Meta-analytic Estimates of Interview Criterion-related Validity:

A Qualitative Assessment

M. Ronald Buckley

Craig J. Russell

University of Oklahoma

Prepared for: Eder, R. W., & Harris, M. M. The Employment Interview: Theory, Research, and Practice, Sage, 1999.

Meta-analytic Estimates of Interview Criterion-related Validity:

A Qualitative Assessment

Research on the interview process has been conducted for many years and by many sage researchers. In fact, few selection techniques have received the attention which has been afforded the interview process (Eder & Buckley, 1988). Qualitative literature reviews conducted starting with Wagner (1949) consistently concluded that interviews yield unimpressive reliability and criterion-related validity relative to other selection techniques. In spite of this well-corroborated finding, Guion and Gibson (1988, p. 367) reported that criterion-related validation "research on interviewing continues, whether in desperation or hope." Management officials continue to treat the interview as an especially important component to selection systems and it continues to be the most frequently used selection device. This led Guion and Gibson (1988) to conclude, that "repeatedly discouraging summaries of their reliabilities and validities (have not) deterred the use of interviews." (p. 367)

The thrust of literature reviews changed considerably due to a qualitative review conducted by Harris (1989). Harris concluded that the interview was probably much more valid than previous reviewers had led us to believe. In addition, he suggested "there is a strong need to develop and apply theories in order to attain greater understanding of the interview process and outcomes" (p. 720) and that meta-analysis should be used to summarize past research. While the former has received scant attention, researchers have responded to the latter suggestion. During the last decade seven reviews of interview criterion-related validity have appeared, all of which contained meta-analyses of the empirical literature. The primary purpose of this chapter is to answer the question "What insight has been forthcoming from application of meta-analytic techniques to empirical interview research?" A secondary purpose is to explore directions for future research.

WHAT IS META-ANALYSIS AND WHAT DOES IT TELL US?

We will first present a brief review of meta-analysis procedures and how they differ from qualitative reviews before summarizing findings reported in the seven meta-analyses reported in the literature.

What is meta-analysis? Meta-analysis is a family of procedures designed to examine statistical effects reported across independent primary research studies. "Primary" research is simply research conducted on the phenomena of interest (e.g., interviews), while secondary meta-analytic research is conducted on some statistic of interest generated by primary research studies. Meta-analyses generically seek to 1) derive the best point estimate of the statistic of interest (e.g., interview criterion-related validity estimates, ρ_{xy}), 2) partition the observed variation in the statistic across studies into portions attributable to random sampling error, statistical artifacts, etc., and 3) derive the credibility interval of that statistic. The credibility interval is similar to confidence intervals, though the estimate of standard error has been corrected for sampling error.

For example, a meta-analyst may have 50 studies reporting the Pearson product moment correlation between candidates' overall interview ratings and subsequent job performance ratings, capturing the strength of the linear relationship between the two (r_{xy}). Each study's r_{xy} represents the best estimate of the true population correlation ρ_{xy} for that study. However, if the studies were drawn from a single population (i.e., there is only one true value of ρ underlying each study's results), some combination of these 50 r_{xy}'s should provide a more accurate estimate of ρ. Schmidt and Hunter (1990) derived an average r_{xy} weighted by sample size across studies such that:

$$\overline{r_{xy}} = \sum_{i=1}^{k} n_i r_i \Big/ \sum_{i=1}^{k} n_i ,$$

$$s_r^2 = \sum_{i=1}^{k} [n_i \{r_i - \bar{r}\}^2] \Big/ \sum_{i=1}^{k} n_i ,$$

and k = the number of studies. Schmidt and Hunter (1990) also noted that the expected variance in r_{xy} across studies due to random sampling error is:

