What Would You Do—Really or Ideally? Constructs Underlying the Behavior Description Interview and the Situational Interview in Predicting Typical Versus Maximum Performance

Ute-Christine Klehe
Universiteit van Amsterdam

Gary Latham
University of Toronto

A predictive validation study was conducted to determine the extent to which the behavior description (BDI) and situational (SI) interviews predict typical versus maximum performance. Incoming MBA-students (*n* = 79) were interviewed regarding teamplaying behavior. Four months later, peers within study groups anonymously evaluated each person's typical teamplaying behavior, whereas other peers within project groups anonymously evaluated each person's maximum teamplaying behavior. Both the BDI and the SI predicted typical performance. The SI also predicted maximum performance. Implications and directions for future research are discussed.

The method most frequently used to select employees is the employment interview (Dipboye, 1994; Schuler, Frier, & Kauffmann, 1993). Structured interviews, such as the behavior description interview (BDI; Janz, 1982) and the situational interview (SI; Latham, Saari, Pursell, & Campion, 1980), are developed to take into account practicality, legal defensibility (Latham & Finnegan, 1993; Terpstra, Mohamed, & Kethley, 1999), and criterion-related validity (Huffcutt & Arthur, 1994). Numerous studies have investigated diverse aspects of the interviews' content

Correspondence should be sent to Ute-Christine Klehe, Programmagroep A&O Psychologie, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: u.klehe@uva.nl

(e.g., Huffcutt, Conway, Roth, & Stone, 2001) and structure (e.g., Taylor & Small, 2002), as well as characteristics of the interviewers (e.g., Huffcutt & Woehr, 1999). Relatively little research, however, has been conducted on the criteria that these interviews predict (e.g., Huffcutt, Conway, Roth, & Klehe, 2004), although the "criterion problem" remains a vexing issue in personnel selection (Viswesvaran & Ones, 2005). A further limitation of the extant research is the lack of knowledge as to the construct(s) that structured interviews assess (e.g., Van Iddekinge, Raymark, Eidson, & Attenweiler, 2004). This study attempts to fill these gaps by examining the literature on structured interviews with regard to typical versus maximum performance.

When selecting employees, organizations can strive to select candidates who can or candidates who will do the best job. This distinction between a person's maximum and typical performance (Sackett, Zedeck, & Fogli, 1988) has wide-reaching implications for organizations. Yet, organizations make huge financial investments in their selection procedures without taking into account which of these two aspects of performance they are trying to predict (Guion, 1991).

In the following sections, we outline (a) the distinction between typical and maximum performance, (b) its underlying theory, and (c) how the distinction is relevant to personnel selection. We then (d) describe the two interview formats used in this study, the BDI and SI, and on the basis of empirical research and theory we investigated whether these two types of interviews have different relationships with a person's behavior under typical versus maximum performance conditions, and (e) why the respective findings are important for both theory and practice in personnel selection.

TYPICAL AND MAXIMUM PERFORMANCE

Cronbach (1960) classified tests into two broad categories, tests of maximum and tests of typical performance. The former term designates tests that seek to assess how much or how well people can perform at their best, such as measures of cognitive ability and proficiency in terms of knowledge or skill (e.g., reading French). The distinguishing feature of these measures is that people must want to do well, and that they are encouraged to do so to earn the best score they can. In contrast, typical performance, Cronbach argued, is what people are likely to do in a given situation. "Tests of typical performance are used to investigate not what the person can do but what he does" (Cronbach, 1960, p. 31). Cronbach further stated the following: "The test of a suitable employee is whether she maintains that courtesy in her daily work even when she is not 'on her best behavior'" (p. 31). Cronbach concluded with the observation that a test of ability has relatively little practical value for predicting a person's characteristic behavior.

In applying Cronbach’s classification to job performance criteria, Sackett et al. (1988) argued that typical performance occurs when people are not aware that they are being evaluated, when they are not instructed to do their “very best,” and when their performance is assessed over an extended period of time. Maximum performance, in contrast, occurs when people know that their performance is being evaluated, when they receive instructions to exert great effort, and when the duration is short enough to enable performers to remain focused on the task. Sackett et al. (1988) further stated that in practice, typical and maximum performance represent a continuum rather than a dichotomy, which renders any comparison between them relative.

The differentiation between typical versus maximum performance mirrors the distinction between a person’s motivation and ability (Sackett & Larson, 1990). Although acknowledging that job performance is a function of both an individual’s ability and motivation (Locke, Mento, & Katcher, 1978; Maier, 1955), Sackett and his colleagues (DuBois, Sackett, Zedeck, & Fogli, 1993; Sackett & Larson, 1990) argued that in situations requiring maximum performance, motivation is constrained to be high. This is because people know that they are being monitored for only a brief period of time, and hence they accept instructions to exert a great deal of effort. “Unless one is inviting disciplinary action, one has little choice but to expend effort on the task in question … (and) accepting these instructions (to focus full attention on optimal performance) leads to a high level of effort” (DuBois et al., 1993, p. 206). Thus Sackett et al. (1988) concluded that differences among people on a measure of maximum performance reflect primarily differences in their ability.

Typical performance, they said, is a person’s normal behavior under everyday work conditions. Choice, level, and persistence of effort are relatively unconstrained by the organization in the day-to-day work setting. Thus, typical performance is largely a function of a person’s motivation. Typical performance reflects what a person “chooses to do” in drawing on ability to execute a task. Janz (1989, p. 164) summarized this distinction as follows: “Maximum performance focuses on competencies, whereas typical performance focuses on choices.” The tasks used where this distinction between typical and maximum performance has been studied include sensory motor, interpersonal, and administrative tasks in field and laboratory settings (e.g., DuBois et al., 1993; Klehe & Anderson, 2004; Ployhart, Lim, & Chan, 2001; Sackett et al., 1988).

Although rarely studied by industrial and organizational psychologists, the distinction between typical and maximum performance has wide-reaching implications for work settings in general (e.g., Dewberry, 2001; Sackett & Larson, 1990; Smith-Jentsch, Salas, & Brannick, 2001; Thorsteinson & Balzer, 1999) and for personnel selection in particular (e.g., Ackerman & Humphreys, 1990; Arvey & Murphy, 1998; Borman, 1991; Klehe & Anderson, 2005). Guion (1991) proposed that the relatively low correlation between measures of typical and maximum performance on the same task (e.g., Ployhart et al., 2001; Sackett et al., 1988) explains in part the low criterion-related validity coefficients for many predictors of job performance. Similarly, Campbell (1990) argued that hiring people on the basis of a predictor score of maximum performance could be one cause for the weak relationship that is subsequently found between results of that selection decision and a person's typical performance on the job. Boudreau (1991) pointed out that a mismatch between a predictor (e.g., of maximum performance) and a criterion measure (e.g., typical performance) adversely affects the results of utility analyses. A utility analysis will be biased if the dollar value is based on a maximum performance measure while one tries to predict an individual's typical performance, and vice versa. Consequently, both researchers and practitioners need to make explicit which aspect of performance they wish to predict, and which type of performance a given predictor actually predicts (Guion, 1998).

