A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview*

Willi H. Wiesner†

Department of Management, Concordia University, Montreal, Quebec H4B 1R6

Steven F. Cronshaw

Department of Psychology, University of Guelph, Guelph, Ontario, Canada

A meta-analysis of the employment interview was carried out to investigate the impact of interview format (individual vs. board interviews) and interview structure (unstructured vs. structured) on the validity of interviews. A thorough review of the unpublished and published literature worldwide yielded 150 usable validity coefficients for the meta-analysis. Contrary to the predominantly pessimistic views of previous researchers, the interview was found to be a generally good selection instrument. These findings suggest that the 'received doctrine' of interview invalidity is false. However, interview structure moderated predictive validity coefficients to a considerable extent. In fact, structured interviews produced mean validity coefficients twice as high as unstructured interviews. Although considerable variance in structured interviews remained unaccounted for after adjustment for statistical artifacts, all of the variation in observed validity coefficients for unstructured interviews was accounted for. It was concluded that a number of social psychological processes examined in previous interview research would have little effect in moderating the validity coefficients of the unstructured interview. The results also suggest that higher validity coefficients are associated with more reliable interviews and the use of formal job-analytic information in developing interview questions. Implications for research and practice in personnel psychology are explored.

The employment interview is a tenaciously popular but controversial selection method. In particular, reviewers of the interviewing literature repeatedly question the predictive validity of the employment interview (Arvey & Campion, 1982; Mayfield, 1964; Milne, 1967; Reilly & Chao, 1982; Rodger, 1952; Rowe, 1981; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949; Webster, 1982; Wright, 1969). These reviewers of the interviewing literature have, however, aggregated interview validity coefficients in a subjective fashion, referred to as a ' narrative ' review by Hunter, Schmidt & Jackson (1982), rather than utilizing an empirical approach.

Specifically, these reviews do not adjust interview validity coefficients for differing sample sizes in each of the studies nor do they adjust for statistical artifacts such as sampling error, measurement error and restriction or range. As a result, conclusions drawn about the validity of interviews from narrative reviews may be misleading.

Hunter *et al.* (1982) have advocated the use of meta-analytic techniques as a method of aggregating a number of validity studies while accounting for the above-mentioned statistical artifacts. The purpose of this paper is to resolve some long-standing controversies in the interviewing literature by first presenting a model of the employment interview as a predictor of employment outcomes and then testing this model by subjecting a very large dataset of interview validity coefficients to meta-analytic procedures.

A model of employment interview validity

Researchers generally agree that the validity of the employment interview is poor, although they differ in their definitions of what constitutes an employment interview. Worse yet, some researchers maintain that the dynamics underlying the employment interview create as many different interviews as there are numbers of interviewer and interviewee interactions (Schmitt, 1976; Wagner, 1949; Zedeck, Tziner & Middlestadt, 1983). This view, perpetuated by the mixed validity results found in the empirical literature, would not allow us to make any confident predictions about interview validity in given circumstances even if employment interviews overall had some predictive validity.

The dual 'received doctrines' (Barrett, 1972) of poor interview validity and differential validity due to idiosyncratic interview process are now widely accepted among industrial/organizational (I/O) psychologists even though the employment interview continues to enjoy ubiquitous (but apparently inexplicable) popularity among practitioners. On careful examination, the empirical bases for these two received doctrines of the interview are surprisingly thin.

This study empirically tested both received doctrines of the interview. First, a large dataset of interview validity coefficients was quantitatively combined to assess whether or not interviews generally have poor validity. Second, a model of the interview was developed which collapses broad categories of 'idiosyncratic' events or behaviours into equivalence classes which summarize and simplify the interview process along two basic dimensions which should largely determine interview validity (Dubin, 1976). This conceptual model of the interview was then tested.

Before detailing the interview model which we are proposing, we offer a conceptual definition of the employment interview. The employment interview is an interpersonal interaction of limited duration between one or more interviewers and a job-seeker for the purpose of identifying interviewee knowledge, skills, abilities and behaviours that may be predictive of success in subsequent employment. The operational indicators of this success include criteria of job performance, training success, promotion and tenure.

