The Impact of Job Complexity and Study Design on Situational and Behavior Description Interview Validity

Allen I. Huffcutt*
Bradley University

James M. Conway
Central Connecticut State University

Philip L. Roth
Clemson University

Ute-Christine Klehe
Universiteit van Amsterdam

The primary purpose of this investigation was to test two key characteristics hypothesized to influence the validity of situational (SI) and behavior description (BDI) structured interviews. A meta-analysis of 54 studies with a total sample size of 5536 suggested that job complexity influences the validity of SIs, with decreased validity for high-complexity jobs, but does not influence the validity of BDIs. And, results indicated a main effect for study design across both SIs and BDIs, with predictive studies having 0.10 lower validity on average than concurrent studies. Directions for future research are discussed.

Introduction

Situational interviews (SIs) and behavior description[1] interviews (BDIs) have emerged as the premier formats for conducting modern structured interviews (Campion, Palmer, & Campion, 1997; Harris, 1989; Motowidlo, 1999). The main difference between them is that the former focuses on evaluation of intended reactions to hypothetical job situations (Latham, Saari, Pursell, & Campion, 1980) while the latter focuses on evaluation of reactions in actual situations from each candidate's past relevant to the target position (Janz, 1982). Both are typically developed from critical incidents and often can be written to assess the same job characteristics (Campion, Campion, & Hudson, 1994). Both also have a theoretical basis, namely goal setting (i.e., that intentions are the immediate precursor of a person's actions; Latham, 1989) and behavioral consistency (i.e., that the past is the best predictor of the future; Janz, 1989).

Research suggests that these interview formats provide good validity for predicting a variety of criteria such as educational grades, task performance, and/or organizational citizenship behavior. In addition to a number of individual studies published over the last two decades (e.g., Janz, 1982; Johnson, 1990; Latham & Saari, 1984; Latham & Skarlicki, 1995; Morgeson, Reider, & Campion, 2002; Orpen, 1985; Weekley & Gier, 1987), there have been two relatively recent meta-analyses conducted specifically on SIs and/or BDIs. Latham and Sue-Chan (1999) found a mean-corrected validity of .47 across 20 SI studies. Similarly, Taylor and Small (2002) found a mean-corrected validity of .45 across 30 SI studies and .56 across 19 behavior description studies.

However, a major limitation of the current literature is that very little is known about the factors that influence (i.e., moderate) the validity of these structured interviews. One potentially important factor is the complexity of the job. There appears to be an implicit assumption in much of the interview literature that SIs and BDIs work equally well for all types of jobs, an assumption that has recently come into question (Pulakos & Schmitt, 1995; Huffcutt, Weekley, Wiesner, & DeGroot, 2001).

Another potentially important moderating factor is the design of the study, namely whether it was conducted in a predictive or in a concurrent manner. Similar to job

*Address for correspondence: Allen Huffcutt, Department of Psychology, Bradley University, Peoria, IL 61625, USA. E-mail: huffcutt@bradley.edu

complexity, there appears to be an implicit assumption in much of the interview literature that the two designs yield equivalent validities (see Barrett, Phillips, and Alexander (1981) for a discussion of the equivalency of these designs in general selection). However, as outlined later (e.g., Hough, 1998; Salgado & Moscoso, 1995; Schmitt, Gooding, Noe, & Kirsch, 1984), there is reason to suspect that predictive and concurrent designs do not yield equivalent results for SIs and BDIs, specifically that predictive validities may be lower than concurrent validities.

Clearly, researchers and organizational decision-makers need to know the extent to which job complexity and study design influence SI and/or BDI outcomes. Without such knowledge, some developers conducting their own validation study may be unpleasantly surprised when they find lower validity than expected. In these cases, knowledge of moderating influences might have allowed them to make different decisions to improve validity, including using one structured format rather than the other format. In situations where a validity study is not conducted locally, relying on mean estimates from other analyses could result in inaccurate estimates of validity, even inappropriate decisions regarding what selection methods to use.

The primary purpose of this investigation was to evaluate the moderating influence that job complexity and study design have on the validity of SIs and BDIs. We believe that understanding the influence of these moderators is an important and necessary step in the continued evolution of structured interviewing methodology. We begin with a more in-depth discussion of the potential influence that job complexity and study design could have on validity.

Moderators of SI and BDI Validity

Job Complexity

Several researchers have suggested that the validity of SIs may be affected by the complexity of the job, particularly that SIs may be less effective for positions of higher complexity. Pulakos and Schmitt (1995) interviewed federal investigative agents with an SI and a BDI and found considerably lower validity for the situational portion. Offering a potential explanation for this unexpected effect, they reported that some candidates thought through every possible contingency when answering the situational questions while other candidates gave responses that were more superficial. Yet, because the latter answers were still essentially correct according to the rating benchmarks, the candidates engaging in more complex thought did not necessarily receive higher ratings. Pulakos and Schmitt did not experience such difficulties with the behavior description portion of their interview. Similarly, in an analysis of retail district managers, Huffcutt *et al.* (2001) found considerably higher validity for the behavior description portion than for the situational portion.

However, the above evidence represents only two studies and additional SI and BDI studies are available which involve a position of higher complexity. Accordingly, a meta-analytic test could more conclusively confirm or disconfirm whether SIs are less effective for complex positions, and that is one of the main purposes of this investigation. We will argue a little later that BDIs should not be as susceptible to job complexity effects.

There are at least two conceptual/methodological reasons to suspect a moderating influence from job complexity on SIs. The first is based on Pulakos and Schmitt's (1995) reported difficulty regarding scoring of responses. It is possible that the standard situational scoring system, namely a five-point scale with relatively brief benchmarks at the one, three, and five points, may be perfectly fine for low- and medium-complexity positions. But, unless carefully pretested and refined, it may not be detailed enough to differentiate thoroughly among the more intricate and detailed answers given by candidates for more complex positions. It is our observation that a number of situational studies available in the interview literature did not do extensive pretesting.

