A tale of two formats: Direct comparison of matching situational and behavior description interview questions

Satoris S. Culbertson [a,*], William S. Weyhrauch [b], Allen I. Huffcutt [c]

[a] University of Portland, Pamplin School of Business, Portland, OR 97210, United States
[b] Kansas State University, Department of Psychological Sciences, 492 Bluemont Hall, Manhattan, KS 66506-5302, United States
[c] Bradley University, Department of Psychology, 73 Bradley Hall, 1501 W. Bradley Avenue, Peoria, IL 61625, United States

ARTICLE INFO

ABSTRACT

The primary purpose of this meta-analysis was to examine the psychometric properties of situational interview (SI) and behavior description interview (BDI) questions written to assess the same set of job attributes. Using a final dataset of 29 coefficients ($N = 8148$), we found an observed mean correlation of 0.40 (0.47 corrected) between construct-matched SI and BDI questions. In terms of moderators, even lower correspondence resulted when (1) both question types had lower internal consistency, (2) there were more questions per format, (3) probing was not allowed in either format or when allowed in one format but not in the other, and (4) the purpose of the interview was for research (vs. employment). Given that their correspondence can be quite low under some moderator combinations, an important implication is that SI and BDI questions should not automatically be assumed to be interchangeable, even when written deliberately to assess the same attributes, and that incremental validity is very possible. Further, results suggest that SIs have higher overall mean validity compared to BDIs for predicting job performance (0.23 vs. 0.18, respectively). Results also indicated a slightly stronger relationship with cognitive ability measures for BDIs than SIs (0.11 vs. 0.09, respectively).

One of the most consistent findings in the employment interview literature is the importance of structure. Whereas unstructured interviews leave the content and the evaluation process largely up to the discretion of the interviewer, structured interviews involve a high degree of standardization across applicants in terms of both the questions asked and the scoring procedures used (Campion, Palmer, & Campion, 1998). The rationale for incorporating structure into the interview process is that it reduces procedural variability across applicants (Huffcutt & Arthur, 1994), which results in higher reliability (Conway, Jako, & Goodman, 1995; Huffcutt, Culbertson, & Weyhrauch, 2013), better predictive validity (Huffcutt, Culbertson, & Weyhrauch, 2014a; Wiesner & Cronshaw, 1988), and lower subgroup differences (Huffcutt & Roth, 1998).

The two most common structured interview formats are the situational interview (SI) and the behavior description interview (BDI). SIs require applicants to indicate what actions they would take in a series of hypothetical job-related situations (Latham, Saari, Pursell, & Campion, 1980). These questions are based on the premise that the intentions applicants state in their answers are predictive of their future workplace behavior (see Locke & Latham, 1990). Conversely, BDIs require applicants to describe actual experiences from their past that demonstrate important job attributes (Janz, 1982) and are grounded in the notion that past behavior is the best predictor of future behavior (see Janz, 1989).

* Corresponding author.
E-mail addresses: culberts@up.edu (S.S. Culbertson), wsweyhra@ksu.edu (W.S. Weyhrauch), huffcutt@fsmail.bradley.edu (A.I. Huffcutt).

Many interview developers use SI and BDI questions interchangeably, often within a single interview. Unfortunately, the degree of correspondence of these two question types has yet to be determined. Even when written deliberately to assess the same job dimensions using the same base of critical incidents, it is unclear whether the assessed constructs are the same and whether the interview scores are predictive of job performance to the same extent. Indeed, there are at least seven reasons to suspect that they may not be.

First, there appears to be differential influence from moderators. For example, research has suggested that SIs are less valid for more complex jobs, while BDIs are not influenced by job complexity (Huffcutt, Conway, Roth, & Klehe, 2004). Second, interviewees tend to utilize different impression management tactics, namely ingratiation with SIs and self-enhancement with BDIs (e.g., Ellis, West, Ryan, & DeShon, 2002). Third, per their original methodology, there is differential use of probing such that probing is inherent with BDIs but not with SIs. Fourth, the context is clearly specified in SIs questions, whereas choice of context is dictated by the interviewees in BDIs.[1]

Fifth, there may be method effects unique to each format, such as a "storytelling" influence with BDIs (see Bangerter, Corvalan, & Cavin, 2014; Huffcutt, Culbertson, & Goebl, 2015). Sixth, from a neuroscience perspective, the mental processes behind SIs and BDIs appear to be different with SIs centered in the prefrontal region and BDIs driven by the hippocampus in the temporal arm of the brain (Huffcutt, 2016). Seventh, empirically speaking, a handful of primary studies reported a surprisingly low SI-BDI correlation, including 0.15 in sample of district managers (Huffcutt, Weekley, Wiesner, Degroot, & Jones, 2001) and 0.20 in a sample of German managers (Klehe, König, Richter, Kleinmann, & Melchers, 2008). Taken together, these seven reasons suggest that SIs and BDIs may be not be entirely interchangeable, an important consideration for a number of reasons including predictive validity, incremental validity, and perhaps even adverse impact.

The first purpose of this investigation was to explore the overall correspondence between matching SI and BDI questions collapsing across moderators. Indeed, if their correspondence is high, then including both types of questions would be unnecessary and the choice of which type to use would largely become a matter of preference. Conversely, if their correspondence is not overly high, then it would seem prudent to view them as distinct methods. In this case, the ratings from these two types of questions should be kept separate and not combined (as is common practice in the literature).

The second purpose of the current study was to explore moderators that could potentially drive the correspondence between SIs and BDIs either higher or lower. For our examination, we selected moderators that should theoretically impact the correspondence between matching SI and BDI questions. As explained in more detail below, we examined four such factors in this investigation. To our knowledge, these four factors have yet to be explored empirically in any previous SI-BDI meta-analysis. Understanding the influence of moderators is crucial to the process of elucidating the nomological boundaries of these modern structured interviewing techniques.

Finally, the third purpose of this study was to assess the relationship between the degree of SI-BDI correspondence and criterion-related validity. Although research suggests that both SI and BDI formats have relatively high overall validity (Huffcutt et al., 2004; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Taylor & Small, 2002), there is limited understanding of factors that can influence (moderate) their validity. The correspondence between the two formats may be one such moderator. When the correspondence between SI and BDI questions are relatively high, a general interviewee performance factor could be operating, and that could influence validity either higher or lower. On the other hand, a lower correspondence suggests measurement of different constructs, and when SI and BDI ratings are combined, a richer and diverse assessment could result. In short, there appears to be insufficient justification to predict the direction of the outcome.

