Journal of Applied Psychology
2004, Vol. 89, No. 3, 553–561

Copyright 2004 by the American Psychological Association
0021-9010/04/$12.00 DOI: 10.1037/0021-9010.89.3.553

A Counterintuitive Hypothesis About Employment Interview Validity and Some Supporting Evidence

Frank L. Schmidt and Ryan D. Zimmerman
University of Iowa

This study found mixed support for the hypothesis that the difference in criterion-related validity between unstructured and structured employment interviews is due solely to the greater reliability of structured interviews. Using data from prior meta-analyses, this hypothesis was tested in 4 data sets by using standard psychometric procedures to remove the effects of measurement error in interview scores from correlations with rated job performance and training performance. In the 1st data set, support was found for this hypothesis. However, in a 2nd data set structured interviews had higher true score correlations with performance ratings, and in 2 other data sets unstructured interviews had higher true score correlations. We also found that averaging across 3 to 4 independent unstructured interviews provides the same level of validity for predicting job performance as a structured interview administered by a single interviewer. Practical and theoretical implications are discussed.

Employment interviews are part of the hiring process for virtually all jobs. Only reviews of resumes and application blanks are used more frequently in selection (Ash, 1982; Dipboye, 1994, 1997; Dipboye & Jackson, 1999). As a result, industrial/organizational psychologists have devoted much research to the validity of employment interviews.

It is now well established that structured employment interviews have higher criterion-related validity for predicting job performance than do unstructured interviews. This finding has been consistent across three major meta-analyses (Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Wiesner & Cronshaw, 1988). There are several possible factors that could contribute to higher validity for structured interviews, and it is not yet clear what the contribution of each is. First, it could be hypothesized that structure leads to the identification and specification of more valid constructs. Structured interviews are typically based on systematic information about job requirements, and this information is used to create the questions making up the structured interview. Second, it might be argued that structure may lead to more valid measurement of the specified constructs. Different interviewers (and different interviews) include the same questions, and these questions are recorded and scored in the same manner (Campion, Pursell, & Brown, 1988), thus enhancing construct measurement over and above the effect on validity of the increased reliability expected from increasing structure. (However, one could argue, as a reviewer did, that this effect would be due solely to the resulting increase in reliability.) Third, structure appears to lead to more reliable measures. There is considerable evidence that structured interviews are more reliable than are unstructured interviews (Conway, Jako, & Goodman, 1995; McDaniel et al., 1994), and it is well known that, other things being equal, higher reliability leads to higher validity and lower reliability leads to lower validity (Nunnally & Bernstein, 1994).

From the point of view of ordinary plausibility, it appears that all of these factors contribute to the higher validity of structured interviews for predicting job performance. However, according to the principle of parsimony (Occum's razor), simpler scientific theories and explanations should be preferred over more complex ones. In particular, if any one of these three factors alone can be shown to account for the higher validity of structured interviews for predicting job performance, then the principle of parsimony would indicate that the other factors are not needed in the explanation. In this connection, the obvious factor to examine initially is the reliability difference. We know that, other things being equal, more reliable predictors are more valid, and this fact raises the possibility that when two measures differ in validity, this difference is due to the difference in reliability. Given the appropriate validity and reliability estimates, this hypothesis can be tested in a straightforward manner by using the principles and methods of measurement theory (e.g., Nunnally & Bernstein, 1994; Schmidt & Hunter, 1996). If this hypothesis is not fully supported, the findings can nevertheless be expected to be useful, because they would document the degree to which the validity difference is accounted for by reliability differences and would reveal the size of the validity difference controlling for the effects of reliability differences. This adjusted difference then estimates the size of the validity difference produced by the remaining factors and therefore provides a platform for research on those factors. This is consistent with recommendations in the literature. For example, Huffcutt, Conway, Roth, and Stone (2001) have called for research to estimate the relative contributions of reliability and content to the validity of structured and unstructured interviews.

Regardless of whether the measurement-error hypothesis is supported, the reliability and hence the validity of both unstructured and structured interviews can be increased by increasing the number of (independent) interviews each candidate undergoes. Many employers at present have applicants interview with multiple interviewers, with the resulting interview evaluations being pooled

Frank L. Schmidt and Ryan D. Zimmerman, Department of Management and Organizations, Tippie College of Business, University of Iowa.

Correspondence concerning this article should be addressed to Frank L. Schmidt, Tippie College of Business, University of Iowa, Iowa City, IA 52242. E-mail: frank-schmidt@uiowa.edu

or averaged. Standard measurement procedures (described below) can be used to determine interview reliability and validity for any specified number of interviews, allowing one to determine, for example, the number of unstructured interviews needed to produce validity equal to that of a single structured interview. These procedures are the same as those used to increase the reliability, and hence validity, of a test by lengthening the test.

In summary, the first purpose of this article was to investigate whether the higher reliability of structured interviews can account for their higher criterion-related validity for the prediction of job performance. A second purpose of this article was to determine the number of unstructured interviews that can be expected to yield validity equal to that of a single structured interview. To our knowledge this question has not been examined in the literature.