Addressing Construct Validity

Binning and Barrett (1989, p. 480) noted that during the theory-building process regarding a selection procedure's validity, the researchers "assume that two of the three inferences [predictor construct to predictor measure, predictor construct to performance construct, performance construct to performance measure] are correct and this, combined with empirical evidence of inference 1 [predictor measure to performance measure], allows a valid conclusion regarding the remaining inference."

Hence the arguments for differentiating typical from maximum performance are also applicable for identifying the constructs assessed by a selection procedure such as the BDI and the SI for predicting a person's performance: First, Sackett et al.'s (1988) arguments and the results of empirical research (Klehe & Anderson, 2004; McCloy, Campbell, & Cudeck, 1994) suggest that motivation influences typical performance more strongly than it influences maximum performance on the same task; and, facets of ability influence maximum performance more strongly than they influence typical performance (inference predictor construct to performance construct). Second, a manipulation of typical and maximum performance in terms of (a) instructions to exert effort, (b) perceived evaluation of one's performance, and (c) duration of the task, while (d) holding the relevance and observability of the respective behaviors constant across conditions, permits an appropriate assessment of the typical versus maximum performance domains (inference performance measure to performance construct; DuBois et al., 1993; Sackett et al., 1988). Finally, one may assess the empirical link between a predictor and a typical versus a maximum performance measure of the same task. Taken together, these three links (predictor construct to criterion construct, criterion construct to criterion measure, predictor measure to criterion measure) allow drawing the last inference, the link between the predictor measure and its underlying dimensions

(Binning & Barrett, 1989). In short, if the respective predictor succeeds better at predicting maximum performance than it does at predicting typical performance, then the predictor is primarily a measure of an individual's task-related abilities (DuBois et al., 1993). If, however, the predictor turns out to be a significantly better predictor of typical performance than of maximum performance, then it is primarily a measure of an individual's motivation. To date, no study has employed the distinction between typical and maximum performance to analyze which of these two constructs, ability or motivation, explains a selection procedure's validity. The purpose of this study was to do so with regard to two structured selection interviews, the BDI and SI.

STRUCTURED SELECTION INTERVIEWS

Hardly anyone would doubt the prevalence of interviews for selecting job candidates (Dipboye, 1994; Schuler et al., 1993), or the importance of structuring an interview to ensure that it has reasonable psychometric properties (e.g. Conway, Jako, & Goodman, 1995; Huffcutt & Arthur, 1994; Wiesner & Cronshaw, 1988). Two frequently used structured interviews are the SI (Latham et al., 1980) and the BDI (Janz, 1982). The SI consists of job-related dilemmas derived from a job analysis, the critical incident technique (CIT; Flanagan, 1954), and asks interviewees to describe what they would do in each situation. Interviewees' answers are then compared to a preestablished scoring guide for each question, which illustrates outstanding, acceptable, and unacceptable responses. The BDI asks each applicant the same questions asked of subject matter experts (SMEs) in conducting the CIT, namely: What were the circumstances in which you demonstrated this behavior in the past? What exactly did you do? What was the outcome? Is there someone who can verify this information?

Structured interviews in general, and SIs and BDIs in particular, have yielded high criterion-related validity coefficients for a wide range of job positions, performance criteria, and demographic groups (Huffcutt et al., 2004; Latham & Sue-Chan, 1999; Taylor & Small, 2002). A question that has yet to be resolved, however, is why these interview techniques have high criterion related validity. Currently, they suffer from the same "validity paradox" as assessment centers (Arthur, Woehr, & Maldegen, 2000; Klimoski & Brickner, 1987; Sackett & Dreher, 1982; Schneider & Schmitt, 1992). Although consistently demonstrating content- and criterion-related validity, evidence of their construct-related validity regarding the abilities, skills, and personality tendencies for which they had been developed to assess is low (Conway & Peneno, 1999; Huffcutt, Weekley, Wiesner, Degroot, & Jones, 2001; Schuler & Funke, 1989; Van Iddekinge et al., 2004). Mitchell (1985) has argued that only by focusing on the construct(s) being measured by a predictor can we understand a predictor's criterion-related validity. Moreover, identifying the underlying construct assessed by a predictor is important for increasing incremental validity (e.g., when combining structured interviews with other predictors).

The assumption underlying the SI is that it assesses intentions (e.g., Latham, 1989; Latham et al., 1980). As noted earlier, inherent in each SI question is a dilemma for which the ideal answer is not evident. The purpose of the dilemma is to minimize a socially desirable response to an interviewer's question (Latham & Skarlicki, 1995; Maurer, Sue-Chan, & Latham, 1999). Intentions, a core variable in social cognitive theory (Bandura, 1986), are defined as "a representation of a future course of action to be performed … a proactive commitment to bringing them (future actions) about" (Bandura, 2000, p. 5). Intentions are generally viewed as the direct motivational instigator of behavior (Fishbein & Ajzen, 1975; Lewin, 1951; Locke & Latham, 1990; Ryan, 1970).

Similarly, the underlying assumption of the BDI is that interviewees' answers "reveal specific choices applicants have made in the past, and the circumstances surrounding those choices" (Janz, 1989, p. 159). Based on the assumption that past behavior is the best predictor of future behavior, Janz (1989, pp. 159–160) argued that "the more long-standing the applicant's behavior pattern in the past, the more likely it will predict behavior in the future."

Janz (1989) stated that the SI is more likely an assessment of maximum rather than typical performance because it invites "ideal" responses. The focus in the SI is on the future, "what would you do?" Giving an "ideal" answer may be more an indicator of a person's knowledge than motivation. In contrast to the SI, the sole focus of the BDI is on "what have you done?" Janz argued that this focus is likely to tap primarily motivation rather than a person's ability. However, in answering an interviewer's question, an interviewee may choose to describe an incident from the past where maximum performance was exhibited on the job. There is as yet no empirical research on whether typical versus maximum performance is being assessed in either one or both of these interview formats.

More recently, Van Iddekinge, Raymark, and Roth (2005) outlined several features of structured interviews which support the notion that both of these interview formats assess an interviewee's motivation. Specifically, the diverse cognitive demands placed on interviewees during the interview likely mitigate them giving socially desirable responses (e.g., the challenge of conceiving reply-distortions that are consistent with what the receiver might already know, time constraints, and the need to maintain an ongoing positive interaction with an interviewer). In addition, a meta-analysis suggests that structured interviews do in fact assess a "generalized motivation factor" (Huffcutt, Roth, & McDaniel, 1996, p. 470). This, of course, does not imply that an interviewee's motivation and ability are unrelated. Latham (1989, p. 175) acknowledged the following: "It is likely that these intentions are affected by, or are related to, certain cognitive abilities and sociability skills." However, Latham's (1989) argument regarding the SI, and Janz's (1989) argument regarding the BDI suggest that these two structured interviews primarily assess a person's motivation and hence primarily predict an interviewee's typical performance on a day-to-day basis. If this is correct, the arguments presented regarding the relationships between motivation and ability on the one hand and typical versus maximum performance on the other hand lead to the following hypotheses:

H1–H2: Both the (H1) SI and (H2) the BDI are significantly better predictors of typical rather than of maximum performance.