While not a direct focus of this investigation, our third purpose has the added benefit of providing an updated analysis of mean SI and BDI validity. Large-scale meta-analyses of interview validity rarely provide specific and separate analyses of SI and BDI validity (i.e., Huffcutt et al., 2004; McDaniel et al., 1994). The one meta-analysis that focused directly on SI and BDI validity had a relatively small number of studies (Taylor & Small, 2002), due mainly to a tendency for researchers to combine question types into overall interview scores (e.g., Campion, Pursell, & Brown, 1988; Morgeson, Reider, & Campion, 2005[2]).

Our reanalysis is advantageous and timely for two reasons. One is that more studies have become available since these meta-analyses (particularly Taylor & Small). Second, while all of these meta-analyses used as a "between-study" design, we limited our dataset to studies with matching SI and BDI questions. Using a "within-study" design provides more direct control for study-to-study differences (e.g., study design, interview design, individual interviewee and interviewer differences) and should result in more accurate mean estimates.

Finally, although again not a direct focus on our investigation, we incorporated an analysis of the correlation between cognitive ability and SIs and BDIs respectively. Although one meta-analysis has already done this (Huffcutt, Roth, & McDaniel, 1996), more studies have become available since it was conducted. Further, this analysis was between-study, and once again, our study is within. Understanding the degree of covariation between SIs and BDIs and cognitive ability is crucial for the practice of selection given that both are top predictors of job performance and that there is the potential to maximize incremental validity if their relationship is not overly high.

---

[1] To illustrate, consider the classic Weekley and Gier (1987) SI question for a mall jewelry store position where a customer becomes irate because his watch is not back from the repair shop and is overdue. In contrast, a reasonable BDI version would be something like "Tell me about a time when you had to deal with an irate person over something that was out of your control," which could elicit an incredibly wide range of contexts including work, social, political, athletic, hobby, and religious.

[2] In the case of Morgeson et al. (2005), we were able to obtain the separate SI and BDI data directly from the authors upon request.

1. Moderators of SI–BDI correspondence

As noted earlier, we chose moderators that should impact the degree of correspondence between matching SI and BDI questions. We were able to identify four such moderators, including internal consistency of the question types, the number of questions per format, the extent to which probing was allowed in one or both formats, and the purpose of the interview (research vs. employment). Each of these moderators is discussed in more detail below.

1.1. Internal consistency

There is the suggestion in the literature that there may be a common underlying factor that influences interview performance (e.g., Huffcutt et al., 2001; see also Huffcutt et al., 2015). Indeed, some researchers have found that a single factor underlies interview ratings (e.g., Pulakos & Schmitt, 1995). Further, general factors have been found in other rating-based selection areas including assessment centers (see Bowler & Woehr, 2006) and performance ratings (see, Viswesvaran, Schmidt, & Ones, 2005).

If the internal consistency of both sets of questions is high, there is a greater possibility that they are tapping into this general factor, which should result in a higher degree of correspondence. Conversely, if their respective internal consistencies are low, then a lower degree of correspondence could result. Lower internal consistency suggests the presence of multiple constructs embedded in the ratings, and the probability that both types of questions contain the same set of constructs is not overly high. A third possibility is that internal consistency is relatively high for one type of question but lower for the other. Here again a lower degree of consistency would be expected.

1.2. Interview length

Recent research with assessment centers (Kuncel & Sackett, 2014) indicates dimensional construct variance is considerably <50% of the total variance when there are fewer than eight exercises. This finding is in line with the theory of composites (Ghiselli, Campbell, & Zedeck, 1981), which states that correlated variance will increase while uncorrelated variance will decrease as the number of multiple measures of the same construct increases. With regard to interviews, this would suggest that there is considerable room for the general interview performance factor to operate when there are a limited number of questions, but dimensional variance should increase when there are more questions. As Huffcutt et al. (2015) noted, "Given the common practice of including only one or two BDI questions per dimension, it would not be unexpected to find that a method (general performance) factor captures mores variance than the specific dimensions the questions were intended to assess" (pp. 34–35). Thus, the degree of correspondence should be inversely related to the number of questions in each interview.

1.3. Interview purpose

Here we assessed whether the interviews were administered for research purposes (e.g., publication, validation) or actual employment selection. There is reason to believe that SI-BDI correspondence is lower when interviews are conducted for employment selection. Given the high-stakes nature of actual selection, candidates tend to be highly motivated to present themselves in the best possible light and not necessarily as they actually are in reality. As such, the differential use of impression management tactics noted earlier (i.e., ingratiation with SIs and self-enhancement with BDIs; Ellis et al., 2002) could be intensified, thereby driving the two formats further apart. Furthermore, because of the verifiable nature of BDIs, interviewees are less likely to engage in impression management tactics when responding to BDI questions in actual selection than they are for the more hypothetical (and therefore less verifiable) SI questions (Levashina & Campion, 2007). Thus, we expect lower SI-BDI correspondence when interviews are conducted for employment purposes. In line with this, Huffcutt et al. (2004) found higher mean validity when SI and BDI interviews were conducted in a concurrent rather than predictive format, lending indirect support to this notion.

1.4. Probing

Given the unique focus in BDIs on the description of past experiences, interviewers often expect more extensive and detailed responses that include a thorough description of the situation, tasks, actions taken, and results of those actions (Tross & Maurer, 2008). When interviewers do not feel that sufficient detail is provided, or when they believe that clarification is needed, they may probe for further information. The original BDI format outlined by Janz (1982) allowed for extensive probing, a practice that has continued in some (but not all) BDI studies that followed. In contrast, only intended behaviors are stated in SIs, thus lessening the need for clarification (Roulin, Bangerter, & Levashina, 2015). The original SI methodology (Latham et al., 1980) did not allow for any probing whatsoever.

The BDI process has already been described as a form of storytelling (Bangerter et al., 2014). With additional probing, there is greater opportunity for interviewees to enhance and elaborate on their presentation of past experiences even more, potentially enhancing the influence of verbal-related constructs. Thus, the degree of correspondence should be lowest when BDIs with probing are compared to SIs without probing. In contrast, correspondence should relatively higher when probing is allowed with both types or no probing is allowed at all.