Determining the Reliability of Interviews

Because a number of different sources of measurement error must be calibrated, determining the reliability of interviews is not necessarily simple or easy. The appropriate estimate of the reliability of interview scores from a single interviewer is the correlation between two different interviewers who interview the same sample of individuals on two different occasions. This procedure controls for all sources of measurement error in the interview scores: random response error on the part of the interviewees, transient error in the responses of the interviewees, and conspect error (disagreement between the raters in how they score or weight the interviewees' responses; Conway et al., 1995; DeShon, 1998; Schmidt & Hunter, 1996, Scenarios 5, 12, 19, and 20; Schmidt, Le, & Ilies, 2003). As noted by McDaniel et al. (1994), many of the interview reliability coefficients in the literature are computed by correlating the interview scores produced by two interviewers present at the same interview (such as in a panel interview). These values are inflated for two reasons. First, each interviewer is observing the same random response errors of the interviews—because they are observing the same responses. Hence interviewee random response error is not fully controlled (i.e., random response error is correlated across interviewers and inflates the correlation between them). Second, there is no control for transient error in interviewees' responses. Because of the transient effects of moods or feelings on particular occasions, interviewees may respond somewhat differently to the same questions on different days. These variations are not part of what the interview is intended to measure and so must be considered measurement error (Becker, 2000; Conway et al., 1995; DeShon, 1998; Schmidt & Hunter, 1999; Schmidt et al., 2003). (Note that there is no need to include a control for transient error on the part of the interviewers because transient error is by definition uncorrelated across interviewers and hence does not inflate reliability estimates.) In fact, the only kind of measurement error that such reliability estimates fully control for is conspect error (scorer disagreement); this is why such estimates are sometimes called "conspect reliabilities" (Schmidt & Hunter, 1996).

Comparisons show the extent to which such reliability estimates are inflated. For structured interviews, McDaniel et al. (1994) found that the average conspect reliability was .84, whereas Conway et al. (1995) reported a mean value of .66 when reliability was properly estimated as described here. For unstructured interviews, mean conspect reliability was .68 (McDaniel et al., 1994), and the mean of appropriate reliability estimates was .37. In connection with the present study, the reliability estimates of the sort presented by Conway et al. provide the needed reliability values.

Estimating the Validity of Interviews

In testing the hypothesis described above, accurate estimates of interview validity are equally important. Almost all interview validity studies focusing on job performance use supervisory ratings of (overall) job performance as the criterion. These ratings are typically produced by one supervisor and are the sum of ratings across multiple dimensions of job performance. It is well known that such ratings have limited reliability (Rothstein, 1990; Viswesvaran, Ones, & Schmidt, 1996), but meta-analyses correct for the biasing effect of unreliability in the ratings (and for range restriction). An equally important consideration with respect to job performance ratings is the distinction between the use of administrative and research ratings. Administrative ratings are part of the regular performance appraisal process in the organization; they are used administratively for salary determination, promotion decisions, and other purposes and are typically obtained from the employees' personnel files when the validity study is conducted. Research ratings are gathered specifically and solely for purposes of the validity study, and the rating supervisors know that such ratings will not be used in any personnel decisions. There is considerable evidence that administrative ratings are influenced by factors extraneous to judgments of actual job performance, and it has generally been found that validity estimates are smaller (and hence downwardly biased) when administrative (vs. research) ratings are used (Wherry & Bartlett, 1982). Other studies have shown that administrative ratings suffer from more leniency and exhibit more halo than do ratings collected solely for research purposes (Sharon & Bartlett, 1969; Taylor & Wherry, 1951; Veres, Feild, & Boyles, 1983; Warmke & Billings, 1979). A recent meta-analysis (Jawahar & Williams, 1997) has shown that the leniency difference is quite large (.40 to .50 standard deviations).

The only meta-analysis that has reported interview validities separately for job performance ratings gathered solely for research purposes is by McDaniel et al. (1994). Hence these validity estimates assume particular importance here. However, as discussed later, we also examined meta-analysis results based solely on administrative ratings. Although this data set produces less accurate estimations of criterion-related validity, it could be argued that the effect of reliability on validity should be proportionally the same in such a data set. As indicated earlier, our central focus was on validity for predicting job performance; however, at the request of a reviewer we also examined available meta-analytic data on the validity of interviews for predicting performance in training programs. Interviews are less frequently used for this purpose, but one meta-analysis (McDaniel et al., 1994) did report validities for training data.

To summarize, the primary purpose of this article was to test the following hypothesis:

> *Hypothesis 1:* The superior validity of structured interviews in comparison with unstructured interviews for predicting job performance can be accounted for by the differences in the reliability of the two types of interviews.

The second purpose of this article was to determine the number of unstructured interviews that can be expected to yield validity equal to that of a single structured interview. This question, like our hypothesis about interview validity, has not been examined in the literature.

Method

Validity Data

The Wiesner and Cronshaw (1988) meta-analysis of interview validity did not report validity results separately by the criterion types of job performance and training performance; in that study these two types of data were combined. Because validities may be different for these two types of criteria, they should typically be analyzed separately. The Marchese and Muchinsky (1993) meta-analysis did not separate results for structured interviews from unstructured interviews. The Schmidt and Rader (1999) meta-analysis reported no validity estimates for unstructured interviews. In addition, the structured interview they studied was very different from traditional structured interviews. We therefore did not examine validity estimates from these three meta-analyses. McDaniel et al. (1994) used standard criteria to classify interviews as either unstructured or structured. As did Schmidt and Hunter (1998), we emphasized validity estimates for each interview type on the basis of job performance ratings used for research purposes only, for the reasons described earlier. On the basis of data provided by M. A. McDaniel (personal communication, August 12, 2002), the interview studies in the research-only category represented a broad range of job complexity levels. According to the Dictionary of Occupational Titles classification of the jobs examined in each study, 35% of the jobs in the structured-interview, research-only category were in the top third of job complexity for jobs in the U.S. economy, and 32% were in the bottom third. For unstructured interviews, 46% of the jobs were in the top third of job complexity for all jobs in the U.S. economy, with 54% in the bottom third. For this criterion measure, McDaniel et al. reported mean corrected validities of .38 for unstructured interviews and .51 for structured interviews. We also included the validity data from McDaniel et al. for administrative-only ratings. As stated earlier, we also included data for the criterion of training performance. The reported mean-corrected validities of the administrative-only ratings were .41 for unstructured interviews and .37 for structured interviews. The validities for training performance were .36 for unstructured interviews and .34 for structured interviews.

