-.04	-.18	-.24	-.01	-.23	-.05	.02	-.05
Konsensorientierung (Ko)	3.63	0.62	0.952	.325					-.07	-.21	-.21	-.40	-.06	-.18	-.15	-.27	-.29	-.20	-.12	-.15	-.01	-.18	-.09	.10	.12
Aufgeschlossenheit und Neugier (AN)	4.82	0.58	1.475	.026						.37	.35	.21	.14	.41	.44	.36	.43	.22	.09	.17	-.05	.23	.09	.21	.01
Führungskompetenz																								
Leadership (Ls)	4.41	0.70	1.323	.060							.67	.53	.07	.42	.58	.37	.72	.61	.37	.56	.25	.53	.47	.20	.16
Steuerungsvermögen (Sv)	4.40	0.82	1.502	.022								.48	.19	.34	.62	.26	.62	.61	.42	.47	.17	.48	.23	.28	.27
Führungswille und Machtmotivation (FM)	3.92	0.96	0.980	.292									.15	.38	.42	.39	.54	.47	.24	.41	.09	.42	.47	.05	.10
Souveränität (So)	3.28	0.74	0.999	.271										.18	.09	.27	.09	.10	.26	.19	-.41	.49	-.22	.46	-.08
Unternehmerische Kompetenz																								
Ganzheitlich-strategische Denkweise (GD)	4.32	0.73	1.324	.060											.34	.46	.52	.43	.25	.32	.01	.39	.29	.18	.00
Kundenorientierung (Ku)	4.77	0.61	1.794	.003												.29	.57	.53	.23	.37	.09	.38	.26	.18	.14
Mut und Risikobereitschaft (MR)	4.12	0.83	1.379	.045													.40	.33	.14	.29	-.10	.43	.21	.03	-.19
Eigeninitiative (Ei)	4.61	0.62	1.388	.042														.52	.42	.53	.28	.50	.43	.24	.19
Markt- und Wettbewerbsorientierung (MW)	3.98	0.94	0.879	.423															.35	.43	.12	.44	.36	.17	.12
Ergebnisorientierung																								
Arbeitsdisziplin (Ad)	3.95	0.85	1.155	.139																.64	.24	.57	.23	.41	.37
Ausdauer und Belastbarkeit (AB)	4.16	0.79	1.302	.067																	.27	.57	.38	.31	.30
Sorgfalt (Sf)	3.75	0.84	1.268	.080																		-.05	.39	.03	.47
Erfolgszuversicht (Ez)	3.93	0.90	1.364	.049																			.20	.36	.08
Leistungsstreben und Erfolgsmotivation (LE)	4.12	0.78	0.914	.373																				-.03	.17
Integrität & Verlässlichkeit																									
Ehrlichkeit (Eh)	4.13	0.75	0.753	.623																					.36
Regelbewusstsein (Rb)	4.16	0.73	0.963	.312																					

Untersuchungsgruppe 1A, $N = 405$; M: Gruppen-Mittelwert der Skalenmittelwerte, SD: Gruppen-Standardabweichung der Skalen, K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

B.6 Faktorenanalyse der Skalen des ITB-PESA

Mit den Facetten des ITB-PESA wurde eine Hauptachsenanalyse mit anschließender Oblimin-Rotation ($delta = 0$) durchgeführt. Dem *Minimum-Average-Partial*-Test (MAP-Test, Velicer, 1976) und Scree-Test zufolge sowie basierend auf bisherigen Befunden (Beermann, 2011) wurden vier Faktoren extrahiert. Erläuterungen, warum sich nur vier schwer interpretierbare Faktoren extrahieren lassen, finden sich bei Beermann (2011) sowie Beermann und Heilmann (2014). Es folgt die Mustermatrix.

Tabelle B.6 - 1: Mustermatrix einer Hauptachsenanalyse der Facetten des ITB-PESA

Skala	Faktor 1	Faktor 2	Faktor 3	Faktor 4
Leadership	**.84**	-.11	.15	.08
Eigeninitiative	**.80**	-.11	.15	-.01
Kommunikationsvermögen	**.78**	.04	.03	.07
Kundenorientierung	**.70**	-.02	.08	.20
Führungswille und Machtmotivation	**.69**	-.05	-.08	-.19
Steuerungsvermögen	**.67**	.03	.26	.10
Mark- und Wettbewerbsorientierung	**.66**	-.04	.09	-.05
Ganzheitlich-strategische Denkweise	**.58**	.07	-.07	-.06
Mut- und Risikobereitschaft	**.55**	.19	-.28	-.19
Aufgeschlossenheit und Neugier	**.54**	.14	-.07	.29
Kontaktfreude	**.51**	**.37**	-.05	-.02
Leistungsstreben und Erfolgsmotivation	**.50**	**-.41**	.08	-.13
Erfolgszuversicht	**.43**	**.42**	.24	**-.36**
Konsensorientierung	**-.34**	.03	.22	.27
Souveränität	-.01	**.86**	.12	-.16
Sorgfalt	.15	**-.57**	**.43**	-.02
Geselligkeit	**.31**	**.53**	.10	**.33**
Regelbewusstsein	-.01	-.16	**.69**	.04
Arbeitsdisziplin	.20	.16	**.59**	**-.35**
Ehrlichkeit	.02	**.43**	**.56**	.06
Ausdauer und Belastbarkeit	**.43**	.05	**.44**	**-.34**
Einfühlungsvermögen	.12	-.02	-.04	**.64**

Untersuchungsgruppe 1A, $N = 405$; Ladungen über $|r| = .3$ sind fett gedruckt.

Korrelationen zwischen Faktoren: $r_{12} = .18$, $r_{13} = .21$, $r_{14} = -.17$, für alle anderen Korrelationen gilt $|r| < .05$.

B.7 Histogramme für die Kriterien in Studie 1

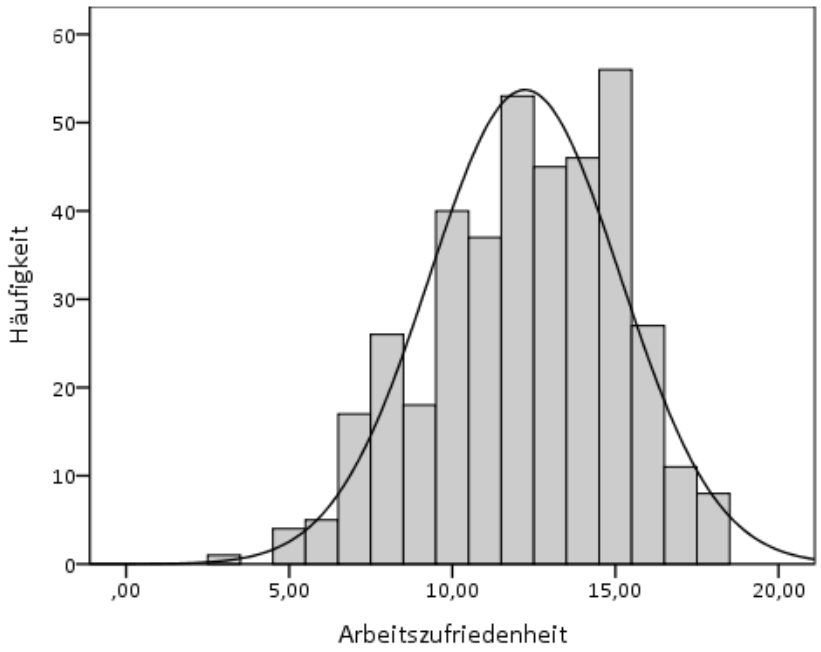

Abbildung B.7 - 1: Histogramm für das Kriterium Arbeitszufriedenheit
Untersuchungsgruppe 1B, $N = 394$

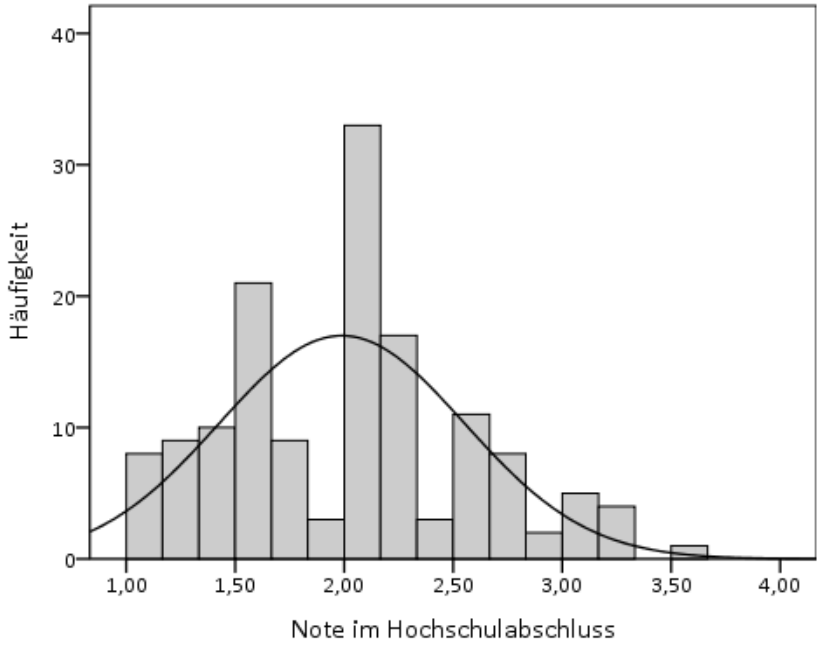

Abbildung B.7 - 2: Histogramm für das Kriterium Note im Hochschulabschluss
Untersuchungsgruppe 1C, $N = 144$

Anhang C Zur Alpha-Adjustierung

In diesem Abschnitt wird die in der vorliegenden Arbeit verwendete Korrektur des Alpha-Fehlerniveaus beschrieben. Im Folgenden werden zunächst Signifikanztests eingeführt. Daran anschließend werden das Problem der Kumulierung des Fehlers 1. Art bei multiplem Testen skizziert und die Bonferroni-Korrektur einschließlich der Erweiterungen von Holm (1979) und Hochberg (1988) vorgestellt. Schließlich wird die Methode von Holland und Copenhaver (1988) und darauf aufbauend die in dieser Arbeit verwendete Methode erläutert.