The other possible reason could relate to the "richness" of the questions themselves. The critical incidents collected for complex positions are likely to be more complicated and multi-faceted than those for less complex positions. Similar to the scoring issue noted above, it could be that the standard process of turning critical incidents into hypothetical scenarios works reasonably well for positions of low and medium complexity. However, unless carefully pretested, the subtle details and underlying dynamics may not always be adequately captured when incidents from more complex positions are turned into hypothetical questions.

In contrast, there is much less reason to suspect that BDI validity would be moderated by job complexity. The complexities and dynamics involved in critical situations for more complex jobs should have a better opportunity to emerge because the candidates present the situations rather than the situations being read from prepared questions. Moreover, in cases where a situation presented by an interviewee is not clear, the interviewer is usually encouraged to probe further (Janz, 1989). While such probing may take extra time, it could, if consistently applied, ensure that interviewers and interviewees share a common understanding of that situation and its inherent complexities and outcomes. We therefore predicted that the validity of SIs would show at least some decrement for positions of higher complexity while the validity of BDIs would be relatively unaffected by the complexity of the job.

Study Design

There appeared to be a common opinion in earlier selection literature that concurrent designs were not as desirable as predictive designs. Barrett, Phillips, and Alexander (1981) summarized the four main criticisms of the concurrent

design that were thought to make it inferior (i.e., missing people, restriction of range, motivation/demographic differences between applicants and incumbents, and confounding by job experience), and argued that these criticisms were not as problematic as previously thought. Further, they cited empirical evidence showing similar validity coefficients for a major test of cognitive ability, the General Aptitude Test Battery (see Bemis, 1968).

In a subsequent and large-scale empirical investigation, Schmitt *et al.* (1984) found that concurrent validities across several types of predictors (not including interviews) were actually higher on average than predictive validities. Specifically, they found that the concurrent studies in their analysis had a mean validity .04 higher than the predictive studies where the predictor was not used in selection and .08 higher than the predictive studies where the predictor was used in selection (see Schmitt *et al.*, 1984, p. 412). They hypothesized that, with many predictors, the indirect restriction assumed to occur in concurrent designs (through attrition and promotion) may in fact be less severe on average than the direct restriction of range that commonly occurs with predictive designs. In a more recent investigation, Hough (1998) found that the mean validity for concurrent personality inventory studies was .07 higher than the mean validity for predictive studies.

Unfortunately, design differences have rarely been studied in relation to structured employment interviews. In the only known meta-analysis, Salgado and Moscoso (1995) found that the mean estimated validity of structured behavior interviews (including both situational and behavior description studies combined) was .08 higher for the 15 concurrent studies in their data set than for the 10 predictive studies in their data set (see also Salgado, 1999). As discussed in the next section, considerably more studies have become available since their meta-analysis, which gave us a larger base upon which to derive estimates and also provided a better opportunity to do separate SI and BDI analyses.

Conceptually, there is reason to expect higher validity for concurrent studies than for predictive studies for both SIs and BDIs. Unlike with mental ability testing, being an incumbent in a structured interview could have a direct influence on the predictor–criterion relationship. In particular, it is likely that candidates base at least some of their responses on situations and experiences from their current position, and supervisors in turn may base their performance ratings on the same set of behaviors. The result of operating from a common base of behaviors should be a higher validity correlation.

Another possible explanation for higher concurrent validity is differences in criterion reliability as a function of the length of the rater–ratee acquaintance. Rothstein (1990), for example, found that criterion reliability increased as the rater knew the ratee for increased lengths of time. It is quite possible that raters know the ratees for longer periods of time in concurrent studies than in predictive studies, which in turn could also increase validity.

In summary, there are two potential factors that could increase validity for concurrent studies relative to that for predictive studies: operating from a common base of behaviors and longer rater–ratee acquaintances. Accordingly, we predicted that SI and BDI studies utilizing a concurrent design would have higher overall validity than studies utilizing a predictive design, perhaps even a greater difference than those reported by Schmitt *et al.* (1984) and Hough (1998) for other predictors.

Method

Search for Studies

We conducted an extensive search to locate SI and BDI studies for our investigation. Studies included in previous interview meta-analyses were identified (Huffcutt & Arthur, 1994; Huffcutt & Woehr, 1999; Latham & Sue-Chan, 1999; McDaniel *et al.*, 1994; Taylor & Small, 2002; Wiesner & Cronshaw, 1988). Issues from 1999 to the present of the *Journal of Applied Psychology* and *Personnel Psychology* were examined to locate any interview studies that were published after those in the meta-analyses listed above. Conference programs from 1999 to 2002 from the *Society for Industrial and Organizational Psychology* and the *Academy of Management* were similarly checked to find any more recent studies. As additional measures, the databases *PsychLit* and *ABI-INFORM* were reviewed, an internet search was made using the engine, *Google*, and supplemental inquiries were made to prominent researchers in the interview area to obtain any additional studies not included in the above sources.

Two main criteria guided our search. First, the criterion in the studies had to reflect overall job performance in an actual position in business and industry (either applicants or incumbents) or in a professional training program that included duties in an on-the-job setting (e.g., medical school resident). Although the two meta-analyses described earlier included studies with other criterion such as academic grades and organizational citizenship behavior (OCB),[2] we believed it important and advantageous to focus only on overall job performance. Doing so strengthens the ability to make generalizations to the prediction of job performance, and allows better comparison with the results of other predictors that have been validated relative to overall job performance (see Borman, 1991; Hunter & Hunter, 1984; Schmidt & Hunter, 1998). Thus, we excluded several studies involving university students in an academic program where the criterion was grades (e.g., Schuler & Funke, 1989, second study; Schuler & Prochaska, 1990; Sue-Chan, Latham, & Evans, 1995), several other studies involving performance in training without corresponding duties in the actual job setting (Huffcutt *et al.*, 2001, study 1; Mosher, 1991), and one study where

the criterion was organizational citizenship behavior (Latham & Skarlicki, 1995).

Second, studies had to be true SIs or BDIs where all of the questions were of the intended type (either situational or behavior description) and responses were scored accordingly to established guidelines. We excluded two of Kennedy's (1986) studies because, after obtaining an actual copy of her dissertation, we discovered that less than half of the questions in these two studies involved hypothetical situations where candidates were asked what they would do. Similarly, we excluded Campion, Pursell, and Brown (1988) study because their interview included other types of questions such as job knowledge and worker requirements. We also eliminated the second study in Latham and Saari (1984) because it appeared that the interviewers did not use the established rating scales but rather just asked the questions and then made subjective ratings after the interviews. (We did use the data from the follow-up to the second study where the interview was used correctly; see Latham & Saari, 1984, p. 573.)