The meta-analysis by Huffcutt and Arthur (1994) did not distinguish between studies based on administrative ratings and studies using ratings gathered only for research purposes, thus their validity estimates may be less accurate. More important, that meta-analysis was based only on validity studies for entry-level jobs, making the focus somewhat narrower, given that interviews are used in hiring for jobs at all levels. Despite these limitations, these data should be useful in testing our hypothesis, because measurement error would be expected to affect validities in a similar manner in these data sets.

Huffcutt and Arthur (1994) classified interviews into four levels of structure (Campion, Palmer, & Campion, 1997; van der Zee, Bakker, & Bakker, 2002), but even they admitted that although their classification scheme is an improvement over traditional dichotomization of interview structure, their method is still problematic. Without a standardized method of classifying interview structure into multiple levels, comparisons can be hampered by the fact that different authors classify the same study into different structure "levels." For the purposes of this study and for ease of comparison, apparently similar levels of structure are combined as described below.

Reliability Data

Only one meta-analysis (Conway et al., 1995) has been reported that presents unbiased estimates of interview reliability as described earlier (i.e., reliability estimates that control for interviewee random response error and transient error). These reliability estimates were obtained by correlating the interview scores assigned by two different interviewers who interviewed the same group of individuals on 2 different days. In their meta-analysis, Conway et al. classified interviews into five levels of structure, from least structured (Level 1) to most structured (Level 5). The five mean reliability estimates reported, from least to most structured, were .37, .61, .56, .66, and .59.

Data Analysis

Standard corrections for measurement error were used to estimate true score correlations between interview scores and job performance ratings. Because the true validity estimates reported in McDaniel et al. (1994) and Huffcutt and Arthur (1994) have already been corrected for criterion unreliability and range restriction, it was necessary to correct only for predictor (interview) unreliability to estimate true score correlations. True score correlations indicate what the mean validity would be if both the unstructured and structured interviews were perfectly reliable, hence partialing out the reliability difference. To the extent that true score correlations are the same for structured and unstructured interviews, the hypothesis that the higher reliability of structured interviews accounts for their higher criterion-related validity for job performance is supported. If the hypothesis is not fully supported, this analysis indicates the degree to which the validity difference is due to the reliability difference.

In an attempt to reveal whether findings were robust, the McDaniel et al. (1994) validity values were corrected for measurement error with three different analyses. In the Conway et al. (1995) reliability data, Structure Levels 2 and 3 appeared to be intermediate levels of structure that were neither structured nor unstructured. Structure Level 4 appeared to correspond most closely with the usual definition of a structured interview, and we matched this value to the McDaniel et al. structured classification in the first analysis. Structure Level 5 appeared to be very similar to Level 4; as a result, we also looked at the results obtained when the average reliability for Structure Levels 4 and 5 was used in the measurement-error correction for the McDaniel et al. validity for the structured interview (Alternative Analysis 1). A second alternative analysis was also carried out. On the basis of several aspects of interview structure, Conway et al. reported a mean reliability of .34 for low-structure interviews and a mean reliability of .67 for high-structure interviews. These reliability values were also used to correct the unstructured and structured interview validities, respectively, from McDaniel et al. (Alternative Analysis 2).

The structure level for validity in Huffcutt and Arthur's (1994) study and the structure levels for reliability in Conway et al.'s (1995) study were matched as follows. Structure Levels 3 and 4 from Huffcutt and Arthur were very similar and yielded similar mean validities (.56 and .57). We averaged the validities for these two structure levels to produce three levels of structure. (The results were the same if one dropped Level 4 as redundant with Level 5.) Structure Level 1 from Huffcutt and Arthur corresponded to Structure Level 1 from Conway et al.; both matched the usual definition of an unstructured interview. Structure Level 2 from Huffcutt and Arthur was closest to Structure Level 3 from Conway et al., and Structure Level 3 from Huffcutt and Arthur corresponded to Structure Level 4 from Conway et al. An alternative analysis was conducted in which the reliability used to correct the validity at this level was the average of the Conway et al. Levels 4 and 5 reliabilities, and the results were similar. A second alternate analysis was conducted in which the four structure levels for validity in Huffcutt and Arthur's study were matched to the first four structure levels of Conway et al.'s study. As will be seen later, these variations in analyses did not affect the conclusions drawn.

Regardless of the reliability of either type of interview, interview reliability can be increased by increasing the number of interviewers. That is, interview reliability can be increased by having multiple interviewers interview each applicant and then averaging the interview scores across interviewers. (Of course, as noted earlier, the different interviewers must interview the candidates on different occasions. This rules out panel interviews.) This is the same process that is used to increase the reliability of job performance ratings by increasing the number of raters (Guilford, 1954) or to increase the reliability (and hence validity) of a scale or measure by increasing the number of items. As noted earlier, the reliability of interview evaluations produced by a single interviewer is the average correlation between pairs of independent interviewers who interview the same individuals on different occasions (as reported by Conway et al., 1995). This value can be entered into the Spearman–Brown formula to compute the reliability of interview ratings based on any number of independent interviewers. Increases in reliability resulting from increases in the number of interviewers result in increases in validity, other things being equal. The simplest way to compute the validity obtained with any number of interviewers is to multiply the square root of that reliability times the true score correlation between interview scores and job performance. That is, one need only attenuate the validity that would exist with perfect reliability to the value that one would observe with a specific level of reliability, that is,

$$\hat{\rho}_{x_n y_\infty} = \sqrt{r_{xx_n}}\,\hat{\rho}_{x_\infty y_\infty}, \qquad (1)$$

where $\hat{\rho}_{x_n y_\infty}$ is the operational true validity given n interviewers, r_{xx_n} is the reliability obtained from n interviewers, and $\hat{\rho}_{x_\infty y_\infty}$ is the true score correlation between interview scores and job performance.