Typical versus maximum performance is on a continuum. Although variance in maximum performance can largely be accounted for by what performers "can do," that is, by variance in their task-related abilities (DuBois et al., 1993), variance in typical performance on the same task is additionally explained by what the same performers "choose to do," that is, by variance in their motivation. Consequently, if the SI and the BDI are assessments of what individuals "choose to do" (their motivation), the inclusion of these interviews should add incremental validity to the prediction of typical performance, after maximum performance (an assessment of what performers "can do") has been accounted for.

H3–H4: Both the (H3) SI and (H4) BDI add incremental validity to the prediction of typical performance after accounting for individuals' maximum performance on the same task.

The importance of testing these hypotheses is at least threefold. First, the continuum between typical and maximum performance has received far more attention conceptually than empirically (Klehe & Anderson, 2005). Second, the logic underlying typical versus maximum performance suggests the use of this continuum to discover constructs underlying a selection procedure's validity. This study is the first to do so. Third, previous studies have analyzed a construct hypothesized to be assessed by a structured selection interview by correlating scores in an interview with an external assessment of the hypothesized construct. Typically this was done without testing whether the construct itself explained the interview's criterion-related validity (e.g., Conway & Peneno, 1999; Moscoso, 2000; Salgado & Moscoso, 2002; Van Iddekinge et al., 2004). This study is the first to test the hypotheses of the two developers of these two interview techniques, namely Janz (1989) and Latham (1989). They stated that their respective interview techniques, BDI and SI, assess the direct motivational antecedents of what individuals "will do" when they encounter a situation on the job similar to the one described during the interview. Support for Janz and Latham's hypotheses would strengthen the case for the interviews' incremental validity for ability and knowledge tests.

METHOD

Participants

The sample consisted of incoming students enrolled in a full-time MBA program at a large Canadian University. Seventy-nine (47%) out of 167 students were interviewed.[1] The interviewees were a representative sample of the incoming students in that 37% were women compared to 31% in the cohort. Their average age was 28 years (*SD* = 3.6) with 4.9 years of work experience (*SD* = 3), compared to 28 years (*SD* = 3.5) with 4.7 years work experience (*SD* = 3) in the cohort. Similarly, the GMAT scores of the sample participants (mean = 660, *SD* = 41.10) and the cohort group (mean = 669.89, *SD*= 53.9) did not differ ($p > .05$).

This sample was selected for six reasons: (a) The performance criterion was teamplaying behavior, a core competency of managerial careers (Allred, Snow, & Miles, 1996; Brodbeck et al., 2000) and an MBA education. (b) This criterion is frequently assessed in structured selection interviews (Huffcutt, Conway, et al., 2001). Hence its use in this study facilitates comparisons with previous studies. (c) The students had been admitted to the MBA program on factors unrelated to measures of teamplaying behavior, or their performance on the SI and the BDI items. Thus, restriction of range was unlikely to attenuate validity coefficients. (d) The participants had not yet entered the organization at the time of the interviews. This was important to minimize any contamination due to the socialization process (Louis, 1980) or decreased motivation during the interview (Arvey, Strickland, Drauden, & Martin, 1990). (e) The interviews were conducted within a narrow time period, which ensured that the duration between the assessment of the predictors and the criterion remained constant across interviewees. (f) As is described later, conducting this study in an MBA program enabled a parallel assessment of typical and maximum performance, as well as a manipulation check regarding these two measures.

Procedure

Predictors. Four 2nd-year MBA students serving as SMEs worked with the first author to collect 24 critical incidents on teamplaying behavior during the MBA program. On the basis of these incidents, they developed 16 SI and BDI questions that were as comparable as the differing question formats would allow. SMEs developed parallel scoring guides with behavioral benchmarks for outstanding (5), acceptable (3), and unacceptable (1) answers. Although a scoring guide is

[1] The sample size would likely have been higher if the business school had not switched its e-mail system during the time period in which this validation study occurred. Outgoing and incoming e-mails were lost without signaling error messages. In addition, the catastrophe of September 11, 2001, caused several interviews to be cancelled. This sample was used in an early study by Klehe and Latham (2005).

not a requirement of the BDI (Janz, 1989), previous research indicates that using a scoring guide improves the validity of BDIs (Taylor & Small, 2002).

A pilot test was conducted with 31 management students who did not take part in the validation study. Following the pilot test, the SMEs, together with eight doctoral students in human resource management, revised the scoring guide and eliminated the questions where they could not reach consensus on how to score interviewee answers, as well as questions for which there was little or no variance in the interviewees' responses. This resulted in nine SI and nine comparable BDI questions (see the Appendix for an example).

At the beginning of the interviews, interviewees were told that they would be asked questions about difficult situations that they are likely to encounter during the MBA program. The order of SI and BDI questions was randomized across participants. Consistent with SI procedures (Latham, 1989), no probing or prompting was done in SIs. For BDIs, this is not the case, as BDIs ask interviewees about past situations that they should have encountered (Janz, 1989). A pilot study was conducted to eliminate questions tapping areas that the interviewees had not experienced in their past. In the actual interview, interviewees who responded that they had never experienced a given situation were encouraged to recall instances similar to the one that the interviewer wished to probe, regardless of whether the situation was in a school, work, or private context (Campion, Palmer, & Campion, 1997; Janz, 1982; Orpen, 1985). In the rare instance that participants could still not report an incident, interviewees were assigned a score of 1 (unacceptable).

Each participant was interviewed by the first author and a doctoral student, who had been trained as an interviewer, but who was blind to the purpose of the study. This second interviewer read the questions to the participant, and both interviewers independently recorded and scored the responses on the basis of the scoring guide. Following each interview, the two interviewers discussed the scoring of each response on which they differed by more than 1 point on the 5-point scale.

Qualitative data suggest that the interviewees treated the interview process seriously. For the majority of them, the interview was their first personal contact with the business school. Most of them entered the interview in business attire and engaged the interviewers in a friendly yet formal manner. Approximately two thirds of the interviewees contacted the first author after the interview to obtain feedback on their interview performance.[2]

Criteria. Both the typical and maximum performance measures were team-playing behavior in an MBA program. Both were assessed through anonymous peer appraisals on the same behavioral observation scales (BOS; Latham &

[2] The interviewees were aware that this feedback would only be given after completion of their first term in the MBA program. This was done to ensure that feedback on the interview had no effect on participants' typical and maximum performance.

