Further analysis of employment interview validity: a quantitative evaluation of interviewer-related structuring methods

ALLEN I. HUFFCUTT[1]* AND DAVID J. WOEHR[2]

[1]*Department of Psychology, Bradley University, Peoria, Illinois 61625, U.S.A.*
[2]*Department of Psychology, Texas A&M University, College Station, Texas 77845, U.S.A.*

Summary   The purpose of this investigation was to analyze the relationship between four interviewer-related factors and the validity of the employment interview. Using a regression approach to meta-analysis, an analysis of 120 interview studies with a total sample size of 18,158 suggested that: (1) training should be provided to interviewers regardless of whether the interview itself (i.e. the questions and rating scales) is structured; (2) the same interviewer should be used across all applicants, especially when the interview itself is not highly structured; and (3) using a panel of interviewers does not contribute to validity, and may actually have a detrimental effect. Results for the fourth interviewer factor, taking notes during the interview, were inconclusive. Limitations and directions for future research are noted. Copyright © 1999 John Wiley & Sons, Ltd.

Introduction

Much of the recent meta-analytic research regarding the employment interview has focused on the structure of the interview itself. For example, Huffcutt and Arthur (1994) found that using preestablished questions and more detailed and sophisticated scales to evaluate responses were associated with higher interview validity. McDaniel, Whetzel, Schmidt and Maurer (1994) found that interviews that used a specific type of job-related questions (i.e. situational) achieved higher overall validity. Relatedly, Wiesner and Cronshaw (1988) found that interviews were more valid when questions were based on information from a job analysis.

Relatively little meta-analytic research has focused on the interviewers themselves. For example, it is possible that training interviewers could improve interview validity (Dipboye and Gaugler, 1993; Dougherty, Ebert and Callender, 1986). Such training might establish a more systematic framework, thereby reducing differences among interviewers and increasing consistency across applicants. Moreover, training might improve the skills of the interviewers in

capturing and using information, regardless of whether they use a systematic framework. But, as Dipboye (1992) noted, while training has considerable potential, its overall effect on outcome measures such as validity has not been clearly established.

Similarly, using the same interviewer or panel of interviewers across all applicants could improve validity. Research suggests that interviewers can vary widely in their approach, use of information, and rating style (Dougherty et al., 1986; Zedeck, Tziner and Middlestadt, 1983). Having different applicants interviewed by different interviewers makes it more challenging to form accurate employment decisions and may lead to underestimation of the true validity of the interview (Dreher, Ash and Hancock, 1988). Standardizing interviewers should establish a more consistent testing framework, thereby increasing the accuracy of the interview process.

Lastly, given limitations in human information processing capabilities, having interviewers takes note during the interview may help them to make better decisions (Macan and Dipboye, 1986; Schuh, 1980). For example, the capacity of short-term memory seems to be about seven pieces of information (see Solso, 1991), suggesting that interviewers, no matter how much material is covered, are likely to base their evaluations on a limited amount of information. Taking notes during the interview should mitigate limitations in human memory and result in greater and more accurate recall of information.

The only interviewer characteristic which has been analyzed meta-analytically is the use of a panel of interviewers versus individual interviewers, and results have been inconclusive. For low-structure interviews, Wiesner and Cronshaw (1988) found that panel interviews had considerably higher validity than interviews conducted individually while McDaniel et al., 1994) found no appreciable difference. For high-structure interviews, McDaniel et al. found that panel interviews actually had *lower* overall validity while Wiesner and Cronshaw found no real difference. (Both studies defined overall structure somewhat generally in terms of questions and rating scales).

Conceptually, there is reason to believe that using a panel of interviewers could enhance validity. For example, from a psychometric perspective, combining multiple ratings may increase interview reliability much like adding additional items tends to increase the reliability of psychological tests (Ghiselli, Campbell and Zedeck, 1981). Another possible benefit is that knowing other interviewers are involved could increase feelings of accountability, which can increase rating accuracy (Mero and Motowidlo, 1995).

The purpose of this investigation was to empirically analyze the relationship between four interviewer-related factors (i.e. training, standardization, note-taking, and use of a panel format) and the validity of the employment interview. We first assessed the individual relationship between each factor and interview validity, to determine whether implementing any one of the factors by itself would improve validity. Many organizations find themselves in a position where they would like to do something to improve their interviews, but are unsure as to what to do. Establishing that these methods are effective would provide organizations with viable options, ones which have solid empirical support. Such information would be particularly useful for organizations which choose not to structure the interview itself, either because they lack (or are unwilling to invest) the resources to develop a structured format or because they prefer to keep the interview more open (e.g. so that interviewers can 'sell' candidates on their company).

