Journal of Applied Psychology
1994, Vol. 79, No. 2, 184-190

0021-9010/94/$3.00

Hunter and Hunter (1984) Revisited: Interview Validity for Entry-Level Jobs

Allen I. Huffcutt and Winfred Arthur, Jr.

The present investigation provides a reanalysis of the employment interview for entry-level jobs that overcomes several limitations of J. E. Hunter and R. F. Hunter's (1984) article. Using a relatively sophisticated multidimensional framework for classifying level of structure, the authors obtained results from a meta-analysis of 114 entry-level interview validity coefficients suggesting that (a) structure is a major moderator of interview validity; (b) interviews, particularly when structured, can reach levels of validity that are comparable to those of mental ability tests; and (c) although validity does increase through much of the range of structure, there is a point at which additional structure yields essentially no incremental validity. Thus, results suggested a ceiling effect for structure. Limitations and directions for future research are discussed.

Hunter and Hunter's (1984) meta-analysis comparing 11 alternative predictors of job performance for entry-level positions continues to be widely cited in both research literature (e.g., Rudner, 1992; Terpstra & Rozell, 1993) and personnel textbooks (e.g., Aamodt, 1991; Cascio, 1991; Muchinsky, 1993) as evidence of the superiority of mental ability tests over other predictors, such as the interview. Specifically, Hunter and Hunter found ability tests to be the best overall predictor of job performance (as measured by supervisory rating criteria), with a mean validity of .53. In contrast, they found much lower overall validities for the other predictors, including a mean validity of only .14 for the interview (see Hunter & Hunter, 1984, Table 9, p. 90).

However, certain methodological problems with Hunter and Hunter's (1984) analyses diminish the ability to make a meaningful comparison between ability tests and the interview. First, although all predictors were corrected for sampling error and criteria unreliability, only ability tests were corrected for range restriction (Roth & Campion, 1992). Hunter and Hunter acknowledged that restriction in range was a potential problem for the interview (see p. 79), and thus their mean validity is likely to be an underestimate. Second, only 10 interview validity coefficients were analyzed, a limited sample at best in comparison with the 425 coefficients analyzed for ability tests. Such a small number of study coefficients is problematic because the obtained results may not be representative of the larger population of validity coefficients, a condition known as second-order sampling error (Hunter & Schmidt, 1990). Finally, as we discuss later, more recent research has suggested that major methodological differences in the format of interviews, particularly the level of structure, moderate validity (e.g., Wiesner & Cronshaw, 1988). With ability tests, the measurement process (and constructs measured) varies only slightly across different tests (McDaniel, Schmidt, & Hunter, 1988; Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990), thus making it unlikely that process variables will substantially influence test validity.

More recent meta-analytic research has, in fact, suggested a much higher overall validity for the interview than the .14 found by Hunter and Hunter (1984). For example, Wiesner and Cronshaw (1988) found a mean corrected validity of .47 across 150 interview validity studies involving all types of criteria. McDaniel, Whetzel, Schmidt, and Maurer (1991) analyzed 106 studies and found a corrected mean validity of .45 for job performance criteria. Finally, Marchese and Muchinsky (1993) found a mean corrected validity of .38 across 31 interview studies. All three meta-analyses included corrections for range restriction.

These studies also suggested that structure moderates the validity of the interview. For example, Wiesner and Cronshaw (1988) found that the mean validity of structured interviews was twice that of unstructured interviews, specifically, .62 versus .31. Marchese and Muchinsky (1993) found a correlation of .45 between interview validity and structure (i.e., structured vs. unstructured), by far the largest effect on interview validity among the six characteristics they studied (e.g., single vs. multiple interviewers and occupation). McDaniel et al. (1991) found a smaller difference between structured and unstructured interviews (.47 vs. .40), which narrowed even further when only job-related interviews were considered (.46 vs. .42).

In summary, more recent meta-analytic research has suggested different conclusions regarding the validity of the interview than those found by Hunter and Hunter (1984). Specifically, this research suggests that interviews can reach a level of validity that is both useful and comparable to many predictors that have traditionally been considered best. In addition, much of this research suggests that structure is an important moderator of interview validity.