Wexley, 1977, 1994) toward the end of the students' first term in the MBA program. Peer appraisals were used because peers have many opportunities to observe each other during their work (Love, 1981; Mumford, 1983). BOS were used because they are content-valid, correlate with hard criteria, and have high interobserver reliability (e.g., Latham & Wexley, 1977, 1994). They are associated with high clarity and user acceptance (Tziner & Kopelman, 1988), high subsequent performance, and satisfaction with the appraisal process (Tziner, Kopelman, & Livneh, 1993).

The BOS used for the assessments of both typical and maximum performance were developed by Sue-Chan and Latham (2004) to assess teamplaying in a comparable MBA program. Fourteen behavioral items such as "coordinates upcoming work with group members who are involved with the work," "knows content of aspects of the project completed by other group members," "acts as a mediator to resolve conflicts among groups," and "motivates teammates to produce extremely high standards of work," are answered on a 5-point Likert-type scale, ranging from 0 (*almost never*) to 4 (*almost always*). Sue-Chan and Latham (2004) found that the BOS correlates significantly with students' grade point average in the MBA program ($r = .61$, $p < .05$). This study used 13 of the 14 items, excluding the item "knows how to use software necessary for completing an assignment," as it was not relevant in this context.

As previously noted, typical and maximum performance are on a continuum, making comparisons between the two relative. Prior research has employed relatively strong manipulations, with assessments of maximum performance lasting minutes (Klehe & Anderson, 2004; Sackett et al., 1988; Smith-Jentsch et al., 2001), hours (Kirk & Brown, 2003), or days (Ployhart, Lim, & Chan, 2001). Very short observation time periods, especially when assessing sociopsychological variables such as teamplaying, are vulnerable to at least two threats to validity relative to longer observation periods (Cascio, 1998). First, different dimensions of performance may arise when a person's behavior in the typical performance condition includes tasks that are not relevant during a short-term maximum performance assessment (e.g., preparing for group meetings). As a result, the tasks used to assess typical performance are no longer the same as those used to assess maximum performance. Second, if the assessment of maximum performance is based on far fewer observations than the assessment of typical performance, the comparability of reliability estimates between the two assessments is not equivalent (Klehe & Anderson, 2005).

Consequently, this study employed a relatively weak manipulation to increase the reliability of the assessments of maximum performance. Rather than assess maximum performance over 2 days, as was done by Ployhart et al. (2001), we assessed maximum performance over a 5-day period. This was done with the knowledge that a weak manipulation, 5 days, as opposed to a strong manipulation, a few minutes (e.g., Klehe & Anderson, 2004), might produce relatively small statistical effects. This limitation is offset by the fact that, as Prentice and Miller (1992) pointed out, the inferences drawn from a weak manipulation are more likely to generalize to other settings than are those made on the basis of strong manipulations.

Maximum performance in this study was defined as the arithmetic mean of scores given anonymously to a participant by peers during a 5-day team project. During this project, teams of five to six students had to analyze a case study and develop a written set of recommendations based on theory as well as develop an action plan on how to implement those recommendations. Teams had been formed by the lecturer of that course based on the requirements that students should be as diverse as possible regarding their professional background, race, gender, and percentage of people whose first language is English. No two members of the same study group (discussed later) could be assigned to the same project group. The project accounted for 25% of the students' final grade in the course in which the present authors were not involved. Five of the 25% was determined by peer evaluations of a student's teamplaying behavior.

The rationale for using project performance as an assessment of maximum performance is threefold. First, consistent with Sackett et al.'s (1988) requirement for high awareness of one's evaluation by others throughout the evaluation period, the course instructor informed the students of the ongoing peer assessment of teamplaying performance throughout the project. Second, the instructor explained to the students the necessity to perform at one's best, and reminded them that their teamplaying performance during the project would influence their course grade. Course grades are crucial in a first year MBA program to attain a relevant summer job. Hence the students were under high pressure to perform at their best. These factors satisfy Sackett et al.'s second requirement for maximum performance. Third, the project lasted only 5 days. Relative to the assessment of typical performance throughout the entire semester, the students could focus their attention on performing well on this project. The project was not unlike those that confront people in the public and private sectors. The participants were confronted with a pressing problem. The causes of the problem were initially vague until the situation was analyzed in detail. The participants were required to quickly formulate directions for future actions within a relatively short time period.

Typical performance was defined as the arithmetic mean of scores given to a student by peers within the person's study-group, which in total consisted of five to six individuals. They too had been assigned to groups by the school based on the requirements that students within a group should be as diverse as possible regarding professional background, race, gender, and percentage of speakers whose first language is English. Study groups work together throughout the term on accomplishing group assignments, analyzing and discussing case studies and their possible solutions, and making presentations.

Typical performance was not assessed until the end of the students' first term in the MBA program, but before they had received their grades. This measure satisfies Sackett et al.'s (1988) definition of typical performance in that (a) the assessment represented a person's mean performance over an extended time period, namely 4 consecutive months as opposed to 5 days; (b) students were not aware during this time period that they would be evaluated by their peers at the end of the term; and (c) the students were not given any instructions by the faculty during that period to maximize their effort in their respective study groups.

This operationalization of typical versus maximum performance allowed for comparable assessments on these two criteria. Both assessments were based on appraisals by peers who had extensive contact with one another in their project and study groups. In both settings, peers within groups depended on each other's ability and motivation while working on similar tasks. Peers were thus the optimal source for assessing a participant's teamplaying behavior. Both assessments were conducted in temporal proximity to each other, namely at the end of the participants' first term in the MBA program. In both cases, peers evaluated participants on exactly the same behaviors using the same appraisal scale. To minimize the possibility that the assessments of maximum performance could influence the assessments of typical performance, no two appraisers from the same study group (typical performance) were members of the same project group (maximum performance). In addition, because each appraiser made both typical and maximum appraisals with the exact same scale (although on different targets—the members of their study groups in the typical and the members of their project group in the maximum performance condition), differences in maximum versus typical performance ratings cannot be influenced by differences in appraisers across the two conditions.

Manipulation checks. A limitation of earlier studies of typical and maximum performance (Klehe & Anderson, 2004; Ployhart et al., 2001; Sackett et al., 1988) is the lack of a manipulation check. Immediately after the respective assessments of maximum and typical performance in this study, the participants indicated the following on a scale from 0 (*almost never*) to 4 (*almost always*): (a) how often they had been aware that they were being evaluated, (b) how often they had done their very best to contribute to the group's performance, and (c) how focused they had been on their group's work.