We propose a model where interview validity (i.e. criterion-related validity) is a function of two major factors. The first factor is *interview reliability*. This proposition comes from psychometric theory. Since reliability places an upper limit on validity

(e.g. Nunnally, 1978), more reliable interviews should achieve better prediction of job success. Specifically, board interviews should be more reliable than individual interviews, particularly when two or more independent ratings are collapsed statistically into a composite. By extension, board interviews should also be more valid than individual interviews. Many researchers through the years have suggested that board interviews may indeed be more valid (Anstey & Mercer, 1956; Arvey & Campion, 1982; Hollingworth, 1922; Mayfield, 1964; Oldfield, 1941; Pursell, Campion & Gaylord, 1980).

An additional hypothesis is suggested in the interview literature. Board interview ratings that are made independently by board members and then statistically combined should be more reliable (and valid) than a single consensus rating reached through group discussion. Several researchers (Landy, 1976; Pursell *et al.*, 1980; Rothstein & Jackson, 1980) suggest that statistically combined ratings are, in fact, more valid.

The second factor suggested by the interview literature to influence or moderate the validity of interviews is *interview structure*: that is, the structured interview should be more predictive of job success than the unstructured interview. Numerous reviewers in fact recommend structuring the interview as one of the best ways of increasing interview validity (Arvey & Campion, 1982; Mayfield, 1964; Milne, 1967; Rodger, 1952; Rowe, 1981; Schmitt, 1976; Ulrich & Trumbo, 1965; Wagner, 1949; Wright, 1969). Improved validity of structured interviews could be produced by increased reliability of the interviewer ratings, greater job-relatedness of interview questions, or a combination of these two factors.

Although a number of reviewers of literature on interviewing contend that the structured interview should be more predictively valid than the unstructured interview, they have been more reluctant to suggest that the board interview is predictively superior to the individual interview (e.g. Arvey & Campion, 1982; Pursell *et al.*, 1980). The implication is that interview structure is likely to be a stronger moderator of interview validity than interview format (board vs. individual interviews). We examined the effects of both moderators (i.e. interview format and structure) on the validity of interviews in this study.

Study hypotheses

We derived the following two hypotheses based on the above model of interview validity.

Hypothesis 1. The predictive validity of different types of interviews would be ordered from least to greatest as follows (1 is least, 4 is greatest):

1. unstructured individual interviews (least reliable, least job-related)
2. unstructured board interviews
3. structured individual interviews
4. structured board interviews (most reliable, most job-related)

This hypothesis of course predicted that unstructured interviews would have lower validity than structured interviews when unstructured and structured interviews

were collapsed over individual and board types. The validity of board interviews would be higher than the validity of individual interviews where these were collapsed over unstructured and structured types. Within the four *a priori* interview types, the amount of variance in true validity coefficients would be minimal or non-existent if the effects of ' idiosyncratic ' social psychological processes on interview validity were to be ruled out.

Hypothesis 2. Board interview desciions based on the statistically combined scores of individual interviews would be more predictively valid than board decisions based on group consensus.

Method

Literature search

Extensive traditional and computer searches of the published literature worldwide dealing with the validity of the employment interview were carried out. North American, Australian, British, Dutch, French, German, Israeli and Swedish sources were investigated. Efforts were also made to obtain as many unpublished studies as possible by searching references in published works and contacting various academic, government and corporate researchers in the United States, Canada, Australia and the United Kingdom. A complete list of the studies used in the meta-analyses is available from the first author upon request.

We did not limit entry of studies by job type, date, country of origin, type of criterion used, use of ancillary data, design (concurrent vs. predictive) or type of correlation because there was no *a priori* rationale for expecting any of these variables to moderate the validity of interviews. However, we coded the studies for preliminary moderator analyses using the seven variables listed above to confirm that these variables would not moderate the validity of interviews. We ran a series of meta-analyses for these other potential moderator variables (detailed results are available from the first author). These analyses allowed us to rule out the moderating effects of all seven variables on interview validity, after controlling for interview format and structure, for three of the four *a priori* interview types identified in hypothesis 1. For the remaining interview type (the structured individual interview), differences possibly due to moderators were found for only three of the above seven variables. We therefore moved on to our major analyses in this study with some assurance that other moderators would not confound our analyses examining the impact of interview format and structure on criterion-related validity.

Decision rules for coding studies into meta-analysis

Six *a priori* decision rules were established to regulate the admission of research data into this study. The overall rationale was to include only those studies that occurred in or very closely approximated actual employment settings and in which interview validity was reported separately from other predictors. The strategy implemented by the six decision rules was designed to enhance the generalizability of the results to employment settings and to prevent contamination of validity coefficients by predictors other than the interview.