However, there are at least four reasons why another meta-analysis of the employment interview is warranted. First, results

Allen I. Huffcutt, Department of Psychology, Bradley University; Winfred Arthur, Jr., Department of Psychology, Texas A&M University.

We sincerely thank three anonymous reviewers for their thoughtful, insightful comments and suggestions on drafts of this article. We also thank Mike McDaniel for his collaborative efforts in locating and analyzing interview validity studies.

Correspondence concerning this article should be addressed to Allen I. Huffcutt, Department of Psychology, Bradley University, Peoria, Illinois 61625.

of the more recent interview meta-analyses, although meaningful in their own right, are not directly comparable to Hunter and Hunter's (1984) finding of .53 as the mean validity for mental ability tests because both non-entry-level jobs and criteria other than supervisory ratings were included in their analysis. Given the frequency with which the mean validity of .53 for ability tests continues to be cited, it seems desirable to have corresponding figures for the interview that are directly comparable and can be cited concurrently.

Second, there appears to be some inconsistency regarding the extent to which structure moderates validity. Both Wiesner and Cronshaw (1988) and Marchese and Muchinsky (1993) found structure to be the single most influential factor affecting interview validity. McDaniel et al. (1991), on the other hand, found a considerably smaller difference. Thus, at a general level, further meta-analytic assessments of the degree to which structure moderates the validity of the employment interview seem warranted in an attempt to clarify this relationship.

Third, it is not clear how validity varies according to the amount of structure. All three of the more recent interview meta-analyses were limited to a simple structured versus unstructured categorization of interview studies. Yet a review of the literature suggests a diversity of approaches to structuring the interview. Some approaches appear to represent various intermediate levels of structure (e.g., Arvey, Miller, Gould, & Burch, 1987; Ghiselli, 1966; Johnson, 1990). Ghiselli (1966), for example, standardized the topical areas to be covered, used a general set of questions to guide discussion within each area, and made one global rating of each applicant after the interview. Other approaches, such as the situational interview (Latham, Saari, Pursell, & Campion, 1980), are highly structured in that questions are precisely specified and responses to each question are individually scored. Given these diverse operationalizations, it appears that structure is far more complex than can be represented by a dichotomous distinction. The relatively sophisticated multidimensional classification scheme presented in this investigation permits an assessment of the relative effects of varying amounts of structure on validity.

Fourth, there seems to be no information on how much structure is enough structure. As Daniel and Valencia (1991) noted, "it is generally recognized that structured interviews are preferable to informal ones, but no consensus exists about how much structure is needed" (p. 128). Determining whether interview validity continues to increase across the entire range of structure or asymptotes at some point (i.e., a ceiling effect) has important ramifications for both practice and research. For instance, the point of asymptote, if there is one, might be highly relevant to practitioners because higher levels of structure typically involve increased development time and concomitant costs.

Consequently, our primary purpose in the present investigation was to provide a reanalysis of the employment interview in an attempt to address these four issues directly. To provide comparable estimates for the interview, it was necessary to keep our meta-analysis compatible with Hunter and Hunter's (1984). Accordingly, we limited our meta-analysis to include only entry-level jobs and excluded jobs representing promotion within a company or trade certification; we used only studies in which the criteria for performance were supervisory ratings; and we corrected for sampling error, criterion unreliability, and range restriction (see Hunter & Hunter, pp. 89–90). In addition, we attempted to base our analysis on a relatively large number of studies to avoid second-order sampling effects (see Hunter & Schmidt, 1990). Because most hiring occurs with entry-level positions and supervisor ratings are by far the most common criteria (Hunter & Hunter, 1984), structuring our meta-analysis in such a manner ensured compatibility with Hunter and Hunter's while still allowing the other issues to be addressed.

As noted earlier, a more sophisticated framework for classifying interview structure was used in the present investigation. This framework, details of which are provided in the Method section, was based on progressively higher levels of two dimensions of structure: standardization of (a) interview questions and (b) response scoring. Other potential moderators of interview validity (e.g., individual vs. panel interviews) were not analyzed, primarily because structure is generally recognized as the largest moderator of interview validity (Harris, 1989; Marchese & Muchinsky, 1993; Wiesner & Cronshaw, 1988). Second, further subgrouping the data set to look at additional moderators would have invited second-order sampling effects (Hunter & Schmidt, 1990). The anticipated effect of collapsing across other potential moderators was an increase in the residual population variance.