$$s_e^2 = (1 - \bar{r}^2)^2 k \Big/ \sum_{i=1}^{k} n_i$$

Given that $\sigma_r^2 = \sigma_\rho^2 + \sigma_e^2$, one can estimate the true variance in ρ by subtracting σ_e^2 from both sides of the equation, or: $\sigma_\rho^2 = \sigma_r^2 - \sigma_e^2$. If in fact there is only one population value of ρ_{xy} underlying interview-criterion relationships in these k = 50 studies, then in the absence of other statistical artifacts, one would expect $\sigma_e^2 = 0$. Note that if in fact one value of ρ does exist, random chance would dictate that 50% of the estimates of σ_e^2 would be positive and 50% would be negative. Interpretations of the presence or absence of multiple ρ based on estimates of σ_ρ^2 are sometimes referred to as the "residualization" approach. Corrections for other statistical artifacts (range restriction in the criterion, measurement error, etc.) may also be performed, though most influence $\bar{r}$ directly.

Differences from Qualitative Reviews. Qualitative literature reviews conducted on selection interview validities, like many qualitative reviews, can be plagued by a number of problems not found (or less severe) in meta-analysis. A sampling of these problems include:

1. The "file drawer" problem (Glass, McGaw, & Smith, 1981). Qualitative literature reviews emphasized published research and typically did not sample unpublished research. Second order sampling error may bias results to the extent that unpublished studies were characterized by very different results (i.e., estimates of population criterion-related validity, ρ) from those obtained in published studies. Further, contrary findings may have been overlooked, especially if published in lower tier or obscure journals.
2. Most qualitative reviews merely report results of research absent any criticism of the research. All data may not have been collected in an appropriate, non-confounded fashion leading to unambiguous interpretation. Some would argue that qualitative reviews inherently focus on the more seminal and higher quality empirical work and are, thus, adequate representations of the empirical research. We would argue that this assigns considerable latitude to the discretion of the qualitative reviewer--a relatively unpredictable phenomenon.
3. The link between research findings and researcher characteristics is frequently ignored. Evidence suggests primary investigator decision processes are influenced by investigator values, attitudes, beliefs, knowledge, skills, ability, etc. (Russell, Settoon, McGrath, Blanton, Kidwell, Lohrke, Scifires, & Danforth. 1994; Sherwood & Nataupsky,

1968). Researchers conducting qualitative reviews are equally likely to make choices to ignore or include primary research in an unsystematic or biased fashion.

4. In data collected from actual interview situations measures of both the predictor variable(s) and the criterion variable(s) have error associated with them. Further, relationships between predictor variables and criterion variables typically are influenced by the range restriction that occurs in concurrent validity designs --- theoretically candidates actually hired tend to receive higher ratings. It is difficult to qualitatively "add together" and synthesize diverse findings when studies are characterized by differing degrees of predictor/criterion unreliability and range restriction.
5. Finally, none of the qualitative literature reviews examined whether differences in sample size accounted for variation in criterion-related validities reported across studies. Perplexing “mixed” results may in fact reflect variation in r_{xy} due to differences in sample size that, in turn, influence the precision with which each study estimates ρ.

Perplexing "mixed" results may in fact reflect variation in r_{xy} due to sampling error that, in turn, influence the precision with which each study estimates ρ. Inferences drawn by qualitative reviews will be spurious to the extent that they fail to consider the possibility that mixed results are due to sampling error.

Meta-analyses of Interview Criterion-related Validities. In order to minimize the aforementioned problems and heeding Harris' (1989) recommendation, seven researchers have performed meta-analyses of interview criterion-related validities. These results aid our insight into interview-criterion relationships by circumventing limitations noted with qualitative reviews, most notably, sampling error's ability to obscure the true strength of latent predictor-criterion relationships.

Table 1 contains a summary of meta-analytic results reported in the literature. Some of the meta-analyses overlap meaningfully in their sampling of primary studies. Choice of variables on which to subgroup studies in search of potential moderators did not appear to be theory driven. Most subgrouping variables seemed to be selected on the basis of convenience (i.e., the variable was noticed and subsequently coded when validities were "harvested" from the primary research studies) or some a priori methodological reason.