2. Method

2.1. Identification of studies

To find studies with matching SI and BDI questions, we first searched numerous databases (e.g., *PsycINFO*, *ABI/INFORM*, *JSTOR*, *ERIC*, *CINAHL*, *ProQuest*, *Dissertations Abstracts*) using such keywords as *employment interview*, *situational interview*, *behavior description interview*, and related variants. Then we searched conference programs from 2003 to the present for SIOP and Academy of Management. We limited our search of conference presentations and papers to this timeframe based on the assumption that most interview studies appearing prior to 2003 would likely be represented in the search for published studies, or obtained in our later request for unpublished papers and datasets. After that, we reviewed references of prior employment interview meta-analyses (e.g., Huffcutt & Arthur, 1994; McDaniel et al., 1994; Taylor & Small, 2002). Finally, we solicited unpublished manuscripts and data from researchers, with no restrictions on when the studies were conducted. Based on these efforts, we believe the studies obtained for use in the current meta-analysis are comprehensive in scope and appropriately representative of the field. Table 1 shows descriptive information for the studies used in the current analyses.

2.2. Inclusion criteria

The first set of screening criteria was that a study must have included both SI and BDI questions with reasonable evidence that the questions were written to assess the same set of job dimensions (i.e., they were matched) and provided a correlation between the two sets of overall ratings. When matched questions were written but a correlation was not provided, the authors were

Table 1
Overall dataset.

Study	Sample	n	Location	Publication type	Interview length	SI α	BDI α	r	Probing
Bosshardt (1994)	Electricians	73	USA	Dissertation	24			0.38	BDI only
Campion, Campion, and Hudson (1994)	Pulp mill employees	70	USA	Article	30			0.73	Neither
Conway and Peneno (1999)	Resident assistants	137	USA	Article	14			0.85	Both
Faulkner (1999)	Customer service, light industrial, and administrative temp applicants	149	USA and Canada	Dissertation	4			0.54	Both
Hollwitz (1998)	Undergraduates	50	USA	Dissertation	20	0.633	0.635	0.57	Both
Huffcutt et al. (2001) Study 1	Naval military officers	59	Canada	Article	12	0.4	0.65	0.32	BDI only
Huffcutt et al. (2001) Study 2	District managers	93	USA	Article	20	0.38	0.55	0.15	Neither
Klehe and Latham (2006)	MBA students	79	Canada	Article	18	0.5	0.71	0.48	BDI only
Klehe et al. (2008) Study 1	Managers	64	Germany	Article	24	0.79	0.63	0.21	BDI only
Klehe et al. (2008) Study 2	Managers	110	Germany	Article	24			0.31	BDI only
Kluemper (2006)	Undergraduates in job strategies seminar	30	USA	Dissertation	10	0.95	0.93	0.64	Both
Kluemper (2006)	Youth treatment worker	81	USA	Dissertation	10	0.92	0.92	0.41	Both
Krajewski, Goffin, McCarthy, Rothstein, and Johnston (2006)	Forestry managers	157	Canada	Article	12	0.89	0.89	0.41	BDI only
Lin and Adrian (1993)	School cafeteria manager	33	USA	Conference	8			0.69	Neither
Little, Schoenfelt, and Brown (2000)	Home improvement retail employees	98	USA	Conference	10	0.88	0.89	0.72	Neither
Melchers et al. (2008)	Recent university grads	110	Germany	Conference	24			0.27	Both
Menkes (2002)	Graduate students in business	66	USA	Dissertation	8	0.48	0.68	0.12	Neither
Morgeson et al. (2005)	Mill employees	90	USA	Article	28			0.50	Neither
Mussel, Behrmann, and Schuler (2008)	Police officers	178	Switzerland	Conference	20	0.76	0.77	0.54	Neither
Sue-Chan, Latham, and Evans (1995)	Student nurses	32	Canada	Conference	18	0.66	0.51	0.5	Both
Ingold, Kleinmann, König, Melchers, & Van Iddekinge (2015)	Graduate students	108	Switzerland	Article	12	0.50	0.55	0.43	Neither
Heimann, Ingold, & Kleinmann (2016)	Interview trainees	223	Switzerland	Conference	30	0.73	0.74	0.43	Neither
Huffcutt, 2016	Resident assistants	54	USA	Unpublished	20	0.76	0.59	0.63	Both
Mussel (2007)	Unclear/Unknown	137	Germany	Dissertation	17	0.72	0.77	0.30	Unclear
Mussel (2007)	Unclear/Unknown	145	Germany	Dissertation	17	0.71	0.73	0.27	Unclear
Mussel (2007)	Unclear/Unknown	182	Germany	Dissertation	10	0.8	0.8	0.35	Unclear
Mussel (2007)	Unclear/Unknown	182	Germany	Dissertation	10	0.78	0.84	0.36	Unclear
Mussel (2007)	Unclear/Unknown	676	Germany	Dissertation	20	0.71	0.75	0.28	Unclear
Mussel (2007)	Unclear/Unknown	4682	Germany	Dissertation	15	0.61	0.74	0.31	Unclear

contacted to request the correlation and any other missing information. For example, we were able to obtain the separate SI and BDI data directly from Morgeson et al. (2005). If the correlation was unavailable, then those studies were removed. A considerable number of studies were removed from consideration based on these two criteria. Examples include Gibb and Taylor (2003) and Roth, Van Iddekinge, Huffcutt, Eidson, and Schmit (2005). Another screening criterion was that the interview occurred in a real interview setting and reflected actual (or potential) employment. Additional studies were excluded based on this criterion as well. To illustrate, Sears and Rowe (2003) and George (2006) used scripted interview responses and study participants rated these responses.

For some criteria, we were able to eliminate studies on the basis of details provided in the title or the abstract. For others studies, our decision to include them was based on a careful inspection of their reported methodology. Application of these screening criteria resulted in a dataset of 29 correlations between matching SI and BDI questions, and 14 studies with separate criterion-related validity reported for both type.

2.3. Coding of study information

The first two authors independently coded each study and jointly resolved discrepancies by consensus. The studies were coded on the necessary statistics including correlations between matching SI and BDI questions, total sample size, and coefficient alpha for each format. We examined Cronbach's (1951) coefficient alpha as the measure of interrelatedness of the interview questions. Despite potential problems with using a cutoff score for acceptability (Cortina, 1993; Schmitt, 1996), many researchers nonetheless consider 0.70 or higher to be an acceptable level of alpha. As such, this is the threshold that we used when separating interviews into high versus low internal consistency.