In summary, our objectives in this article were (a) to present a reanalysis of the relationship between the interview and job performance for entry-level positions and (b) to use a more systematic and sophisticated operationalization of interview structure to explore the nature of the relationship between structure and the validity of the employment interview.

Method

Interview Data Set

The data set analyzed consisted of 114 interview validity coefficients, all of which involved entry-level jobs and supervisory rating criteria and could be classified as to their level of structure with respect to standardization of interview questions and response scoring. These studies encompassed both published and unpublished research; our sources were journal articles (n = 45), dissertations (n = 30), technical reports (n = 26), master's theses (n = 8), unpublished or submitted manuscripts (n = 2), books (n = 2), and conference papers (n = 1). Although a majority of the studies were North American, five were conducted in other countries (Ghana, England, The Netherlands, Israel, and Australia). A complete listing of the sources for these studies is available from Allen I. Huffcutt.

Although there was some overlap between our studies and those analyzed in other interview meta-analyses, this overlap tended to be relatively low. In total, there were 84 references from which the 114 validity coefficients were taken. Of these 84 references, 31 were also included in Wiesner and Cronshaw's (1988) list, 45 were in McDaniel et al.'s (1991) list, and 7 were in Marchese and Muchinsky's (1993) list. Expressed as a percentage of references common with our list, the amount of overlap with these lists was 37%, 54%, and 8%, respectively. (Overlap with Hunter & Hunter, 1984, could not be calculated because primary study references were not provided, although with only 10 validity coefficients there is likely to be very little overlap.) The relatively low overlap between our data set and those of previous meta-analyses can be attributed to several possible reasons, including (a) inclusion of only entry-level studies; (b) limiting studies to those with supervisory rating criteria; (c) the use of different decision rules for determining which studies and

study coefficients to include; and (d) the inclusion of more recent studies, in the case of Wiesner and Cronshaw (1988). Thus, our meta-analysis appears to represent a reasonably independent assessment of the interview.

Location of Interview Studies

An extensive literature search was conducted to identify interview validity studies that involved both entry-level positions and on-the-job supervisory performance rating criteria. The search process started with the reference lists from Wiesner and Cronshaw's (1988) and McDaniel et al.'s (1991) meta-analyses and then was expanded to include computer searches on literature databases such as PsycLIT and Dissertation Abstracts International and manual searches with journals such as *Public Personnel Review* that are not indexed in computerized databases. References from collected studies were also searched for additional studies, and a number of prominent researchers in the interview area were contacted directly regarding recent unpublished studies.

We used a number of decision rules to determine which study coefficients would be retained for analysis. To avoid duplication, we retained only one validity coefficient for each unique sample of subjects. Thus, coefficients representing an overall evaluation on both the interview and job performance were used, if available; otherwise, the coefficients representing the individual component–dimension ratings were averaged (e.g., Barrett, Svetlik, & Prien, 1967). When alternative performance-rating criteria were reported—either from different supervisors or different rating instruments—the composite criteria were retained, if provided; otherwise, the alternative coefficients were averaged. When multiple interviews were conducted using the same sample and the same interview format, we kept the coefficient for the first interview (e.g., Janz, 1982). Although few in number, when coefficients were presented for the same criteria collected at different time periods, the longest time period was used because this represented the most stable relationship between the interview and job performance. The only exceptions to this rule were when the interview consisted of distinct and structurally different parts for which separate validity coefficients were reported (e.g., Kennedy, 1986) and when the same subjects were interviewed twice by structurally different methods (e.g., Gillies, 1988).[1]

As a result of the search and application of these decision rules, we obtained 130 interview validity coefficients, all of which involved entry-level positions and supervisory performance rating criteria. Two of these coefficients were subsequently dropped because they involved a procedure known as the "extended interview," which involves extensive contact with the applicant, including assessment-center-type exercises (Vernon, 1950; Wilson, 1948). Such a technique goes well beyond the boundaries of a typical interview. Four additional coefficients from a laboratory study were dropped because of concerns over generalizability (Heneman, Schwab, Huett, & Ford, 1975). Lopez's (1966) data were omitted because the interview was designed to capture objective biographical information. Finally, Freeman, Manson, Katzoff, and Pathman's (1942) study was not used because part of the interview involved a portable test apparatus, which was used to induce stress. Such a procedure also was not representative of a typical interview. One hundred twenty-two validity coefficients remained after elimination of these 8 coefficients.