Using these methods, we computed the validity of both unstructured and structured interviews for different numbers of independent interviewers, using the McDaniel et al. (1994) data based on research ratings only. The resulting figures were used to answer the following question: How many independent unstructured interviews must be conducted and averaged to produce validity equal to that of a structured interview administered by one interviewer? That is, is it feasible to equate operational validities of unstructured and structured interviews simply by administering more unstructured interviews to each applicant? This question can have a positive answer even if measurement-error differences fail to explain the validity differences between structured and unstructured interviews based on one interviewer. As noted earlier, the second purpose of this article was to determine whether the use of a small number of independent unstructured interviews yields validity comparable to that obtained for a single administration of a structured interview. However, equal validity does not imply that the sum of several unstructured interviews measures the same combination of constructs as those measured by a single structured interview. This point is developed further in the Discussion section.

Results

Tables 1–3 show the analysis based on the McDaniel et al. (1994) validity data. The data in Table 1 provide perhaps the best test of the measurement-error hypothesis because not only are the reliabilities used of the appropriate type, but the validities are also the most appropriate, being based on job performance ratings gathered solely for research purposes. The second column in Table 1 shows the mean validity estimates from McDaniel et al. The third column shows the corresponding reliability estimates from Conway et al. (1995), and the fourth column shows the true score correlations. As can be seen, the true score correlations were identical to two decimal places, indicating that reliability differences fully accounted for the validity difference.

Table 1
Test of the Measurement-Error Hypothesis Using Research-Only Ratings Validity Data From McDaniel et al. (1994)

Level of structure	Validity ($\hat{\rho}_{xy_\infty}$)	Interview reliability (r_{xx})	True score correlation ($\hat{\rho}_{x_\infty y_\infty}$)
Unstructured	.38	.37	.63
Structured	.51	.66	.63

Note. Criteria are supervisory ratings of overall job performance made for research purposes only. Reliabilities are from Conway et al. (1995); see text for details. There are 36 studies (N = 3,069) of structured interviews and 9 studies (N = 531) of unstructured interviews.

Table 2
Test of the Measurement-Error Hypothesis Using Administrative-Only Ratings Validity Data From McDaniel et al. (1994)

Level of structure	Validity ($\hat{\rho}_{xy_\infty}$)	Interview reliability (r_{xx})	True score correlation ($\hat{\rho}_{x_\infty y_\infty}$)
Unstructured	.41	.37	.67
Structured	.37	.66	.45

Note. Criteria are supervisory ratings of overall job performance made for administrative purposes only. Reliabilities are from Conway et al. (1995); see text for details. There are 50 studies (N = 8,155) of structured interviews and 24 studies (N = 4,259) of unstructured interviews.

The first alternative analysis (not shown in Table 1) yielded essentially the same conclusion. When instead of using the .66 reliability estimate for Structure Level 5 of Conway et al. (1995), we used the average of the reliabilities of Structure Levels 5 and 6 as the reliability of structured-interview scores, we obtained a reliability of .63 and a true score correlation of .64, essentially identical to the .63 shown in Table 1.

The second alternative analysis (again, not shown in Table 1) of the M. A. McDaniel et al. (1994) validities in Table 1 also yielded essentially the same results. On the basis of several aspects of interview structure, Conway et al. (1995) computed the mean reliabilities of low and high structured interviews as .34 and .67, respectively. When these reliability figures were used in Table 1 to eliminate the effects of measurement error (instead of the .37 and .66 shown), the resulting true score correlations were .65 for unstructured interviews and .62 for structured interviews, leading to the same conclusion as did the other two analyses.

Table 2 is analogous to Table 1 except that it presents data for administrative-only ratings. These data were anomalous in that unstructured interviews appeared to have higher operational validity (.41) than did structured interviews (.37). After correction for interview reliability, this difference became even larger. Possible reasons for these results are explored in the Discussion section. Although technically the results do not support our main hypothesis, in light of the findings we could restate the hypothesis as follows: After measurement error in the predictor is corrected for, unstructured interviews will have true score correlations with the criteria at least as large as those of structured interviews.

Table 3 shows estimated true score correlations for both types of interviews with training performance criteria. As with the results obtained from the administrative-only data set, this data set appeared to be anomalous in that initial mean validities were higher for unstructured than structured interviews. It was not clear at the time of the McDaniel et al. (1994) study, nor is it clear now, why structured and unstructured interviews should have essentially equal validity (.34 vs. .36) for training criteria, given the superior

Table 3
Test of the Measurement-Error Hypothesis Using Validity Data for Training Performance From McDaniel et al. (1994)

Level of structure	Validity ($\hat{\rho}_{xy_\infty}$)	Interview reliability (r_{xx})	True score correlation ($\hat{\rho}_{x_\infty y_\infty}$)
Unstructured	.36	.37	.59
Structured	.34	.66	.42

Note. Criteria are measures of performance in training programs. Reliabilities are from Conway et al. (1995); see text for details. There are 25 studies (N = 3,576) of structured interviews and 30 studies (N = 47,576) of unstructured interviews.

validity of structured interviews for job performance criteria. However, some contributing factors are discussed later in this article.