The six decision rules were as follows. First, only studies where face-to-face interactions occurred between interviewer and interviewee in an employment context were admitted for analysis (i.e. no laboratory studies were admitted). Second, only studies incorporating a work-related criterion were included. Third, studies in which the interviewers influenced or had prior knowledge of criterion ratings were eliminated. Fourth, assessment centre studies were excluded if interviewers observed participants in assessment centre exercises prior to conducting the interview. Otherwise the likelihood of interviewers' ratings being contaminated by their involvement in assessment centre procedures was deemed too great. Fifth, studies in which validity coefficients were selectively reported for only some of the dimensions tapped by the interview (i.e. usually the largest coefficients) were not included. Sixth, the overall interview rating provided by interviewers was used whenever possible as this is the rating typically used to make the hiring decision. When the interview was scored along a number of dimensions but an overall rating was not provided, an average of the validity coefficients given for each of the dimensions was used.

Well over 200 criterion-related validity coefficients were found and evaluated using the above-mentioned decision rules. The use of these decision rules reduced the number of usable validity coefficients to 151. About 90 per cent of the eliminations resulted from the application of the first and second decision rules.

One of the remaining studies was based on a sample of 37 862 interviews, representing more than 40 per cent of the total sample size. In accordance with Hunter *et al.*'s (1982) recommendations, this study was omitted from the meta-analysis leaving 150 validity coefficients although the meta-analytic results were virtually the same when this study was included as when it was omitted.

Coding of a priori *moderator variables*

The studies admitted for analysis were coded along the dimensions corresponding to the *a priori* moderator variables examined in this investigation. Where a study provided insufficient data to locate it clearly along a certain dimension, it was omitted from analyses of the moderating effect of that particular dimension. For example, if it was not clear whether the interview was unstructured or structured in a particular study, that study was not used in the comparison of the validities of unstructured and structured interviews.

Interview format. The *individual interview* consisted of one interviewer who interviewed and rated a single applicant at a time. This was the most commonly used form of interview in the studies examined. In the *board interview*, two or more interviewers interviewed and rated each applicant. The board interview further consisted of two other *a priori* types: interview boards where ratings of candidates made by board members were statistically combined through *averaging* and interview boards where ratings were obtained through member *consensus*.

Interview structure

Unstructured interview was coded where interviewers followed a free interview format without predetermined questions and/or where no rating scales where used. In

particular, if the interviewers made a global, subjective rating of the interviewee at the end of the interview rather than combining scores based on individual questions, the interview was considered unstructured. *Structured interview* was coded where the interview corresponded to 'a series of job-related questions with predetermined answers that are consistently applied across all interviews for a particular job' (Pursell *et al.*, 1980, p. 908). An additional requirement for coding to the structured interview category was that rating scales were completed during the interview based on the answers to each of the questions posed and that these ratings for each of the questions for each interviewer were combined to arrive at an overall interview rating.

Inter-rater reliability of coding methodology

A random sample of 30 studies was coded by a second rater according to interview format and interview structure using the coding strategy developed by the authors. Rater agreement between the second rater and first author, used as the measure of inter-rater reliability, was assessed in two ways: percentage agreement and the kappa statistic (Cohen, 1960). Percentage inter-rater agreement was 100 per cent for inter-view format ($\kappa = 1.00$) and 87 per cent for interview structure ($\kappa = 0.81$). Inter-rater agreement for the additional variables examined ranged from 93 to 100 per cent (κ range 0.89–1.00). Disagreements were resolved by discussing the studies with the second rater and arriving at a consensus regarding the appropriate coding. All the above indices of inter-rater agreement compare favourably with the percentage agreement reported by Bullock & Svyantek (1985) which ranged between 72.2 and 88.9 per cent.

Analyses

Data from the collected studies were analysed using the meta-analytic procedures developed by Hunter *et al.*, (1982). In this study, the uncorrected and corrected mean validities of the combined distribution of frequency-weighted validity coefficients were examined to assess overall interview validity. The effects of *a priori* moderators on interview validity coefficients were tested by performing moderated meta-analyses on subsets of the overall validity distribution sorted on the basis of interview format, structure and averaging/consensus.