Signifikanztests

Mit Signifikanztests wird überprüft, wie wahrscheinlich es ist, dass sich ein Effekt (z. B. Mittelwerts-Unterschied zwischen zwei Bedingungen, Zusammenhang zwischen zwei Merkmalen) in den Daten zeigt, obwohl er in der Gesamtpopulation nicht besteht. Das Ergebnis eines Signifikanztests, das Signifikanzniveau, entspricht einer bedingten Wahrscheinlichkeit: der Wahrscheinlichkeit zu folgern, es gibt in der Gesamtpopulation einen Effekt, unter der Bedingungen, dass dieser nicht existiert (Fehler 1. Art). Per Konvention liegt die Toleranzgrenze für den Fehler 1. Art (Alpha-Fehler) in der psychologischen Forschung bei $\alpha = .05$. Liegt das Signifikanzniveau bei $p = .05$ oder darunter, so gilt es als hinreichend unwahrscheinlich, dass der gefundene Effekt zufällig zustande gekommen ist: Man nimmt an, dass er für die Gesamtpopulation gilt.

Multiples Testen und die Kumulierung des Fehlers 1. Art

Wenn zum Überprüfen einer Hypothese mehrere ($i = 1, \dots, k$) Signifikanztests durchgeführt werden, ist die Wahrscheinlichkeit, dass einer der Tests $p_i \leq .05$ ergibt, größer als $p = .05$ und somit nicht klein genug, als dass ein Effekt in der Gesamtpopulation angenommen werden kann. Daher wird das α-Niveau für die einzelnen Tests zu α_i (bzw. α') korrigiert.

Die Bonferroni-Korrektur

Zu den populärsten Korrekturmethoden zählen Bonferroni-Korrekturen (Lix & Sajobi, 2010), von denen die klassische Methode – $\alpha_i = \frac{\alpha}{k}$ für alle i (Holland & Copenhaver, 1988; Lix &

Sajobi, 2010) – sehr konservativ ausfällt. Bei diesem Verfahren ist die Wahrscheinlichkeit dafür, dass ein beliebiger Test zufällig (d. h. unter der Bedingung, dass in der Gesamtpopulation kein Effekt vorliegt) $p_i \leq \alpha_i$ ergibt, in jedem Fall kleiner als α. Dies trifft auch dann zu, wenn die Ereignisse $p_i \leq \alpha_i$ für alle Tests disjunkt sind, d. h. wenn für jeden Test nur dann $p_i \leq \alpha_i$ gelten kann, wenn für alle andere Tests $p_i > \alpha_i$ gilt.

Sequentielle Verfahren als Erweiterung der klassischen Bonferroni-Korrektur

Holm (1979) schlägt ein sequentielles Verfahren zur Erweiterung der klassischen Bonferroni-Korrektur (Bonferroni-Holm-Korrektur) vor: Bei k Signifikanztests gilt der Test mit dem niedrigsten p-Wert ($i = 1$) als signifikant, wenn $p_1 \leq \frac{\alpha}{k}$. Fällt dieser Test nicht signifikant aus, gelten alle Tests als nicht signifikant. Bei einem signifikanten Ergebnis wird der Test mit dem zweitniedrigsten p-Wert ($i = 2$) überprüft; dieser ist bei $p_2 \leq \frac{\alpha}{k-1}$ signifikant. Fällt der Test nicht signifikant aus, gelten alle weiteren Tests ebenfalls als nicht signifikant. Bei einem signifikanten Ergebnis wird der Test mit dem drittniedrigsten p-Wert ($i = 3$) überprüft; dieser ist bei $p_3 \leq \frac{\alpha}{k-2}$ signifikant, usw.

Inhaltlich wird beim Verfahren Holms – nachdem ein oder mehrere Tests signifikant ist bzw. sind und die Entscheidung bei diesem Test für die Alternativhypothese (für einen Effekt) getroffen wurde – bei den verbleibenden Tests erneut geprüft, ob diese alpha-korrigiert signifikant sind.

Nach Hochbergs (1988) sequentiellem Verfahren wird zuerst der Signifikanztest mit dem höchsten p-Wert ($i = 1$) betrachtet: Wenn $p_1 \leq \alpha$, dann gelten alle Tests als signifikant. Bei $p_1 > \alpha$ wird der Signifikanztest mit dem zweithöchsten p-Wert ($i = 2$) betrachtet: Dieser und alle weiteren Tests gelten als signifikant, wenn $p_2 \leq \frac{\alpha}{2}$. Ist er nicht signifikant, wird der Signifikanztest mit dem dritthöchsten p-Wert ($i = 3$) betrachtet. Dieser und alle weiteren Tests gelten als signifikant, wenn $p_2 \leq \frac{\alpha}{3}$, usw.

Inhaltlich wird beim Vorgehen Hochbergs zunächst ein Test (mit dem höchsten p-Wert) überprüft. Ist dieser nicht signifikant, wird ein weiterer Test geprüft. Da damit allerdings die Ergebnisse eines zweiten Tests betrachtet werden, muss das Alpha-Niveau soweit korrigiert

werden, dass die Wahrscheinlichkeit, dass mindestens ein Test unter die konventionelle Grenze fällt, wiederum $\alpha = .05$ entspricht. Dies wird jeweils mit der klassischen Bonferroni-Korrektur sichergestellt. Wenn ein Test signifikant ausfällt, dann wird für diesen die Alternativhypothese angenommen und für jeden weiteren Test, der stets einen kleineren p-Wert aufweist als der vorherige Test, wird das Ergebnis ebenfalls signifikant.

Die Korrektur von Holland und Copenhaver (1988)

Holland und Copenhaver (1988) schlagen eine Alternative zur Bonferroni-Methode vor: Sie gehen von der Annahme aus, dass im Falle zufälliger Effekte (d. h. bei Gültigkeit der Nullhypothese in der Gesamtpopulation) die Ergebnisse mehrerer Signifikanztests unabhängig voneinander sind. Entsprechend ist den Autoren zufolge die Wahrscheinlichkeit, dass keiner von k Tests signifikant wird, $P = (1 - \alpha')^k$. α' ist das Signifikanzniveau der einzelnen Tests. Daraus abgeleitet ist die Wahrscheinlichkeit, dass mindestens ein Test signifikant wird, also das globale Alpha-Fehlerniveau: $\alpha = 1 - (1 - \alpha')^k$. Aus dieser Formel berechnen Holland und Copenhaver das Signifikanzniveau für die einzelnen k Tests, α'. Für das globale Alpha-Fehlerniveau $\alpha = .05$ ergibt sich:

$$\alpha' = 1 - \sqrt[k]{.95}$$

Anlass für eine neue Korrekturmethode

Die klassische Bonferroni-Korrektur, die Bonferroni-Holm-Korrektur sowie die von Holland und Copenhaver (1988) vorgeschlagene Korrektur sind sehr konservativ. Z. B. würde man sich allen drei Korrekturen zufolge bei einer Reihe von k Tests, deren Signifikanzniveau jeweils bei $p = .049$ liegt, jeweils für die Nullhypothese entscheiden. Dies trifft auch auf das Verfahren von Holland und Copenhaver zu, wenn man es auf die Weise erweitert, wie Holm (1979) die Bonferroni-Korrektur erweitert hat. Das Verfahren von Hochberg (1988) wäre in einigen Fällen ebenfalls zu konservativ: Wenn beispielsweise von k Tests einer $p = .051$ und $k - 1$ Tests $p = .049$ ergeben, fällt die Entscheidung nach Hochberg ebenfalls jeweils für die Nullhypothese. Dies trifft auch auf eine Verbesserung des Verfahrens mit der von Holland und Copenhaver (1988) vorgeschlagenen Methode anstelle der Bonferroni-Methode zu.

Wie die Beispiele zeigen, verlangt die Auswahl einer Alpha-Korrektur stets sorgfältiges Abwägen zwischen der Kontrolle des Fehlers 1. Art und der Aufrechterhaltung der Teststärke, die in den Beispielen beeinträchtigt ist. Die im Folgenden geschilderte Methode ist zweifelsfrei liberaler als die genannten Verfahren, sie zielt also darauf ab, die Teststärke zu maximieren; d. h. mit der neu entwickelten Methode sollen Effekte, die in der Grundgesamt vorliegen, leichter aufgedeckt werden. Zugleich soll der Fehler 1. Art konstant $\alpha = .05$ betragen.

Simultane Korrektur für verschiedene Anzahlen signifikanter Tests

Die vorliegende Methode kann als Erweiterung des Verfahrens von Holland und Copenhaver (1988) gesehen werden. In deren Verfahren wird berechnet, wie unwahrscheinlich das Ergebnis *eines* Signifikanztests ausfallen muss, damit das globale Signifikanzniveau, $\alpha = .05$, unterschritten wird. Der Einfachheit halber sei fortan die Gleichung für die Gegenwahrscheinlichkeit ($1 - \alpha = .95$) aufgeführt:

$$(1 - \alpha')^k = .95 \qquad (1)$$

Durch Auflösen dieser Gleichung nach α' erhält man das entsprechende Signifikanzniveau. Als Erweiterung lässt sich die Frage stellen: Wie hoch muss das Alpha-Niveau sein, unter das die p-Werte von zwei Tests fallen müssen, damit die Nullhypothese für diese beiden Tests hinreichend unwahrscheinlich ist? Zur Beantwortung dieser Frage wird wiederum die Gleichung für die Gegenwahrscheinlichkeit aufgestellt. Zur linken Seite in Gleichung (1) wird die Wahrscheinlichkeit hinzuaddiert, dass genau einer der k Tests signifikant wird. Diese ist:

$$P = \binom{k}{1} \cdot (1 - \alpha')^{k-1} \cdot \alpha' \qquad (2)$$

Erläuterung: Die Wahrscheinlichkeit, dass von $k - 1$ Tests keiner signifikant wird, entspricht $(1 - \alpha')^{k-1}$; diese wird multipliziert mit der Wahrscheinlichkeit, dass ein weiterer Test signifikant wird, α'. Da nun jeder beliebige der k Tests signifikant werden kann, wird dieses Produkt mit dem Binomialkoeffizienten multipliziert: Denn unter k Tests gibt es $\binom{k}{i} = \frac{k!}{(k-i)!i!}$ Möglichkeiten i Tests auszuwählen, im vorliegenden Fall: $\binom{k}{1} = k$.