In addition, we examined whether the behaviors assessed by the BOS were of comparable relevance in the typical and maximum performance conditions to ensure comparable dimensionality of performance (Cascio, 1998). We also examined the ease with which the behaviors could be observed in the typical and maximum performance situations to ensure comparable reliability of performance observation (Cascio, 1998). For this purpose, a different group of MBA students from the same MBA program, who had not been involved in the validation study, evaluated the BOS items on relevance and ease of observation in either their study groups (n = 51), or during their 5-day project group (n = 53). Relevance and observability were assessed on a 5-point Likert-type scale, ranging from 0 (*not at all*) to 4 (*very much*). The power of detecting a medium effect size on the rating of the 13 BOS items in the typical and maximum performance situation is .80 (Cohen, 1988, 1992).

RESULTS

Manipulation Checks

To determine whether the peer assessments of typical and maximum performance are distinct constructs, the BOS were subjected to an exploratory factor analysis. Bartlett's test of sphericity was significant ($\chi^2 = 3906$, $df = 1326$, $p < .001$). Hence the data did not represent an identity matrix and could be factor analyzed. A Scree plot revealed two factors that explained 68.1% of the variance. These two factors were extracted with principal component analysis and rotated oblimin. All assessments of maximum performance loaded .75 or higher on factor 1, but negligibly on factor 2. All assessments of typical performance loaded .70 or higher on factor 2, but negligibly on factor 1 (see Table 1). The correlation between the two factors was significant ($r = .30$, $p < .01$).

Independent two tailed t tests on the relevance of the 13 behaviors of the BOS ($M = 3.09$, $SD = .63$ vs. $M = 3.17$, $SD = .41$; $t = .70$, *ns*), as well as their ease of observation ($M = 2.71$, $SD = .58$ vs. $M = 2.83$, $SD = .58$; $t = 1.12$, *ns*), revealed no significant differences between the typical and maximum performance conditions, respectively.

The participants in both the typical and the maximum performance group rated the relevance ($t = 4.10–18.09$, $p > .01$) and the ease of observation ($t = 2.39–16.84$, $p > .05$) of all items to be significantly greater than 2, the graphic midpoint of the Likert-type scale. This indicates that the items were relevant and observable in both conditions, suggesting comparable dimensionality of performance and reliability of observation (Cascio, 1998).

Finally, to test whether the two performance measures assessed typical versus maximum performance, the means of the three manipulation-check items provided by participants directly following their peer assessments were compared via a paired t test. Participants perceived the maximum performance condition ($M = 3.16$, $SD = .53$) as significantly more "maximum" than the typical performance condition ($M = 2.90$, $SD = .56$; $t = 3.47$, $p < .001$). In sum, these results indicate that (a) the assessments of typical and maximum performance represent two interrelated factors with (b) no differences regarding the relevance or observability of the respective behaviors. Yet, participants (c) perceived the maximum condition as

TABLE 1
Exploratory Factor Analysis: Factor Loadings
for Peer Assessments of Typical and Maximum Teamplaying Performance

Item	*Factor 1 (Maximum)*	*Factor 2 (Typical)*
Maximum 1	.83	.13
Maximum 2	.82	.04
Maximum 3	.83	.01
Maximum 4	.83	–.07
Maximum 5	.84	.03
Maximum 6	.85	–.03
Maximum 7	.84	.06
Maximum 8	.89	–.02
Maximum 9	.78	–.15
Maximum 10	.87	.00
Maximum 11	.82	.09
Maximum 12	.75	.15
Maximum 13	.84	.04
Typical 1	.00	.86
Typical 2	–.06	.83
Typical 3	–.13	.83
Typical 4	–.12	.87
Typical 5	.10	.80
Typical 6	.12	.81
Typical 7	.00	.87
Typical 8	.07	.85
Typical 9	.00	.72
Typical 10	.12	.70
Typical 11	.08	.73
Typical 12	.06	.76
Typical 13	–.08	.82
Eigenvalues	11.73	5.99
% variance explained	45.10	23.02
Cumulative variance explained	45.10	68.12

Note. $N = 162$.

more “maximum” than the typical performance condition. Thus, the manipulation of typical versus maximum performance in this study was successful.

Reliability

The interrater reliability was .90 for the SI and BDI, respectively. The interrater reliability of typical as well as maximum performance was calculated using intraclass correlations (Cohen, Cohen, West, & Aiken, 2003). The single measure intraclass correlation was .67 for typical, and .71 for maximum performance. Cronbach’s alpha was .96 for both typical and maximum performance, .50 for the

SI and .71 for the BDI. Although the SI's internal consistency was lower than that of the BDI, this is an acceptable level for structured interviews. Conway, Jako, and Goodman (1995) found that average interitem correlations in interviews diminish with interview structure, particularly when the interviews are based on a job analysis. Descriptive statistics and correlations between variables are reported in Table 2.

PREDICTING TYPICAL AND MAXIMUM PERFORMANCE

Two approaches were followed to test the predictive validity of the BDI and SI of typical and maximum performance. First, the predictive validity coefficients of the two structured interviews for typical and maximum performance were compared (H1, H2). Second, multiple regression was used to examine whether the interviews accounted for incremental validity in the prediction of typical performance after accounting for variance in maximum performance (H3, H4).

Predictive Validity

To control for interviewer effects, the data were examined separately in terms of the validity coefficients obtained by the interviewer who was aware of the purpose of this study, versus the validity coefficients obtained by the interviewer who was blind to the hypotheses. Regardless of the interview format, the two interviewers did not differ in the prediction of participants' typical or maximum performance (t_W = .16–.80; *df* = 76, *ns*). Similarly, there was no difference regarding the average score of the interview questions (*t* = .60–1.13, *ns*) or (β = .02–.09, *ns*) regardless of whether a SI or a BDI question was presented first. Mahalanobis D^2 (Tabachnick & Fidell, 2001) revealed the absence of outliers.

The SI was a significant predictor of both typical (*r* = .41, *p* < .01; 90% confidence interval [CI] = .25–.55) and maximum performance (*r* = .25, *p* < .05; 90% CI = .07–.41). Several procedures exist for comparing dependent correlations (Hotelling, 1940; Meng, Rosenthal, & Rubin, 1992; Olkin, 1967; Williams, 1959). Because these procedures led to the same conclusions in all of the analyses presented later, the results are reported only for the procedure outlined by Williams (1959).

A one-tailed *t* test, consistent with H1, revealed that the predictive validity coefficient of the SI for typical performance was marginally higher than for maximum performance (t_W =1.32; *df* = 72, *p* < .10). The predictive validity coefficient of the BDI was significant for typical (*r* = .34, *p* < .01; 90% CI = .17–.50), but not maximum performance (*r* = .11, *ns*; 90% CI = –.08–.29). The difference between these two validity coefficients was significant in the predicted direction (t_W =1.72; *df* = 72, *p* < .05), supporting H2.