Then, we assessed whether each factor contributed to interview validity over and above structuring of the interview itself (i.e. question and response evaluation standardization). For organizations that do use a structured interview format such as a situational interview (Latham, Saari, Pursell and Campion, 1980) or a behavior description interview (Janz, 1982), it becomes important to assess whether interviewer enhancements have incremental validity over and above interview structure.

Method

Location of studies/decision rules

Through an extensive search we compiled a data set of 123 interview validity studies, both published and unpublished. These studies all involved on-the-job performance criteria and represented a wide range of organizations, applicants, and job types. A listing of the sources for these studies can be obtained from the first author.

Coding of interview characteristics

The uncorrected validity coefficient and sample size were recorded as listed in the studies. In cases where multiple validity coefficients were reported for the same study, we employed three main decision rules to determine which coefficient to retain. Some studies reported correlations between various interview dimensions and job performance. In these cases we selected the correlation that represented the overall interview evaluation, if one was provided. If one was not available then we averaged the correlations for the individual interview dimensions. Similarly, some studies reported correlations between interview ratings and various aspects of performance. In these cases we selected the correlation that represented overall job performance if one was available; otherwise, we averaged the correlations for the individual aspects. Finally, a few studies reported correlations between interview ratings and more than one performance evaluation, either from different supervisors or different rating instruments. Here, we averaged the resulting correlations.

The four moderator variables were all coded dichotomously. Specifically, studies were coded as: (0) if training was not provided or (1) if interviewers were trained; (0) if different interviewers interviewed different applicants or (1) if the same interviewer or interview panel interviewed all applicants; (0) if the study made no mention of information being formally recorded from the interview or (1) if information was formally recorded; and finally, (0) if interview ratings were formed from one interviewer or (1) if interview ratings were formed from multiple interviewers.

With regard to interviewer training, we did not differentiate among the types of training provided. While differentiating among types of rater training has been shown to be important in performance appraisal (Woehr and Huffcutt, 1994), there simply was not enough information provided in our interview studies to make clear distinctions. Thus, our coding reflected the general effectiveness of providing interviewers with some type of training as part of the interview development process.

Given that one purpose of this investigation was to assess the incremental validity of interviewer factors, it was also necessary to record the level of interview structure for each study. Structure was coded using the framework developed by Huffcutt and Arthur (1994), who identified four overall levels of interview structure based on various combinations of standardization with respect to questions and response evaluation. Level 1 included no constraints on the questions and general summary rating(s). Level 2 was characterized by the use of some formal structure on the questions, and/or use of more differentiated rating scales to evaluate responses. Level 3 represented a higher level of structure on both questions and response evaluation, but still involved variability in the interview process from applicant to applicant. Finally, level 4 involved asking all applicants essentially the same questions with no follow-up or probing, and scoring each response individually.

Two things should be noted with respect to the coding process. First, lower levels of the characteristics were coded as lower numbers, higher levels were coded as higher numbers (e.g. no training = 0; training = 1). Such a scheme simplifies interpretation of results because positive effects from a factor (e.g. training increases validity) and then come out positive statistically (e.g. positive regression coefficients).

Second, a conservative strategy was utilized for coding. That is, unless information was provided in a study to warrant coding of a higher level, the lowest level of each moderator variable was assumed (e.g. no training). Such a strategy has been used effectively in other interview analyses (Conway, Jako and Goodman, 1995) and is particularly advantageous when using regression methodology. Most of the studies provided sufficient information to determine the specific level of each characteristic, and therefore conservative coding was necessary only in a limited number of cases.

In order to verify the accuracy of the coding process, 30 studies were independently coded. Half of these studies were coded by the second author while the other half were coded by another researcher with no vested interest in the study. The correlations between codings by the first author and codings by a another rater were in the 0.8 to 0.9 range for all variables.

Preliminary analyses

Prior to examining the relationship between interview validity and the various interview factors, we conducted three preliminary analyses. First, we assessed whether the variables had sufficient variability to allow meaningful analyses. Results are presented in Table 1. As indicated, the variables in general appeared to have adequate representation at all their various levels.