Interview Structure Classification

Huffcutt (1992) defined *structure* as the reduction in procedural variability across applicants, which can translate into the degree of discretion that an interviewer is allowed in conducting the interview. Such a definition suggests that interview structure is both continuous and multidimensional in nature. At an operational level, there are two dimensions of structure directly relating to degree of discretion in the conduct of the interview, namely, standardization of (a) interview questions and (b) response scoring.

Huffcutt's (1992) review of a large number of primary studies suggested that question standardization could be adequately described by four progressively higher levels of structure. Level 1 was characterized by an absence of formal constraints, the typical unstructured interview. Level 2 was characterized by limited constraints, typically standardization of the topical areas to be covered (e.g., Ghiselli, 1966). Level 3 was characterized by prespecification of the questions, although applicants were not asked the exact same questions because of the use of different interview forms or allowing interviewers to choose among alternative questions and to probe responses to the specified questions (e.g., Janz, 1982). Level 4 involved complete standardization: Applicants were asked the exact same questions, and no deviation or follow-up questioning was permitted (e.g., Latham et al., 1980).

This review also suggested that response-scoring standardization could be adequately described by three progressively higher levels of structure. Level 1 was typified by the formation of a single overall evaluation based on total interview information (e.g., Ghiselli, 1966). Level 2 was distinguished by the formation of multiple evaluations along preestablished criteria, such as job dimensions or traits (e.g., Janz, 1982). Finally, Level 3 was distinguished by the evaluation of applicant responses to each individual question according to preestablished benchmark answers (e.g., Latham et al., 1980).

So, for the current analyses, structure was defined in terms of the standardization of interview questions and the standardization of response scoring. Specifically, we coded each interview validity study according to the level of structure along these two dimensions, which together formed a number of unique combinations of structure. In classifying studies by structure, we did not find sufficient information to classify eight studies—one from Darany (1971) and seven from Maurer (1983)—and so we eliminated these. This resulted in a final data set of 114 interview validity coefficients. A graphic representation of the classification framework along with the number of validity coefficients and total sample sizes of the various levels and combinations of structure are presented in Figure 1.

To assess the reliability of the coding process, we both independently coded 20 randomly selected studies. Interrater reliability was .99 ($p < .001$) for sample size, .95 ($p < .001$) for the validity coefficient, .99 ($p < .001$) for question standardization, and .78 ($p < .001$) for response scoring. The relatively low reliability for response scoring was due primarily to Latham and Saari's (1984) situational interview study, in which interviewers, rather than recording and scoring responses, simply asked questions and then made global assessments at the end of the interview; the study was coded as a Level 3 (as designed) by one rater and a Level 1 (as implemented) by the other. The latter rating was used in the analyses. Excluding this study for response scoring resulted in an interrater reliability of .95 ($p < .001$). In summary, these results indicated that study features could be reliably coded. (The remainder of the studies were coded by Allen I. Huffcutt.)

As shown in Figure 1, there appears to be a tendency for researchers who standardize questions to also standardize the scoring of responses. Mayfield (1964) noted such collinearity in his earlier review of the interview. Thus, many of the cells representing high standardization on one dimension and low standardization on the other dimension had few, if any, validity coefficients. Two of these cells, Levels 1 and 2 of questions standardization with Level 3 of response scoring, in fact represent combinations of structure that are theoretically possible but highly unlikely ever to be observed in practice. This observation is further supported by

[1] The combined sample size from nonunique studies comprised less than 5% of the total sample size of the final data set. The analyses reported later in this article were rerun, excluding these studies, and the results were essentially identical.