TABLE 2
Means, Standard Deviations, Internal Consistencies and Intercorrelations Among Study Variables

	n	*M*	*SD*	*Age*	*Gender*	*SI*	*BDI*	*SI and BDI*	*Typical Performance*	*Maximum Performance*
Age	167	28.18	3.48	—						
Gender[a]	167	0.31	.46	–.03	—					
SI	79	3.00	.47	–.15	.07	(.50)				
BDI	79	2.68	.59	–.18	.01	.48**	(.71)			
SI and BDI combined	79	2.84	.46	–.19	.04	.82**	.89**	(.69)		
Typical performance	162	3.18	.54	–.24**	.12	.41**	.34**	.43**	(.96)	
Maximum performance	167	3.37	.45	–.26**	.08	.25*	.11	.20	.34**	(.96)

Note. SI = Situational interview; BDI = behavior decripton interview. Internal consistencies are reported in the diagonal.
[a]male = 0, female = 1, correlations are point biserial, because the variable is dichotomous.
**p* < .05. ***p* < .01.

Multiple Regression

Next, we used regression to test the incremental validity of the two structured interviews for predicting typical performance beyond the prediction of maximum performance (see Table 3). The difference between typical and maximum performance is due primarily to a person's motivation (e.g., McCloy et al., 1994; Sackett et al., 1988). Hence, any variable adding incremental validity to the prediction of typical performance beyond that which has been accounted for by an assessment of maximum performance on the same task is likely to be motivational in nature.

The results revealed that both the SI ($\beta = .35$, $p < .01$) and the BDI ($\beta = .30$, $p < .01$) added significant variance to the prediction of typical performance after taking into account maximum performance as a predictor. These findings support H3 and H4. The structured interviews primarily assess an interviewee's task-related motivation. In addition, the inclusion of the SI into the regression on typical performance lowered the unique impact of maximum performance on typical performance from $\beta = .34$, $p < .01$ to $\beta = .25$, $p < .05$. A Sobel test revealed that this effect was significant ($z = 2.24$, $p < .05$). This suggests that the SI accounted for variance common to performance under both typical and maximum performance conditions. Given that the main component common across typical and maximum performance on the same task is a person's task related ability (Sackett et al., 1988), this finding suggests that the SI assessed those aspects of an individual's task related knowledge or ability that is relevant to teamplaying under both typical and maximum performance conditions.

DISCUSSION

The results of this study suggest that the BDI assesses primarily an interviewee's motivation rather than ability. The SI too appears to assess a person's motivation. In addition, it accounted for variance due to ability in typical as well as maximum performance. Furthermore, the SI accounted for incremental validity in the prediction of typical performance beyond that which had been accounted for by maxi-

TABLE 3
Typical Performance Regressed on Interviews and Underlying Constructs

	Situational Interview				*Behavior Description Interview*			
Variable	*F*	*df*	R^2	β	*F*	*df*	R^2	β
Maximum performance	9.48**	73	.11	.34**	9.48**	73	.11	.34**
Maximum performance	10.74**	72	.23	.25*	9.30**	72	.21	.30**
Interview				.35**				.30**

*$p < .05$. **$p < .01$.

mum performance. In short, the SI not only measured what interviewees were able to do, but even more so what interviewees chose to do on a day-to-day basis.

The significance of this study is fivefold in that it addresses empirical, theoretical, methodological, and practical considerations. The introduction of the typical or maximum performance distinction appeared in the literature years ago (Sackett et al., 1988). Although it is acknowledged to be important (e.g., Borman, 1991; Boudreau, 1991; Campbell, 1990; Guion, 1991, 1998; Sackett & Larson, 1990), it has generated very little empirical research. This study contributes to the emerging empirical findings regarding the importance of distinguishing between typical and maximum performance. Both the observations of the participants, as well as the factor analysis, suggest that typical and maximum performance are distinct although related constructs. Moreover, the predictive validation and regression analyses show that different conclusions are reached when one performance measure is used rather than the other. Despite the few preceding empirical studies, this appears to be a relatively robust finding given the diversity of the populations from which the samples were drawn, namely, grocery store workers, military personnel, participants in a laboratory experiment, and in this study, MBA students, as well as the different countries in which these studies were conducted, namely, Canada, the Netherlands, Singapore, and the United States, and as well as the different criterion measures that have been used, namely, speed of processing grocery store items, transformational leadership, performance on an administrative task, and team-playing behavior in an MBA program.

Building on the knowledge gathered regarding the changing role of ability and motivation during typical and maximum performance (Klehe & Anderson, 2004; Sackett et al., 1988), this study is the first to draw on the distinction between typical and maximum performance for delineating the basic nature of the constructs underlying the criterion-related validity of the BDI and SI.

Of further theoretical as well as practical significance is that a person's motivation can be assessed by focusing on either past choices or future intentions in an interview setting. The validity coefficients regarding the SI for both typical and maximum performance provide support for Latham's (1989) assertion that the motivation voiced in a SI may well be influenced by an interviewee's ability. Furthermore, only the SI predicted performance in both performance domains. This may be due to the situational specificity of the SI.[3] The SI captures situations likely to be encountered in the respective performance domain. This situational specificity likely explains the predictive validity even when the internal consistency of answers is low. In the context of this study, this effect may have been particularly pronounced due to the clearly defined context of the study (teamplaying within a relatively standardized MBA program). The BDI, in contrast, asks participants to

[3] We thank an anonymous reviewer for raising this suggestion.

recall situations matching the major characteristics of the question. It does not, however, impose the same situational boundaries interviewees will encounter on the job. That is, if applicants can report a suitable situation from their past, this situation and their reaction to it will be additionally influenced by the context (e.g., by their past organization having a different organizational climate than the current one). Thus, the suitability of the match achieved between reported and predicted situation may depend not only on participants' understanding of the question, but also on their prior experience in similar performance contexts (or the absence thereof) and their capability to remember such situations. This suggests that these variables may serve not as predictors accounting for the BDI's criterion-related validity (testable via mediation analyses), but as moderators influencing validity.

A methodological contribution of this research to the study of typical and maximum performance is that it is the first to use manipulation checks for (a) ensuring that typical and maximum performance conditions differ on the three dimensions outlined by Sackett et al. (1988), and (b) for ensuring that the dimensions of typical and maximum performance are comparable (Cascio, 1998). This was indicated in this study by the comparability on the relevance of the behaviors evaluated, and the reliability of the performance observations (Cascio, 1998) as indicated by the ease with which the behaviors were observed in the typical and maximum performance situations, respectively. These manipulation checks are important for strengthening the confidence that can be placed in the inferences that are drawn.

Finally, the practical significance of this study is that it shows how the predictive validity of these two interviews are affected by the use of typical versus maximum measures of a person's performance. For jobs demanding typical as well as high maximum performance (e.g., fire fighter, ambulance driver), because of the stressful short-duration circumstances, the SI appears to be a measure of choice.