Die Summe aus Gleichung (1) und Gleichung (2) ergibt die Wahrscheinlichkeit dafür, dass bei maximal einem Test α' unterschritten wird. Schließlich sind beide Summanden, (i) „kein p-Wert liegt unter α'“ und (ii) „genau ein p-Wert liegt unter α'“, disjunkte Ereignisse. Die Summe soll wieder $.95$ ergeben:

$$(1-\alpha')^k + \binom{k}{1} \cdot (1-\alpha')^{k-1} \cdot \alpha' = .95 \qquad (3)$$

Die Erweiterung lässt sich fortführen: Wie hoch muss das Alpha-Niveau sein, unter dass die p-Werte von drei Tests fallen müssen, damit die Nullhypothese für diese drei Tests hinreichend unwahrscheinlich ist? Wiederum lassen sich die Gleichung für die Gegenwahrscheinlichkeit – der p-Wert von maximal zwei Tests fällt unter α' – aufstellen und auf diese Weise α' bestimmen. Zur linken Seite in Gleichung (3) wird die Wahrscheinlichkeit hinzuaddiert, dass bei genau zwei der Tests α' unterschritten wird:

$$(1-\alpha')^k + \binom{k}{1} \cdot (1-\alpha')^{k-1} \cdot \alpha' + \binom{k}{2} \cdot (1-\alpha')^{k-2} \cdot \alpha'^2 = .95 \qquad (4)$$

Auf der linken Seite von Gleichung (4) ist nun die Wahrscheinlichkeit dafür abgebildet, dass maximal zwei Tests signifikant werden. Ein Ereignis, bei dem mindestens drei Tests signifikant werden, also bei $p_i \leq \alpha'$ liegen, ist gleich der Gegenwahrscheinlichkeit von $\alpha = .05$. Die Formel in Gleichung (4) lässt sich fortsetzen bis auf der linken Seite die Wahrscheinlichkeit aufgeführt ist, dass maximal $k-1$ Tests α' unterschreiten. Für $m < k$ lässt sich die Formel verallgemeinern zu:

$$(1-\alpha')^k + \binom{k}{1} \cdot (1-\alpha')^{k-1} \cdot \alpha' + \binom{k}{2} \cdot (1-\alpha')^{k-2} \cdot \alpha'^2 + \cdots + \binom{k}{m} \cdot (1-\alpha')^{k-m} \cdot \alpha'^m = .95 \qquad (5)$$

Auf der linken Seite in Gleichung (5) ist die Wahrscheinlichkeit aufgeführt, dass bei maximal m Tests α' unterschritten wird.

Unter Berücksichtigung der Gleichungen (1) bis (5) wird nun folgende Methode zur Alpha-Korrektur vorgeschlagen:

(i) Zunächst werden die ($i = 1, \ldots, k$) Tests – wie bei den Verfahren nach Holm (1979) und Hochberg (1988) – nach ihrem Signifikanzniveau geordnet. Dem Test mit dem geringsten

Signifikanzniveau wird $i = 1$ zugeordnet, dem Test mit dem höchsten Signifikanzniveau wird $i = k$ zugeordnet.

(ii) Bei Tests, deren Signifikanzniveau $p = .05$ übersteigt, wird die Entscheidung zugunsten der Nullhypothese getroffen.

(iii) Unter den verbleibenden n Tests wird zunächst der Test mit dem höchsten Signifikanzniveau ($i = n$) betrachtet. Dessen p-Wert wird verglichen mit dem aus Gleichung (5) gewonnenen α' für $m = n - 1$. Falls $p_n \leq \alpha'$, gelten alle weiteren ($i = 1, \dots, n$) Tests als signifikant. Bei $p_n > \alpha'$ wird der p-Wert des „$n - 1$“-ten Signifikanztests betrachtet.

Liegt der p-Wert des „$n - 1$“-ten Signifikanztests unter α', das mit Gleichung (5) für $m = n - 2$ ermittelt wurde, sind alle weiteren ($i = 1, \dots, n - 1$) Tests signifikant. Bei $p_{n-1} > \alpha'$ wird der p-Wert des „$n - 2$“-ten Signifikanztests betrachtet, usw. Das Verfahren ist abgeschlossen, wenn entweder $p_i \leq \alpha'$ (in diesem Fall ist für den i-ten und die weiteren $i - 1$ Tests die Alternativhypothese anzunehmen) oder wenn der Test mit dem niedrigsten p-Wert überprüft wurde. Für den Test mit dem niedrigsten p-Wert reduziert sich die Gleichung zum Ermitteln von α' auf Gleichung (1).

Da die Auflösung der algebraischen Gleichung (5) sehr komplex und in einigen Fällen unmöglich ist, wurden für die vorliegende Arbeit die α'-Werte näherungsweise, aber mit ausreichender Genauigkeit, durch Einsetzen in einer Excel-Tabelle bestimmt.

Tabelle C - 1 verdeutlicht exemplarisch für die Tests zur Überprüfung des Einflusses von Variabilität auf die Retestreliabilität der Skalen des ITB-PESA, inwieweit sich die referierten und das vorliegend beschriebene Verfahren unterscheiden. Aufgeführt sind die Signifikanztests sortiert nach ihrer Größe und die jeweils korrigierten Alpha-Niveaus.

Tabelle C - 1: Vergleich der berichteten Alpha-Korrekturen am Beispiel der Tests zur Überprüfung des Einflusses von Variabilität auf die Retestreliabilität der Skalen des ITB-PESA

Signifikanztest		Bonferroni-Korrekturen					nach Holland und Copenhaver (1988)					*hier vorgestellt*	
		klassisch		sequentiell			klassisch		sequentiell				
					Holm	Hochberg				Holm	Hochberg		p_i
i	p	α_i	$p_i \leq \alpha_i$	α_i	$p_i \leq \alpha_i$	$p_i \leq \alpha_i$	α_i	$p_i \leq \alpha_i$	α_i	$p_i \leq \alpha_i$	$p_i \leq \alpha_i$	α_i	$\leq \alpha_i$
22	.971	.0023		.0500			.0023		.0500			.0500	
21	.909	.0023		.0250			.0023		.0253			.0500	
20	.894	.0023		.0167			.0023		.0170			.0500	
19	.843	.0023		.0125			.0023		.0127			.0500	
18	.764	.0023		.0100			.0023		.0102			.0500	
17	.575	.0023		.0083			.0023		.0085			.0500	
16	.514	.0023		.0071			.0023		.0073			.0500	
15	.453	.0023		.0063			.0023		.0064			.0500	
14	.407	.0023		.0056			.0023		.0057			.0500	
13	.382	.0023		.0050			.0023		.0051			.0500	
12	.370	.0023		.0045			.0023		.0047			.0500	
11	.354	.0023		.0042			.0023		.0043			.0500	
10	.321	.0023		.0038			.0023		.0039			.0500	
9	.305	.0023		.0036			.0023		.0037			.0500	
8	.277	.0023		.0033			.0023		.0034			.0500	
7	.273	.0023		.0031			.0023		.0032			.0500	
6	.062	.0023		.0029			.0023		.0030			.0500	
5	.034	.0023		.0028			.0023		.0028			.0500	X
4	.026	.0023		.0026			.0023		.0027			.0500	X
3	.015	.0023		.0025			.0023		.0026			.0382	X
2	.003	.0023		.0024			.0023		.0024			.0162	X
1	<.001	.0023	X	.0023	X	X	.0023	X	.0023	X	X	.0023	X

Untersuchungsgruppe 1D, $N = 93$; X bedeutet $p_i \leq \alpha_i$.

Die Nullhypothese wird nach der hier berichteten Alpha-Korrektur bei deutlich mehr Tests abgelehnt als bei den referierten Verfahren (und die Entscheidung wird deutlich häufiger zugunsten der Alternativhypothese getroffen). Dass bereits ab vier Test mit einem p-Wert von $\alpha' = .05$ die Nullhypothese für diese vier Test abgelehnt wird, erscheint auf den ersten Blick sehr liberal. Allerdings liegt die Wahrscheinlichkeit, dass mindestens 4 der 22 Tests per Zufall diese Grenze erreichen oder unterschreiten, nur bei $p = .0222$.

In der vorliegenden Arbeit werden neben dem hier neu eingeführten Verfahren jeweils auch die Ergebnisse nach der Bonferroni-Holm-Korrektur (Holm, 1979) berichtet.

Anhang D Ergänzungen zu den Ergebnissen in Studie 1

D.1 Analyse der Inter-Item-SD der Skalen des ITB-PESA

Tabelle D.1 - 1: Analyse der Inter-Item-SD im ITB-PESA

	Inter-Item-SD						korr. Inter-Item-SD		
	M	SD	K-S-Test		a		K-S-Test		a
Kompetenzbereich und Skala			Z	p		R^2	Z	p	
Soziale Kompetenz									
Kontaktfreude	1.19	0.42	1.186	.120	.46	.180	1.290	.072	.62
Kommunikationsvermögen	1.07	0.40	1.981	.001	.45	.235	1.680	.007	.57
Geselligkeit	1.19	0.38	0.922	.363	.52	.254	0.935	.347	.68
Einfühlungsvermögen	1.17	0.41	1.170	.129	.42	.080	0.808	.531	.54
Konsensorientierung	1.40	0.34	0.735	.652	.53	.019	0.716	.685	.62
Aufgeschlossenheit und Neugier	0.94	0.39	1.729	.005	.35	.230	1.700	.006	.52
Führungskompetenz									
Leadership	0.99	0.35	1.612	.011	.53	.184	1.442	.031	.58
Steuerungsvermögen	0.82	0.40	1.954	.001	.42	.215	1.688	.007	.49
Führungswille und Machtmotivation	1.03	0.37	1.392	.041	.52	.131	1.811	.003	.61
Souveränität	1.19	0.39	1.149	.143	.52	.072	0.996	.275	.66
Unternehmerische Kompetenz									
Ganzheitlich-strategische Denkweise	1.25	0.41	0.882	.418	.45	.225	0.963	.311	.61
Kundenorientierung	0.98	0.44	1.972	.001	.35	.280	1.379	.045	.55
Mut und Risikobereitschaft	1.19	0.38	1.306	.066	.50	.169	1.524	.019	.65
Eigeninitiative	0.93	0.33	1.940	.001	.46	.135	2.058	<.001	.57
Markt- und Wettbewerbsorientierung	1.17	0.46	1.303	.067	.51	.190	1.294	.070	.55
Ergebnisorientierung									
Arbeitsdisziplin	1.17	0.34	0.729	.663	.42	.246	0.782	.573	.63
Ausdauer und Belastbarkeit	1.08	0.38	1.040	.230	.50	.095	1.034	.235	.59
Sorgfalt	1.26	0.39	0.966	.308	.41	.227	0.952	.327	.62
Erfolgszuversicht	1.09	0.37	0.936	.344	.43	.154	1.136	.151	.57
Leistungsstreben und Erfolgsmotivation	1.15	0.40	1.320	.061	.48	.206	1.130	.155	.64
Integrität & Verlässlichkeit									
Ehrlichkeit	1.28	0.36	1.178	.125	.52	.162	1.080	.194	.66
Regelbewusstsein	1.12	0.39	1.462	.028	.47	.160	1.408	.038	.61