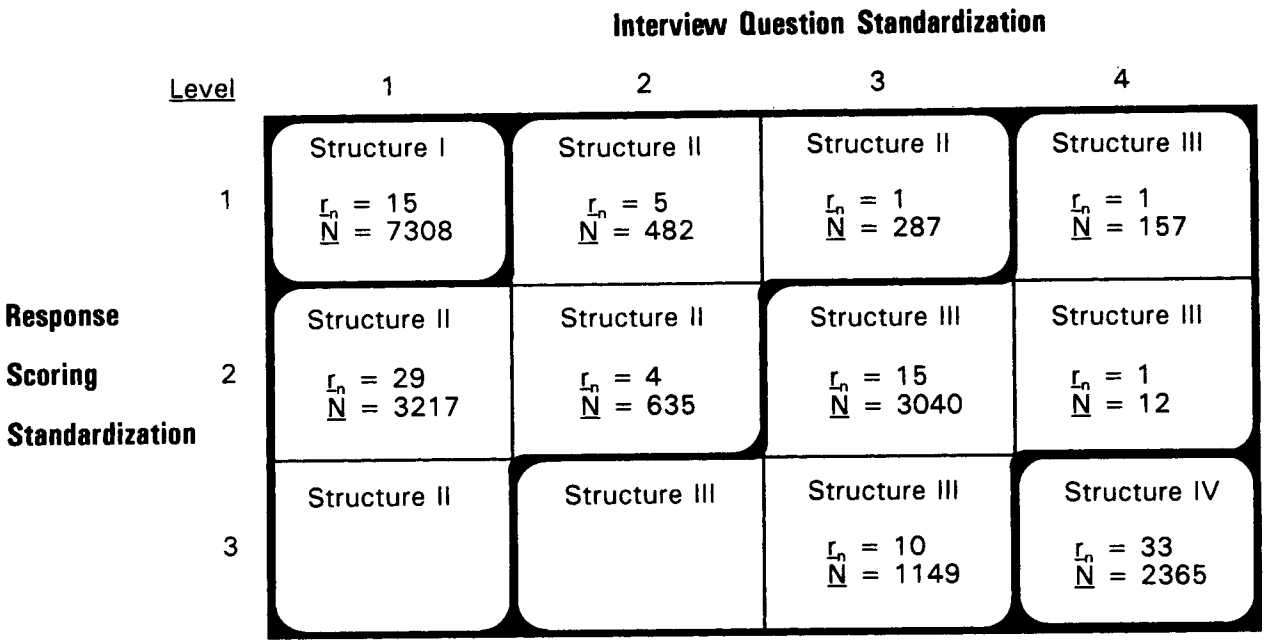

Figure 1. Classification scheme used to differentiate interview studies by level of structure. r_n = number of validity studies; N = total sample size. Level 1 question standardization was no constraints; Level 2 was limited constraints, typically on the topical areas; Level 3 was precise specification of questions from which interviewers could choose or follow-up; Level 4 was asking the exact same questions with no choice or follow-up. Level 1 response scoring was a global assessment; Level 2 response scoring was assessment along multiple established criteria; Level 3 was evaluation of each individual response according to preestablished answers.

the fact that no studies were found that could be classified into these cells.

Artifact Information

Consistent with Hunter and Hunter (1984), we used a value of .60 for criterion unreliability because research has suggested that this is the upper limit for interrater reliability, even for experienced raters (King, Hunter, & Schmidt, 1980; Rothstein, 1990). Artifact information for range restriction was collected directly from the 114 interview studies in the final data set. In total, it was possible to collect 15 unique values of u the ratio of the standard deviation of the restricted sample to the standard deviation of the unrestricted sample. The mean of these 15 range restriction ratios was .74 (*SD* = .16).

Analyses of Structure

The first analysis conducted was a meta-analysis of all 114 interview validity studies collectively. The mean validity from such an analysis provided an estimate of the overall validity of the interview for entry-level jobs that was directly comparable to the mean validity of .53 found by Hunter and Hunter (1984) for mental ability tests. The meta-analytic procedures used were those outlined by Hunter and Schmidt (1990); the actual computations were performed using a SAS (SAS Institute, 1990) PROC MEANS program developed by Huffcutt, Arthur, and Bennett (1993). The sample-size-weighted mean correlation and the sample-size-weighted variance were computed for the observed (uncorrected) coefficients. The proportion of observed variance attributable to sampling error and study-to-study differences in the level of range restriction was calculated and removed. The mean validity and residual variance were then corrected for criterion unreliability and range restriction, resulting in an estimate of the correlation between interview ratings and job performance in the population (ρ) and the corresponding variance of this relationship.