Potential Limitations and Directions for Future Research

Our conclusions regarding the constructs underlying the SI and the BDI are arguably tentative due to the fact that no direct measure of ability or motivation was included in the study. Future research should do so as typical and maximum performance lie on a continuum.

A difficulty in comparing typical and maximum performance is establishing parallel situations that do not differ practically from each other beyond the three dimensions outlined by Sackett et al. (1988). Prior studies have manipulated typical versus maximum without taking into account potential covariates of their manipulation, such as variations in customer demands during typical but not maximum performance conditions (Sackett et al., 1988), or recruits' focus on military knowledge and physical development during the typical versus their focus on leadership abilities during the maximum performance condition (Ployhart et al., 2001, p.

822). Concerns regarding the comparability of typical and maximum performance are also justified in this study. For example, the MBA students might have interacted differently, or to a lesser extent with each other in their project group as opposed to their study group. However, the manipulation checks and the reliability coefficients of the assessments of typical and maximum performance suggest that this was not a problem in this study. In addition, the assessments of both typical and maximum performance were highly reliable.

Arguably, the results obtained from students performing in an MBA program may not generalize to other work settings. However, the MBA students in this study had an average of 5 years of work experience in a vast array of positions. Arguably, these results may generalize to work settings because of the importance of teamplaying in many work contexts. The items that comprised the teamplaying scale used in this study are likely to be applicable across organizational settings (e.g., "meets deadlines," "pays attention during group meetings"). Finally, there was a high level of task involvement, responsibility, and interdependence among the people who participated in this study that is similarly required in many work settings (Druskat & Wolff, 1999). Nevertheless, the participants in this study had already entered (chosen) this business school, and wanted to make a positive first impression. In other contexts, a jobseeker who already has several job offers may not be highly motivated to perform well, although the situation calls for maximum performance.

Future research should not only replicate these results, but also examine the degree to which they are content dependent. The items that constitute a written intelligence, interest, or personality test are invariant. In contrast to these pencil and paper tests, the questions that constitute the BDI and SI reflect the interests of different interviewers in terms of the information that they desire from interviewees. Although the current content domain of team playing is likely to be primarily a function of performers' ongoing motivation (e.g., Cronbach, 1960), there is nothing to preclude interview questions that aim primarily at a person's knowledge or ability (e.g., "The light on one panel is amber; the light on the other panel appears to have changed to a light red. What would you do in this situation?" "The validity coefficient for a test of emotional intelligence is .09 (with n = 800 and p < .05) for a specific position. What would you recommend to the client?").

The results of this study provide additional support for the contention that the typical structured interview taps primarily a person's motivation (Van Iddekinge, Raymark, & Roth, 2005). This finding likely reflects the fact that interviewers typically seek this information. Moreover, the job analysis on which the BDI and SI are derived, namely, the CIT, yield information on critical behaviors that require ability (maximum performance) and desire (typical performance) to demonstrate on the job. The SI taps these two dimensions of performance.

ACKNOWLEDGMENTS

This article is based on the first author's doctoral dissertation that was conducted under the supervision of the second author. In addition to the dissertation committee, consisting of Richard Arvey, Maria Rotundo, and Glen Whyte, we thank Neil Anderson, Peter Heslin, Martin Kleinmann, and Cornelius König for their helpful comments on an earlier draft of this article.

REFERENCES

Ackerman, P. L., & Humphreys, L. G. (1990). Individual differences in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 1* (2nd ed., pp. 223–282). Palo Alto, CA: Consulting Psychologists Press.

Allred, B. B., Snow, C. C., & Miles, R. E. (1996). Characteristics of managerial careers in the 21st century. *Academy of Management Executive, 10,* 17–27.

Arthur, W., Woehr, D. J., & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical reexamination of the assessment center construct-related validity paradox. *Journal of Management, 26,* 813–835.

Arvey, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology, 49,* 141–168.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43,* 695–716.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (2000). Social cognitive theory: An agentic perspective. *Annual Review of Psychology, 52,* 1–26.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74,* 478–494.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 2* (2nd ed., pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.

Boudreau, J. W. (1991). Utility analysis for decisions in human resource management. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 2* (2nd ed., pp. 621–745). Palo Alto, CA: Consulting Psychologists Press.

Brodbeck, F. C., Frese, M., Akerblom, S., Audia, G., Bakacsi, G., Bendova, H., et al. (2000). Cultural variation of employeeship prototypes across 22 European countries. *Journal of Occupational & Organizational Psychology, 73,* 1–29.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 1* (2nd ed., pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50,* 655–702.

Cascio, W. F. (1998). *Applied psychology in human resource management* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80,* 565–579.

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology, 13,* 485–506.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Oxford, England: Harper & Row.

Dewberry, C. (2001). Performance disparities between Whites and ethnic minorities: Real differences or assessment bias? *Journal of Occupational and Organizational Psychology, 74,* 659–673.

Dipboye, R. L. (1994). Structured selection interviews: Why do they work? Why are they underutilized? In N. Anderson & P. Herriot (Eds.), *International handbook of selection and assessment* (pp. 455–473). New York: Wiley.

Druskat, V. U., & Wolff, S. B. (1999). Effects and timing of developmental peer appraisals in self-managing work groups. *Journal of Applied Psychology, 84,* 58–74.

DuBois, C. L. Z., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Further exploration of typical and maximum performance criteria: Definitional issues, prediction, and White–Black differences. *Journal of Applied Psychology, 78,* 205–211.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading, MA: Addison-Wesley.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327–358.

Guion, R. M. (1991). Personnel assessment, selection, and placement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 2* (2nd ed., pp. 327–397). Palo Alto, CA: Consulting Psychologists Press.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hotelling, H. (1940). The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics, 11,* 271–283.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79,* 184–190.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U.-C. (2004). Evaluation and comparison of the situational and behavior description interview formats. *International Journal of Selection and Assessment, 12,* 262–273.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86,* 897–913.

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology, 81,* 459–473.

Huffcutt, A. I., Weekley, J. A., Wiesner, W. H., Degroot, T. G., & Jones, C. (2001). Comparison of situational and behavior description interview questions for higher-level positions. *Personnel Psychology, 54,* 619–644.

Huffcutt, A. I., & Woehr, D. J. (1999). Further analysis of employment interview validity: A quantitative evaluation of interviewer-related structuring methods. *Journal of Organizational Behavior, 20,* 549–560.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67,* 577–580.

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158–168). Thousand Oaks, CA: Sage.

Kirk, A. K., & Brown, D. F. (2003). Latent constructs of proximal and distal motivation predicting performance under maximum test conditions. *Journal of Applied Psychology, 88,* 40–49.

Klehe, U.-C., & Anderson, N. (2004, April). *Motivation and ability during typical and maximum performance. A test of the underlying assumptions.* Paper presented at the 19th annual meeting of the Society of Industrial and Organizational Psychology, Chicago.