Untersuchungsgruppe 1A, $N = 405$

Inter-Item-SD: intraindividuelle Standardabweichung pro Skala, korr. Inter-Item-SD: intraindividuelle Standardabweichung pro Skala, korrigiert um Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts

M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, R^2: Anteil durch den Item-Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts aufgeklärter Varianz der Inter-Item-SD, a: Ladung auf dem Faktor der jeweiligen Faktorenanalyse

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

Für die korrigierten Inter-Item-SD werden weder Gruppen-Mittelwert noch -Standardabweichung berichtet, da es sich um Residuen handelt, deren Gruppen-Mittelwert jeweils Null ist. Die Standardabweichung ergibt sich aus der Standardabweichung für die Inter-Item-SD und R^2.

D.2 Analyse der Inter-Item-SD der Skalen des HEXACO-PI-R

Tabelle D.2 - 1: Analyse der Inter-Item-SD im HEXACO-PI-R

	Inter-Item-SD						korr. Inter-Item-SD		
	M	SD	K-S-Test		a		K-S-Test		a
Skala			Z	p		R^2	Z	p	
Ehrlichkeit-Bescheidenheit	1.34	0.38	0.670	.760	.49	.325	0.712	.692	.71
Emotionalität	1.43	0.30	0.920	.366	.66	.091	0.688	.732	.71
Extraversion	1.22	0.35	0.927	.356	.45	.227	1.245	.090	.69
Verträglichkeit versus Ärger	1.35	0.30	1.068	.204	.68	.035	0.714	.688	.72
Gewissenhaftigkeit	1.23	0.31	0.739	.646	.45	.180	0.763	.605	.66
Offenheit für Erfahrungen	1.43	0.36	0.877	.425	.50	.287	0.676	.751	.69

Untersuchungsgruppe 1A, $N = 405$

Inter-Item-SD: intraindividuelle Standardabweichung pro Skala, korr. Inter-Item-SD: intraindividuelle Standardabweichung pro Skala, korrigiert um Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts

M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, R^2: Anteil durch den Item-Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts aufgeklärter Varianz der Inter-Item-SD, a: Ladung auf dem Faktor der jeweiligen Faktorenanalyse

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

Für die korrigierten Inter-Item-SD werden weder Gruppen-Mittelwert noch -Standardabweichung berichtet, da es sich um Residuen handelt, deren Gruppen-Mittelwert jeweils Null ist. Die Standardabweichung ergibt sich aus der Standardabweichung für die Inter-Item-SD und R^2.

D.3 Scree-Plots für die Faktorenanalysen der korrigierten Inter-Item-SD

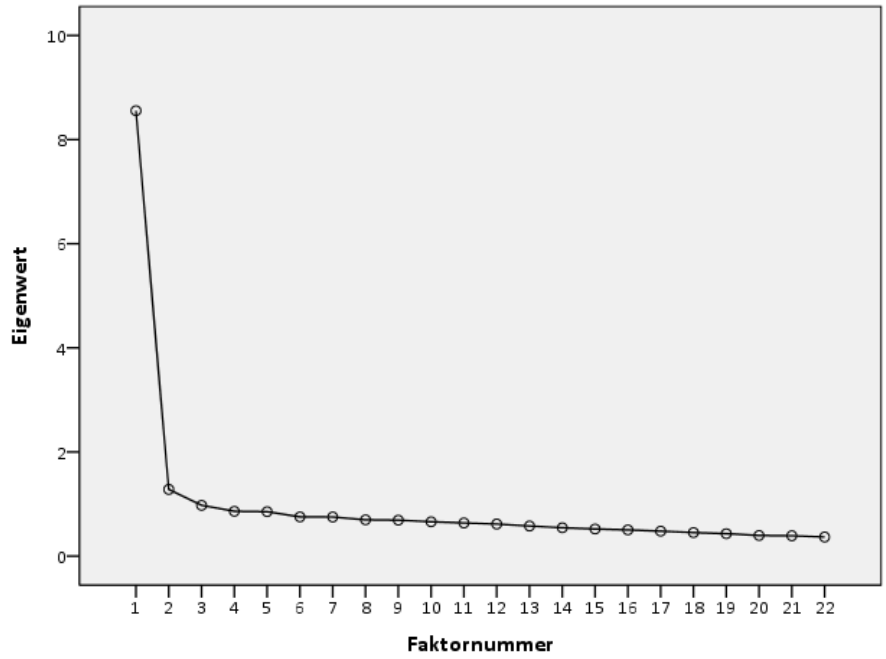

Abbildung D.3 - 1: Scree-Plot zur Faktorenanalyse der korrigierten Inter-Item-SD der Skalen des ITB-PESA Untersuchungsgruppe 1A, $N = 405$

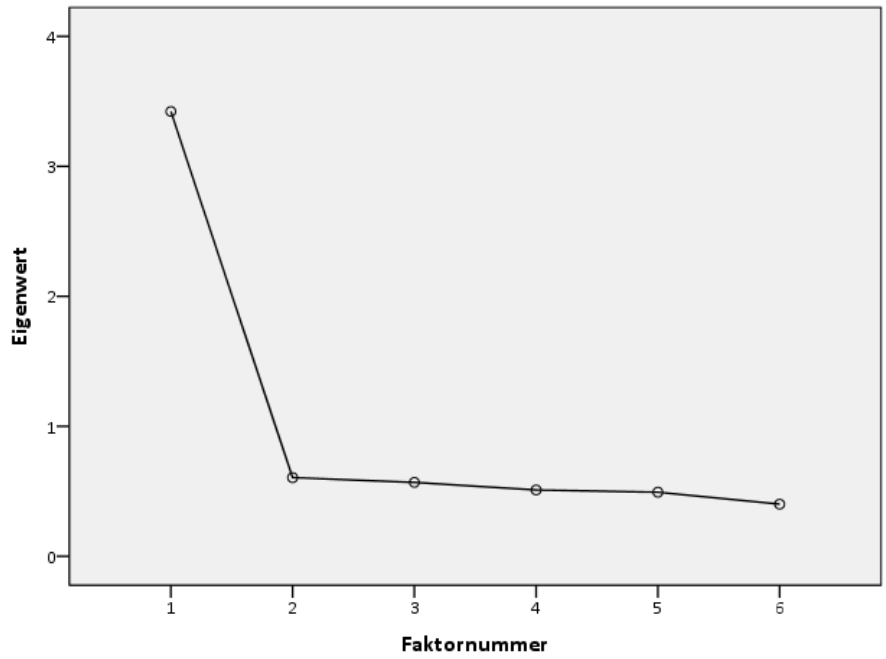

Abbildung D.3 - 2: Scree-Plot zur Faktorenanalyse der korrigierten Inter-Item-SD der Skalen des HEXACO-PI-R Untersuchungsgruppe 1A, $N = 405$

D.4 Korrelationen zwischen den korrigierten Inter-Item-SD der Skalen des ITB-PESA

Tabelle D.4 - 1: Korrelationen zwischen den korrigierten Inter-Item-SD der Skalen des ITB-PESA

	Korrelation zur korrigierten Inter-Item-SD der Skala																				
Korrigierte Inter-Item-SD der Skala	Kv	Gs	Ev	Ko	AN	Ls	Sv	FM	So	GD	Ku	MR	Ei	MW	Ad	AB	Sf	Ez	LE	Eh	Rb
Soziale Kompetenz																					
Kontaktfreude	.33	.44	.32	.46	.26	.36	.32	.40	.43	.38	.30	.36	.33	.32	.42	.40	.39	.34	.44	.39	.36
Kommunikationsvermögen (Kv)		.34	.31	.34	.30	.41	.23	.33	.34	.40	.26	.39	.40	.27	.40	.33	.34	.36	.32	.37	.34
Geselligkeit (Gs)			.44	.46	.38	.42	.29	.44	.51	.38	.39	.42	.34	.31	.43	.40	.39	.37	.46	.44	.42
Einfühlungsvermögen (Ev)				.37	.29	.30	.17	.35	.35	.34	.23	.36	.30	.17	.34	.35	.39	.32	.33	.41	.31
Konsensorientierung (Ko)					.33	.37	.28	.35	.42	.36	.32	.45	.32	.30	.39	.33	.42	.38	.34	.40	.36
Aufgeschlossenheit und Neugier (AN)						.32	.23	.20	.35	.35	.34	.34	.42	.28	.27	.31	.28	.24	.35	.39	.32
Führungskompetenz																					
Leadership (Ls)							.39	.34	.33	.40	.42	.33	.34	.36	.26	.36	.37	.27	.40	.36	.26
Steuerungsvermögen (Sv)								.24	.27	.40	.40	.23	.34	.46	.35	.24	.34	.21	.28	.28	.26
Führungswille und Machtmotivation (FM)									.46	.36	.30	.42	.34	.32	.40	.44	.40	.39	.39	.40	.38
Souveränität (So)										.42	.29	.43	.33	.36	.42	.38	.37	.43	.41	.49	.44
Unternehmerische Kompetenz																					
Ganzheitlich-strategische Denkweise (GD)											.37	.36	.45	.43	.36	.32	.36	.31	.34	.33	.33
Kundenorientierung (Ku)												.38	.34	.40	.34	.27	.38	.26	.33	.34	.32
Mut und Risikobereitschaft (MR)													.39	.35	.40	.39	.44	.41	.50	.44	.40
Eigeninitiative (Ei)														.30	.34	.39	.30	.29	.33	.35	.35
Markt- und Wettbewerbsorientierung (MW)															.32	.25	.37	.34	.36	.32	.39
Ergebnisorientierung																					
Arbeitsdisziplin (Ad)																.40	.40	.45	.40	.43	.39
Ausdauer und Belastbarkeit (AB)																	.36	.37	.41	.43	.33
Sorgfalt (Sf)																		.35	.42	.39	.35
Erfolgszuversicht (Ez)																			.36	.39	.37
Leistungsstreben und Erfolgsmotivation (LE)																				.45	.41
Integrität & Verlässlichkeit																					
Ehrlichkeit (Eh)																					.47
Regelbewusstsein (Rb)																					