Both uncorrected and corrected mean validity coefficients are reported in our meta-analyses. Adjustments were made to the uncorrected validity coefficients for both range restriction and criterion unreliability. The adjustment for *range restriction* was determined as follows. About 10 per cent of the validity coefficients collected were accompanied by sufficient information to determine the ratio of interviewees hired (here the average hiree-to-interviewee ratio was 0.33). These data were also used to establish an assumed artifact distribution for direct range restriction. The adjustment for *criterion unreliability* was determined by collecting criterion reliability data from validity studies reporting such data. Ten studies reported criterion reliabilities with an average reliability coefficient of 0.75. These reliability data were also used to establish an assumed artifact distribution. What is important is that this mean reliability is higher than the $r = 0.60$ reported by Schmidt & Hunter (1977) for the reliability of criteria. The use of $r = 0.75$, therefore, yields a more conservative

estimate of corrected validity coefficients than Schmidt & Hunter's suggested correction for criterion unreliability.

Results

Test of a priori *hypotheses*

Before testing hypotheses 1 and 2, a meta-analysis was run on the entire dataset of interview validity coefficients without regard to *a priori* moderators (see first line of Table 1). The uncorrected ($r = 0.26$) and corrected ($r = 0.47$) mean validity coefficients for the combined interview dataset reported here refute the doctrine of interview invalidity. These mean values are much larger than interview validity coefficients reported by previous researchers (e.g. Hunter & Hunter, 1984; Maurer & Russell, 1985). However, it is important to note that the estimated true validity coefficients of the interviews included in the overall sample range from -0.08 to 1.00 within a 95 per cent confidence band. This result suggests that, while interviews *on average* achieve acceptable validity coefficients, particular interviews can vary greatly in their predictive efficiency. This finding was not unexpected. However, it remains to be established whether the model of interview validity developed here can substantially explain these differences in predictive validity among interviews.

The moderator meta-analyses reported in the remainder of Table 1 directly test the two hypotheses generated by our conceptual model of interview validity. As predicted in hypothesis 1, the unstructured individual interview has the least predictive validity of all interview types (see second row of Table 1). The unstructured board interview has the next highest mean validity and differs significantly from the unstructured individual interview ($z = 6.61$, $P < 0.001$; McNemar, 1969, p. 158). Moreover, in support of hypothesis 1, the structured individual interview is a significantly better predictor than the unstructured board interview ($z = 16.46$, $P < 0.001$). Contrary to hypothesis 1, however, structured board interviews do not predict any better than structured individual interviews. Therefore, the general pattern of mean validity coefficients predicted by hypothesis 1 is borne out with the above exception.

All the variance in both unstructured individual and unstructured board interviews was explained by statistical artifacts. Therefore, the uncorrected and corrected mean validity coefficients reported here represent population coefficients for these two kinds of interviews. Because 100 per cent of the variance in unstructured interviews was explained by statistical artifacts, other moderators do not have room to exert their influence on unstructured interviews generally. Therefore, the impact of other ' idiosyncratic ' moderators is largely ruled out. Considerable variance was left unexplained, however, in the distribution of validity coefficients for structured individual and structured board interviews, suggesting the operation of further moderators on these types of interviews.

Hypothesis 2 was then tested. Because insufficient variation remained in the unstructured interview category to justify further moderator analyses, the moderator effects of statistical and consensus decision processes were examined only for structured board interviews (see Table 1). Hypothesis 2 was not supported for the structured board interview. Contrary to our prediction, the board ratings reached by

Table 1. Meta-analyses of the predictive validity of *a priori* interview types

Interview source	Total sample size[a]	Mean validity coefficient[b]	Observed variance in sample correlations	Residual variance unexplained by statistical artifacts[c]	Per cent of variance explained by statistical artifacts[c]	95 per cent confidence interval of validity coefficient
All studies[d]	51 459 (150)	0.47 (0.26)	0.093	0.080	14	$0.08 \leqslant P \leqslant 1.00$
Unstructured individual	2 303 (19)	0.20 (0.11)	0.013	−0.014	208	—
Unstructured board	3 134 (19)	0.37 (0.21)	0.013	−0.008	159	—
Structured individual	7 873 (32)	0.63 (0.35)	0.070	0.051	27	$0.18 \leqslant P \leqslant 1.00$
Structured board	2 104 (15)	0.60 (0.33)	0.043	0.017	61	$0.34 \leqslant P \leqslant 0.86$
Averaging	260 (6)	0.41 (0.23)	0.122	0.051	58	$0.03 \leqslant P \leqslant 0.86$
Consensus	1 694 (7)	0.64 (0.35)	0.018	−0.002	110	—