Insert Table 1 About Here

A number of patterns in ρ and $\bar{r}$ are of interest. First, note that, with the exception of a portion of Hunter and Hunter's (1984) results, all estimates of ρ or $\bar{r}$ are approximately .20 or higher. Hunter and Hunters' (1984) findings have been previously criticized by Roth and Campion (1992) for failure to correct for range restriction and potential second order sampling error (i.e., too few primary research studies). This suggests interview evaluations are not the poor predictors they are made out to be in qualitative reviews. We would label this level of criterion-related validity "moderate" in comparison to meta-analytic estimates in the .30-.40 range reported for cognitive ability tests, biodata information inventories, and assessment centers (Gaugler, Rosenthal, Thornton, & Benson, 1987; Russell, C.J. & Kuhnert, K. W., 1992).

Second, the presence of "structure" in the interview appears to coincide with meaningful increases in interview criterion-related validity. Huffcutt (1992) defined structure as a reduction in procedural variability across applicants. Huffcutt and Arthur (1994) content-analyzed structure in terms of scoring and question standardization to create four a priori levels. At one extreme, Level I structure was characterized by no constraints in question standardization and global assessment. In contrast, Level IV structure was characterized by requiring the exact same questions to be asked with no deviations or custom follow-up questions and scoring of each individual response using pre-established answers.

Results summarized in Table 1 suggest validity increases from .20 to .57, a net gain of .37, in using Level I versus Level IV interview structure. These findings underscore and expand on the prior observation on absolute levels of interview criterion-related validities. The two most comprehensive meta-analyses reported by Huffcutt and Arthur (1994) and McDaniel et al (1994) suggest criterion-related validities for structured selection interviews to be "large," i.e., greater than .40.

Do Meta-analytic Findings Suggest the Need for More Interview Research?

A number of authors have argued that meta-analytic results such as those reported in Table 1 resolve almost all ambiguity regarding criterion-related inferences one might wish to draw about a selection procedure (Schmidt, 1996). Hunter and Schmidt (1990) argued that when variance due to statistical artifacts constitutes a large portion of observed variance in criterion-related validities across studies (most notably variance due to random sampling error, i.e., $\sigma_\rho^2 = \sigma_r^2 - \sigma_e^2$), one can assume subjects in the studies were drawn from a single population characterized by one value of ρ. The common decision heuristic is to assume “validity generalizes” if statistical artifacts (e.g., sampling error, range restriction, measurement error) explain at least 75% of the observed variance in effects sizes (note, no probability distribution of $\bar{r}$ is derived in this literature, so no probabilistic inferences about $\bar{r}$ can be drawn, Thomas, 1989). As noted above, this has come to be known as the residualization approach.

If one adopts this perspective, Table 1 results permit at least three inferences. First, as noted above, structure in interview format is important, moderating average effect sizes obtained ($\bar{r}$). Second, when structured, interviews do yield validities comparable to estimates of $\bar{r}$ reported for cognitive ability tests, biographical information inventories, or assessment centers (Russell & Kuhnert, 1992). However, it is important to note that too few studies simultaneously examined various interview formats and alternate selection technologies to estimate, for example, interviews unique contribution to criterion-related validity in the presence of cognitive ability tests, scored biographical information inventories, or assessment center ratings. Third, with the exception of interview format and type of performance criterion (cf McDaniel et al., 1994), all variation in observed criterion-related validities appears to be due to statistical artifacts (i.e., sampling error).

Again, adopting this residualized view of meta-analytic results, Table 1 yields implications for practice that deviate greatly from conclusions drawn by qualitative reviews. However, this implication will not change practice as firms have happily continued to use interviews in spite of conclusions drawn from prior qualitative reviews (note that a number of primary researchers, e.g., Latham and colleagues, have made compelling arguments for structured interview formats for over 15 years, Latham & Saari, 1984; Latham, Saari, Pursell, & Campion, 1980).   Guion and Gibson recommended a discontinuation of interview research due to low criterion-related validities almost 10 years ago. Should we draw the same conclusion on the basis of meta-analysis findings indicating generalizable high validities for structured interviews? Alternatively, have important questions gone unanswered?