In addition, we coded whether probing was allowed, the number of questions for each format, and the interview purpose (research vs. employment). With regard to interview purpose, there is potential for this distinction to be confounded given that research settings often involve college students whereas employment settings typically involve job applicants or current employees. As such, it may be better to think of this as a measure of realism, with research being less realistic compared to selection purposes. In addition, given Kuncel and Sackett's (2014) finding that it would likely take eight assessment center exercises to reach a point where dimensional variance becomes the majority source of variance, we coded our interview studies as having either fewer than eight questions or having eight or more questions.[3]

Lastly, we weighted the studies by the square root of their sample size in order to allow larger studies to carry more weight in the analyses (as is typical in meta-analytic work), while not allowing any single study to dominate the results. This approach and weighting scheme has been utilized in prior employment interview meta-analyses; see Huffcutt et al. (2014a) for a recent example. In addition, we provide all results both with and without the largest two samples (both from Mussel, 2007; $n = 676$ and $n = 4682$) to further assess their influence on the results. References for the final data sources used are marked by an asterisk (*) in the reference section. Among the 29 total coefficients, 10 were drawn from published articles, 5 from conference papers, and 12 from unpublished sources including dissertations.

3. Results

Our final dataset included 29 samples with a total sample size of 8148. Sample sizes ranged from 30 to 4682, with a mean of 281.0 and a median of 98. Across all samples, there was a moderate correlation of 0.40 between ratings for matching SI and BDI questions, which increased to 0.47 after correction for unreliability in interview scoring. The correction was based on a recent estimate of 0.73 for the inter-rater reliability of structured interviews (Huffcutt et al., 2014a), and the corrected value provides information on their construct (rather than operational) association. In most cases, our analysis treats the BDI and SI as separate interviews, for the purpose of comparison. However, in correcting for interview unreliability, we determined it would be inappropriate to conduct a double-correction (one for each interview). So in this case, they are treated as a single interview.

Sampling error accounted for only 9.4% of the observed variance among correlations, suggesting the presence of one or more moderator variables. Regarding the first potential moderator, internal consistency, the reported coefficients alpha in our dataset ranged from 0.38 to 0.95, with an average value of 0.71. As expected, when the reported alphas for both the SI and BDI in a sample met the baseline standard for acceptable measurement (≥ 0.70), the correlation between the two question types was stronger ($\bar{r} = 0.40$; $k = 11$) than when one or both coefficients alpha did not meet the 0.70 standard ($\bar{r} = 0.34$, $k = 10$). The correlation between the average alpha for each study $[(SI_\alpha + BDI_\alpha)/2]$ and SI–BDI correspondence was fairly high ($r = 0.41$). In summary, as expected, the samples in our dataset indicated that greater internal consistency were associated with greater correspondence between construct-matched SI and BDI questions. These and all other study results are presented in Tables 2–5.

For the second moderator, interview lengths in our dataset ranged from four total questions to 30, with an average (and median) value of 17.0 questions. Counter to our expectations, samples with longer interviews (containing >17 questions, or approximately 7.5 questions per method) did not differ in terms of SI–BDI correspondence ($\bar{r} = 0.40$, $k = 14$) compared to samples

---

[3] We openly acknowledge that the number of interview question needed to reach 50% dimensional construct variance may not be the same as the number needed with assessment centers. Nonetheless, we could not find any studies that addressed this issue with interviews, and thus, by fault, strived to use eight as well. Our interview length metric was calculated as the total number of questions across both interview types, because in some cases they were unbalanced. We grouped the samples by the median of 17 total questions, or 7.5 questions per method, which is conceptually in line with the 8-question logic presented here.

Table 2
Moderators of SI-BDI correspondence estimates.

Moderators	$\bar{r}_{SI-BDI}$	k	N	SD	SD_{res}	$CI_{95\%}$	$CI_{80\%}$
Internal consistency							
Alphas >0.7	0.40 (0.47)	11	2089	0.13	0.11	0.19–0.61	0.26–0.54
Either alpha <0.7	0.34 (0.40)	10	5287	0.13	0.12	0.10–0.58	0.19–0.50
Interview length							
>17 questions	0.40 (0.47)	14	1902	0.16	0.14	0.13–0.67	0.23–0.58
≤17 questions	0.40 (0.47)	15	6246	0.17	0.16	0.08–0.72	0.20–0.61
Interview context							
Research	0.41 (0.48)	16	1340	0.17	0.14	0.13–0.69	0.23–0.60
Employment	0.40 (0.47)	13	6808	0.16	0.15	0.10–0.70	0.20–0.60
Response probing							
Both with probing	0.55 (0.64)	8	643	0.18	0.16	0.60–0.50	0.35–0.76
Neither with probing	0.47 (0.55)	9	959	0.2	0.18	0.11–0.83	0.24–0.71
Only BDI probing	0.36 (0.42)	6	542	0.08	#	0.36–0.36	0.36–0.36
Overall correspondence	0.40 (0.47)	29	8148	0.16	0.16	0.10–0.71	0.21–0.60

Note. Estimates of SI-BDI correspondence are corrected for unreliability in the predictor (interview reliability, 0.73). The corrected values are in parentheses. #The residual variance in this case was negative, making the SD_{res} a non-real number.

with 17 or fewer questions, ($\bar{r} = 0.40$, $k = 15$). Thus, it appears that overall interview length does not impact SI–BDI correspondence.

Regarding interview context, our dataset contained 16 samples with interviews conducted for research purposes and 13 samples conducted for actual employment purposes. Although the results were in line with our expectations that SI–BDI correspondence would be lower when interviews are conducted for employment purposes, the SI–BDI correspondence was not substantially higher among the research group ($\bar{r} = 0.41$) than the employment group ($\bar{r} = 0.40$). Rather, they were arguably equal in terms of SI–BDI correspondence. This is not altogether surprising in hindsight given that the interviews conducted for research and employment were quite similar on average with regard to length (18 and 15 questions, respectively) and internal consistency ($\alpha = 0.70$ and 0.73, respectively). Worthy of note is that the direction of the interview context moderator is different (the SI–BDI correspondence is much higher for employment settings) when the largest two samples in the dataset are included (see Tables 2–5).

The final moderator, question structure, also revealed some compelling differences. Our dataset, by its nature, consisted of interviews with a fairly high level of structure. The main point of difference lay in whether or not probing was allowed, in one or both interview types. Out of the 23 samples in our dataset with sufficient information provided in this regard, 8 allowed probing for both SI and BDI formats, 6 allowed probing for BDI questions only, and 9 allowed no probing for either question type. As anticipated, SI–BDI correspondence was lowest when probing was allowed for BDI questions only ($\bar{r} = 0.36$), higher when probing was not allowed in either case ($\bar{r} = 0.47$), and highest when probing was allowed in both cases ($\bar{r} = 0.55$).