Lastly, we excluded three additional studies even though they met the above criteria. We did not use the study of convenience store cashiers by Hays (1990) because, at the time the performance criteria were collected (90 days after hire), only 38 of the original 104 cashiers were still employed. In fact, many appear to have left shortly after being hired, often within the first week. This resulted in fairly extreme and atypical range restriction on the criterion. We also did not use the study by Lin and Adrian (1993) because it involved internal promotion and the interviewers were given a copy of the candidates' most recent performance appraisal right before the interview. Lastly, we did not use the study by Kleinmann and Deller (1993) because the subjects were people from within the organization who volunteered to be interviewed but were not in the intended position nor did they plan to apply for it, and performance appraisals from their current positions were used as the validation criteria.

As a result of this search we were able to locate 54 usable employment interview studies with a total sample size of 5536. Citations for these studies are included in the general list of references, identified by an asterisk. The studies included a wide range of sources, job types, companies, and products. In terms of format, 32 of these studies were situational, with a total sample size of 2815, and 22 were behavior description, with a total sample size of 2721.

Comparison with Other SI and BDI Meta-Analyses

Latham and Sue-Chan (1999) analyzed 20 studies in their meta-analysis of SI validity, and 13 of those studies were also included in our data set. In regard to the seven studies in their data set that we chose to exclude, five involved criteria other than overall job performance such as university grades (e.g., Schuler & Prochaska, 1990) and organizational citizenship behavior (Latham & Skarlicki, 1995), one had atypical problems with the sample (Hays, 1990), and another involved several different types of questions (e.g., job knowledge, worker requirements) in addition to situational (Campion *et al.*, 1988). Through our extensive search we were able to find an additional 19 SI studies, giving us a total of 32 SI studies.

Taylor and Small's (2002) meta-analysis included analysis of both SI and BDI validity. They had 30 SI studies in their data set, 23 of which were included in our data set. Of the remaining seven studies in their data set that we chose to exclude, two involved criteria other than overall job performance such as organizational citizenship behavior (e.g., Latham & Skarlicki, 1995), three were not true SIs in that they included other types of questions (Campion *et al.*, 1988; Kennedy, 1986, studies 1 and 2), and another was the Hays (1990) study mentioned above. The final study we excluded was the original second study in Latham and Saari (1984) where the interviewers did not use the rating scales provided. Through our search we were able to find an additional nine SI studies, giving us a total of 32 SI studies.

Taylor and Small (2002) also had 19 BDI studies, 17 of which were in our data set. The two studies from their data set which we did not use were Latham and Skarlicki (1995), where the criterion was organizational citizenship behavior, and Latham and Saari (1984), where the majority of questions were not behavior description in nature but rather appeared to be Ghiselli-type (1966) questions such as "Why did you want this job?". Through our search we were able to find an additional five BDI studies, giving us a total of 22 BDI studies.

As evident from the above discussion, our data set differed from Latham and Sue-Chan (1999) and Taylor and Small (2002) in four ways: (1) it was larger than either of these two data sets, (2) we limited it to studies where the criterion was overall job performance, (3) we excluded several individual studies because they contained other types of questions, and (4) we excluded several other studies for methodological shortcomings. As a result we believe that our studies are a better representation of SI and BDI validity and that estimates derived from them should provide more accurate information for personnel decision-makers, especially if they want to compare them with estimates for other predictors that have been validated relative to overall job performance. Moreover, our capability to perform meaningful moderator analyses was enhanced because we controlled for extraneous variance resulting from inclusion of heterogeneous criteria and inclusion of studies with non-situational or behavior description questions.

Coding of Study Variables

All validity correlations were recorded as observed without correction for artifacts such as criterion unreliability. Job complexity was coded using the three-level framework outlined by Hunter, Schmidt, and Judiesch (1990) where

unskilled and semi-skilled jobs such as toll-booth collector, receptionist, and mail-sorter are coded as low complexity, jobs such as skilled trades, first-line supervisors, and lower-level administrators are coded as medium complexity, and higher-level positions such as managers, physicians, and engineers are coded as high complexity. Their framework is based on Hunter's (1980) original system for coding the information processing requirements of positions and combines various levels of data and/or things from the *Dictionary of Occupational Titles* (US Department of Labor, 1977). Other variables, including sample size and study format (predictive vs. concurrent), were coded as provided in the studies.

Both the first and the second authors independently coded all study variables in order to increase the reliability of the coding process and to allow assessment of inter-rater reliability. Any disagreements were discussed and resolved. In a limited number of studies where consensus was not reached because the information provided was not sufficiently clear, we contacted the authors for clarification. The inter-rater reliability correlations were .99 for the validity coefficients, .87 for job complexity, and 1.00 for study design (all $p < .0001$). These values suggest that the study variables were reliably coded.

Overall Analysis of Situational and Behavior Description Validity

To get an idea of the overall effect across studies features, we first computed the mean validity correlation for the SI and BDI studies in our data set. Similar to Huffcutt, Roth, and McDaniel (1996), we employed a modified version of sample weighting. As in that study, the concern was that a handful of large-sample studies would dominate the results, particularly so with the moderator analyses described later where the number of studies in some of the categories was relatively small. Huffcutt *et al.* (1996) used a three-point framework where the largest study was weighted three times the smallest based on naturally occurring groupings of sample size. We used a very similar system here where studies with sample sizes less than 40 were weighted 1.0 ($k = 15$), studies with sample sizes between 40 and 75 were weighted 1.5 ($k = 15$), studies with samples sizes between 76 and 110 were weighted 2.0 ($k = 13$), studies with sample sizes between 111 and 200 were weighted 2.5 ($k = 7$), and, finally, studies with sample sizes greater than 200 were weighted 3.0 ($k = 4$). Such a system retained the advantage of allowing studies based on a larger sample to contribute more to the results, yet limited their contribution to no more than three times that of any other study.