We feel neither individual primary research or meta-analytic conclusions drawn from secondary analyses address a number of important issues. The remainder of this chapter will outline why suspension of primary research activities would be premature and a programmatic guide to future research efforts.

Why Meta-analytic Results Do Not Tell Us All We Need to Know. A lack of agreement is starting to emerge concerning conclusions drawn from the residualized view of meta-analytic results (i.e., where inferences are drawn from $\bar{r}$ and σ_{ρ}^{2} after adjustments for statistical artifacts). James, Demaree, Mulaik, and Ladd (1992) demonstrated that meta-analysis systematically disguises true moderator effects when those moderators covary with statistical artifacts controlled for in performing the meta-analysis (cf. Burke, Rupinski, Dunlap, & Davison, 1996). Burke et al. tested James et al.'s notion and failed to find support. Unfortunately, the Burke et al. test meta-analyzed relationships observed between measures of job satisfaction and job performance. Specifically, previous meta-analytic estimates of job satisfaction-job performance correlations are extremely low (e.g., Iaffaldano & Muchinsky, 1985, reported $\bar{r}$ = .05 for pay satisfaction). If a true moderator is present, it would have to be of the strongest order, where a slight majority of moderator levels yield observations characterized by progressively stronger positive satisfaction-performance relationships, while the remaining moderator levels would yield observations characterized by progressively stronger negative relationships. Russell and Gilliland

(1995) demonstrated how meta-analytic evidence of moderator presence does not necessarily yield any insight into moderator processes. Just because the effect size covaries with some variable (e.g., degree of interview structure) does not mean that variable is the true cause of moderation.

Further, and perhaps most disturbing, are results suggesting that meaningful moderator effects can be present even when meta-analysis suggests statistical artifacts account for all variance in effect sizes. This could mean, for example, that even though meta-analyses suggest almost all variance in effect sizes for structured interviews is due to sampling error, some situational variable could cause wide swings in observed criterion-related validities. In this regard, one line of inquiry is of particular interest to inferences regarding interview criterion-related validity. Specifically, a number of literatures have shown that the purpose or source of motivation driving decision situations greatly influences both cognitive processes and decision outcomes. For example, Longenecker, Sims, and Gioia (1987) in a qualitative analysis demonstrated that performance appraisal purpose was related to appraisal outcome. Using interview decision environments, Adkins, Russell, and Werbel (1994) reported congruence between recruiter and applicant work values was related to recruiter assessments of general employability and organizational "fit." Hence, prior findings suggest interview motivation remains a prime moderator candidate of interview validities in spite of meta-analytic findings.

Perhaps most troublesome in light of the meta-analytic results reported above, empirical evidence suggests key research decisions are influenced by researcher source of motivation in conducting the study to begin with. For example, Russell et al. (1994) found that characteristics of investigators publishing criterion-related studies in the Journal of Applied Psychology and Personnel Psychology between 1965 and 1992 predicted size of criterion-related validity reported (Russell et al. only corrected for sampling error while controlling for type of predictor, criterion, job type, and study design). It is very easy to become cynical about applied social science when Russell et al. report $\bar{r}$ = .218 and .331, respectively, for investigators employed in academic settings attempting to test or develop some theory of performance prediction versus investigators employed in industry attempting to document compliance with EEO guidelines. At this point, there is no reason to believe researchers examining interview criterion-related validity are immune to these influences.