Untersuchungsgruppe 1A, $N = 405$

D.5 Modifikationsindizes für die konfirmatorische Faktorenanalyse der korrigierten Inter-Item-SD der Skalen des ITB-PESA

Abgebildet sind für Untersuchungsgruppe 1A die Modifikationsindizes $M.I > 4$ für die konfirmatorische Faktorenanalyse der korrigierten Inter-Item-SD der Skalen des ITB-PESA. Links aufgeführt werden die Pfade, für die zweite Zeile der Tabelle bedeutet dies beispielsweise: Ein Einfügen des Pfads zwischen den Fehlerfaktoren der korrigierten Inter-Item-SD der Skalen „Steuerungsvermögen“ und „Markt- und Wettbewerbsorientierung“ führt zu einer Verringerung des χ^2-Wertes um 31.00 (mittlere Spalte). Die Skalen-Mittelwerte der beiden Skalen korrelieren zu $r_{(405)} = .61$.

Tabelle D.5 - 1: Modifikationsindizes zur konfirmatorischen Faktorenanalyse der korrigierten Inter-Item-SD der Skalen des ITB-PESA

Pfade zwischen Fehlerfaktoren der korrigierten Inter-Item-SD der Skalen	Modifikations-index	Korrelation zwischen den Skalen-MW
Steuerungsvermögen ↔ Markt- und Wettbewerbsorientierung	31.00	.61
Steuerungsvermögen ↔ Kundenorientierung	15.71	.62
Aufgeschlossenheit und Neugier ↔ Führungswille und Machtmotivation	14.07	.21
Einfühlungsvermögen ↔ Markt- und Wettbewerbsorientierung	13.43	-.04
Aufgeschlossenheit und Neugier ↔ Eigeninitiative	12.98	.43
Ganzheitlich-strategische Denkweise ↔ Eigeninitiative	11.48	.52
Leadership ↔ Arbeitsdisziplin	11.07	.37
Ganzheitlich-strategische Denkweise ↔ Markt- und Wettbewerbsorientierung	10.23	.43
Leadership ↔ Steuerungsvermögen	10.05	.67
Steuerungsvermögen ↔ Ganzheitlich-strategische Denkweise	9.71	.34
Leadership ↔ Kundenorientierung	9.58	.58
Kundenorientierung ↔ Markt- und Wettbewerbsorientierung	8.94	.53
Leadership ↔ Regelbewusstsein	8.42	.16
Steuerungsvermögen ↔ Mut und Risikobereitschaft	8.07	.26
Arbeitsdisziplin ↔ Erfolgszuversicht	7.90	.57
Mut und Risikobereitschaft ↔ Leistungsstreben und Erfolgsmotivation	7.33	.21
Kontaktfreude ↔ Konsensorientierung	6.51	-.27
Kommunikationsvermögen ↔ Leadership	6.18	.65
Kommunikationsvermögen ↔ Eigeninitiative	6.15	.65
Einfühlungsvermögen ↔ Steuerungsvermögen	6.14	.04
Ganzheitlich-strategische Denkweise ↔ Ehrlichkeit	6.08	.18
Souveränität ↔ Kundenorientierung	6.01	.09
Ehrlichkeit ↔ Regelbewusstsein	5.98	.36
Markt- und Wettbewerbsorientierung ↔ Ausdauer und Belastbarkeit	5.83	.43
Führungswille und Machtmotivation ↔ Ausdauer und Belastbarkeit	5.31	.41
Geselligkeit ↔ Einfühlungsvermögen	5.25	.29
Geselligkeit ↔ Souveränität	4.45	.50
Konsensorientierung ↔ Leistungsstreben und Erfolgsmotivation	4.42	-.09
Geselligkeit ↔ Markt- und Wettbewerbsorientierung	4.39	.19

Untersuchungsgruppe 1A, $N = 405$

D.6 Analyse der Extremwerthäufigkeiten für die Skalen des ITB-PESA

Tabelle D.6 - 1: Analyse von ERS für die erste Erhebung in Studie 1 (ITB-PESA)

Kompetenzbereich und Skala	Zahl der Items	M	SD	K-S-Test		a
				Z	p	
Soziale Kompetenz						
Kontaktfreude	8	1.99	1.85	3.917	<.001	.58
Kommunikationsvermögen	8	2.02	1.83	3.827	<.001	.72
Geselligkeit	10	2.48	2.07	3.122	<.001	.62
Einfühlungsvermögen	8	1.52	1.58	4.030	<.001	.56
Konsensorientierung	10	2.00	1.82	3.426	<.001	.69
Aufgeschlossenheit und Neugier	8	2.57	2.02	3.299	<.001	.63
Führungskompetenz						
Leadership	10	2.15	2.12	4.013	<.001	.81
Steuerungsvermögen	10	1.86	2.47	5.460	<.001	.68
Führungswille und Machtmotivation	10	1.99	2.30	3.954	<.001	.65
Souveränität	9	1.44	1.63	4.822	<.001	.66
Unternehmerische Kompetenz						
Ganzheitlich-strategische Denkweise	8	2.29	1.86	3.385	<.001	.70
Kundenorientierung	8	2.61	2.08	2.742	<.001	.72
Mut und Risikobereitschaft	9	2.14	1.90	3.661	<.001	.69
Eigeninitiative	9	2.12	2.05	4.124	<.001	.82
Markt- und Wettbewerbsorientierung	9	2.26	2.29	3.930	<.001	.65
Ergebnisorientierung						
Arbeitsdisziplin	11	2.12	2.13	3.809	<.001	.67
Ausdauer und Belastbarkeit	10	1.98	2.16	4.008	<.001	.72
Sorgfalt	9	1.89	1.82	4.400	<.001	.59
Erfolgszuversicht	9	1.58	1.81	4.572	<.001	.68
Leistungsstreben und Erfolgsmotivation	9	1.90	1.90	4.388	<.001	.67
Integrität & Verlässlichkeit						
Ehrlichkeit	10	2.57	2.18	3.211	<.001	.61
Regelbewusstsein	8	1.52	1.66	4.636	<.001	.63

Untersuchungsgruppe 1A, $N = 405$

M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, a: Ladung auf dem Faktor der Faktorenanalyse

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

D.7 Analyse der Extremwerthäufigkeiten für die Skalen des HEXACO-PI-R

Tabelle D.7 - 1: Analyse von ERS für die erste Erhebung in Studie 1 (HEXACO-PI-R)

Skala	Zahl der Items	M	SD	K-S-Test		a
				Z	p	
Ehrlichkeit-Bescheidenheit	16	6.47	3.71	1.684	.007	.62
Emotionalität	16	3.84	2.74	2.609	<.001	.64
Extraversion	16	4.55	3.39	2.619	<.001	.67
Verträglichkeit versus Ärger	16	3.19	2.53	3.115	<.001	.71
Gewissenhaftigkeit	16	4.17	3.21	2.841	<.001	.73
Offenheit für Erfahrungen	16	6.33	3.45	1.837	.002	.68

Untersuchungsgruppe 1A, $N = 405$

M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, a: Ladung auf dem Faktor der Faktorenanalyse

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

D.8 Scree-Plots für die Faktorenanalysen der Extremwert-häufigkeiten

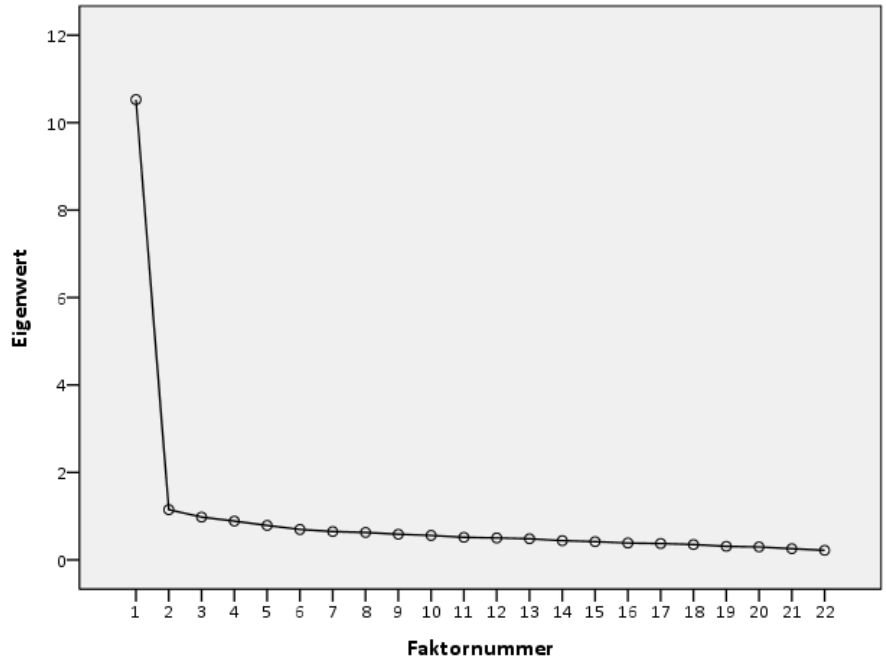

Abbildung D.8 - 1: Scree-Plot zur Faktorenanalyse der Extremwerthäufigkeiten auf den Skalen des ITB-PESA *Untersuchungsgruppe 1A,* $N = 405$