Second, we analyzed the data set for possible outliers. Given the size of the data set and the diversity of studies, a few anomalous coefficients could be expected over and above the normal variation caused by sampling error. Possible causes for such coefficients include unusually high but unreported levels of an artifact such as range restriction or criterion unreliability, transcription or computation error, or some highly unusual characteristic of the study or the study subjects (Huffcutt and Arthur, 1995). We computed a DFFITS statistic, one of a class of regression outlier procedures known as influence statistics (Freund and Littell, 1991), for each study. We then compared these values to the critical value of 0.57 computed using the formula provided by Belsley, Kuh and Welsch (1980). Three studies greatly exceeded the critical DFFITS, with values of 0.93, 0.93, and 0.89. These studies were removed, leaving 120 studies in the final data set.[1]

Third, we assessed the level of intercorrelation (i.e. multicollinearity) among the study variables. A high level of intercorrelation among variables can make it difficult to isolate the unique effects of each variable (Cohen and Cohen, 1983; Freund and Littell, 1991). To assess multicollinearity, we formed a matrix of simple correlations among the study variables. In addition, we computed the variance inflation factor (VIF) for each variable (Freund and Littell, 1991), which shows the combined overlap between one predictor and all other predictors. As Myers (1990) noted, VIF values exceeding 10 may be cause for serious concern. Results are

---

[1] We ran the analyses both with and without the outliers and the results were very similar. For example, the correlation between training and validity changed by only 0.03, while the correlation between interviewer standardization changed by only 0.01.

Table 1. Preliminary assessment of the moderator variables: analysis of variability

Moderator variable	Number	Percentage
Interviewer training		
No	71	57.7
Yes	52	42.3
Interviewer standardization		
No	100	81.3
Yes	23	18.7
Note-taking		
No	55	44.7
Yes	68	55.3
Interviewer format		
Individual	70	56.9
Panel	53	43.1
Level of structure		
Level 1	12	9.8
Level 2	42	34.1
Level 3	33	26.8
Level 4	36	29.3

Table 2. Preliminary assessment of the moderator variables: analysis of multicollinearity

Moderator variable	T	IS	NT	IF	LS	VIF
Interviewer training (T)	1.00	0.23	0.34	−0.06	0.43	1.26
Interviewer standardization (IS)		1.00	0.33	−0.26	0.44	1.45
Note-taking (NT)			1.00	−0.07	0.68	2.02
Interviewer format (IF)				1.00	0.17	1.31
Level of structure (LS)					1.00	2.72

The two-tailed probability of chance was less than 0.0001 for all correlations 0.4 and above, less than 0.001 for correlations in the 0.3 range, and less than 0.01 for correlations in the 0.2 range. The two-tailed probability for the correlation of 0.17 was 0.06, while the probability for the correlation of −0.07 was 0.45. VIF is the variance inflation factor, and represents the overall overlap between one independent variable and all of the other independent variables. All correlations are based on a sample size of 120.

presented in Table 2. As indicated, there was minimal overlap among the interviewer factors. The highest correlation was 0.34, between training and note-taking. Thus, there appeared to be relatively little overlap among the interviewer factors.

However, there was more overlap between the interviewer factors and the level of interview structure. Note-taking and interview structure correlated 0.68, indicating a fairly strong tendency for researchers who structured the questions and rating scales to also build systematic note-taking procedures into the process (e.g. the situational interview; see Latham et al., 1980). This level of intercorrelation made it somewhat difficult to isolate the unique effect that note-taking has on validity in our investigation. Training and standardization both correlated in the 0.4 range with interview structure, indicating a minor tendency for researchers who structured the interview to also train and standardize the interviewers. While these variables may also have carried some

structure information, the magnitude of the effect should have been considerably less. Use of a panel of interviewers appeared to be more independent of the level of interview structure.

Meta-analytic methodology

The traditional approach to meta-analysis is to compute sample-weighted averages and variances for each analysis. For example, we could have computed the mean validity coefficient for the studies where interviewers were not trained and compared it to the mean validity coefficient for the studies where interviewers were trained. Hunter and Schmidt (1980) described an alternate approach to meta-analysis, one based on regression/correlation methodology. In this approach, the validity coefficients are regressed directly on moderator variables such as training and standardization. The main advantage of the regression approach is that a number of moderator variables can be evaluated simultaneously, including interaction effects. While not as common, several studies have employed some form of regression methodology to summarize a body of research (e.g. Marchese and Muchinsky, 1993; Schwab, Olian-Gottlieb and Heneman, 1979). Given that we had five moderator variables in our investigation (i.e. the four interviewer factors and structure) and possible interactions among these variables, we decided to use the regression approach.