The results obtained using interview validity estimates from Huffcutt and Arthur (1994) are shown in Table 4. The top part of Table 4 shows the first analysis described in the Method section, the analysis using three levels of structure. In this analysis, corrections for the effects of reliability differences did not equalize correlations with job performance. Even at the true score level correlations were larger for more structured interviews. This picture did not change when instead of using the .66 reliability for the highest level of structure in the top part of Table 4, we used the average of the reliabilities of the two highest levels of structure from Conway et al. (1995). Doing this produced a reliability estimate of .63 and a true score correlation of .71—not materially different from the .69 shown in Table 4.

The bottom part of Table 4 shows the results when four levels of structure were used. Again, the last column showing the true score correlations indicates that reliability differences did not appear to account for the validity differences: Except for the highest level of structure (Level 4), true score correlations increased in size with increasing levels of interview structure. Again, this picture did not change when in the last entry of the third column we substituted the mean reliability across Structure Levels 5 and 6 of Conway et al. (1995). Doing this yielded a reliability of .63 (vs. the .66 shown) and a true score correlation of .72 (vs. the .70 shown). This value was not appreciably lower than the .75 for Structure Level 3.

Table 4
Test of the Measurement-Error Hypothesis Using Validity Data From Huffcutt and Arthur (1994)

Level of structure	Validity ($\hat{\rho}_{xy_\infty}$)	Interview reliability (r_{xx})	True score correlation ($\hat{\rho}_{x_\infty y_\infty}$)
Using three levels of structure			
1	.20	.37	.33
2	.35	.56	.47
3	.56	.66	.69
Using four levels of structure			
1	.20	.37	.33
2	.35	.61	.45
3	.56	.56	.75
4	.57	.66	.70

Note. Validities are for entry-level jobs only, and the job performance ratings used as criteria include both administrative and research ratings. Reliabilities are from Conway et al. (1995); see text for details. Structure Level 1: 15 studies (N = 7,308); Structure Level 2: 39 studies (N = 4,621); Structure Level 3: 27 studies (N = 4,358); Structure Level 4: 33 studies (N = 2,365).

Hence data from Huffcutt and Arthur (1994) did not support the hypothesis that the superior validity of more structured interviews can be explained by their higher reliabilities. Because Huffcutt and Arthur did not separate research-only and administrative performance ratings, an argument could be made that to provide a direct comparison with the McDaniel et al. (1994) data, one would need to use McDaniel et al.'s data set combining both types of performance ratings. Two notes are warranted before considering this data set. First, because this combined data set is not independent of the results presented in Tables 1 and 2, we do not present these data in a separate table. Second, despite being a combination of the research-only and administrative performance ratings, a single large study (study n = 4,105 out of total N= 8,895) not included in either of the separate analyses is added to the data set for unstructured interviews, because the performance ratings could not be clearly classified into either research only or administrative in nature. However, with these caveats, McDaniel et al.'s combined job performance data set is considered for comparative purposes. For this combined job performance criteria category, the operational validities were .33 for unstructured interviews and .44 for structured interviews. After interview unreliability was corrected for, the true score validity was the same (.54) for both types of interviews. This true score correlation value was attenuated from the .63 in Table 1 to .54 because of inclusion of the administrative job performance ratings.

Table 5 shows the reliability and validity of unstructured and structured interviews as a function of the number of independent interviews used. These calculations were based on the data in Table 1. Training performance data were not examined given that in the analysis shown in Table 3, unstructured interviews had operational validities at least as large as those of structured interviews. As expected, validity increased with the number of inter-

Table 5
Mean Reliability and Validity of Unstructured and Structured Interviews for Varying Numbers of Interviewer-Occasion Combinations Based on Data From Table 1

	Unstructured interview		Structured interview	
No. of interviewers[a]	Validity ($\hat{\rho}_{xy_\infty}$)	Reliability[b] (r_{xx})	Validity ($\hat{\rho}_{xy_\infty}$)	Reliability[b] (r_{xx})
1	.38	.37	.51	.66
2	.46	.54	.56	.79
3	.50	.64	.58	.85
4	.53	.70	.59	.88
5	.55	.75	.60	.91
6	.56	.78	.60	.92
7	.56	.80	.61	.93
8	.57	.82	.61	.94

[a] Each interviewer interviews each candidate separately, with the final interview score for each candidate being the sum (or average) of the scores given by each interviewer. [b] Computed using the Spearman–Brown formula (Nunnally & Bernstein, 1994).

views for both structured and unstructured interviews. However, even with eight interviews the validity of the unstructured interview was still not equal to that of the structured interview (.57 vs. .61). Nevertheless, the validity difference did become considerably smaller as the number of interviews increased. With only one interview, the validity of the structured interview was 34% larger than that of the unstructured interview. With eight interviews (arguably near the maximum any employer would normally be willing to use), the validity of the structured interview was only 7% larger.

More important, it can also be seen in Table 5 that the validity of the unstructured interview based on three to four interviews by different interviewers was the same as that of the structured interview administered by one interviewer. By increasing the number of interviews (and interviewers), one can raise the validity of the unstructured interview to that of the structured interview. The implications of this are discussed in the next section.