[a]The number of coefficients contributing to the total sample is given in parentheses.
[b]Means were corrected for direct restriction of range and criterion unreliability; uncorrected values are given in parentheses.
[c]Variance in corrected sample correlations was adjusted for the following statistical artifacts: sampling error, differences in range restriction and differences in criterion reliability.
[d]Not all interviews were classifiable according to structure and format; the subset sample sizes therefore do not sum to the total for all studies.

consensus were more valid than the statistical combination of independent ratings made by board members using structured interviews. The amount of explained variance increased considerably in the statistical vs. consensus moderator analyses over variance explained in overall distribution of structured board validity coefficients. These variance results should be interpreted cautiously, however, because the total sample size for structured boards using a statistical combination of board member ratings is quite small.

Hypothesis 1 further predicted that structured interviews would predict better than unstructured interviews when these two interview sources were collapsed over interview format. Similarly, hypothesis 1 predicted that board interviews should predict better than individual interviews. As Table 2 demonstrates, the prediction for a difference by interview structure was strongly supported whereas the difference by interview format was not. In fact, the structured interviews had mean validity coefficients twice those of unstructured interviews. The advantage for structured interviews in terms of dollar utility (Cronshaw & Alexander, 1985; Cronshaw, Alexander, Wiesner & Barrick, 1987; Schmidt, Hunter, McKenzie & Muldrow, 1979) over unstructured interviews should be very large. However, no overall advantage of board over individual interview was found despite that fact that the very large sample sizes reported here yield sufficient statistical power to detect even small differences in mean validity between the individual and board interviews.

Post hoc analyses

The theory developed in this study suggests that interview validity is largely determined by interview reliability. The study hypotheses are, in fact, partially based on the assumption that certain types of interviews (e.g. board, structured) are more reliable and hence more valid. However, the hypothesized linkage between reliability and validity can be directly tested. We further addressed the reliability issue by first searching all studies in the meta-analysis dataset for interview reliability coefficients. These reliability coefficients were then aggregated and mean reliability coefficients computed within subsamples classified according to interview format, interview structure and statistical/consensus decision (see Table 3).

As expected on the basis of the *a priori* model, the reliability of boards was higher than the reliability of individual interviews. However, the absolute difference in mean reliability coefficients was only 0.07. This reliability difference was accompanied by a non-significant difference in validity coefficients between individual and board interviews. The difference in reliability between unstructured and structured interviews was considerably greater. The mean reliability of the structured interview was 0.21 higher than the unstructured. Therefore, it is probable that the higher mean validity coefficients for the structured interview are at least partly accounted for by the improved reliability of structured over unstructured interviews. Reliability coefficients were higher for consensus board decisions than statistical combination of board ratings (although these results should again be interpreted with caution because the total sample size for statistical combination is low). This reliability result is again contrary to our hypothesis, but is consistent with the finding that the consensus board

Table 2. Meta-analyses of the predictive validity of interviews by structure and format

Interview source[a]	Total sample size[b]	Mean validity coefficient[c]	Observed variance in sample correlations	Residual variance unexplained by statistical artifacts[d]	Per cent of variance explained by statistical artifacts[d]	95 per cent interval of validity coefficient
Unstructured	5 518 (39)	0.31 (0.17)	0.024	0.000	99	$0.27 \leqslant P \leqslant 0.34$
Structured	10 080 (48)	0.62 (0.34)	0.064	0.043	32	$0.21 \leqslant P \leqslant 1.00$
Individual	30 100 (77)	0.44 (0.25)	0.097	0.085	12	$-0.13 \leqslant P \leqslant 1.00$
Board	17 253 (55)	0.44 (0.25)	0.040	0.027	34	$0.12 \leqslant P \leqslant 0.76$

[a]Not all interviews were classifiable according to structure and format; the subset sample sizes therefore do not sum to the total for all studies given in Table 1.
[b]The number of coefficients contributing to the total sample is given in parentheses.
[c]Means were corrected for direct restriction of range and criterion unreliability; uncorrected values are given in parentheses.
[d]Variance in corrected sample correlations was adjusted for the following statistical artifacts: sampling error, differences in range restriction.