Our dataset also allowed us to conduct a very direct comparison of the criterion-related validity of each question format. There were 14 samples that provided a separate validity estimate for each question type. In line with standard practice, the validity estimates have been corrected for unreliability in the criterion but not the predictor. Given concerns that the common estimate

Table 3
Validity estimates and cognitive ability.

	$\bar{r}$	k	N	SD	SD_{res}	$CI_{95\%}$	$CI_{80\%}$
*Criterion validity (corrected with 0.52)**							
SI	0.23 (0.32)	14	6355	0.12	0.13	−0.01–0.47	0.07–0.38
BDI	0.18 (0.25)	14	6355	0.11	0.10	−0.02–0.37	0.05–0.31
Overall	0.20 (0.28)	14	6355	0.10	0.09	0.02–0.39	0.08–0.32
SI-BDI < 0.4^	0.17 (0.24)	5	5583	0.07	0.06	0.04–0.29	0.08–0.25
SI-BDI ≥ 0.4^	0.26 (0.36)	9	772	0.11	0.04	0.18–0.33	0.21–0.31
*Criterion validity (corrected with 0.71)**							
SI	0.23 (0.27)	14	6355	0.12	0.13	−0.01–0.47	0.07–0.38
BDI	0.18 (0.21)	14	6355	0.11	0.10	−0.02–0.37	0.05–0.31
Overall	0.20 (0.24)	14	6355	0.10	0.09	0.02–0.39	0.08–0.32
SI-BDI < 0.4^	0.17 (0.20)	5	5583	0.07	0.06	0.04–0.29	0.08–0.25
SI-BDI ≥ 0.4^	0.26 (0.31)	9	772	0.11	0.04	0.18–0.33	0.21–0.31
Cognitive ability							
SI	0.09	13	6559	0.13	0.12	−0.14–0.33	−0.06–0.25
BDI	0.11	13	6559	0.15	0.14	−0.16–0.38	−0.07–0.28
Overall	0.10	13	6559	0.13	0.13	−0.15–0.35	−0.06–0.26

Note. ^The overall mean for correspondence between matched SI and BDI interviews is 0.4. *The correction for criterion unreliability is performed here twice, using the traditional 0.52 estimate and a more conservative 0.71 estimate, based on Huffcutt et al. (2014b). The corrected values are in parentheses.

Table 4
Moderators of SI-BDI correspondence estimates (without extreme N studies).

Moderators	$\bar{r}_{SI-BDI}$	k	N	SD	SD_{res}	$CI_{95\%}$	$CI_{80\%}$
Internal consistency							
Alphas >0.7	0.43 (0.50)	10	1413	0.13	0.10	0.24–0.62	0.30–0.55
Either alpha <0.7	0.37 (0.43)	9	605	0.18	0.12	0.12–0.61	0.21–0.53
Interview length							
>17 questions	0.43 (0.50)	13	1226	0.16	0.13	0.17–0.70	0.26–0.60
≤17 questions	0.44 (0.51)	14	1564	0.19	0.17	0.11–0.79	0.23–0.67
Interview context							
Research	0.41 (0.48)	16	1340	0.17	0.14	0.13–0.69	0.23–0.60
Employment	0.47 (0.55)	11	1450	0.18	0.17	0.15–0.80	0.26–0.68
Response probing							
Both with probing	0.55 (0.64)	8	643	0.18	0.16	0.60–0.50	0.35–0.76
Neither with probing	0.47 (0.55)	9	959	0.2	0.18	0.11–0.83	0.24–0.71
Only BDI probing	0.36 (0.42)	6	542	0.08	#	0.36–0.36	0.36–0.36
Overall correspondence	0.44 (0.51)	27	2790	0.16	0.18	0.13–0.75	0.24–0.64

Note. Estimates of SI-BDI correspondence are corrected for unreliability in the predictor (interview reliability, 0.73). The corrected values are in parentheses. #The residual variance in this case was negative, making the SD_{res} a non-real number.

of 0.52 for the reliability of the criterion (job performance ratings; as suggested by Viswesvaran, Ones, & Schmidt, 1996) results in an overestimate of criterion-related validity (see LeBreton, Scherer, & James, 2014), we opted to use the more conservative estimate of 0.71 that was suggested by Huffcutt, Culbertson, and Weyhrauch (2014b). Nevertheless, we also present validity estimates using the more traditional 0.52 correction (see Tables 3 and 5).

Results of our analyses revealed that SI scores demonstrated overall higher validity than BDI scores ($\bar{r} = 0.27$ vs. 0.21 for the conservative correction; $\bar{r} = 0.32$ vs. 0.25 for the traditional correction), with an overall combined validity of 0.24 (conservative correction; 0.28 for traditional correction). Using the conservative correction, combining validity and correspondence, average mean interview validity was considerably higher ($\bar{r} = 0.31$, $k = 9$) in samples with a higher SI–BDI correspondence (above the median correspondence value of 0.40) than in samples with a lower SI–BDI correspondence ($\bar{r} = 0.20$, $k = 5$). Similar results emerged when using the traditional correction, with higher validity resulting for higher SI–BDI correspondence ($\bar{r} = 0.36$, $k = 9$) versus lower SI–BDI correspondence ($\bar{r} = 0.24$, $k = 5$). Worthy of noting is that the levels of SI and BDI validity were lower when the largest two samples were excluded from the analyses (see Tables 3 and 5).

Several of the studies in our dataset also included measures of cognitive ability, so we examined the average relationship with cognitive ability for each of the formats. The results indicated generally low correlations with cognitive ability overall ($\bar{r} = 0.10$), and only very slightly weaker relationship for SI ($\bar{r} = 0.09$) than BDI ($\bar{r} = 0.11$).

Table 5
Validity estimates and cognitive ability (without extreme N studies).