Similar to the sample-weighted approach, a correction is made to the results of a regression meta-analysis to account for the influence of various statistical artifacts. In particular, the validity coefficients are affected by sampling error and study-to-study differences in the levels of range restriction and criterion unreliability. These artifacts create instability in the validity coefficients, artificially reducing the magnitude of the relationships with the moderator variables. Such instability is analogous to the effects of measurement error on a more traditional dependent measure such as performance ratings. Using the methodology outlined by Hunter and Schmidt (1990), we estimated that the reliability of validity coefficients was 0.64. Therefore, the correction factor for interview validity coefficients was 0.80 (i.e. the square root of the reliability). Details on the computation of this value are provided in Appendix 1.

Accordingly, we ran all of our analyses twice. We first ran the analyses without the correction for reliability of the validity coefficients. Then, we ran the analyses with the correction for reliability. Here, we formed various correlation matrices (as described in the next section), corrected these matrices by dividing all correlations involving the validity coefficients by 0.80, and input the results into a multiple regression program developed by Hunter (1992) specifically to handle corrected correlation matrices.

Analysis of each interviewer factor by itself

For the uncorrected analyses we computed the simple correlation between each interviewer factor and validity. For the corrected analyses we formed two-variable intercorrelation matrices between each factor and validity, divided the correlation between the factor and validity by 0.80, and input the results in the multiple regression program described above.

For the sake of completeness, we computed the simple correlation between interview structure and validity as well. This correlation provided information relevant to the next set of analyses, where we assessed the incremental validity of interviewer factors. In addition, it provided verification of the findings of previous research that interview structure does make a strong contribution to interview validity.

Analysis of the incremental validity of each interviewer factor

For the uncorrected analyses, we formed a regression model in SAS (SAS Institute, 1990) for each interviewer factor where the validity coefficients were regressed on both interview structure and that factor. Including the appropriate option statement in the model allowed us to obtain the semi-partial correlation for each factor. Semi-partial correlations indicate the unique contribution of a variable over and above the other variable(s) in the model (Cohen and Cohen, 1983) and represent a direct measure of incremental validity.

For the corrected analyses, we formed three-variable intercorrelation matrices which included the validity coefficients, interview structure, and an interviewer factor. These matrices were corrected and input into the multiple regression program. The semi-partial correlation for each interviewer factor was computed from the output.

Results

Simple correlations between each of the interviewer factors and interview validity are presented in Table 3. As shown, three of the four interviewer factors had a positive and significant relationship with interview validity. Training had the strongest relationship among the interviewer factors, with a corrected correlation of 0.41, while note-taking and standardization had corrected correlations in the 0.3 range.[2] Interviewer format had a very small and negative correlation with validity, and there was roughly a 68 per cent probability that this relationship was due to chance. Interview structure had a corrected correlation of 0.63 with validity.

Semi-partial correlations for the interviewer factors are shown in Table 4. As indicated, training appeared to make a modest contribution to validity over and above the effects of

Table 3. Simple correlations between the study factors and interview validity

Study factor	r	r_c	Prob
Interviewer training	0.33	0.41	0.0001
Interviewer standardization	0.25	0.31	0.0033
Note-taking	0.29	0.36	0.0007
Interviewer format	−0.04	−0.05	(0.6790)
Level of structure	0.50	0.63	<0.0001

r is the simple correlation between that factor and interview validity, r_c is the simple correlation corrected for sampling error in the validity coefficients, and Prob is the one-tailed probability that the uncorrected simple correlation occurred by chance. The probability for interviewer format (as shown in parentheses) was left as a two-tailed probability since the relationship was not in the predicted direction. All correlations are based on a sample size of 120.

[2] For each analysis we report the one-tailed probability that the predicted relationship could have occurred by chance. We obtained these probabilities from the uncorrected analyses, although it makes relatively little difference since the probabilities for the corrected values are very similar in most cases. The reason the two probabilities are similar is that while the correction increases the magnitude of the correlation coefficient thereby increasing the numerator of the test statistic, it also increases the standard error thereby increasing the denominator of the test statistic. For example, with the correlation between training and validity, the value of the test statistic for the uncorrected analysis differed by only 0.01 from that for the corrected analysis.