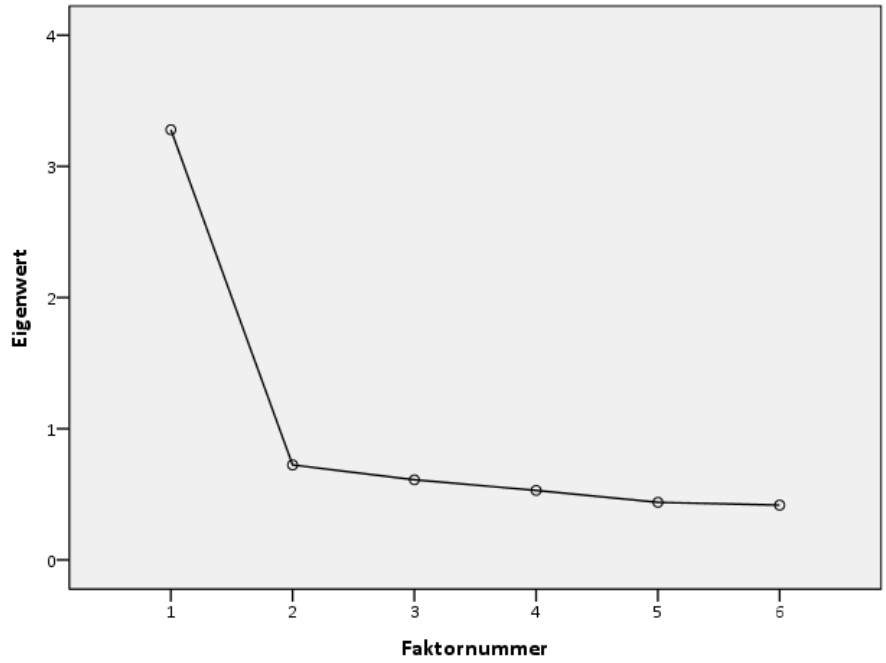

Abbildung D.8 - 2: Scree-Plot zur Faktorenanalyse der Extremwerthäufigkeiten auf den Skalen des HEXACO-PI-R Untersuchungsgruppe 1A, $N = 405$

Anhang E Ergänzungen zu den Messungen in Studie 2

E.1 Histogramme für das Kriterium in Studie 2

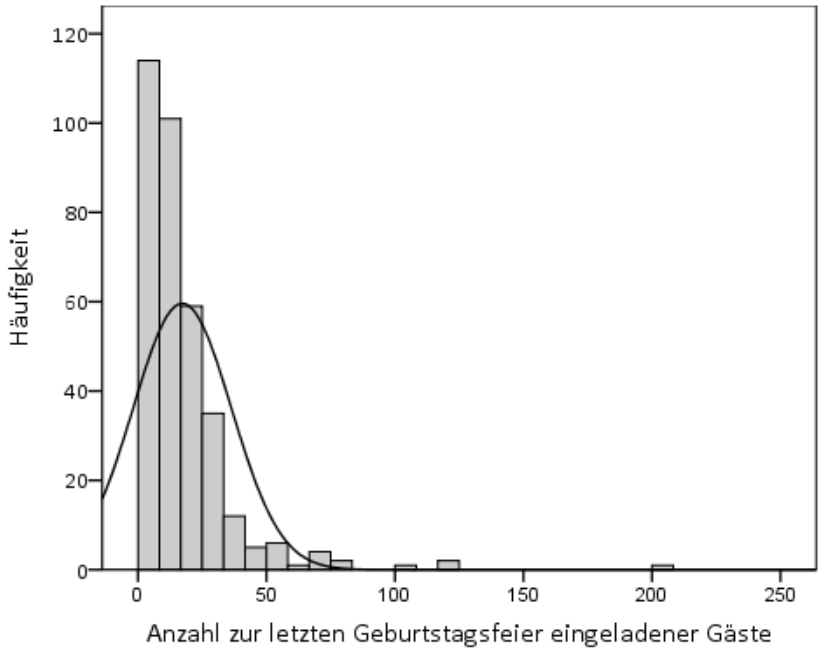

Abbildung E.1 - 1: Histogramm zur Anzahl der zur letzten Geburtstagsfeier eingeladenen Gäste
Untersuchungsgruppe 2B, $N = 343$

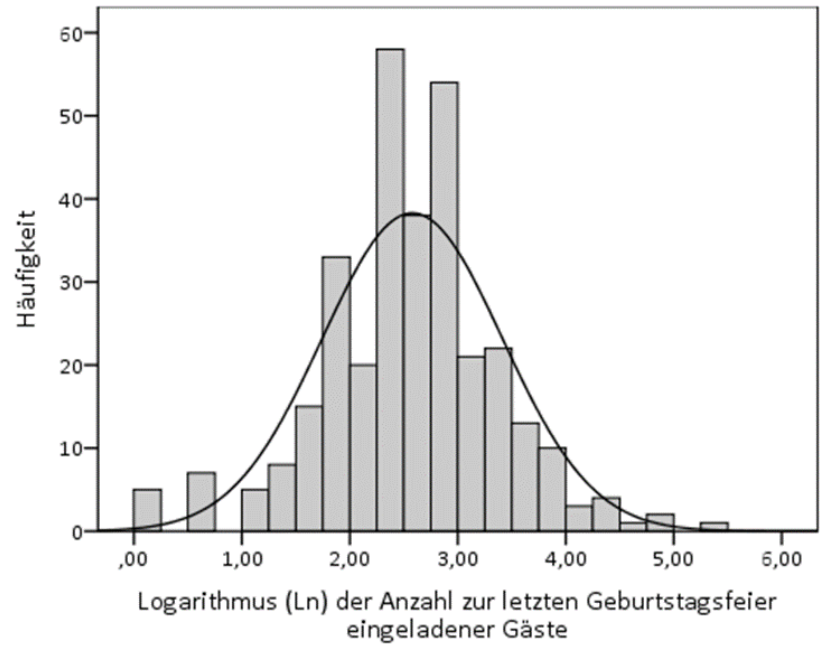

Abbildung E.1 - 2: Histogramm zum Logarithmus (Ln) der Anzahl der zur letzten Geburtstagsfeier eingeladenen Gäste
Untersuchungsgruppe 2B, $N = 319$

E.2 Skalen des ITB-PESA und Item-Beispiele

Tabelle E.2 - 1: Skalen der Vertriebsversion des ITB-PESA und Items mit der jeweils höchsten Trennschärfe im Auswahlkontext

Skala	Item mit höchster Trennschärfe (Polung) (r_{it})
Kontaktfreude	Im Kontakt mit Fremden finde ich ohne Probleme ein ergiebiges Gesprächsthema. (+) (.53)
Kommunikationsvermögen	Es fällt mir leicht, auch trockene Sachverhalte unterhaltsam zu präsentieren. (+) (.54)
Geselligkeit	Ich bin gerne unter Leuten. (+) (.43)
Einfühlungsvermögen	Ich bin sehr empfänglich für die Gefühle anderer. (+) (.42)
Erfolgszuversicht	Nach Misserfolgen bin ich einige Zeit entmutigt. (-) (.49)
Aufgeschlossenheit und Neugier	Dinge, die ich noch nicht richtig verstanden habe, lassen mir keine Ruhe. (+) (.45)
Eigeninitiative	Wenn ich auf etwas Neues stoße, versuche ich mehr darüber zu erfahren. (+) (.57)
Leistungsstreben und Erfolgsmotivation	Wenn ich etwas erreicht habe, bin ich nicht lange zufrieden und versuche, in Zukunft noch mehr zu erreichen. (+) (.56)

Untersuchungsgruppe 2A, $N = 367$

Tabelle E.2 - 2: Skalen der Vertriebsversion des ITB-PESA und Items mit der jeweils höchsten Trennschärfe im Nicht-Auswahl-Kontext

Skala	Item mit höchster Trennschärfe (Polung) (r_{it})
Kontaktfreude	Es fällt mir leicht, andere anzusprechen. (+) (.71)
Kommunikationsvermögen	*Es fällt mir leicht, auch trockene Sachverhalte unterhaltsam zu präsentieren.* (+) (.52)
Geselligkeit	Manche halten mich für einen Einzelgänger / eine Einzelgängerin. (-) (.54)
Einfühlungsvermögen	*Ich bin sehr empfänglich für die Gefühle anderer.* (+) (.55)
Erfolgszuversicht	Auch bei schwierigen Projekten bin ich mir sicher, dass ich sie erfolgreich abschließen werde. (+) (.68)
Aufgeschlossenheit und Neugier	Mich faszinieren Menschen, die „anders“ und ungewöhnlich sind. (+) (.41)
Eigeninitiative	Wenn ich Gegebenheiten für verbesserungswürdig halte, dann packe ich zu und ändere etwas. (+) (.53)
Leistungsstreben und Erfolgsmotivation	*Wenn ich etwas erreicht habe, bin ich nicht lange zufrieden und versuche, in Zukunft noch mehr zu erreichen.* (+) (.55)

Untersuchungsgruppe 1A, $N = 405$

kursiv: Im Auswahl- und Nicht-Auswahl-Kontext hat dasselbe Item die höchste Trennschärfe.