Then we computed the variance across the SI and BDI studies using the same weighting scheme described above and estimated the percent of that variance that was attributable to sampling error.[3] Hunter and Schmidt (1990) suggest that if at least 75% of the observed variance is accounted by sampling error, moderator variables are most likely not present, or if present, have minimal influence. We followed their guideline in determining whether other factors (e.g., job complexity, study design) moderated SI and/or BDI validity.

After that we formed confidence intervals around the observed mean correlations using the methodology outlined by Viswesvaran, Ones, and Schmidt (1996). They formed 80% intervals around the mean correlations in their meta-analysis of criterion reliability values, and we did likewise because they appear to provide an adequate representation of sampling error variability for selection-related meta-analyses. Specifically, we took the mean correlation ± 1.28 times the square root of the observed variance divided by the number of coefficients (Osburn & Callender, 1992; Viswesvaran *et al.*, 1996). In regard to interpretation, there is an 80% probability that the mean correlation from an entirely different set of studies (e.g., 32 other SI studies) would fall within this interval.

Finally, we corrected the observed mean validity estimates for two artifacts: measurement error in the criterion and range restriction in the interview ratings (Hunter & Schmidt, 1990). Our data set contained nine estimates of criterion interrater reliability, and the mean of these estimates was .72. This value is somewhat higher than the frequently cited value of .52 reported by Viswesvaran *et al.* (1996). There are two reasons for our more conservative estimate. First, in some of these studies fairly elaborate and sophisticated performance evaluation instruments were created as part of the validity study, and these would be expected to have higher reliability than standard administrative ones used by many organizations. Second, in over half of these studies performance ratings were obtained from multiple raters, and these ratings were subsequently combined for the validity analysis. In these cases we used the reliability estimates representing all raters involved rather than using the reliability of a single rater, as the former appeared more representative of the actual reliability.

Our data set also contained three estimates of range restriction, specifically the ratio of the restricted to unrestricted standard deviation, which is typically designated as "u". The mean value for u in these three studies was .70. This value is slightly lower than the .74 value found by Huffcutt and Arthur (1994) but higher than the .61 value found by Salgado and Moscoso (2002); given its position between these two more established estimates, we went ahead and used the value of .70 for the range restriction ratio. Corrections for both criterion unreliability and interview range restriction were performed using the methodology outlined by Hunter and Schmidt (1990).

Analysis of Moderator Variables

First, we analyzed the influence that job complexity had on SI and BDI validity. Here we separated the SI studies according to whether the position involved was low, medium, or high complexity and then computed the mean

Table 1. Overall validity of situational and behavior description interviews and analysis of job complexity

	k	TSS	$\bar{r}$	$CI_{80\%}$	Var(*r*)	Var(*e*)	PVA (%)	$\bar{r}_{(c)}$
Situational interviews	32	2815	.26	.23–.29	.0213	.0100	47	.43
Low complexity	13	1320	.27	.21–.33	.0257	.0085	33	.44
Medium complexity	13	714	.31	.27–.35	.0111	.0152	100	.51
High complexity	6	781	.18	.11–.25	.0191	.0072	38	.30
Behavior description interviews	22	2721	.31	.28–.34	.0149	.0067	45	.51
Low complexity	4	835	.30	.24–.36	.0100	.0040	40	.48
Medium complexity	10	903	.32	.27–.37	.0126	.0090	71	.51
High complexity	8	983	.31	.25–.37	.0203	.0067	33	.51

In the heading above "*k*" refers to the number of studies, "TSS" is the total sample size for those studies, "$\bar{r}$" is the mean uncorrected validity correlation, "$CI_{80\%}$" is the 80% confidence interval around that mean, "Var(*r*)" is the observed variance in the validity correlations, "Var(*e*)" is the variance expected from sampling error, "PVA" is the percent of variance attributable to sampling error, and "$\bar{r}_{(c)}$" is the mean validity correlation corrected for range restriction in the interview ratings and unreliability in the criterion.

validity, observed variance, sampling error variance, confidence interval, and corrected validity separately for each level using the methodology described above. Then we separated the BDI studies into the same three levels and performed the same computations.

To analyze the moderating effect of study design, we separated the SI and BDI studies, respectively, as to whether they were predictive or concurrent and performed the same analyses as with job complexity. Then, we combined the SI and BDI studies and did a final analysis where we compared the predictive and concurrent designs across both types of interviews.

Results

Results for the analyses of overall SI and BDI validity are presented in Table 1. Across the 32 situational studies with a total sample size of 2815, the mean observed validity correlation was .26 and the observed variance was .0213. Estimated sampling error variance was .0100, suggesting that only 47% of the observed variance was attributable to sampling error. Thus, the presence of moderator variables appeared likely. After correction for criterion unreliability and range restriction in the interview ratings, the mean correlation increased to .43.

Across the 22 behavior description studies with a total sample size of 2721, the mean observed validity correlation was .31 and the observed variance was .0149. Estimated sampling error variance was .0067, suggesting that only 45% of the observed variance was attributable to sampling error. Thus, the presence of moderator variables also appeared likely. After correction for criterion unreliability and range restriction in the interview ratings, the mean correlation increased to .51.

Results for the analyses of job complexity are also presented in Table 1. For SIs, the mean validity across the 13 low-complexity studies with a total sample size of 1320 was .27 (.44 corrected), while the mean validity across the 13 medium-complexity studies with a total sample size of 714 was .31 (.51 corrected). In contrast, the mean validity across the six high-complexity SI studies with a total sample size of 781 was only .18 (.30 corrected). Moreover, the 80% confidence interval for the high-complexity SI studies did not overlap at all with the one for the medium-complexity SI studies and overlapped only slightly with the confidence interval for the low-complexity SI studies. Consistent with our first hypothesis, these findings suggest that SIs may not be as effective for positions of higher complexity, although the modest number of studies for the high-complexity category ($k = 6$) makes this conclusion at least somewhat tentative.