Discussion

Findings from this study do not appear to be fully adequate to answer definitively the question of whether reliability differences account for the mean validity difference between unstructured and structured interviews, but they are sufficient to demonstrate the importance of this hypothesis and the need for further research on it. Despite (and perhaps because of) the surprising and counterintuitive nature of the findings from the McDaniel et al. (1994) data set in Table 1, it is important from a scientific perspective to give these results appropriate consideration. Research suggests that even scientists unfortunately tend to seek out and give credibility to those findings that support their preexisting beliefs while showing a strong tendency to reject findings that are inconsistent with their prior beliefs (Lord, Ross, & Lepper, 1979; Mahoney, 1977). Studies by Mahoney and associates found that many scientists have a strong preference for confirmatory studies and often do not recognize the scientific value of disconfirmatory findings (Mahoney & DeMonbreun, 1977; Mahoney & Kimper, 1976). Mahoney and DeMonbreun (1977) found that scientists rarely used disconfirmatory logic. Weimer (1977) argued that the logic of science should be more falsificational and that disconfirmatory evidence is more informative than confirmatory results. Perhaps the best interpretation of the mixed results in this study is that they indicate that there is evidence that is countertheoretical to the prevailing beliefs about the structured interview and that additional research should be conducted to understand this issue more fully.

The results of this study are mixed. The findings from what we consider to be the most appropriate data set, shown in Table 1, support the hypothesis that the higher reliability of structured interviews explains the superior validity of structured interviews in comparison with unstructured interviews. The data in Table 1 are based on a broad spectrum of jobs and on job performance ratings obtained solely for research purposes. In addition, the data in Table 2 (based on administrative performance ratings) and Table 3 (based on training performance criteria) suggest that after measurement error is controlled for the validity of unstructured interviews is at least as high as that of structured interviews (but see discussion below). The data in these three tables are inconsistent with the data from Huffcutt and Arthur (1994), shown in Table 4. The Huffcutt and Arthur data do not support the measurement error hypothesis: Correlations with job performance ratings remain higher for structured interviews even after correction for interview reliability. The most direct comparison is between the McDaniel et al. (1994) combined performance rating data (as discussed in the Results section) and the Huffcutt and Arthur data in Table 5, because both data sets are based on combined administrative and research-only job performance ratings. Both data sets are based on large sample sizes; the total sample size for the McDaniel et al. combined performance-rating data set is 20,696, and the total sample size for Table 4 is 16,287. The McDaniel et al. data is based on a broad spectrum of jobs, where the Huffcutt and Arthur data include only entry-level jobs. This is a limitation for the Huffcutt and Arthur data, but it is not clear why this limitation would cause results to be different from those for the wider spectrum of jobs represented in the McDaniel et al. data set. We can conclude only that our findings are mixed: There is some support for the measurement-error hypothesis, but there was a substantial set of data in which it was not supported.

The data in Table 2 (the data based solely on administrative ratings of job performance) are anomalous, in that the higher reported mean operational validity for unstructured interviews (.41 vs. .37 for structured interviews) is contrary to the accepted and strongly supported conclusion that structured interviews have higher validity for predicting job performance. It is possible that this result is due to second-order sampling error (Hunter & Schmidt, 2004). In any event, given the anomalous reversal of the initial validities, it is clear before the correction for interview reliability that the true score correlations will be larger for unstructured than structured interviews. This is also the case for the results for training criteria reported in Table 3. Also, as noted earlier, the central focus of this article was on job performance; the analysis of training criteria was included at the suggestion of reviewers. (On the basis of the relative numbers of cumulative validity studies, one would conclude that interviews are less frequently used to predict training performance.) McDaniel et al. (1994) commented on the anomalous validity results in both Tables 2 and 3 but could find no explanation for them other than second-order sampling error. Because these two data sets are anomalous, we conclude, as indicated above, that the critically important results are those in Tables 1 and 4, and that these tables, taken together, present a mixed picture of support for the measurement-error hypothesis. However, in light of the counterintuitive nature of our hypothesis, a finding of mixed support (along with the results from Tables 2 and 3 showing higher true score correlations for unstructured interviews) must be regarded as remarkable and surprising. Our findings indicate that this question deserves further research, something that would otherwise not have been suspected.

If the measurement-error hypothesis is supported in future research, the implications will be important. The procedures used in creating structured interviews are designed not only to increase reliability but also to better assess the determinants of job success. The rationale is that job analysis can be used to identify job requirements, and on the basis of this information, questions and scoring methods can be created that better assess these required knowledges, skills, and abilities (KSAs) than are possible with unstructured interviews (see, e.g., Campion et al., 1988, and Dipboye & Gaugler, 1993). Although unstructured interviewers typically have some familiarity with job requirements, unstructured interviews, being less systematic and less standardized, are assumed to assess required KSAs less well and are assumed to omit at least some important KSAs altogether. It is clear that this conceptualization, that is, the rationale for the structured interview, implies that the validity difference between structured and unstructured interviews is not due solely to the difference in reliability but is due in significant part to better procedures for identifying job requirements and better procedures for measuring those requirements. The findings in this study suggest the possibility that these latter two factors may in fact not play a role. This possibility deserves further research.

Issues in This Research

Certain questions raised during the review process for this study are important for the light they shed on relevant conceptual issues. The first such issue is whether a finding of equal true score correlations implies that the constructs measured by structured and unstructured interviews must be the same. It was stated during the review process that if this were the case, then the findings would be implausible, because it seems highly likely that structured interviews measure constructs different from those assessed by unstructured interviews. First, it is important to remember that interviews measure a combination of different constructs—very likely a fairly large number of different constructs. It is possible for different combinations of multiple constructs to have the same true score correlation with job performance ratings. That is, the fact that the two construct combinations (mixtures) have the same true score correlation with job performance ratings does not necessarily imply that their true score correlation with each other is 1.00. Equal true score correlations with job performance criteria do not imply that exactly the same combination of constructs is being measured.