The ideal analysis to assess how validity changes with amount of structure would have been to perform a separate meta-analysis for each unique combination of structure shown in Figure 1. Such an analysis would provide insight into how mean validity varies with increased structure along each dimension separately and whether there is an interactive effect between the dimensions. Unfortunately, as is apparent in Figure 1, the number of studies in the data set did not permit such an analysis. The low number of studies in some of the individual cells would most likely have resulted in second-order sampling effects (Hunter & Schmidt, 1990).

Accordingly, a decision was made to collapse the cells into four progressively higher combinations (levels) of structure. The lowest level of structure, Structure 1, was distinguished by no formal structure, namely, no constraints on the questions with only a single overall evaluation (Level 1 questioning with Level 1 scoring). The next level of structure, Structure 2, was characterized by the use of some formal structure, namely, Level 1 questioning with Level 2 scoring, Level 2 questioning with Level 1 scoring, Level 2 questioning with Level 2 scoring, Level 3 questioning with Level 1 scoring, and Level 1 questioning with Level 3 scoring. Here the levels of the two dimensions sum to either three or four. Structure 3 represented a high level of structure but still involved some variability in the process and included Level 3 questioning with Level 2 scoring, Level 3 questioning with Level 3 scoring, Level 4 questioning with Level 2 scoring, Level 4 questioning with Level 1 scoring, and Level 2 questioning with Level 3 scoring. Here the two dimensions sum to either five or six. Finally, the highest level of structure, Structure

4, was characterized by asking all applicants the same questions with no deviation or probing and scoring of each individual response according to benchmark answers (Level 4 questioning with Level 3 scoring).

Forming these four progressively higher levels of structure maximized the available data while still allowing the research questions raised earlier to be addressed. Specifically, such combinations permitted an evaluation of the overall extent to which structure moderates validity and an assessment of the extent to which varying levels of structure are related to interview validity. Obviously, because structure is best conceptualized as varying on a continuum, neither the structural framework presented nor the collapsing into four progressively higher levels of structure are a perfect representation of structure. Nonetheless, they represent a significant improvement over the structured versus unstructured categorization used in previous research. Meta-analyses for each of these four levels of structure were similarly carried out as we described earlier.

Results

The original results of Hunter and Hunter (1984) for all 11 of the predictors they compared are presented in Table 1. The mean sample-size-weighted correlation across all 114 validity coefficients in the current investigation was .22 (*SD* = .14). For the four levels of structure, the sample-size-weighted correlations were .11 (*SD* = .04), .20 (*SD* = .11), .34 (*SD* = .12), and .34 (*SD* = .17) for Level 1 to Level 4 structures, respectively. Correcting for criterion unreliability and range restriction resulted in estimates of rho ranging from .20 to .57. These estimates of rho along with their population standard deviations, are reported in the lower half of Table 1.

Table 1
Results From Hunter and Hunter (1984) and From the Current Meta-Analysis: Prediction of Supervisory Rating Criteria With Entry-Level Jobs

Predictor	Validity *M*	Validity *SD*	No. of studies	Total subjects
Hunter and Hunter (1984)[a]				
Ability composite	.53	.15	425	32,124
Job tryout	.44	—	20	—
Biographical inventory	.37	.10	1	4,429
Reference check	.26	.09	10	5,389
Experience	.18	—	425	32,124
Interview	.14	.05	10	2,694
Training and experience ratings	.13	—	65	—
Academic achievement	.11	.00	11	1,089
Education	.10	—	425	32,124
Interest	.10	.11	3	1,789
Age	−.01	—	425	32,124
Current investigation				
All interviews	.37	.24	114	18,652
Structure 1	.20	.08	15	7,308
Structure 2	.35	.18	39	4,621
Structure 3	.56	.20	27	4,358
Structure 4	.57	.28	33	2,365

Note. Dashes indicate that these data were not provided by Hunter & Hunter.

[a] Data are from "Validity and Utility of Alternative Predictors of job performance," by J. E. Hunter & R. F. Hunter, 1984, *Psychological Bulletin, 96*, p. 90. Copyright 1984 by the American Psychological Association. Reprinted by permission.