For BDIs, the mean validity across the three complexity levels was .30 for low complexity ($k = 4$, $N = 835$), .32 for medium complexity ($k = 10$, $N = 903$), and .31 for high complexity ($k = 8$, $N = 983$). The corrected values were .48, .51, and .51, respectively. The very consistent nature of these findings suggests that the validity of BDIs is not moderated by job complexity, although the modest number of low-complexity studies ($k = 4$) makes the finding for this category more tentative.

Results for study design are shown in Table 2. The mean validity across the 12 predictive studies (both SI and BDI combined) with a total sample size of 1573 was .23 (.38 corrected). In contrast, the mean uncorrected validity across the 41 concurrent studies (both SI and BDI combined) with a total sample size of 3838 was .30 (.48 corrected), an uncorrected difference of .07 and a corrected difference of .10.

As a further analysis of study design, we removed the six situational studies involving a high complexity position. All six of these studies were concurrent, and, as noted above in the job complexity results, this cell had noticeably lower validity. Accordingly, there was a possible confound on the study design results from the job complexity results. After

Table 2. Analysis of study design: predictive vs. concurrent

	k	TSS	$\bar{r}$	$CI_{80\%}$	Var(*r*)	Var(e)	PVA (%)	$\bar{r}_{(c)}$
Study Design – Overall								
Predictive	12	1573	.23	.18–.28	.0166	.0069	42	.38
Concurrent	41	3838	.30	.27–.33	.0188	.0088	47	.48
Concurrent adjusted	35	3097	.33	.30–.36	.0154	.0091	59	.53
Situational interviews								
Predictive	10	884	.25	.19–.31	.0221	.0101	46	.41
Concurrent	22	1931	.27	.23–.31	.0210	.0101	48	.44
Concurrent adjusted	16	1150	.31	.27–.35	.0169	.0115	68	.50
Behavior description interviews								
Predictive	2	689	.20	.06–.34	.0006	–	–	.33
Concurrent	19	1947	.34	.31–.37	.0138	.0077	56	.54

All terms and symbols in the heading are the same as defined in Table 1. Error variance and percent-of-variance accounted for are not shown for the predictive behavior description interviews category because there were only two studies. The concurrent adjusted category shows the results with the six situational studies involving a high-complexity position removed.

removing these six studies, which removed the confounding effect of job complexity, the resulting mean validity across the remaining 35 concurrent studies with a total sample size of 3097 was .33 (.53 corrected), an uncorrected difference of .10 and a corrected difference of .15 from the predictive results. This finding is labeled "Concurrent adjusted" in Table 2. As shown, the 80% confidence intervals for the predictive and concurrent-adjusted designs do not overlap. Thus, our second hypothesis that concurrent studies would have higher overall validity than predictive studies, also appears to be supported.

Results for analysis of study design within each type of interview are as follows. For the SI studies, the mean validity across the 10 predictive SI studies with a total sample size of 884 was .25 (.41 corrected) while the mean validity across the 22 concurrent SI studies with a total sample size of 1931 was .27 (.44 corrected). After removing the six high-complexity studies, the mean validity for the remaining 16 concurrent studies with a total sample size of 1150 was .31 (.50 corrected), an uncorrected difference of .06 from the predictive mean (.09 corrected difference). For the BDI studies, the mean validity across the two predictive BDI studies with a total sample size of 689 was .20 (.33 corrected) while the mean validity across the 19 concurrent BDI studies with a total sample size of 1947 was .34 (.54 corrected), an uncorrected difference of .14 (.21 corrected difference). However, this difference should be interpreted with extreme caution given the low number of predictive BDI studies ($k = 2$).

Discussion

The primary purpose of this investigation was to assess the influence that job complexity and study design have on validity estimates of overall job performance for SIs and BDIs. Our results suggest that job complexity has an influence on the validity of SIs and that study design has an influence on the validity of both SIs and BDIs. Clearly it would appear that these factors need to be taken into account when making selection decisions or inappropriate decisions and/or loss of performance could result.

In regard to job complexity, our results suggest some caution when using SIs for high-complexity positions. For these jobs the mean uncorrected validity was estimated to be .18 for SIs, which is noticeably lower than the uncorrected estimate of .31 found for BDIs. For positions of medium and low complexity, however, our results indicate that both SIs and BDIs provide comparable validity. In fact, the mean uncorrected validity for both types of interviews across both low- and medium-complexity positions ranged from only .27 to .32. Accordingly, it would appear that validity tends to be very robust in these situations, and not particularly dependent on the type of job involved or the structured interview format utilized.

A key question that emerges from the above discussion is why SIs may not work as well for positions of higher complexity. One strong possibility is that it is more difficult to develop SIs for these types of jobs. It may be somewhat tricky, for example, to write hypothetical questions that capture the intricacies and subtle dynamics often found in critical situations for these types of positions and/or to develop scoring anchors that allow relatively sophisticated differentiation of responses. While such methodological concerns could be alleviated through careful pretesting and refinement of both the questions and the scoring anchors, it is our observation that such pretesting and refinement is not done consistently in the studies available in the interview literature.

Another possibility is that the candidates for these types of positions, because of their greater faculties, may have more of a tendency to report what they believe the interviewers want to hear and not necessarily what they would actually do in the situations presented. Kleinmann and Klehe (2001), for example, found that interviewees who recognized the underlying dimensions being assessed by SI questions received higher overall ratings. Latham and his colleagues (Latham & Sue-Chan 1999; Latham & Skarlicki, 1995) in fact emphasize that situational questions should include some type of dilemma that forces the respondents to choose between equally desirable courses of action in order to reduce the possibility of socially desirable responding. It is our observation that a majority of the situational studies in the current interview literature include questions that do not have an obvious dilemma. One important direction for future research therefore is to assess the effectiveness of including a dilemma in reducing socially desirable responding and in improving validity.

Regardless of the underlying reasons, differences in validity as a function of job complexity could have important financial and operational implications. Assuming uncorrected validity estimates for high-complexity jobs of .18 for SIs and .31 for BDIs, a standard deviation in performance (SD_y) of $32,300 (e.g., regional sales manger from Burke & Frederick, 1984), that 50 individuals are selected a year, and that individuals stay with the organization for 5 years, the increase in performance from using a BDI rather than a SI over this timeframe would be estimated at $1,049,750. This analysis of course assumes that costs of development and administration for these interviews are equivalent, which is a reasonable assumption. Our point here is not to make the most financially accurate estimate of utility, but rather to illustrate how choosing the most appropriate interviewing approach might help organizations to function more effectively.