Table 3. Mean reliability coefficients classified by interview format, interview structure and consensus/averaging combination of data

Interview variable	Number of studies	Total sample size	Mean reliability coefficient
Interview format			
Individual	20	5561	0.78[b]
Board	11	1909	0.85[b]
Interview structure			
Unstructured	9	629	0.61[b]
Structured	25	6936	0.82[b]
Consensus/averaging combination of data			
Consensus	5	1594	0.84[a]
Averaging	3	84	0.74[a]

[a]Mean reliability coefficients differ at $P < 0.05$ using the test for differences between uncorrelated groups of correlation coefficients (McNemar, 1969, p. 158).
[b]Mean reliability coefficients differ at $P < 0.001$.

decisions yielded higher validity coefficients. The higher validity of consensus board decisions may therefore also be partially due to improved reliability.

In summary, the pattern of the reliability results demonstrated that, where validity coefficients differed under a particular moderator, reliability coefficients also differed in the same direction. However, the most direct method of examining the relationship between interview reliability and validity is to correlate them. When this was done interview reliability and validity were correlated at 0.48 in those studies where both coefficients were available ($n = 40$). This finding again demonstrates that higher reliability was associated with higher interview validity as psychometric theory would predict. If the correlation is squared to yield the coefficient of determination ($r^2 = 0.23$), this result suggests that 23 per cent of the variance in interview validity is explained by differences in interview reliability.

Another assumption underlying the study hypotheses was that higher predictive validities would be associated with more highly job-related interview questions. An important requirement in assuring job-relatedness of the interview questions is in turn the conduct of a formal job analysis on which the interview questions are based. Therefore, the match of interview questions with job requirements should be best, and predictive validity coefficients should be highest, where interview questions are reported to be based on a formal job analysis. This *post hoc* prediction is made only for structured interviews because the unstructured interview follows a free format where, by definition, the extent of relationship with job requirements is unknown.

In a *post hoc* analysis, individual structured interviews were coded as based on formal job analysis information, on an 'armchair' or informal assessment of job

content, or as ' unknown ' where the information was not available to determine the interview content/job requirement match. Structured board interviews were not included in this analysis because they had already been subdivided into statistical and consensus decision processes and to break them down further with reference to job analysis information would have increased the likelihood of incurring second-order sampling error. We expected individual structured interviews based on formal job analysis to predict better than the other two individual structured interview types. This expectation was clearly supported by the results reported in Table 4. However, the percentage of variance accounted for by statistical artifacts increased over the combined distribution of structured individual interview validity coefficients only for the ' armchair ' job analysis. Therefore, further moderator effects on the validity of structured individual interviews, besides type of job analysis, cannot be ruled out. However, on the basis of our findings we concluded that the best prediction achievable for structured interviews would be obtained where structured interview questions are based on a formal job analysis rather than a less systematic assessment of job requirements.

Discussion

This study tested the veracity of two popular ' received doctrines ' of employment interview validity by accumulating and analysing a very large dataset of interview validity coefficients. The meta-analyses conducted here cast serious doubt on both doctrines. Contrary to the belief in the invalidity of interviews, interview validity coefficients were satisfactorily high overall. When the doctrine of idiosyncratic interview validities was examined, all variance in validity coefficients was explained by statistical artifacts for the unstructured interview (although considerable variance was left to be explained in structured individual interviews). By definition, no variance in unstructured interview validity coefficients could remain to be determined by any micro-process social psychological moderators of the type described by reviewers such as Schmitt (1976). Given that the doctrine of idiosyncratic interview validity was developed on the basis of research into the unstructured interview, the viability of this doctrine is seriously questioned by our findings.

The simplified conceptual model of interview validity developed here as an alternative to social psychological theory received partial support. These supportive findings require further summary and clarification. Although the ordering of mean validity coefficients predicted in hypothesis 1 was largely as expected, interview format (i.e. individual vs. board) had less impact in moderating interview validity than the previous literature would suggest. On the other hand, structured interviews had consistently and substantially higher predictive validity coefficients than unstructured interviews. In fact, the validity coefficients of structured interviews, both individual and board, are comparable with the best other predictors available to industrial/organizational psychologists, including mental ability tests (Hunter & Hunter, 1984).