Second, even if the two types of interviews measure different constructs, structured and unstructured interviews are likely to correlate nearly 1.00 at the true score level. This conclusion follows from certain simple statistical principles. Consider an example in which there are 10 different constructs, all of which intercorrelate .50 with each other. Now suppose structured interviews measure only Constructs 1 through 5, whereas unstructured interviews measure only Constructs 6 through 10, so that the two types of interviews measure no constructs at all in common. Yet a simple computation (the correlation of sums) reveals that the correlation between the two interviews is .99. Because they correlate .99, if their reliabilities are the same (or if we look at true score correlations with job performance), their correlations with other variables are likely to be identical or very similar.

Structured and unstructured interviews probably do not measure exactly the same combination of constructs. However, because many constructs are measured—probably more than the 5 in our example—and because correlations among these constructs are typically positive, we have a statistical model for how true score criterion-related correlations can be essentially equal despite the fact that different packages of constructs are being measured by the two types of interviews.

Of course one could ask why the correlations among constructs would be positive. For example, the correlation between cognitive ability and emotional stability may be very low or zero. The key here is that we are dealing not with constructs as theoretically defined but with constructs as assessed by the interviewers. Interviewers, like other evaluators and raters, are affected by halo—which means that the constructs as assessed will be positively correlated. This implies that different packages or combinations of constructs will be highly correlated, as in our example.

A related issue is the question of potential confounding of constructs measured with the effects of measurement error. That is, if we do not first ensure that structured and unstructured interviews are measuring the same constructs before we look at whether controlling for measurement error will equalize their true score correlations with job performance, are we potentially confounding these two effects? Actually, there is no problem of confounding here. The question this study asks is as follows: When one adjusts for the effects of measurement error, are the mean true score correlations of structured and unstructured interviews equal? As shown above, it is not necessary in answering this question that the (combination of) constructs measured by the two types of interviews be identical. That is, this question can be asked and answered regardless of whether the two kinds of interviews measure the same constructs or different constructs. By way of analogy, we can ask whether a brick and a piece of iron have the same weight. The fact that they are made of different substances does not imply that there is any confounding in our answer to this question.

Finally, there is some evidence that the two types of interviews measure similar combinations of constructs. Although Huffcutt et al. (2001) found differences at the microlevel in the constructs measured by structured and unstructured interviews, they found that at the next level of aggregation the two types of interviews measured similar packages of constructs. For example, mental ability was assessed by 19.2% of low-structure interviews and 14.4% of high-structure interviews. Personality was assessed by 36.9% of high-structure interviews and 33.7% of low-structure interviews. Other KSAs were assessed by 11.5% of high-structure interviews and 8.7% of low-structure interviews. These findings suggest a high correlation between the summed scores for the two sets of construct combinations, just as in the case of our model above.

Another issue concerns the standard deviations associated with the estimates of mean true validity used in this study. As one can see in McDaniel et al. (1994) and Huffcutt and Arthur (1994), these standard deviations were neither all zero nor all equal to each other. The question is whether this is relevant to the present study. Analyses of the sort presented in this study are directed to the explanation of differences between means (mean true validities here), and variability within categories is not relevant to the analysis. (for a similar study see Schmidt, Hunter, & Outerbridge, 1986; see also Viswesvaran & Ones, 1995). The research question in this study is as follows: Can measurement error explain the difference in mean true validities between structured and unstructured interviews? Any residual variability that may exist in validities within the types of interview is not a focus of interest. There are a variety of possible potential sources of (nonartifactual) variability of true validities within each interview type, but these are not the focus of interest. In the language of analysis of variance, our interest in this study is between-groups variance, not within-group variance. There is nothing in our analysis that requires the assumption that the validity of either type of interview is constant across all settings, usages, or situations. (The results obtained by McDaniel et al. [1994] and Huffcutt and Arthur [1994] indicate that validity of interviews is generalizable, but this is not the same as concluding that it is completely invariant [Hunter & Schmidt, 2004].)

Increasing the Number of Interviewers

Independent of the issues discussed previously, it will still be the case that the validity of both unstructured and structured interviews can be increased by increasing their respective reliabilities. The results obtained by Conway et al. (1995) indicate that in the case of structured interviews, there may be a limit to how much reliability can be increased by increasing structure. This study found that the highest level of structure (their Structure Level 5) showed slightly lower reliability than the next lower level of structure (.59 vs. .66). This finding of diminishing reliability returns to increasing structure is understandable in light of the fact that increasing structure cannot eliminate interviewee random response error or interviewee transient measurement error. By definition, increased structure cannot be used to increase the reliability of unstructured interviews because the imposition of all but minimal structure causes them to cease being unstructured interviews. Hence the obvious way to increase reliability, and thus validity (ceteris paribus), for unstructured interviews is by increasing the number of interviews given to each applicant, just as one increases the reliability of job performance ratings by increasing the number of raters. The data in Table 5 indicate that administering three to four unstructured interviews and using their average or total score can be expected to yield validity equal to that of a structured interview administered by a single interviewer. Analyses based on validity data other than those used in Table 5 might indicate that this would require more (or less) than three to four unstructured interviews, but there would always be some number of unstructured interviews that would achieve this equality of validities.