In regard to study design, our results clearly suggest that predictive and concurrent designs are not equivalent. We found that situational and behavior description studies conducted with a concurrent design had a mean validity that was .10 higher than the studies conducted with a predictive design using uncorrected values and a mean validity that was .15 higher when comparing corrected values. We believe this finding to be of significance for the science and practice of structured interviewing. Such a finding is particularly noteworthy because it exceeds the differences reported for other predictors (Hough, 1998; Schmitt *et al*., 1984), and contributes to a growing body of evidence, which suggests that for certain predictors (e.g., personality tests, structured interviews) concurrent designs often yield higher validity correlations than predictive designs.

There are two possible explanations for the higher validity of concurrent designs. One is that the opportunity for incumbents in concurrent studies to utilize experiences and knowledge from their present position when answering the interview questions tends to magnify the resulting correlation with job performance. Another important avenue for future research is to determine empirically just how much of the answers provided by incumbents are related to their present position, and the extent to which this influences validity. Our findings, although highly tentative, further suggest that BDIs might be especially vulnerable to effects of study design. Additional predictive BDI studies are needed to verify this pattern.

The other possible explanation is that, as suggested by Schmitt *et al*. (1984), the indirect restriction in range associated with concurrent methodology tends to be less severe than the direct restriction in range caused by predictive methodology. If true, a problem emerges when correcting for artifacts because the data for range restriction typically comes from predictive interview studies and thus may result in overcorrection when applied to concurrent studies. Unfortunately, specific information on the degree of restriction in structured interview scores for concurrent samples does not appear to be available at the present time.

Regardless of the cause, two important implications for the science and practice of structured interviewing emerge from this finding. First, collapsing across predictive and concurrent interview studies, when computing mean validity estimates, may not be advisable, as the resulting estimates would be more heavily influenced by whichever design that is more frequently represented (which in our case was concurrent), thereby making it less accurate for the design with lower representation. Second, predictive validity coefficients will tend to be lower, and thus organizations should exert caution when making selection decisions based on results of a predictive validation study. It is possible, for example, that they might decide to exclude a structured interview entirely if the validity correlation comes out lower than expected.

Along the way to investigating our moderator variables, we also computed overall estimates of validity for both types of structured interviews collapsing across study features such as job complexity and study design. We openly acknowledge that these estimates are more heavily weighted by concurrent studies, and that the corrections for them are based on artifact data from predictive studies. Nevertheless, these are the same conditions under which previous estimates of mean validity for other predictors have been derived (Hunter & Hunter, 1984; Schmidt & Hunter, 1998), and are necessary to maintain compatibility. Our estimates, which we believe to be the best available at the present time, suggest a mean corrected validity of .43 for SIs and a mean corrected validity of .51 for BDIs. We note that these validities were computed only for the criterion of overall job performance and were based only on studies in which all of the questions were situational or behavior description.

The magnitude of these estimates is such that they appear to compare very favorably with the mean validity

found for better predictors such as ability tests and work samples (Hunter & Hunter, 1984; Schmidt & Hunter, 1998). Given the typically low correlation between these structured interviews and mental ability tests (Cortina, Goldstein, Payne, Davison, & Gilliland, 2000; Huffcutt *et al*., 1996), there is a distinct possibility of using both to enhance and maximize validity.

As always, limitations should be noted. First, although larger than other meta-analyses of situational and behavior description validity, our data set, nonetheless, was smaller than meta-analyses of some other selection methods such as ability tests. Part of the reason for the modest size was our restriction that the studies involve actual job performance and part of the reason was that there just are not an overwhelming number of SI and BDI validity studies available. Given our very extensive search, we doubt that many more studies could be found at this point in time. It is also important to note that interviews are often used in the second stage of a selection system after applicants are screened on another test (Roth, Bobko, Switzer, & Dean, 2001), and thus sample sizes are often smaller than corresponding studies with other selection methods. The modest size of our data set was not such a problem for the overall analyses of situational and behavior description validity, but was more of a problem with some of the moderator categories (e.g., predictive BDI) where the number of studies was considerably lower.

Second, we had to use mean criterion reliability and range restriction values to correct the observed validity correlations because only a minority of studies reported this information, a problem that is common in virtually all meta-analyses relating to selection. In regard to range restriction, studies with general mental ability often find greater restriction in range with high-complexity jobs than with low-complexity jobs. It is possible that interviews follow the same pattern. If so, then making one global correction could result in some degree of overestimation of corrected validity for lower-complexity jobs and some degree of underestimation for higher-complexity jobs. Similarly, criterion reliability may be higher in concurrent designs than in predictive designs due to a longer acquaintance between the raters and the ratees, resulting in some degree of overestimation in mean-corrected validity for concurrent studies and underestimation in predictive studies.

Combined, the range restriction and criterion reliability corrections may have caused overestimation of results for concurrent, low-complexity studies and underestimation of results for predictive, high-complexity studies. Unfortunately, no data are available at the present time for estimating such differences in range restriction and criterion reliability empirically, and thus we are left in a position to make uniform corrections and acknowledge the limitations thereof.

Lastly, we did not look at other potential moderator variables in addition to job complexity and study design. It is quite possible that other factors also influence the validity of SIs and BDIs, although they may not have as much influence as job complexity and study design.

Not withstanding these limitations, we believe that this investigation makes a valuable contribution to the interview literature. For one thing, we identified two important moderators of situational and behavior description validity: job complexity and study design. For another, we provide stable estimates of the average validity of SIs and BDIs in predicting overall job performance, something that was not available previously. Lastly, we identified several avenues for future research that could further contribute to our understanding of these unique interviewing approaches.