The test of the second hypothesis produced findings contrary to prediction. In fact, the moderated meta-analysis showed that consensus board ratings were more

Table 4. Meta-analyses of the structured individual interview by job analysis type

Type of job analysis	Total sample size[a]	Mean validity coefficient[b]	Observed variance in sample correlations	Residual variance unexplained by statistical artifacts[c]	Per cent of variance explained by statistical artifacts[c]	95 per cent confidence interval of validity coefficient
Formal job analysis	1 328 (9)	0.87 (0.48)	0.114	0.084	26	$0.30 \leqslant P \leqslant 1.00$
'Armchair' job analysis	4 943 (10)	0.59 (0.35)	0.024	0.012	53	$0.38 \leqslant P \leqslant 0.80$
Unknown job analytic method	1 602 (13)	0.56 (0.31)	0.114	0.086	25	$-0.02 \leqslant P \leqslant 1.00$

[a]The number of coefficients contributing to the total sample is given in parentheses.
[b]Means were corrected for direct restriction of range and criterion unreliability; uncorrected values are given in parentheses.
[c]Variance in corrected sample correlations was adjusted for the following statistical artifacts: sampling error, differences in range restriction and differences in criterion reliability.

predictively valid than statistical combination of independent ratings. A sufficient number of studies was included to allow some confidence in these results (i.e. six coefficients from statistical and seven from consensus boards), although the small total sample size in the statistical board group warrants caution in interpreting the results. Given that our preliminary evidence indicates an advantage for consensus, experimental research is required to determine *why* board consensus ratings are more valid than statistical combination.

The results reported in this study cast considerable doubt on the doctrine of idiosyncratic interview validity, particularly for the unstructured interview. This conclusion must, however, be accompanied by two cautions. First, the limitations on the mean sample sizes and number of studies used in the moderated analyses of the unstructured interview leave open the possibility that other moderator variables could still impact to a limited extent without being detected by the meta-analysis procedure. As Sackett, Harris & Orr (1986) have demonstrated, meta-analysis may fail to detect true moderator effects with small mean sample sizes and few studies. Comparison of the results reported in the first five rows of Table 1 (i.e. meta-analyses of overall validity plus the four major interview types) to Sackett *et al.*'s power tables for the Schmidt–Hunter technique suggest that we had sufficient power (>0.80) to detect a population difference in correlations of 0.20 but insufficient power (<0.50) to find a population difference of 0.10. Therefore, small moderator effects might still occur for unstructured interviews although these effects would be of considerably less practical importance than the 0.31 difference in corrected correlations between unstructured and structured interviews reported in Table 2.

The second caveat to the finding that other moderators do not have an impact on interview validity coefficients concerns the results for structured interviews. For structured interviews, considerable variance was left to be explained after adjustment of the validity distribution for statistical artifacts. However, the division of structured boards into consensus and statistical types accounted for most of the remaining variance for the structured board type. As well, the *post hoc* breakdown of structured individual interviews by job analysis accounted for additional variance over considering structure alone. Therefore, even for structured interviews the doctrine of idiosyncratic interview validity which would postulate a large number of potential moderators is highly questionable.

We suggested in the introduction that the validity coefficients of interviews are moderated by two major factors–interview reliability and interview structure. On this basis we postulated a ranking of interview types in terms of validity. The results of this study supported the contention that reliability and interview structure affect the validity of interviews. First, improved reliability was invariably associated with higher interview validity. Second, interview structure was associated with a considerable increase in predictive validity. Interestingly, use of formal job-analytic information in developing structured individual interviews was also associated with higher validity. This result confirms predictions by a number of reviewers that using job analysis should improve interview validity (Arvey & Campion, 1982; Pursell *et al.*, 1980; Schmitt, 1976).

Based on our findings, we offer the following five recommendations for future I/O research and applied practice:

(1) Researchers should divert their attention from the unstructured to the structured interview. Although the unstructured board interview has modest validity and appears to be preferable to the unstructured individual interview, none of a multitude of other potential moderators identified in previous research (e.g. sex or race of rater/ratee; presence of ancillary information) could have greatly moderated the validity of unstructured interviews in this study. In fact, Webster (1982) indicates that further micro-research into these potential moderators is unlikely to yield any important new data on how to improve interview validity.

(2) Researchers should concentrate on identifying factors moderating the validity of the structured interview with the goal of maximizing the validity of the structured interview. Two promising factors identified in *post hoc* analyses are improvement of interview reliability and the use of formal job-analytic information for developing structured questions.