E.3 Skaleninterkorrelationen und –statistiken zum ITB-PESA

Tabelle E.3 - 1: Kolmogorov-Smirnov-Tests auf Ablehnung der Normalverteilungsannahme und Skaleninterkorrelationen zur Vertriebsversion des ITB-PESA im Auswahlkontext

	K-S-Test		Korrelation zu						
Skala	Z	p	Kv	Gs	Ev	Ez	AN	Ei	LE
Kontaktfreude	1.461	.028	.54	.54	.22	.50	.40*	.44	.32**
Kommunikationsvermögen (Kv)	1.113	.168		.49	.23	.60	.50	.64	.49**
Geselligkeit (Gs)	1.488	.024			.34	.50	.35	.45*	.23**
Einfühlungsvermögen (Ev)	1.039	.230				.07	.37*	.15	.13**
Erfolgszuversicht (Ez)	0.933	.349					.39	.71*	.48**
Aufgeschlossenheit und Neugier (AN)	1.355	.051						.54	.40**
Eigeninitiative (Ei)	1.589	.013							.56**
Leistungsstreben und Erfolgsmotivation (LE)	1.583	.013							

Untersuchungsgruppe 2A, $N = 367$

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

Mittels Fishers Z-Tests wurden die Korrelationen zwischen Auswahl- und Nicht-Auswahl-Kontext verglichen: * $p < .05$, ** $p < .01$

Tabelle E.3 - 2: Kolmogorov-Smirnov-Tests auf Ablehnung der Normalverteilungsannahme und Skaleninterkorrelationen zur Vertriebsversion des ITB-PESA im Nicht-Auswahl-Kontext

	K-S-Test		Korrelation zu						
Skala	Z	p	Kv	Gs	Ev	Ez	AN	Ei	LE
Kontaktfreude	1.032	.237	.51	.56	.22	.48	.26	.38	.13
Kommunikationsvermögen (Kv)	1.465	.027		.40	.19	.59	.43	.60	.31
Geselligkeit (Gs)	1.402	.039			.39	.39	.30	.33	-.07
Einfühlungsvermögen (Ev)	1.073	.200				-.06	.23	.13	-.08
Erfolgszuversicht (Ez)	1.207	.108					.30	.61	.27
Aufgeschlossenheit und Neugier (AN)	1.408	.038						.45	.18
Eigeninitiative (Ei)	1.651	.009							.41
Leistungsstreben und Erfolgsmotivation (LE)	0.920	.366							

Untersuchungsgruppe 1A, $N = 405$

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

Anhang F Ergänzungen zu den Ergebnissen in Studie 2

F.1 Analyse der Inter-Item-SD der Skalen des ITB-PESA

Tabelle F.1 - 1: Analyse der Inter-Item-SD der Skalen der Vertriebsversion des ITB-PESA im Auswahlkontext

	Inter-Item-SD						korr. Inter-Item-SD		
	M	SD	K-S-Test		a		K-S-Test		a
Skala			Z	p		R^2	Z	p	
Kontaktfreude	1.29	0.44**	1.229	.097	.53	.184	1.194	.115	.67
Kommunikationsvermögen	1.09**	0.43	0.946	.333	.53	.256	1.719	.005	.68
Geselligkeit	1.21	0.46**	0.899	.395	.45	.283	0.712	.691	.61
Einfühlungsvermögen	1.49**	0.40	0.519	.950	.37	.060	0.522	.948	.65
Erfolgszuversicht	1.06*	0.38**	0.940	.340	.66	.239	1.428	.034	.71
Aufgeschlossenheit und Neugier	1.01	0.43	1.271	.079	.40	.354	1.800	.003	.64
Eigeninitiative	1.02**	0.43**	1.697	.006	.52	.267	1.361	.049	.65
Leistungsstreben und Erfolgsmotivation	1.13	0.49**	1.372	.046	.45	.258	1.041	.228	.63

Untersuchungsgruppe 2A, $N = 367$

Inter-Item-SD: intraindividuelle Standardabweichung pro Skala, korr. Inter-Item-SD: intraindividuelle Standardabweichung pro Skala, korrigiert um Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts

M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, R^2: Anteil durch den Item-Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts aufgeklärter Varianz der Inter-Item-SD, a: Ladung auf dem Faktor der jeweiligen Faktorenanalyse

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

Unterschiede zwischen Auswahl- und Nicht-Auswahl-Kontext: * $p < .05$, ** $p < .01$ (T-Tests für Mittelwerte, Levene-Tests für den Vergleich von Varianzen). Für die korrigierten Inter-Item-SD werden weder Gruppen-Mittelwert noch -Standardabweichung berichtet, da es sich um Residuen handelt, deren Gruppen-Mittelwert jeweils Null ist. Die Standardabweichung ergibt sich aus der Standardabweichung für die Inter-Item-SD und R^2.

Tabelle F.1 - 2: Analyse der Inter-Item-SD der Skalen der Vertriebsversion des ITB-PESA im Nicht-Auswahl-Kontext

	Inter-Item-SD						korr. Inter-Item-SD		
	M	SD	K-S-Test		a		K-S-Test		a
Skala			Z	p		R^2	Z	p	
Kontaktfreude	1.24	0.39	1.118	.164	.41	.157	1.092	.184	.61
Kommunikationsvermögen	1.21	0.38	1.164	.133	.51	.187	1.053	.218	.64
Geselligkeit	1.24	0.38	1.139	.149	.59	.186	0.913	.376	.68
Einfühlungsvermögen	1.27	0.39	0.942	.338	.52	.137	0.893	.403	.66
Erfolgszuversicht	1.12	0.32	1.068	.204	.55	.102	1.068	.204	.67
Aufgeschlossenheit und Neugier	1.01	0.37	1.561	.015	.43	.213	1.141	.148	.61
Eigeninitiative	1.10	0.34	1.498	.023	.53	.126	1.594	.012	.65
Leistungsstreben und Erfolgsmotivation	1.15	0.39	1.197	.114	.42	.155	1.294	.070	.69

Untersuchungsgruppe 1A, $N = 405$

Inter-Item-SD: intraindividuelle Standardabweichung pro Skala, korr. Inter-Item-SD: intraindividuelle Standardabweichung pro Skala, korrigiert um Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts

M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, R^2: Anteil durch den Item-Mittelwert und das Quadrat des (z-standardisierten) Mittelwerts aufgeklärter Varianz der Inter-Item-SD, a: Ladung auf dem Faktor der jeweiligen Faktorenanalyse

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

Unterschiede zwischen Auswahl- und Nicht-Auswahl-Kontext: * $p < .05$, ** $p < .01$ (T-Tests für Mittelwerte, Levene-Tests für den Vergleich von Varianzen)

Für die korrigierten Inter-Item-SD werden weder Gruppen-Mittelwert noch -Standardabweichung berichtet, da es sich um Residuen handelt, deren Gruppen-Mittelwert jeweils Null ist. Die Standardabweichung ergibt sich aus der Standardabweichung für die Inter-Item-SD und R^2.

F.2 Scree-Plots für die Faktorenanalysen der korrigierten Inter-Item-SD

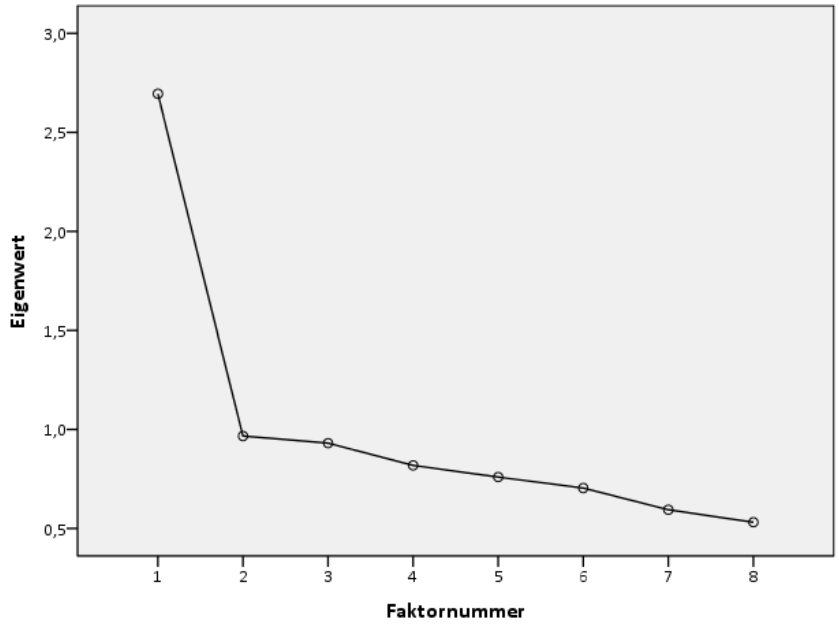

Abbildung F.2 - 1: Scree-Plot zur Faktorenanalyse der korrigierten Inter-Item-SD der Skalen der Vetriebsversion des ITB-PESA im Auswahlkontext

Untersuchungsgruppe 2A, $N = 367$

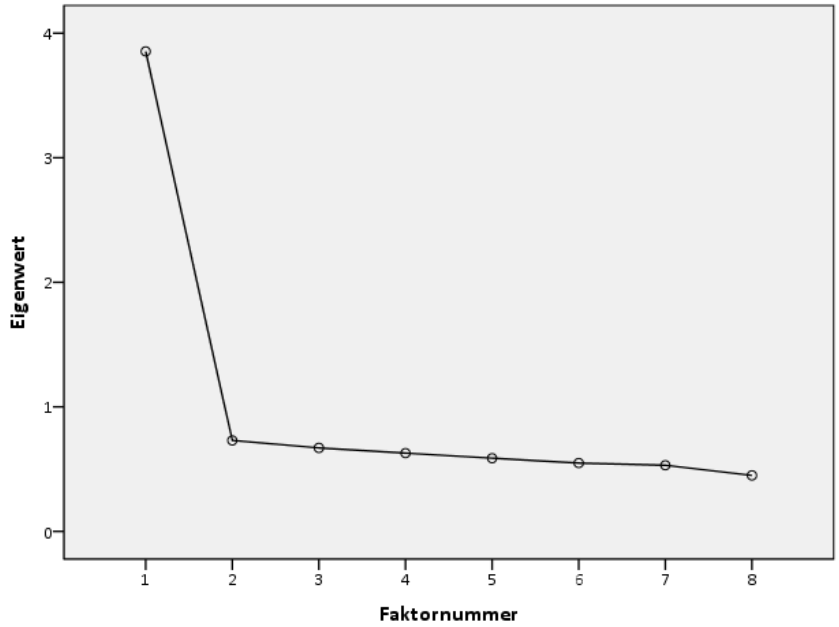

Abbildung F.2 - 2: Scree-Plot zur Faktorenanalyse der korrigierten Inter-Item-SD der Skalen der Vetriebsversion des ITB-PESA im Nicht-Auswahl-Kontext

Untersuchungsgruppe 1A, $N = 405$

F.3 Analyse der Extremwerthäufigkeiten für die Skalen der Vertriebsversion des ITB-PESA

Tabelle F.3 - 1: Analyse der Extremwerthäufigkeiten auf den Skalen der Vertriebsversion des ITB-PESA im Auswahlkontext