Notes

1. These interviews have been published under various labels. For instance, Pulakos and Schmitt (1995) called their past-oriented questions "experience-based" while Motowidlo, Carter, Dunnette, Tippins, Werner, Burnett, and Vaughan (1992) referred to their past-oriented questions collectively as a "structured behavior interview." These formats are highly similar in that they ask candidates to describe situations from their past relevant to the target position, including their reactions to and the outcomes of those situations.
2. Although OCB is related to evaluation of performance (Podsakoff, MacKenzie, Paine, & Bachrach, 2000), it has generally been viewed as behaviors that are more likely to be discretionary and less likely to be formally or explicitly rewarded in the organization (Organ, 1988). Thus, in a study focusing on overall job performance, OCB represents an at least somewhat deficient criterion because it does not cover all relevant aspects of the criterion-space.
3. Other artifacts such as study-to-study differences in criterion reliability and range restriction no doubt contributed to the observed variance as well. However, Hunter and Schmidt (1990) note that the magnitude of these other variances is typically minor in comparison with sampling error. The example provided by Huffcutt, Arthur, and Bennett (1993) illustrates this tenet (see p. 126). Nonetheless, these other sources do exist and as a result our percent-of-variance accounted values are likely to be underestimates.

References

Barrett, G.V., Phillips, J.S. and Alexander, R.A. (1981) Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, **66**, 1–6.

Bemis, S.E. (1968) Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology*, **52**, 240–244.

Borman, W.C. (1991) Job behavior, performance, and effectiveness. In M.D. Dunnette and L.M. Hough (Eds.), *Handbook of*

industrial and organizational psychology (2nd Edn., Vol. 2, pp. 271–326). Palo Alto, CA: Consulting Psychologists Press.

*Bosshardt, M.J. (1992) *Situational interviews versus behavior description interviews: A comparative validity study*. Unpublished doctoral dissertation, University of Minnesota.

Burke, M.J. and Frederick, J.T. (1984) Two modified procedures for estimating standard deviations in utility analyses. *Journal of Applied Psychology*, **69**, 482–489.

*Campion, M.A., Campion, J.E. and Hudson, J.P., Jr. (1994) Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology*, **79**, 998–1002.

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

Campion, M., Pursell, E. and Brown, B. (1988) Structured interviewing: Raising the psychometric properties of the employment interview. *Personnel Psychology*, **41**, 25–42.

Cortina, J.M., Goldstein, N.B., Payne, S.C., Davison, H.K. and Gilliland, S.W. (2000) The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology*, **53**, 325–351.

*Davis, R. (1986) *Personal communication*.

*Delery, J.E., Wright, P.M., McArthur, K. and Anderson, D.C. (1994) Cognitive ability tests and the situational interview: A test of incremental ability. *International Journal of Selection and Assessment*, **2**, 53–58.

*Flint, D. (1995) *A situational interview for the telecommunications industry*. Paper presented at the annual meeting of the Canadian Psychological Association, Charlottetown, PEI.

Ghiselli, E.E. (1966) The validity of a personnel interview. *Personnel Psychology*, **19**, 389–394.

*Gibb, J. and Taylor, P.J. (in-press) Past experience versus situational employment interview questions in a New Zealand social service agency. *Asia-Pacific Journal of Human Resources*.

*Green, P.C., Alter, P. and Carr, A.F. (1993) Development of standard anchors for scoring generic past-behavior questions in structured interviews. *International Journal of Selection and Assessment*, **1**(4), 203–212.

*Grove, D.A. (1981) A behavioral consistency approach to decision making in employment selection. *Personnel Psychology*, **34**, 55–64.

*Hanson, M., Houston, J., Paullin, C. and Dohm (1992) *Untitled report*. Minneapolis: Personnel Decisions Research Institutes.

*Harel, G., Arditi-Vogel, A. and Janz, T. (n.d.) *Comparing the validity and utility of interview versus assessment center ratings*. Unpublished manuscript. [Under review at Managerial Psychology]

Harris, M.M. (1989) Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personnel Psychology*, **42**, 691–726.

Hays, E.J. (1990) *Relationship of a situational interview to the job performance of convenience store clerks*. Unpublished master's thesis, University of North Texas, Denton, TX.

*Hoffman, C.C. and Holden, L.M. (1993) *Dissecting the interview: An application of generalizability analysis*. Paper presented at the 8th Annual Meeting of the Society for Industrial and Organizational Psychology. San Francisco, CA.

Hough, L.M. (1998) Personality at work: Issues and evidence. In M. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Hillsdale, NJ: Erlbaum Assoc. Inc.

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

Huffcutt, A.I., Arthur, W., Jr. and Bennett, W. (1993) Conducting meta-analysis using the Proc Means procedure in SAS. *Educational and Psychological Measurement*, **53**, 119–131.

Huffcutt, A., Roth, P. and 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., Wiesner, W.H. and DeGroot, T. (2001) Comparison of situational and behavior description interview questions for higher-level positions. *Personnel Psychology*, **54**, 619–644.

Huffcutt, A.I. and 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.

Hunter, J.E. (1980) *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. Washington, DC: US Employment Service.

Hunter, J.E. and Hunter, R.F. (1984) Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96**, 72—98.

Hunter, J.E. and Schmidt, F.L. (1990) *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage Publications.

Hunter, J.E., Schmidt, F.L. and Judiesch, M.K. (1990) Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology*, **75**, 28–42.

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 and G.R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158–168). Newbury Park, CA: Sage.

*Johnson, E.K. (1990) *The structured interview: Manipulating structuring criteria and the effects on validity, reliability, and practicality*. Unpublished doctoral dissertation, Tulane University.

*Kennedy, R.L. (1986) *An investigation of criterion-related validity for the structured interview*. Unpublished master's thesis, East Carolina University, Greenville, NC.

Kleinmann, M. and Deller, J. (1993) Das Situative Interview. In A. Gebert and W. Hacker (Eds.), *1. Deutscher Psychologentag* (pp. 336–343). Bonn, Germany: Deutscher Psychologen Verlag GmbH.

Kleinmann, M. and Klehe, U. (2001, September) *Erfolg im Situativen Interview: Verhalten sich Bewerber so wie von ihnen angegeben?* [Success in the situational interview: Do applicants actually do what they said they would do?]. Paper presented at the 2nd Annual Meeting of the Society of Industrial and Organizational Psychology of the German Psychological Association, Nürnberg, Germany.