	$\bar{r}$	k	N	SD	SD_{res}	$CI_{95\%}$	$CI_{80\%}$
*Criterion validity (corrected with 0.52)**							
SI	0.30 (0.42)	12	997	0.15	0.10	0.09–0.50	0.16–0.43
BDI	0.22 (0.31)	12	997	0.13	0.07	0.08–0.36	0.13–0.31
Overall	0.26 (0.36)	12	997	0.11	0.04	0.18–0.34	0.21–0.31
SI-BDI < 0.4^	0.17 (0.24)	3	225	0.08	#	0.17–0.17	0.17–0.17
SI-BDI ≥ 0.4^	0.26 (0.36)	9	772	0.12	0.07	0.13–0.39	0.17–0.34
*Criterion validity (corrected with 0.71)**							
SI	0.30 (0.36)	12	997	0.15	0.10	0.09–0.50	0.16–0.43
BDI	0.22 (0.26)	12	997	0.13	0.07	0.08–0.36	0.13–0.31
Overall	0.26 (0.31)	12	997	0.11	0.04	0.18–0.34	0.21–0.31
SI-BDI < 0.4^	0.17 (0.20)	3	225	0.08	#	0.17–0.17	0.17–0.17
SI-BDI ≥ 0.4^	0.26 (0.31)	9	772	0.12	0.07	0.13–0.39	0.17–0.34
Cognitive ability							
SI	0.08	11	1201	0.18	0.15	−0.20–0.37	−0.10–0.27
BDI	0.11	11	1201	0.20	0.17	−0.23–0.44	−0.11–0.33
Overall	0.10	11	1201	0.18	0.16	−0.21–0.40	−0.10–0.30

Note. ^The overall mean for correspondence between matched SI and BDI interviews is 0.4. #The residual variance in this case was negative, making the SD_{res} a non-real number. *The correction for criterion unreliability is performed here twice, using the traditional 0.52 estimate and a more conservative 0.71 estimate, based on Huffcutt et al. (2014b). The corrected values are in parentheses.

4. Discussion

There tends to be an implicit assumption in the employment interview literature that situational and behavior description questions are simply alternative ways to assess the same job attributes and can be treated as being interchangeable. The results of this investigation suggest that such an assumption is unwarranted. Our findings indicate only a moderate overall association between SI and BDI formats, and further, that under certain interview design conditions, the magnitude of their association can drop noticeably lower. The lowest level of correspondence should occur when an interview is administered for actual employment with less than approximately eight SI and eight BDI questions that have lower respective internal consistencies and probing is not allowed. Conversely, the highest correspondence should occur when there are more than approximately eight SI and eight BDI questions (respectively) that have higher respective internal consistencies and probing is allowed with both types.

We urge researchers and interview developers to refrain from the common practice of treating SI and BDI questions as being interchangeable, and further, from including only one type of question or including both but combining them. It is true that under certain interview design conditions they may achieve a high level of interchangeability, but that is likely to be more of the exception than the rule. Our moderator results should provide preliminary guidance as to when higher correspondence should occur.

The lack of strong overall correspondence between matching SI and BDI questions is interesting theoretically for a number of reasons. It may be that the different question formats induce unique measurement (method) effects. Certainly the presence of a BDI method effect has already been suggested (Huffcutt et al., 2015). The possibility of a unique method effect with SIs has not yet been explored, but is an important avenue for future research. That interviewees tend to use different impression management tactics with each format (see Ellis et al., 2002) could lend support to this notion.

Close scrutiny of the interview design features that tend to result in higher correspondence could provide additional understanding, or at least highlight directions for future research. It is possible that allowing probing with both types induces (or magnifies) a more general interviewee performance factor (see Huffcutt et al., 2001), which in turn increases internal consistencies. Including a larger number of questions may serve to enhance measurement of this factor, although results in the assessment center area would suggest stronger assessment of actual dimensions (see Kuncel & Sackett, 2014).

That a higher correspondence between matching SI and BDI questions was associated with stronger prediction of job performance is particularly interesting. One possibility that would be fruitful for future researchers to explore is whether the skill sets associated with interviewee performance overall, including oral narration ability (see Bangerter et al., 2014), are also beneficial in the workplace and tend to enhance job performance.

Finally, further analysis of the interview design features that tend to result in lower correspondence could also be valuable. Without probing with either format, it may be that the unique measurement (method) influences associated with each type of question kicks in to a stronger degree, thereby driving their correspondence lower. That their internal consistencies were lower should suggest measurement of multiple (and largely different) constructs, but it is uncertain what those constructs are.

Turning to mean validity, a discussion of SI–BDI correspondence raises the question of which format is better for employment purposes. Counter to Taylor and Small (2002), our findings support the higher validity of SI formats over BDI formats in terms of predicting job performance. It is unclear why this is the case though. It could be that the underlying premise behind BDI questions (that the best predictor of future behavior is past behavior) is less accurate than the premise behind SI questions (that intentions are predictive of future behavior). Alternatively, both premises may be correct, but intentions are simply more aligned with future behaviors than are past behaviors.

One direction for future research is to examine additional moderator variables. We excluded other potential moderators from our examination because an a priori reason for them impacting the degree of SI–BDI correspondence was not evident. For example, although many studies examine the impact of individual versus panel interviews, it was unclear why an interview being conducted by a panel or a single interviewer would impact the correspondence between the two types of questions.

In addition to examining other moderator variables, it would be worthwhile to explore additional predictors that may relate to SI–BDI correspondence. We were able to examine the relationship between the SI–BDI correspondence and cognitive ability, but there are numerous predictors that could be informative, including personality and assessment center exercises. Future researchers should explore these other predictors.

Another avenue for future research involves the issue of evaluating SI and BDI correspondence using a multi-trait, multi-method (MTMM) approach. As a reviewer aptly pointed out, our study is limited by the lack of any examination of identifiable differences among constructs. It may be the case that each question type provides a better basis for evaluating some constructs than others (e.g., narration ability). Ideally, we would have been able to separate the construct from the format and evaluate the findings using a MTMM approach. This was not possible in the current study due to the limited availability of information in the primary studies, but we encourage future researchers to explore these issues using the MTMM approach. In doing so, it can be better determined the extent to which our findings are potentially obscured by the collapsing of the interview evaluations across constructs.

Along these lines, future researchers and practitioners would do well to not assume that SIs and BDIs are equivalent simply on the basis of the construct being measured. As Latham and Sue-Chan (1999) point out in their paper on the research comparing BDIs and SIs, just because the same dimensions are measured does not mean that the BDI and SI are truly equivalent in the construct that is measured. Clearly, although the two forms of interviews are developed based on the same set of criteria, there may be differences in the constructs actually measured by the two formats. Our study provides an indirect assessment to this question

in that the relatively low correspondence suggests that matching questions are not parallel and interchangeable and are therefore not capturing the same constructs. Indeed, it would appear from our findings that method effects unique to each format and/or differential influence from moderators may be driving them (and their constructs) apart. Nevertheless, our study is only a start to addressing this issue. Future researchers would do well to explore these issues in greater detail, identifying what each question format is capturing, and identifying ways to make the formats more interchangeable if that is to be the intention of them in practice.