Table 4. Semi-partial correlations between the interviewer factors and interview validity holding level of interview structure constant

Interviewer factor	sr	sr_c	Prob
Interviewer training	0.13	0.16	0.0515
Interviewer standardization	0.03	0.04	0.3421
Note-taking	−0.07	−0.09	(0.3734)
Interviewer format	−0.12	−0.16	(0.1179)

sr is the uncorrected semi-partial correlation between that factor and interview validity, sr_c is the semi-partial correlation corrected for sampling error in the validity coefficients, and Prob is the one-tailed probability that the uncorrected simple correlation occurred by chance. The probabilities for note-taking and interviewer format (as shown in parentheses) were left as a two-tailed probability since their relationships were not in the predicted direction. All correlations are based on a sample size of 120.

interview structure. The direction of the semi-partial correlation was positive, indicating that training improved validity.

The semi-partial correlation for interviewer standardization, while also positive, was minimal in size. In retrospect, this was not surprising given that structure and standardization have a similar effect on the interview process, namely an increase in consistency. Conway *et al.* (1995), for example, found that increasing interview structure was associated with higher interrater reliability. Accordingly, we did an additional analysis to assess whether standardizing interviewers was more effective at lower levels of interview structure than at higher levels. Specifically, we formed the interaction term between standardization and structure, and then assessed its contribution over and above these two main variables (see Cohen and Cohen, 1983; Pedhazur, 1982). The corrected semi-partial correlation for the interaction was −0.16 ($p = 0.05$, one-tailed), suggesting that standardization is somewhat more effective at lower levels of interview structure. (Because interactions tend to correlate highly with their main terms, especially when one of the variables is coded dichotomously, our estimate of the semi-partial correlation for this interaction may be conservative).

The semi-partial correlation for note-taking was small and negative in direction. The negative correlation could have been due to chance, or it could have been due to the intercorrelation between note-taking and the other study variables, particularly interview structure. Myers (1990) noted that incorrect signs are a common outcome when multicollinearity is present.

The semi-partial correlation for interviewer format was also negative. The negative correlation could again have been due to chance, or it is possible that the negative direction could indicate a detrimental effect from using a panel of interviewers. It is interesting to note that had a detrimental effect been predicted, the one-tailed probability of chance would have been 0.0590. And, unlike note-taking, results of the preliminary analyses would suggest that the negative correlation for interviewer format could not be attributed to multicollinearity.

Discussion

Results of this investigation suggest that interviewer factors can contribute to the validity of the employment interview. For organizations that choose not to structure their interviews, either because of preference or because of resources, there now appears to be several interviewer-related

options that they can still use to improve their interview process. For example, providing interviewers with training appears to improve validity. One area for future research could be to isolate the specific mechanism by which training improves validity, whether it is by establishing a more common frame of reference among interviewers, by improving skills, or simply by inducing some variation of the Hawthorne effect (see Muchinsky, 1993). Another avenue could be to investigate the relative effectiveness of different types of training.

Using the same interviewer (or panel of interviewers) across all applicants also appears to improve validity, and this option does not necessarily involve any additional resources. Several authors have noted the considerable differences which can exist among interviewers (Dougherty *et al.*, 1986; Zedeck *et al.*, 1983), and our results suggest that these differences do adversely affect validity. It is important to point out that using the same interviewers does not imply that these interviewers are any more accurate that other interviewers. Rather, the increase in validity probably reflects reduction or removal of the psychometric 'dampening' which occurs when collapsing across different interviewers (see Dreher *et al.*, 1988).

For organizations that do structure their interviewers (i.e. they standardize their questions and rating processes), training interviewers still appears to provide some benefit in terms of validity. It can help familiarize them with the structured interview process, which often is fairly complicated, and can ensure that they learn to use it correctly right from the beginning. Standardizing interviewers does not appear to be as necessary when the interview is structured, probably because consistency (i.e. reliability) has already increased to the point where further improvements would be unlikely. Bosshardt (1992), for example, reported interrater reliabilities of 0.79 and 0.87, respectively, for the behavior description and situational portions of his interview.