The results demonstrate that mean validity generally appears to increase with increasing levels of structure. However, this trend asymptotes at Structure 3, beyond which additional structure yields very little incremental validity ($\Delta r = .01$), thus suggesting the presence of a ceiling effect for structure. Moreover, the validity of the interview at Structure 3 and Structure 4 appears highly comparable to the validity found for ability tests. An unexpected finding was that the population standard deviation increased consistently from Structure 1 to Structure 4.

Discussion

Results of this investigation suggest that the overall validity of the interview for entry-level jobs is much higher than was indicated by Hunter and Hunter (1984) in their analysis. In addition, our relatively sophisticated system of structure classification resulted in several insights into the relationship between level of structure and the validity of the interview. Specifically, these results (a) confirm that structure is a major moderator of interview validity, (b) demonstrate that validity generally increases with increasing structure, and (c) suggest that there is a point beyond which additional structure yields little or no incremental validity, a ceiling effect of structure. Such results are particularly relevant to practitioners, who must weigh cost and development time against anticipated results when deciding how much to structure an interview.

The finding that highly structured interviews can provide essentially the same validity as ability tests is interesting because it has been suggested that such interviews are actually nothing more than verbal ability tests (Campion, Pursell, & Brown, 1988; Wright, Lichtenfels, & Pursell, 1989). Empirical research on this issue, however, appears mixed. Although some studies have found a high correlation between structured interview ratings and scores on ability tests (e.g., Campion et al., 1988), others have found a relatively low correlation (e.g., Bosshardt, 1992; Delery, Wright, Tolzman, & Anderson, 1992). Clearly, assessing the correspondence between structured interviews and ability tests and the conditions that moderate this relationship (e.g., content) is an important avenue for future research.

Results of this investigation may also provide a plausible explanation for the inconsistency among the other meta-analyses noted earlier there regarding the extent to which structure moderates interview validity. For instance, Wiesner and Cronshaw (1988) appeared to have used a very precise and conservative classification scheme to differentiate structured and unstructured interviews. Consequently, they seemed to have captured the more extreme ends of the interview structure continuum and not the intermediate ranges. Thus, their unstructured and structured classification may be similar to our Structure 1 and Structure 4 levels. McDaniel et al.'s (1991) classification scheme, on the other hand, seems much less stringent and more liberal. Thus, there may have been some mixing of studies with intermediate levels of structure in both their structured and unstructured categories, which would narrow the difference between them. On the basis of this scenario, the fact that Wiesner and Cronshaw found a much greater effect for structure becomes more understandable.

An interesting issue that could be addressed in future research is why the population standard deviation increased consistently from Structure 1 to Structure 4. Some residual variability in the population was expected, for several reasons. First, other potential moderator variables, such as the use of an interview panel, were not addressed. Second, other dimensions of interview structure may moderate validity in addition to those analyzed here. Third, variability as a result of differences in the level of criterion unreliability across studies was not removed because a global value of .60 was used. Finally, outlier studies may have been present that might have increased the residual variability (see Orr, Sackett, & DuBois, 1991). Nevertheless, the consistent increase across levels was not expected and warrants further investigation.

Another suggestion for future research concerns the observed ceiling effect for structure. One plausible explanation of this phenomenon is that a well-trained interviewer could get better insights using careful, in-depth probing of a standard question than would otherwise be the case when restricted by completely fixed interview questions with no opportunity for follow-up. Thus, at Structure 3 standardization, in which the interviewer still has some discretion, it is possible for individual differences in interviewing ability to influence the validity of the interview (Dreher, Ash, & Hancock, 1988). Future primary studies could investigate this as an explanation for the observed ceiling effect.

There were several limitations of the present study, and these should be noted. First, the two-dimensional framework presented in this investigation may not fully represent all of the dimensions of interview standardization. Second, the structure classification scheme resulted in a few combinations of structure that, although theoretically possible, are highly unlikely to occur in real interview situations. Third, although the structure classification scheme could have permitted more refined analyses, the number of usable data points required a simplification of the framework. Finally, these results generalize only to entry-level jobs and not to jobs representing promotion within a company or trade certification.