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

*Latham, G.P. and Saari, L.M. (1984) Do people do what they say? Further studies of the situational interview. *Journal of Applied Psychology*, **69**, 569–573.

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

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

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

Lin, T.-R. and Adrian, N. (1993, May) *Multi-method multi-dimension structure interviewing: A field study*. Paper presented at the 8th Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

*Little, J.P., Shoenfelt, E.L. and Brown, R.D. (2000) *The situational versus the patterned-behavior-description interview for predicting customer-service performance*. Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans.

McDaniel, M.A., Whetzel, D.L., Schmidt, F.L. and Maurer, S.D. (1994) The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, **79**, 599–616.

*Morgeson, F.P., Reider, M.H. and Campion, M.A. (2002) *Selecting individuals in team settings: Comparing a structured interview, personality test, and teamwork ability tests*. *Manuscript submitted for publication*.

*Moser, K. and Schuler, H. (2001) Validity of behavioral and situational interview questions: A reanalysis of the Schuler *et al*. (1995) data. Unpublished report. Universität Erlangen-Nürnberg and Universität Hohenheim, Germany.

Mosher, M.R. (1991, June) *Development of a behaviorly consistent structured interview*. Paper presented at the 27th International Applied Military Psychology Symposium, Stockholm, Sweden.

Motowidlo, S.J. (1999) Asking about past behavior versus hypothetical behavior. In R.W. Eder and M.M. Harris (Eds.), *The employment interview handbook* (pp. 179–90). Thousand Oaks, CA: Sage.

*Motowidlo, S.J., Carter, G.W., Dunnette, M.D., Tippins, N., Werner, S., Burnett, J.R. and Vaughan, M.J. (1992) Studies of the structured behavioral interview. *Journal of Applied Psychology*, **77**, 571–587.

*Motowidlo, S.J. and Schmit, M.J. (1997) *Performance assessment and interview procedures for store manager and store associate positions*. Gainesville, FL: University of Florida Human Resource Research Center.

Organ, D.W. (1988) *Organizational citizenship behavior*. Lexington, MA: Lexington Books.

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

Osburn, H.G. and Callender, J. (1992) A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology*, **77**, 115–122.

Podsakoff, P.M., MacKenzie, S.B., Paine, J.P. and Bachrach, D.G. (2000) Organizational citizenship behaviors: A critical review of the theoretical and empirical literature and suggestions for future research. *Journal of Management*, **26**, 513–563.

*Pulakos, E.D. and Schmitt, N. (1995) Experience-based and situational interview questions: Studies of validity. *Personnel Psychology*, **48**, 289–308.

*Robertson, I.T., Gratton, L. and Rout, U. (1990) The validity of situational interviews for administrative jobs. *Journal of Organizational Behavior*, **11**, 69–76.

Roth, P.L., Bobko, P., Switzer, F.S. III and Dean, M.A. (2001) Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. *Personnel Psychology*, **54**, 591–617.

Rothstein, H.R. (1990) Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, **75**, 322–327.

Salgado, J.F. (1999) Personnel selection methods. In C.L. Cooper and I.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology*, *12*, 1–53.

Salgado, J.F. and Moscoso, S. (1995) Validez de las entrevistas conductuales estructuradas [Validity of the behavior structured interview]. *Revista de Psicologia del Trabajo y las Organizaciones*, **11**, 9–24.

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

Schmidt, F.L. and Hunter, J.E. (1998) The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, **124**, 262–274.

*Schmit, M.J. and Motowidlo, S.J. (1995) *Development and validation of a situational interview and personality test for selecting sales associates*. Gainesville, FL: University of Florida Human Resource Research Center.

Schmitt, N., Gooding, R.Z., Noe, R.A. and Kirsch, M. (1984) Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, **37**, 407–422.

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

*Schuler, H., Funke, U., Moser, K. and Donat, M. (1995) *Personalauswahl in Forschung und Entwicklung. Eignung und Leistung von Wissenschaftlern und Ingenieuren*. [Personnel selection in research and development. Aptitudes and performance of scientists and engineers]. Göttingen, Germany: Hogrefe.

*Schuler, H., Moser, K., Diemand, A. and Funke, U. (1995) Validität eines Einstellungsinterviews zur Prognose des Ausbildungserfolgs [Validity of an employment interview for the prediction of training success]. *Zeitschrift für Pädagogische Psychologie* [German Journal of Educational Psychology], *9*, 45–54. (Same study as Schuler & Funke 1989.)

Schuler, H. and Prochaska, M. (1990) *Empirische Untersuchung der Leistungsmotivation in Einstellungsinterviews* [Empirical investigation of achievement motivation in selection interviews]. Unpublished data, Universität Hohenheim, Germany.

*Stohr-Gillmore, M.K., Stohr-Gillmore, M.W. and Kistler, N. (1990) Improving selection outcomes with the use of situational interviews: Empirical evidence from a study of correctional officers for new generation jails. *Review of Public Personnel Administration*, **10**(2), 1–18.

Sue-Chan, C., Latham, G.P. and Evans, M.G. (1995) *The construct validity of the situational interview and patterned behavior description interviews: Cognitive ability, tacit knowledge, and self-efficacy as correlates*. Paper presented at the Annual Meeting of the Canadian Psychological Association, Charlottetown, Prince Edwards Island.

*Tarico, V.S., Altmaier, E.M., Smith, W.L., Franken, E.A. and Berbaum, K.S. (1986) Development and validation of an accomplishment interview for radiology residents. *Journal of Medical Education*, **61**, 845–847.

Taylor, P. and 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.

US Department of Labor, Employment and Training Administration (1977) *Dictionary of occupational titles* (4th Edn.). Washington, DC: Government Printing Office.

*US Office of Personnel Management (1987) *The structured interview*. Office of Examination Development, Alternative Examining Procedures Division. Washington, DC.

Vishwesvaran, C., Ones, D.S. and Schmidt, F.L. (1996) Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, **81**, 557–574.

*Weekley, J.A. and Gier, J.A. (1987) Reliability and validity of the situational interview for a sales position. *Journal of Applied Psychology*, **72**, 484–487.

Wiesner, W. and Cronshaw, S. (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.