The one interviewer factor for which we were not able to do a thorough assessment was note-taking. We found a fairly high correlation between note-taking and interview structure, which made it more difficult to isolate the unique effect of having interviewers systematically take notes. Such intercorrelation is not surprising, since many of the high structure techniques have note-taking procedures built into their process. Thus, the effectiveness of taking notes is an issue left for exploration in future research. Given real limitations in human information processing, taking notes has the potential to improve both the quantity and the quality of the information utilized in the evaluation of applicants and should be thoroughly investigated.

Undoubtedly the most curious implication of our findings is that use of a panel of interviewers does not contribute to the validity of the employment interview, and may actually have a detrimental effect. We consistently found a negative correlation between use of an interview panel and validity, although admittedly the magnitude of the correlation was not always large and there is always the possibility that it was a chance finding. It is possible that facing a group of interviewers is intimidating for some applicants, and may interfere with their responses to the questions. Clearly, this is one issue that should be addressed in future research, especially since the use of an interview panel can be advantageous for other reasons (see Campion and Arvey, 1989).

Moreover, our findings for use of an interview panel contrast with those of Wiesner and Cronshaw (1988), who found significantly higher validity for unstructured interviews when conducted as a panel. Our own follow-up investigation suggested that the discrepancy may be attributable to differences in the data sets. Specifically, Wiesner and Cronshaw included several 'extended interview' studies in their data set, an elaborate process in which the interview is combined with a number of assessment center exercises (e.g. Vernon, 1950; Wilson, 1948). These studies typically were of low structure, used a panel format, and reported unusually high validity. Later meta-analyses like McDaniel *et al.* (1994) excluded such studies on the grounds that they were not representative of a typical employment interview. And, McDaniel *et al.* actually did find

a detrimental effect for structured interviews, which our findings do support, although they found virtually no difference for unstructured interviews.

On a more philosophical note, an interesting question is whether these four interviewer enhancements should be considered as alternate ways to 'structure' an employment interview. There has been some confusion in the literature regarding what constitutes structure, although standardizing the questions and using more sophisticated procedures for evaluating responses appear to be widely accepted as such. Based on a new model of interview structure proposed by Campion, Palmer and Campion (1997), which identifies 15 different ways to improve the interview, these four methods would be considered as structure since they are included in their model. Thus, we consider training, standardizing interviewers, taking notes, and using an interview panel as interviewer-related structuring methods.

Limitations of this investigation should be noted. Meta-analysis as a technique is correlational in nature and drawing causal inferences from the findings may not always be appropriate. Also, since findings represent average trends across a diversity of interview conditions, the outcome for any one individual situation may vary from these trends. In addition, there was some collinearity among the study variables, which is not unexpected when dealing with field rather than experimental studies. As noted above, the main consequence of collinearity in this investigation was the inability to get an accurate assessment of the unique contribution of note-taking.

Nonetheless, we feel that these results do make a significant contribution to the interview literature. At a conceptual level, they identify a separate class of methods for improving the employment interview, namely those methods relating to the interviewers themselves. At a more pragmatic level, they provide direction in terms of how to design an interview to achieve maximum validity.

Lastly, we feel that our investigation makes a contribution to the advance of meta-analytic methodology. As noted earlier, a handful of other studies have used the regression approach to meta-analysis. But, these studies did not correct for artifacts and they did not analyze any interactions. Meta-analyses appear to be growing increasingly complex as more and more moderator variables are being considered, and the regression approach should gain in popularity since it can accommodate more comprehensive and more sophisticated analyses. We hope that our efforts help to establish the regression form of meta-analysis as a viable approach to summarizing research.

References

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*, John Wiley & Sons, New York.

Bosshardt, M. J. (1992). 'Situational interviews versus behavior description interviews: A comparative validity study'. Doctoral dissertation, University of Minnesota, *Dissertation Abstracts International*, **54**, 3882A.

Campion, J. E. and Arvey, R. D. (1989). 'Unfair discrimination in the employment interview', in Eder, R. W. and Ferris, G. R. (Eds) *The Employment Interview: Theory, Research, and Practice*, Sage, Newbury Park, CA, pp. 61–73.

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

Cohen, J. C. and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd edn., Lawrence Erlbaum, Hillsdale, NJ.

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

Dipboye, R. L. (1992). *Selection Interviews: Process Perspectives*, South-Western Publishing, Cincinnati, OH.

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

Dougherty, T. W., Ebert, R. J. and Callender, J. C. (1986). 'Policy capturing in the employment interview', *Journal of Applied Psychology*, **71**, 9–15.