Nonetheless, the results obtained allow us to make two summary conclusions. Specifically, entry-level employment interviews with high levels of structure appear to be highly valid predictors of supervisory ratings of job performance. In addition, although interview validity increases with increasing levels of structure, there seems to be a point at which it asymptotes or, possibly, declines.

References

Aamodt, M. G. (1991). *Applied industrial/organization psychology.* Belmont, CA: Wadsworth.

Arvey, R. D., Miller, H. E., Gould, R., & Burch, P. (1987). Interview validity for selecting sales clerks. *Personnel Psychology, 40,* 1–12.

Barrett, G. V., Svetlik, B., & Prien, E. P. (1967). Validity of the job-concept interview in an industrial setting. *Journal of Applied Psychology, 51,* 233–235.

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

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

Cascio, W. F. (1991). *Applied psychology in personnel management.* Reston, VA: Reston.

Daniel, C., & Valencia, S. (1991). Structured interviewing simplified. *Public Personnel Management, 20,* 127–134.

Darany, T. (1971). *Summary of state police trooper 07 validity study.* Lansing: Michigan Department of Civil Service.

Delery, J. E., Wright, P. M., Tolzman, K., & Anderson, D. C. (1992, May). *Employment tests and the situational interview: A test of incremental validity.* Paper presented at the seventh annual conference of the Society for Industrial and Organizational Psychology, Montreal, Quebec, Canada.

Dreher, G. F., Ash, R. A., & Hancock, P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. *Personnel Psychology, 42,* 691–726.

Freeman, G. L., Manson, G. E., Katzoff, E. T., & Pathman, J. H. (1942). The stress interview. *Journal of Abnormal and Social Psychology, 37,* 427–447.

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

Gillies, T. K. (1988). *The relationship between selection variables and subsequent performance ratings for teachers in an Oregon school district* (Doctoral dissertation, University of Oregon, 1988). *Dissertation Abstracts International, 49,* 3555.

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

Heneman, H. G., Schwab, D. P., Huett, D. L., & Ford, J. J. (1975). Interviewer validity as a function of interview structure, biographical data, and interviewee order. *Journal of Applied Psychology, 60,* 748–753.

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.

Huffcutt, A. I., Arthur, W., Jr., & Bennett, W. (1993). Conducting meta-analysis using the "PROC MEANS" procedure is SAS. *Educational and Psychological Measurement, 53,* 119–131.

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

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

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

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

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

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology, 65,* 507–516.

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

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

Lopez, F. M., Jr. (1966). Current problems in test performance of job applicants: 1. *Personnel Psychology, 19,* 10–18.

Marchese, M. C., & Muchinsky, P. M. (1993). The validity of the em-

ployment interview: A meta-analysis. *International Journal of Selection and Assessment, 1,* 18–26.

Maurer, S. (1983). *Economic analysis of the employment interview.* Unpublished doctoral dissertation, University of Oregon, Eugene.

Mayfield, E. C. (1964). The selection interview—A re-evaluation of published research. *Personnel Psychology, 17,* 239–260.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology, 41,* 283–314.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. (1991). *The validity of employment interviews: A comprehensive review and meta-analysis.* Manuscript submitted for publication.

Muchinsky, P. M. (1993). *Psychology applied to work: An introduction to industrial and organizational psychology* (4th ed.). Pacific Grove, CA: Brooks/Cole.

Orr, J. M., Sackett, P. R., & DuBois, C. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44,* 473–486.

Roth, P. L., & Campion, J. E. (1992). An analysis of the predictive power of the panel interview and pre-employment tests. *Journal of Occupational and Organizational Psychology, 65,* 51–60.

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.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75,* 175–184.

Rudner, L. M. (1992). Pre-employment testing and employee productivity. *Public Personnel Management, 21,* 133–150.

SAS Institute (1990). *SAS language: Reference, version 6* (1st ed.). Cary, NC: Author.

Terpstra, D. E., & Rozell, E. J. (1993). The relationship of staffing practices to organizational level measures of performance. *Personnel Psychology, 46,* 27–48.

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

Wiesner, W., & 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.

Wright, P. M., Lichtenfels, P. A., & Pursell, E. D. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology, 62,* 191–199.

Received August 13, 1992
Revision received August 9, 1993
Accepted August 10, 1993 ■