Russell et al. concluded that "if the universe of all criterion-related validity studies ever conducted were included, (meta-analytic) results can still be influenced by the capabilities and motivational agendas of the original investigators. . ." (1994, p. 169). Russell et al. (1994) suggested that subtle aspects of investigator decision making are influenced by their capabilities and motivational agenda resulting in enhanced or attenuated estimates of r. Wanous, Sullivan, and Malinak (1989) made the same observation about judgment calls made by meta-analytic investigators using archival secondary data. In a similar vein, we find it curious that no mention is made of Sherwood and Nataupsky's (1968) finding that demographic characteristics of primary investigators predicted differences in black-white intelligence test scores reported in published research given recent attention to racial differences in cognitive ability (Herrnstein & Murray, 1995). Again, there is no reason to believe either interviewers or researchers examining interview criterion-related validity are immune to these influences.

Additional Research Needs

Simply stated, meta-analysis cumulates its own set of shortcomings in addition to effect sizes. For example, consider the researcher facing hypothetical choices between expending resources needed to acquire information available from a single primary research effort with sample of size N versus information available from a meta-analysis of k studies where $Sn_i = N$. In the former circumstance, internal and external threats to validity of inferences drawn are well known and documented (Cook & Campbell, 1978). Given a priori theory and/or a body of prior research findings, the investigator can estimate expected effect sizes and derive the probability of Type II error in any parametric statistical inferences. In the latter, meta-analytic circumstance:

1. internal and external threats to inference validity are cumulated across k studies - it is unlikely that these threats to validity are counterbalanced across studies in such a way that they sum to zero;
2. those threats to inference validity are likely not to be independent (reviewers and editors have this nagging tendency to require authors to demonstrate how their research reflects and extends past research -- we cannot think of any behavioral science literature in which large numbers of pure replications are conducted or published); and ,
3. errors due to the over zealous interpretations of meta-analytic results such as those described above (cf. Schmidt, 1992, 1996). These include failure to detect true moderators due to confounding with statistical artifacts and failure to detect true moderators due to over interpretation of the 75% "rule."

In contrast, “critical” tests of competing theoretical predictions in primary research can shed insight without subsequent “sanction” of meta-analytic inferences (Greenwald, 1975). In the presence of prior research findings or strong theory, a priori estimation of expected effect sizes permit researchers to derive estimates of samples sizes (N) required for adequate statistical power for tests of research hypotheses. The specific directions for research outlined below constitute a programmatic effort to leverage existing knowledge to discover facets of interview content, context, or process influencing criterion-related validities. While strong theories of work performance are not available to guide future primary research (Campbell, 1990), a programmatic effort at grounded theory building (Glaser & Strauss, 1967) should permit development of theories or models characterized by strong conceptual and operational definitions (Greenwald, 1975).

Clearly, we need a better understanding of how investigators values, beliefs, and motivation influence outcomes of interview criterion-related validity research. Meta-analyses comparable to those reported by Russell et al. (1994) would be most useful in partitioning sources of variance in interview criterion-related validities. Once identified, investigator teams can be constructed to minimize or control for these influences, thus ensuring any observed variation in future primary research results is due to characteristics of the interview, job, candidate, interviewer, etc., and not some characteristic of the investigator. Latham, Erez, and Locke (1989) powerfully demonstrated how investigator teams can be constructed to resolve, through primary research initiatives, critical key issues and create new knowledge (in the absence of a meta-analysis).

What Should We Do Next?

Meta-analyses results suggest interviews are more valid than originally perceived by qualitative reviews of the literature and that structure is an important asset to an interview. In spite of this we would suggest that meta-analytic results should be examined with caution

- it would be inappropriate to conclude research on interview criterion-related validity and situational moderators should be discontinued.   Using different combinations of variables for every situation, unstructured interviews may yield higher criterion-related validity in some selection contexts. Research should continue to examine situational moderators (and those moderators which may possibly influence research) in theory-based, programmatic efforts to understand the role interview information play in latent models of performance prediction.