Dreher, G. F., Ash, R. A. and Hancock, P. (1988). 'The role of the traditional research design in underestimating the validity of the employment interview', *Personnel Psychology*, **41**, 315–327.

Freund, R. J. and Littell, R. C. (1991). *SAS System for Regression*, 2nd edn., SAS Institute, Cary, NC.

Ghiselli, E. E., Campbell, J. P. and Zedeck, S. (1981). *Measurement Theory for the Behavioral Sciences'*, New York: W. H. Freeman.

Huffcutt, A. I. (1992). 'An empirical investigation of the relationship between multidimensional degree of structure and the validity of the employment interview', Unpublished doctoral dissertation, Texas A&M University, College Station, Texas.

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. and Arthur, W. Jr. (1995). 'Development of a new outlier statistic for meta-analytic data', *Journal of Applied Psychology*, **80**, 327–334.

Hunter, J. E. (1992). *Multiple Regression Program*, Michigan State University, East Lansing, MI.

Hunter, J. E. and Schmidt, F. L. (1990). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Sage Publications, Newbury Park.

Janz, T. (1982). 'Initial comparisons of patterned behavior description interviews versus unstructured interviews', *Journal of Applied Psychology*, **67**, 577–580.

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

Macan, T. M. and Dipboye, R. L. (April, 1986). 'Biases in interviewers' processing of information in the employment interview'. Paper presented at Southwestern Psychological Association Convention, Fort Worth, Texas.

Marchese, M. C. and 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. and Maurer, S. (1994). 'The validity of employment interviews: A comprehensive review and meta-analysis', *Journal of Applied Psychology*, **79**, 599–616.

Mero, N. P. and Motowidlo, S. J. (1995). 'Effects of rater accountability on the accuracy and the favorability of performance ratings', *Journal of Applied Psychology*, **80**, 517–524.

Muchinsky, P. M. (1993). *Psychology Applied to Work*, 4th edn, Brooks/Cole, Pacific Grove, CA.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*, 2nd edn, PWS and Kent Publishing, Boston, MA.

Pedhazur, E. J. (1982). *Multiple Regression in Behavioral Science Research*, 2nd edn, Holt, Rinehart, and Winston, New York.

SAS Institute, Inc. (1990). *SAS Language: Reference, Version 6*, 1st edn, Author, Cary, NC.

Schuh, A. J. (1980). 'Effects of early interruption and note taking on listening accuracy and decision making in the interview', *Bulletin of the Psychonomic Society*, **13**, 263–264.

Schwab, D. P., Olian-Gottlieb, J. D. and Heneman, H. G. III (1979). 'Between subject's expectancy theory research: A statistical review of studies predicting effort and performance', *Psychological Bulletin*, **86**, 139–147.

Solso, R. L. (1991). *Cognitive Psychology*, 3rd edn, Allyn and Bacon, Boston, MA.

Vernon, P. E. (1950). 'The validation of civil service selection board procedures', *Occupational Psychology*, **24**, 75–95.

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.

Wilson, N. A. B. (1948). 'The work of the civil service selection board', *Occupational Psychology*, **22**, 204–212.

Woehr, D. and Huffcutt, A. (1994). 'Rater training for performance appraisal: A quantitative review', *Journal of Occupational and Organizational Psychology*, **67**, 189–205.

Zedeck, S., Tziner, A. and Middlestadt, S. E. (1983). 'Interviewing validity and reliability: An individual analysis approach', *Personnel Psychology*, **35**, 355–370.

Appendix 1: Correcting the Moderator/
Validity Coefficient Correlations for Artifacts

A correction factor was computed using the methodology described by Hunter and Schmidt (1990, p. 116). Specifically, the reliability of the validity coefficients was computed by dividing the population variance of the validity coefficients by the observed variance. The population variance was estimated by subtracting the variances due to sampling error and study-to-study differences in other artifacts from the observed variance. In this investigation, the observed variance was 0.023882, sampling error variance was 0.005874, range restriction variance was 0.002075, and criterion unreliability variance was 0.000547. Range restriction data were actually taken from Huffcutt and Arthur (1994) while criterion unreliability data came from Huffcutt (1992), both of which should be highly representative of the current data set. Subtracting the artifact variances resulted in a population variance of 0.015386 and a reliability estimate of 0.64. The correction factor is the square root of the reliability of the validity coefficients, which is 0.80. All correlations involving the validity coefficients were divided by this correction factor.