This fact has potentially important practical implications. Applicants react more positively to interviewers who are warm, accepting, and socially responsive, and unstructured interviews allow the expression of these qualities better than do structured interviews (Dipboye, 1992 1997). Latham and Finnegan (1993) and Schuler (1993) reported evidence that applicants prefer unstructured interviews to structured interviews. Applicants appear to believe that unstructured interviews provide them with more freedom to describe their strengths and achievements (as they see them), whereas structured interviews restrain what they can discuss in the interview. As a result, they view structured interviews as less fair. Because many (perhaps most) employers have applicants interviewed by multiple interviewers, the use of multiple unstructured interviews could in principle simultaneously allow for high validity and applicant satisfaction with the interview process if interview scores are recorded and combined appropriately across interviews. Interviewee satisfaction with the interview process could be important under tight labor market conditions (Dipboye, 1997; Kohn & Dipboye, 1998). In addition to applicant preferences, there is some evidence that managers who conduct job interviews prefer unstructured interviews to structured interviews (van der Zee et al., 2002).

Future Research

The hypothesis that reliability differences account for the validity differences between more and less structured interviews received some support in this research, although results were mixed. Given the counterintuitive nature of this hypothesis, and given that it is in conflict with traditional assumptions, this hypothesis merits further research in light of the partial support it received in this research. Perhaps the most profitable approach would be to attempt to obtain additional estimates of interview validity for unstructured and structured interviews on the basis of both job performance ratings created solely for research purposes and objective measures of job performance such as work output and/or sales figures. At present the only study reporting interview validities based on such objective job performance measures is Schmidt and Rader (1999).[1] Unfortunately the interview used in that research, although structured in its own way, is not similar to traditional structured interviews. In addition, Schmidt and Rader reported no validities for unstructured interviews for any criterion type. We recommend that new data be gathered that can be used to test the measurement error hypothesis with objective criterion measures of job performance.

[1] In this connection, it is significant that structured interviews did predict objective measures of job performance (e.g., sales). This finding argues against the hypothesis that interview validity in general is caused by correlated "method variance," that is, the hypothesis that positive validities indicate only that subjective impressions of interviewers are correlated with subjective impressions of supervisor ratings of job performance.

References

Ash, R. A. (1982). Comparison of four approaches to the evaluation of job applicant training and work experience. *Dissertation Abstracts International, 42*(11-B), 4606.

Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods, 5,* 370–379.

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

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.

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.

DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods, 3,* 412–423.

Dipboye, R. L. (1992). *Selection interviews: Process perspectives.* Cincinnati, OH: South-Western.

Dipboye, R. L. (1994). Structured and unstructured selection interviews: Beyond the job-fit model. In G. Ferris (Ed.), *Research in personnel and human resources management* (Vol 12, pp. 79–123). Greenwich, CT: JAI Press.

Dipboye, R. L. (1997). Structured interviews: Why do they work? Why are they underutilized? In N. Anderson & P. Herriott (Eds.), *International handbook of selection and assessment* (pp. 455–473). New York: Wiley.

Dipboye, R. L., & Gaugler, B. B. (1993). Cognitive and behavioral processes in the selection interview. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 135–170). San Francisco: Jossey-Bass.

Dipboye, R. L., & Jackson, S. L. (1999). Interviewer experience and expertise effects. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 229–292). Thousand Oaks, CA: Sage.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

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

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

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance purpose effect. *Personnel Psychology, 50,* 905–925.

Kohn, L. S., & Dipboye, R. L. (1998). The effects of interview structure on recruiting outcomes. *Journal of Applied Social Psychology, 28,* 821–843.

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

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1,* 161–175.

Mahoney, M. J., & DeMonbreun, B. G. (1977). *Confirmatory bias in scientists and non-scientists*. Unpublished manuscript.

Mahoney, M. J., & Kimper, T. P. (1976). From ethics to logic: A survey of scientists. In M. J. Mahoney (Ed.), *Scientist as subject* (pp. 187–193). Cambridge, MA: Ballinger.

Marchese, M. C., & Muchinsky, P. M. (1993). The validity of the employment interview: A meta-analysis. *International Journal of Selection and Assessment, 1,* 18–26.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

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.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.

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

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27,* 183–198.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71,* 432–439.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8,* 206–234.

Schmidt, F. L., & Rader, M. (1999). Exploring the boundary conditions for interview validity: Meta-analytic validity findings for a new interview type. *Personnel Psychology, 52,* 445–464.

Schuler, H. (1993). Is there a dilemma between validity and acceptance in the employment interview? In B. Nevo & R. S. Jaeger (Eds.), *Educational and psychological testing: The test taker's outlook* (pp. 239–250). Kirkland, WA: Hogrefe & Huber.

Sharon, A. T., & Bartlett, C. J. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. *Personnel Psychology, 22,* 251–263.

Taylor, E. K., & Wherry, R. J. (1951). A study of leniency in two rating systems. *Personnel Psychology, 4,* 39–47.

United States Department of Labor, Employment and Training Division. (1991). *Dictionary of occupational titles* (4th ed., Vol I & II). Washington, DC: United States Department of Labor.

van der Zee, K. I., Bakker, A. B., & Bakker, P. (2002). Why are structured interviews so rarely used in personnel selection? *Journal of Applied Psychology, 87,* 176–184.

Veres, J. G., Feild, H. S., & Boyles, W. R. (1983). Administrative versus research performance ratings: An empirical test of rating data quality. *Public Personnel Management, 12,* 290–298.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology, 48,* 865–885.

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.

Warmke, D. L., & Billings, R. S. (1979). Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. *Journal of Applied Psychology, 64,* 124–131.

Weimer, W. B. (1977). *Psychology and the conceptual foundations of science*. Hillsdale, NJ: Erlbaum.

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521–551.

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

Received March 4, 2002
Revision received June 13, 2003
Accepted June 23, 2003 ■