We would like to suggest that the meta-analytical database developed to date has limited usefulness in understanding why interviews yield higher criterion-related validities. Average interview criterion-related validities suggest they are useful. Premature infatuation with meta-analytic interpretations may have contributed to a situation where the field relied too heavily on procedures and methods and too little on theory development. Bechtoldt (1959) made an observation in a different measurement context that seems an appropriate response to those who feel meta-analysis is the only means of advancing psychological theory (e.g., Schmidt, 1992): ?To admit ignorance as an (sic) temporary state of science is one thing. To raise vagueness or lack of definition to the central status of a methodological principle is another.? Similarly, in spite of repeated nominations as a ?methodological principle? poised to reveal previously unseen truth and beauty (Schmidt, 1992, 1996), meta-analysis has not advanced our understanding of why interviews work. Marchese and Muchinsky (1993) admonished, "we should resist the temptation to produce singular coefficients with the accompanying appellation that they are estimates of the truth." (p. 25)

The next logical step in understanding interview processes and how they might differ across situational contexts does not involve meta-analysis . What is needed is a return to theory-driven sequence of empirical primary research efforts to incrementally 1) eliminate alternative explanations via “critical” tests (Greenwald, 1975) 2) construct better models of human performance in organizations. While the meta-analyses have given us an idea of where interview research has been it has not facilitated the development of any important theoretical insight concerning the directions in which this research should go.

Promising Moderator Candidates and Questions

We believe that there are a number of moderator variables and research questions. While some moderator candidates were included in previous meta-analyses, we feel they require further investigation. Some of these variables are discussed at greater length in other chapters, others were selected on the basis of our qualitative assessment of the existing literatures in behavioral science. They include:

1. Decision Risk - The costs, both positive and negative, associated with selection decisions. The reader need only consider the large literature on prospect theory (Tversky & Kahneman, 1981) for evidence that perceived likelihood of positive versus negative outcomes influence decision making in fundamentally different ways.
2. Interview Task Clarity - The degree to which the selection task is unambiguous and the interviewer is prepared for the selection task.
3. Interview Purpose/Interviewer Motivation - The degree to which the purpose of the interview or interviewers motivation is related to the selection outcome.
4. Candidate Quality - The true quality of the interviewees certainly influences the dynamics of the interview process. Again, the cognitive psychology literature provides models of circumstances where cues and cue weights are not independent (Nisbett & Ross, 1980) that might guide research on this process in the interview context.

5. Research Study Sample - Real interviewers may well use different judgement criteria than those decision rules used by undergraduate students.
6. Participant Acceptability of the Interview Process - The degree to which practicing managers believe the interview process is an effective methodology to use may influence criterion validities, as may candidate perceptions of the interview.
7. Incremental Validity - An interview is rarely the sole component of the selection process. Investigations of construct domain overlap with other selection technologies are needed.

There are many potential moderators of interview validity. Meta-analyses shed little light on how interview processes influence interview outcomes. Moderators examined within meta-analyses reflect where the field has been - ad hoc, atheoretical examinations of criterion-related validity. Meta-analytic results cause us to echo Harris' (1989) call for theory development, though the best means of doing so seems to be through programmatic primary research.

References

Adkins, C.L., Russell, C.J., & Werbel, J.D. (1994). Judgments of fit in the selection process: The role of work value congruence. Personnel Psychology, 47, 605-623.

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

Bechtoldt, P. (1959). Construct validity: A critique. American Psychologist, 14, 619-629.

Burke, M.J., Rupinski, M.T., Dunlap, W.P., & Davison, H.K. (1996). Do situational variables act as substantive causes of relationships between individual difference variables? Two large-scale tests of “common cause” models. Personnel Psychology, 49, 573-598.

Campbell, J.P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M.D. Dunnette and L. Hough (Eds.), Handbook of Industrial and Organizational Psychology (2nd ed., vol. 1, pp. 687-732). Palo Alto, CA: Consulting Psychologist Press.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A Meta-analysis of Interrater and Internal Consistency reliability of Selection Interviews. Journal of Applied Psychology, 80, 565-579.