There are limitations that must be acknowledged. First, due to the small number of studies in some analyses, caution should be used when interpreting findings. Along these lines, we were unable to conduct analyses that combined the four moderators together to determine the combined effect on SI–BDI correspondence and validity. Future researchers may wish to compare a single interview in which multiple low correspondence features are present (e.g., a large number of questions, probing only for BDI questions) to an interview in which the high correspondence features are present (i.e., few items with consistent probing). In this manner, it may be possible to determine the extent to which these design features interact with one another to impact format correspondence.

Second, as with all meta-analyses, our results are only as good as the primary studies that comprise them. As such, vague or incomplete reporting in primary studies as well as faulty methodology or analyses could limit the validity of our findings. We urge researchers to provide more detailed information when reporting their samples, methods, and other supporting details that would be of use when combining results across studies.

Third, we were limited to exploring four moderators that were consistently present in primary studies and able to be accurately coded. Along these lines, we were at the mercy of the primary studies and the information provided therein when coding for moderators. For example, as two separate reviewers pointed out, applicant experience could impact the extent to which SIs and BDIs correspond, as more experienced applicants may be better able to respond to hypothetical questions compared to inexperienced applicants while applicants at all experience levels may be able to generate answers regarding past behavioral experiences. However, the information available to us prevented us from being able to assess applicant experience as a moderator. Specifically, each study would presumably have a wide array of experience in their sample, and because the study authors rarely provided interview experience as a measured variable in their study, we were unable to include it in our analyses. Thus, not all moderators could be assessed. Future researchers may wish to examine other possible moderators, including design factors (e.g., availability scoring guides, extent to which the interview is based on a detailed job analysis), that may increase or decrease the correspondence between SIs and BDIs, to further determine when it is appropriate to consider the formats interchangeable.

Fourth, our estimates do not include any statistical correction for range restriction, in either the predictor or the criterion. More and better reporting in regard to the type of range restriction operating in each study would be necessary to adequately make such corrections.

Finally, we did not conduct moderator analyses on the predictive validity analyses beyond the SI–BDI correspondence. There may be important moderators worthy of consideration. For instance, Huffcutt et al. (2004) found that job complexity affected SI but not BDI validity. Similarly, Klehe and Latham (2006) found that whereas both SIs and BDIs predicted maximum performance, SIs were better at predicting typical performance. Given that our primary focus was on SI-BDI correspondence, we were not able to address such issues and leave them for future research.

5. Conclusion

The way in which interview questions are posed is, perhaps, the most influential element of interview design and one of the defining ways in which structured interviews differ from unstructured interviews. Despite the abundance of research on SI and BDI formats, combined with their popularity in practice, prior to the current meta-analysis, little was known regarding whether SI and BDI questions written in parallel to assess the same job dimensions correspond sufficiently to consider (and use) them interchangeably and what factors may influence (moderate) both the validity and interchangeability of SIs and BDIs. In the current investigation, we demonstrated a moderate relationship between SI and BDI questions, despite them being based on the same job analyses, written to assess the same constructs, and asked of the same respondents. Furthermore, there are conditions that elevate or diminish this correspondence. It is clear that more research is needed on these two formats to further understand the psychometric properties and underlying phenomena creating these differences.

References

Bangerter, A., Corvalan, P., & Cavin, C. (2014). Storytelling in the selection interview? How applicants respond to past behavior questions. *Journal of Business Psychology*, *29*, 593–604. http://dx.doi.org/10.1007/s10869-014-9350-0.

*Bosshardt, M. J. (1994). Situational interviews versus behavior description interviews: A comparative validity study. Unpublished doctoral dissertation, University of Minnesota.

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, *91*, 1114. http://dx.doi.org/10.1016/j.jvb.2009.03.008.

Campion, M. A., Campion, J. E., & Hudson, J. P., Jr. (1994). Structured interviewing: A note on incremental validity and alternate question types. *Journal of Applied Psychology*, *79*, 998–1002. http://dx.doi.org/10.1037/0021-9010.79.6.998.

*Campion, M. A., Palmer, D. K., & Campion, J. E. (1998). Structuring employment interviews to improve reliability, validity and users' reactions. *Current Directions in Psychological Science*, *7*, 77–82. http://dx.doi.org/10.1111/1467-8721.ep10773001.

Campion, M. A., Pursell, E. D., & Brown, B. K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology*, *41*, 25–42. http://dx.doi.org/10.1111/j.1744-6570.1988.tb00630.x.

*Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology*, *13*(4), 485–506. http://dx.doi.org/10.1023/A:1022914803347.

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. http://dx.doi.org/10.1037/0021-9010.80.5.565.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104. http://dx.doi.org/10.1037/0021-9010.78.1.98.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *12*, 1–16. http://dx.doi.org/10.1007/BF02310555.

Ellis, A. P. J., West, B. J., Ryan, A. M., & DeShon, R. P. (2002). The use of impression management tactics in structured interviews: A function of question type? *Journal of Applied Psychology*, *87*, 1200–1208. http://dx.doi.org/10.1037//0021-9010.87.6.1200.

*Faulkner, R. J. (1999). How age and expertise moderates the relationship between interview type and interview performance: An analysis of response patterns. (Unpublished master's thesis), California State University, Long Beach.

George, E. (2006). Interviewer accuracy across levels of structure in the employment interview. Unpublished doctoral dissertation, Colorado State University.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco, CA: Freeman & Company.

Gibb, J. L., & Taylor, P. J. (2003). Past experience versus situational employment: Interview questions in a New Zealand social service agency. *Asia Pacific Journal of Human Resources*, *41*, 371–383. http://dx.doi.org/10.1177/1038411103041003008.

Heimann, A. L., Ingold, P., & Kleinmann, M. (2016 April). Developing personality-based interviews: Facilitating both construct and criterion. In J. Levashina (Chair) (Ed.), *Asking the right questions: Investigations of structured interview questions*. Anaheim, CA: Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology.

*Hollwitz, J. (1998). Investigations of a structured interview for pre-employment integrity screening. Unpublished doctoral dissertation, University of Nebraska-Lincoln.

Huffcutt, A. I. (2016, April). Brain activity underlying situational and behavior description interviews. In J. Levashina (Chair), Asking the right questions: Investigation of structured interview questions. Symposium to be presented at the annual meeting of the Society for Industrial and Organizational Psychology, Anaheim CA.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*, 184–190. http://dx.doi.org/10.1037/0021-9010.79.2.184.