Skala	Zahl der Items	M	SD	K-S-Test		a
				Z	p	
Kontaktfreude	9	3.40	2.41	2.723	<.001	.79
Kommunikationsvermögen	10	3.75	2.61	2.604	<.001	.87
Geselligkeit	11	5.28	2.82	1.567	.015	.79
Einfühlungsvermögen	10	3.17	2.39	2.986	<.001	.74
Erfolgszuversicht	14	4.75	3.66	2.576	<.001	.83
Aufgeschlossenheit und Neugier	10	4.43	2.71	2.249	<.001	.77
Eigeninitiative	11	4.66	2.90	1.971	.001	.87
Leistungsstreben und Erfolgsmotivation	9	3.11	2.53	2.490	<.001	.75

Untersuchungsgruppe 2A, $N = 367$

M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, a: Ladung auf dem Faktor der Faktorenanalyse

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

Tabelle F.3 - 2: Analyse der Extremwerthäufigkeiten auf den Skalen der Vertriebsversion des ITB-PESA im Nicht-Auswahl-Kontext

Skala	Zahl der Items	M	SD	K-S-Test		a
				Z	p	
Kontaktfreude	9	2.17	1.95	3.679	<.001	.62
Kommunikationsvermögen	10	2.33	1.97	3.592	<.001	.79
Geselligkeit	11	2.75	2.21	3.039	<.001	.67
Einfühlungsvermögen	10	2.25	2.01	3.351	<.001	.69
Erfolgszuversicht	14	2.42	2.49	4.049	<.001	.76
Aufgeschlossenheit und Neugier	10	3.19	2.32	2.469	<.001	.72
Eigeninitiative	11	3.08	2.45	3.030	<.001	.83
Leistungsstreben und Erfolgsmotivation	9	1.75	1.83	4.743	<.001	.62

Untersuchungsgruppe 1A, $N = 405$

M: Gruppen-Mittelwert, SD: Gruppen-Standardabweichung, a: Ladung auf dem Faktor der Faktorenanalyse

K-S-Test: Kolmogorov-Smirnov-Test auf Ablehnung der Normalverteilungsannahme (Ablehnung bei signifikantem Ergebnis), Z: Teststatistik des K-S-Tests, p: Signifikanzniveau des K-S-Tests

F.4 Scree-Plots für die Faktorenanalysen der Extremwerthäufigkeiten

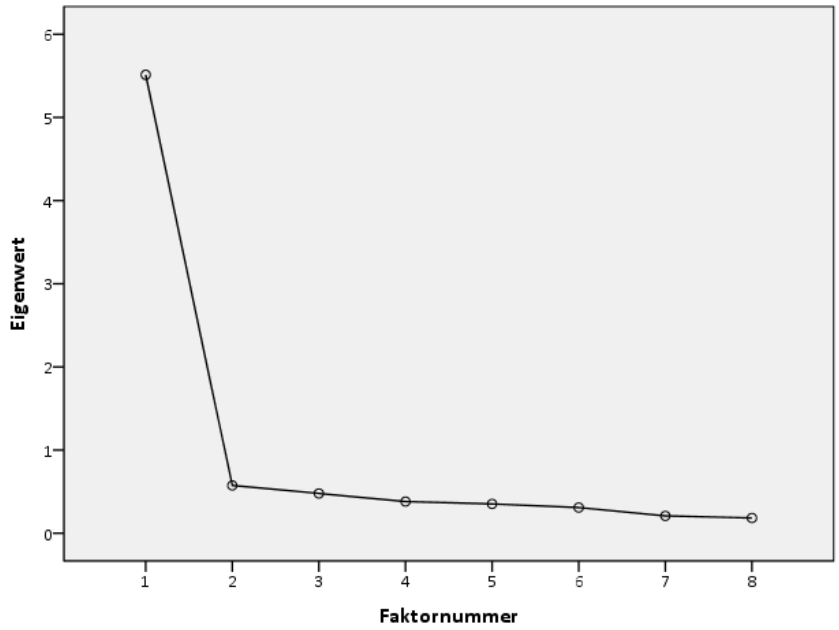

Abbildung F.4 - 1: Scree-Plot zur Faktorenanalyse der Extremwerthäufigkeiten auf den Skalen der Vertriebsversion des ITB-PESA im Auswahlkontext

Untersuchungsgruppe 2A, $N = 367$

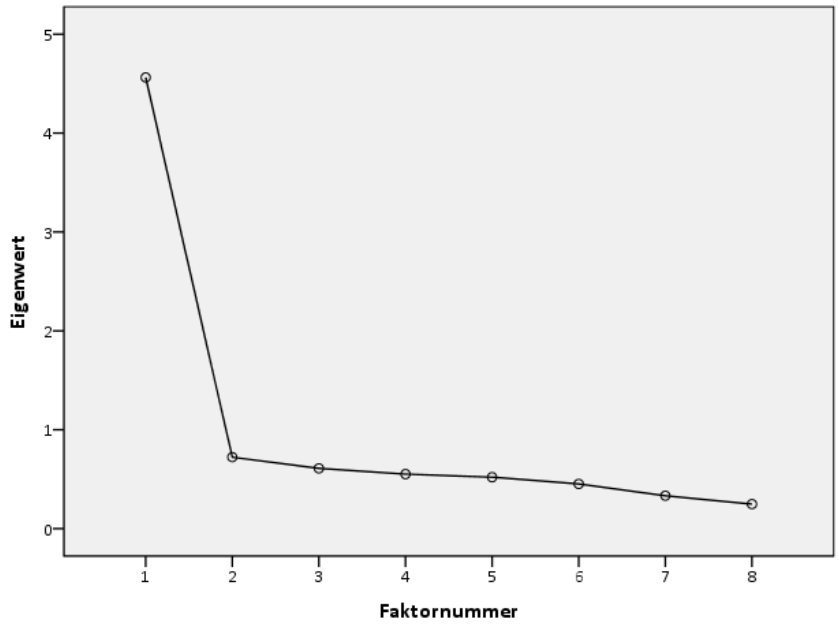

Abbildung F.4 - 2: Scree-Plot zur Faktorenanalyse der Extremwerthäufigkeiten auf den Skalen der Vertriebsversion des ITB-PESA im Nicht-Auswahl-Kontext

Untersuchungsgruppe 1A, $N = 405$

F.5 Faktorenanalysen der Vertriebsversion des ITB-PESA

Tabelle F.5 - 1: Prüfung der Voraussetzungen für eine Faktorenanalyse der Skalen der Vertriebsversion des ITB-PESA

		Bartlett-Test auf Sphärizität			signifikante
Faktor der Skalen	KMO	χ^2	df	p	K-S-Tests
Auswahl	.85	1210.93	28	<.001	4 von 8
Nicht-Auswahl	.79	1076.87	28	<.001	4 von 8

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

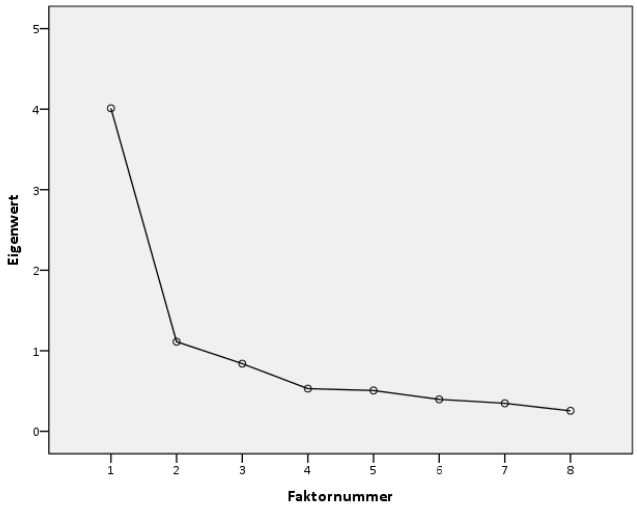

Abbildung F.5 - 1: Scree-Plot zur Faktorenanalyse der Skalen der Vertriebsversion des ITB-PESA im Auswahlkontext

Untersuchungsgruppe 2A, $N = 367$

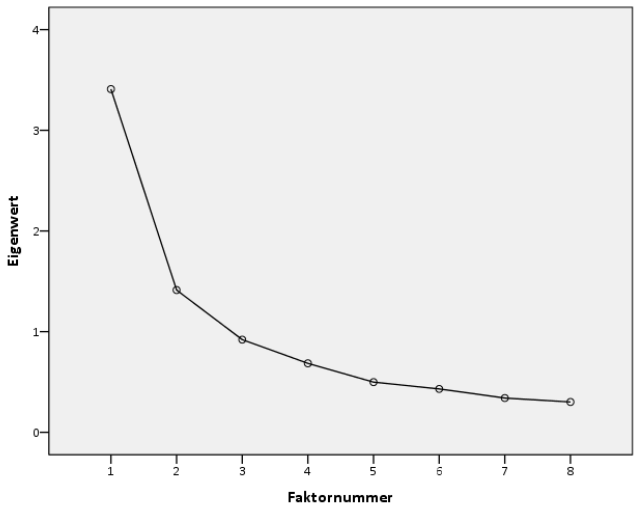

Abbildung F.5 - 2: Scree-Plot zur Faktorenanalyse der Skalen der Vertriebsversion des ITB-PESA im Nicht-Auswahl-Kontext

Untersuchungsgruppe 1A, $N = 405$

Tabelle F.5 - 2: Ladungen der Skalen der Vertriebsversion des ITB-PESA auf dem ersten Faktor einer Hauptachsenanalyse

	Ladungen		Vergleich zw. Auswahl u. Nicht-Auswahl	
Skala	Auswahl	Nicht-Auswahl	Fishers Z	p
Kontaktfreude	.65	.64	0.216	.829
Kommunikationsvermögen	.80	.80	-0.269	.788
Geselligkeit	.63	.56	1.434	.152
Einfühlungsvermögen	.30	.24	0.937	.349
Erfolgszuversicht	.77	.72	1.535	.125
Aufgeschlossenheit und Neugier	.63	.52	2.206	**.027**
Eigeninitiative	.81	.75	2.234	**.026**
Leistungsstreben und Erfolgsmotivation	.58	.32	4.632	**<.001**

Auswahl: Untersuchungsgruppe 2A, $N = 367$; Nicht-Auswahl: Untersuchungsgruppe 1A, $N = 405$

p: Signifikanzniveau zu Fishers Z-Test, p (zweiseitig) < .05 fett gedruckt

Eigenständigkeitserklärung

Ich, Dennis Beermann, geboren am 05.09.1985 in Düsseldorf, 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, 06.02.2015