(3) Industrial/organizational psychologists should discard the 'received doctrine' of interview invalidity. Even the unstructured interview has modest validity which could produce considerable productivity gains over random selection.

(4) Practitioners should use the structured interview (as opposed to the unstructured interview) wherever possible. The structured interview questions should be based on formal job-analytic information and every attempt should be made to maximize the reliability of the structured interview.

(5) When using structured board interviews, the consensus board may be preferable to the statistical combination of individual board member ratings. Further research should be conducted to understand *why* consensus boards seem to yield superior prediction.

Acknowledgement

References

Anstey, E. & Mercer, E. O. (1956). *Interviewing for the Selection of Staff*. London: Allen & Unwin.

Arvey, R. D. & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology*, **35**, 281–322.

Barrett, G. V. (1972). Research models of the future for industrial and organizational psychology. *Personnel Psychology*, **25**, 1–18.

Bullock, R. J. & Svyantek, D. J. (1985). Analyzing meta-analysis: Potential problems, an unsuccessful replication, and evaluation criteria. *Journal of Applied Psychology*, **70**, 108–115.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Cronshaw, S. F. & Alexander, R. A. (1985). One answer to the demand for accountability: Selection utility as an investment decision. *Organizational Behavior and Human Decision Processes*, **35**, 102–118.

Cronshaw, S. F., Alexander, R. A., Wiesner, W. H. & Barrick, M. R. (1987). Incorporating risk into selection utility: Two models for sensitivity analysis and risk simulation. *Organizational Behavior and Human Decision Processes*, **40**, 270–286.

Dubin, R. (1976). Theory building in applied areas. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*, pp. 17–39. Chicago, IL: Rand McNally.

Hollingworth, H. L. (1922). *Judging Human Character*. New York: Appleton.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, **96**(1), 72–98.

Hunter, J. E., Schmidt, F. L. & Jackson, G. B. (1982). *Meta-analysis: Cumulating Research across Studies*. Beverly Hills, CA: Sage.

Landy, F. J. (1976). The validity of the interview in police officer selection. *Journal of Applied Psychology*, **61**(2), 193–198.

Maurer, S. D. & Russell, J. S. (1985). Validity of the employment interview revisited: A meta-analysis of existing research. Unpublished manuscript, University of Kentucky, Department of Management, Lexington and University of Oregon, Graduate School of Management, Eugene.

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

McNemar, Q. (1969). *Psychological Statistics*, 4th ed. Toronto: Wiley.

Milne, G. G. (1967). The interview: Let us have perspective. *Australian Psychologist*, **2**(2), 77–84.

Nunnally, J. C. (1978). *Psychometric Theory*, 2nd ed. New York: McGraw-Hill.

Oldfield, R. C. (1941). *The Psychology of the Interview*. London: Methuen.

Pursell, E. D., Campion, M. A. & Gaylord, S. R. (1980). Structured interviewing: Avoiding selection problems. *Personnel Journal*, **59**, 907–912.

Reilly, R. R. & Chao, G. T. (1982). Validity and fairness of some alternative selection procedures. *Personnel Psychology*, **35**, 1–62.

Rodger, A. (1952). The worthwhileness of the interview. *Occupational Psychology*, **26**, 101–106.

Rothstein, M. & Jackson, D. N. (1980). Decision making in the employment interview: An experimental approach. *Journal of Applied Psychology*, **65**(3), 271–283.

Rowe, P. M. (1981). The employment interview: A valid selection procedure. *Canadian Personnel and Industrial Relations Journal*, **28**(1), 37–40.

Sackett, P. R., Harris, M. M. & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, **71**, 302–310.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, **62**(5), 529–540.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C. & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology*, **64**, 609–626.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology*, **29**, 79–101.

Ulrich, L. & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin*, **63**(2), 100–116.

Wagner, R. (1949). The employment interview: A critical summary. *Personnel Psychology*, **2**, 17–46.

Webster, E. C. (1982). *The Employment Interview: A Social Judgement Process*. Schomberg, Ontario: SIP Publications.

Wright, O. R. (1969). Summary of research on the selection interview since 1964. *Personnel Psychology*, **22**, 391–413.

Zedeck, S., Tziner, A. & Middlestadt, S. E. (1983). Interviewer validity and reliability: An individual analysis approach. *Personnel Psychology*, **36**, 355–370.