Klehe, U.-C., & Anderson, N. (2005). The prediction of typical and maximum performance. In A. Evers, N. Anderson, & O. Smit-Voskuijl (Eds.), *Handbook of personnel selection* (pp. 331–353). Oxford, England: Blackwell.

Klehe, U.-C., & Latham, G. P. (2005). The predictive and incremental validity of the situational and patterned behavior description interviews for team playing behavior. *International Journal of Selection and Assessment, 13,* 108–115.

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40,* 243–260.

Latham, G. P. (1989). The reliability, validity, and practicality of the situational interview. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 169–182). Thousand Oaks, CA: Sage.

Latham, G. P., & Finnegan, B. J. (1993). Perceived practicality of unstructured, patterned, and situational interviews. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 41–55). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65,* 422–427.

Latham, G. P., & Skarlicki, D. P. (1995). Criterion-related validity of the situational and patterned behavior description interviews with organizational citizenship behavior. *Human Performance, 8,* 67–80.

Latham, G. P., & Sue-Chan, C. (1999). A meta-analysis of the situational interview: An enumerative review of reasons for its validity. *Canadian Psychology, 40,* 56–67.

Latham, G. P., & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30,* 255–268.

Latham, G. P., & Wexley, K. N. (1994). *Increasing productivity through performance appraisal* (2nd ed.). Reading, MA: Wesley Publishing Company.

Lewin, K. (1951). *Field theory in social science: selected theoretical papers* (D. Cartwright, Ed.). Oxford, England: Harpers & Brothers.

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance.* Upper Saddle River, NJ: Prentice Hall.

Locke, E. A., Mento, A. J., & Katcher, B. L. (1978). The interaction of ability and motivation in performance: An exploration of the meaning of moderators. *Personnel Psychology, 31,* 269–280.

Louis, M. R. (1980). Surprise and sense making: What newcomers experience in entering unfamiliar organizational settings. *Administrative Science Quarterly, 25,* 226–251.

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology, 66,* 451–457.

Maier, N. R. F. (1955). *Psychology in industry* (2nd ed.). Oxford, England: Houghton Mifflin.

Maurer, S. D., Sue-Chan, C., & Latham, G. P. (1999). The situational interview. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 159–177). Thousand Oaks, CA: Sage.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79,* 493–505.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111,* 172–175.

Mitchell, T. R. (1985). An evaluation of the validity of correlational research conducted in organizations. *Academy of Management Review, 10,* 192–205.

Moscoso, S. (2000). A review of validity evidence, adverse impact and applicant reactions. *International Journal of Selection and Assessment, 8,* 237–247.

Mumford, M. D. (1983). Social comparison theory and the evaluation of peer evaluations: A review and some applied implications. *Personnel Psychology, 36,* 867–881.

Olkin, I. (1967). Correlations revisited. In J. C. Stanley (Ed.), *Improving experimental design and statistical analysis* (pp. 102–128). Chicago: Rand McNally.

Orpen, C. (1985). Patterned behavior description interviews versus unstructured interviews: A comparative validity study. *Journal of Applied Psychology, 70,* 774–776.

Ployhart, R. E., Lim, B. C., & Chan, K. Y. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology, 54,* 809–843.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160–164.

Ryan, T. A. (1970). *Intentional behavior: An approach to human motivation.* Oxford, England: Ronald Press.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401–410.

Sackett, P. R., & Larson, J. R., Jr. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 1* (2nd ed., pp. 419–489). Palo Alto, CA: Consulting Psychologists Press.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73,* 482–486.

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology, 11,* 299–324.

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77,* 32–41.

Schuler, H., Frier, D., & Kauffmann, M. (1993). *Personalauswahl im europäischen vergleich* [Personnel selection: A European comparison]. Göttingen, Germany: Verlag für angewandte Psychologie.

Schuler, H., & Funke, U. (1989). The interview as a multimodal procedure. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 183–192). Thousand Oaks, CA: Sage.

Smith-Jentsch, K. A., Salas, E., & Brannick, M. T. (2001). To transfer or not to transfer? Investigating the combined effects of trainee characteristics, team leader support, and team climate. *Journal of Applied Psychology, 86,* 279–292.

Sue-Chan, C., & Latham, G. P. (2004). The relative effectiveness of external, peer, and self coaches. *Applied Psychology: An International Review, 53,* 260–278.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics.* Needham Heights, MA: Allyn & Bacon.

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75,* 277–294.

Terpstra, D. E., Mohamed, A. A., & Kethley, R. B. (1999). An analysis of federal court cases involving nine selection devices. *International Journal of Selection and Assessment, 7,* 26–34.

Thorsteinson, T. J., & Balzer, W. K. (1999). Effects of coworker information on perceptions and ratings of performance. *Journal of Organizational Behavior, 20,* 1157–1173.

Tziner, A., & Kopelman, R. (1988). Effects of rating format on goal-setting dimensions: A field experiment. *Journal of Applied Psychology, 73,* 323–326.

Tziner, A., Kopelman, R. E., & Livneh, N. (1993). Effects of performance appraisal format on perceived goal characteristics, appraisal process satisfaction, and changes in rated job performance: A field experiment. *Journal of Psychology, 127,* 281–291.

Van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., & Attenweiler, W. J. (2004). What do structured selection interviews really measure? The construct validity of behavior description interviews. *Human Performance, 17,* 71–93.

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing personality with a structured employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology, 90,* 536–552.

Viswesvaran, C., & Ones, D. S. (2005). Job performance: Assessment issues in personnel selection. In A. Evers, N. Anderson, & O. Smit-Voskuijl (Eds.), *Handbook of personnel selection* (pp. 354–375). Oxford, England: Blackwell.

Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61,* 275–290.

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, 21,* 396–399.

APPENDIX
Example for a comparable SI and BDI question

SI Question Your group is working on a very important project. All of you want to achieve a good grade. You have a tight deadline. One member of your group was especially successful in this area the last term. Supported by two other group members, she takes the lead on your group project. She keeps the minutes and controls the flow of information during the discussion. However, you have the strong impression that she only records ideas supportive of her position, and makes decisions on issues without consulting with others. What would you do?

BDI Question Tell me about a time when someone took over the leadership of a group-project, and ignored contributions that were not in accordance with his or her own opinion. What were the circumstances? What exactly did you do? What was the outcome?

Scoring Guide

5 Involve the other group members; ask them their opinions on the topic of discussion; ask everyone to take notes; ask the leader to send her notes for correction and supplementation by the others.

3 Ask someone to take the minutes and send them to the group for comment.

Or: confront the current leader only if I am not satisfied with the direction the project is taking. Do nothing if I think that the solution achieved so far is actually good for the project.

1 Do nothing.

Rating: 5 4 3 2 1