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in employment interview evaluations. *Journal of Applied Psychology*, *83*, 179–189. http://dx.doi.org/10.1037/0021-9010.83.2.179.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U. C. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment*, *12*, 262–273. http://dx.doi.org/10.1111/j.0965-075X.2004.280_1.x.

Huffcutt, A. I., Culbertson, S. S., & Goebl, A. P. (2015). The interactive influence of ambition and sociability on performance in a behavior description interview. *Personnel Assessment and Decisions, 1*, 30–36.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment*, *21*, 264–276. http://dx.doi.org/10.1111/ijsa.12036.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014a). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology. *International Journal of Selection and Assessment*, *22*, 297–309. http://dx.doi.org/10.1111/ijsa.12036.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014b). Multi-stage artifact correction: An illustration with structured employment interviews. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*, 548–553.

Huffcutt, A., Roth, P., & McDaniel, M. (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. http://dx.doi.org/10.1111/j.1744-6570.2001.tb00225.x.

Ingold, P. V., Kleinmann, M., König, C. J., Melchers, K. G., & Van Iddekinge, C. H. (2015). Why do situational interviews predict job performance? The role of interviewees' ability to identify criteria. *Journal of Business and Psychology*, *30*(20), 387–398. http://dx.doi.org/10.1007/s10869-014-9368.

Janz, T. (1982). Initial comparison of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology*, *67*, 577–580. http://dx.doi.org/10.1037/0021-9010.67.5.577.

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). Newbury Park, CA: Sage.

*Klehe, U., & Latham, G. (2006). What would you do — Really or ideally? Constructs underlying the behavior description interview and the situational interview in predicting typical versus maximum performance. *Human Performance*, *19*(4), 357–382. http://dx.doi.org/10.1207/s15327043hup1904_3.

*Klehe, U., König, C. J., Richter, G. M., Kleinmann, M., & Melchers, K. G. (2008). Transparency in structured interviews: Consequences for construct and criterion-related validity. *Human Performance*, *21*, 107–137. http://dx.doi.org/10.1080/08959280801917636.

*Kluemper, D. H. (2006). An examination of ability-based emotional intelligence in the structured employment interview. Unpublished doctoral dissertation, Indiana State University.

*Krajewski, H. T., Goffin, R. D., McCarthy, J. M., Rothstein, M. G., & Johnston, N. (2006). Comparing the validity of structured interviews for managerial-level employees: Should we look to the past or focus on the future? *Journal of Occupational and Organizational Psychology*, *79*, 411–432. http://dx.doi.org/10.1348/096317905X68790.

Kuncel, N. R., & Sackett, P. R. (2014). Resolving the assessment center construct validity problem (as we know it). *Journal of Applied Psychology*, *99*, 38–47. http://dx.doi.org/10.1037/a0034147.

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., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology*, *65*, 422–427. http://dx.doi.org/10.1037/0021-9010.65.4.422.

LeBreton, J. L., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 7*(4), 478–500.

Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: Development and validation of an interview faking behavior scale. *Journal of Applied Psychology*, *92*(1638–1656), 1638. http://dx.doi.org/10.1037/0021-9010.92.6.

*Lin, T. -R., & Adrian, N. (1993, May). Multi-method multi-dimension structure interviewing: A field study. *Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.*

*Little, J. P., Schoenfelt, E. L., & Brown, R. D. (2000, April). The situational versus patterned behavior description interview for predicting customer service performance. *Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, New Orleans.*

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

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*, 599–616. http://dx.doi.org/10.1037/0021-9010.79.4.599.

*Melchers, K. G., König, C. J., Richter, G. M., Harstein, T., Kleinmann, M., & Klehe, U. (2008, April). Effects of assessor rotation on interview and AC construct validity. *Paper presented at the 23rd annual conference of the Society for Industrial and Organizational Psychology, San Francisco.*

*Menkes, J. (2002). What do structured interviews actually measure? A construct validity study. Unpublished doctoral dissertation, Claremont Graduate University.

*Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology*, *58*(3), 583–611. http://dx.doi.org/10.1111/j.1744-6570.2005.655.x.

Mussel, P. (2007). *Die Konstruktvalidität des Multimodalen Interviews (Unpublished doctoral dissertation).* Stuttgart: Universität Hohenheim.

*Mussel, P., Behrmann, M., & Schuler, H. (2008, April). Explaining the psychometric properties of structured and unstructured interviews. *Poster presented at the 23rd annual conference of the Society for Industrial and Organizational Psychology, San Francisco.*

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology*, *48*, 289–308. http://dx.doi.org/10.1111/j.1744-6570.1995.tb01758.x.

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., & Schmit, M. J. (2005). Personality saturation in structured interviews. *International Journal of Selection and Assessment*, *13*, 261–273. http://dx.doi.org/10.1111/j.1468-2389.2005.00323.x.

Roulin, N., Bangerter, A., & Levashina, J. (2015). Honest and deceptive impression management in the employment interview: Can it be detected and how does it impact evaluations? *Personnel Psychology*, *68*, 395–444. http://dx.doi.org/10.1111/peps.12079.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353. http://dx.doi.org/10.1037/1040-3590.8.4.350.

Sears, G. J., & Rowe, P. M. (2003). A personality-based similar-to-me effect in the employment interview: Conscientiousness, affect- vs. competence-mediated interpretations, and the role of the job relevance. *Canadian Journal of Behavioral Science*, *35*(1), 13–24. http://dx.doi.org/10.1037/h0087182.

*Sue-Chan, C., Latham, G. P., & Evans, M. G. (1995). The construct validity of the situational and patterned behavior description interviews: Cognitive ability, tacit knowledge, and self-efficacy as correlates. *Poster session presented at the annual meeting of the Canadian Psychological Association, Charlottetown, Prince Edward Island*.

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 behavior employment interview questions. *Journal of Occupational and Organizational Psychology*, *75*, 277–294. http://dx.doi.org/10.1348/096317902320369712.

Tross, S. A., & Maurer, T. J. (2008). The effect of coaching interviewees on subsequent interview performance in structured experience-based interviews. *Journal of Occupational and Organizational Psychology*, *81*, 589–605. http://dx.doi.org/10.1348/096317907X248653.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574. http://dx.doi.org/10.1037/0021-9010.81.5.557.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*, 108–131. http://dx.doi.org/10.1037/0021-9010.90.1.108.

Weekley, J. A., & Gier, J. A. (1987). Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology*, *72*, 484–487. http://dx.doi.org/10.1037/0021-9010.72.3.484.

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. http://dx.doi.org/10.1111/j.2044-8325.1988.tb00467.x.