Cook, T. & Campbell, D. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago, IL: Rand-McNally.

Eder, R. W., & Buckley, M. R. (1988). The employeemnt interview: An interactionist perspective. In G. R. Ferris and K. M. Rowland (Eds.) Research in Personnel and Human Resources Management (vol 6, pp. 75-108). Greenwich, CT: JAI Press.

Gaugler, B.B., Rosenthal, D.B., Thornton, G.C. III, & Benston, C. (1987). Meta-analysis of assessment center validity. Monograph. *Journal of Applied Psychology, 72*, 493-511.

Glaser, B. G., & Strauss, A. L. 1967. *The discovery of grounded theory: Strategies for qualitative research.* Chicago: Aldine Publishing Co.

Greenwald, A.G. (1975). On the inconclusiveness of “crucial” cognitive tests of dissonance versus self-perception theories. *Journal of Experimental and Social Psychology, 11*, 490-499.

Hernstein

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, TX.

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

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

Iaffaldano, M.T., & Muchinsky, P.M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin, 97*, 251-273.

James, L.R., Demaree, R.G., Mulaik, S.A., & Ladd, R.T. 1992. Validity generalization in the context of situational models. *Journal of Applied Psychology, 77*, 3-14.

Latham, G.P., Erez, M., & Locke, E.A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez-Latham dispute regarding participation in goal setting. Journal of Applied Psychology, 73, 753-772.

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

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

Nisbett, R. & Ross, L. (1980). Human inference: Strategies and shortcoming of social judgment. Englewood Cliffs, NJ: Prentice-Hall, Inc.

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

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.

Russell, C.J. & Gilliland, S.W. (1995). Why meta-analysis doesn't always tell you what the data really mean. Journal of Management, 21, 813-831.

Russell, C.J. & Kuhnert, K.W. (1992). New frontiers in management selection systems: Where measurement technologies and theory collide. Leadership Quarterly, 3, 109-135.

Russell, C.J., Settoon, R.P., McGrath, R.N., Blanton, A.E., Kidwell, R.E., Lohrke, F.T., Scifires, E.L., & Danforth, G.W. (1994). Investigator characteristics as moderators of personnel selection research: A meta-analysis. Journal of Applied Psychology, 79, 163-170.

Schmidt, F.L. (1992). What do the data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. American Psychologist, 47, 1173-1181.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115-129.

Sherwood, J.J. & Nataupsky, M. (1968). Predicting the conclusions of negro-white intelligence research from biographical characteristics of the investigator. Journal Personality and Social Psychology, 8, 53-58.

Tversky, A. & Kahneman, D (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458.

Weisner, W. H., & Cronshaw, S. F. (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.

Wanous, J.P., Sullivan, S.E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. Journal of Applied Psychology, 74, 259-264.

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.

Table 1

Meta-analyses of Interview Criterion-Related Validities

Study	K	ΣN_i	$\bar{r}$	s^2_r	s^2_e	s^2_r	r	
Machese & Muchinsky (1993)								
Subjective Criteria	23	2,290	.248	.025	.006	.036	.368	
Objective Criteria	12	1,875	.287	.024	.003	.034	.391	
Wright, Lichtenfels, & Pursell (1989)	13	827	.260	.028	.014	.014	.340	
Huffcutt & Arthur (1994)								
Structure level I	15	7,308	.200	-	-	.0064	-	
Structure level II	39	4,621	.350	-	-	.0324	-	
Structure level III	27	4,358	.560	-	-	.0400	-	
Structure level IV	33	2,365	.570	-	-	.0784	-	
Hunter & Hunter (1984)								
Reanalysis of Dunnette (1972)	30	-	-	-	-	-	.160	
Reanalysis of Reilly & Chao (1982)		11	-	-	-	-	-	.230

Table 1 continued

Study	K	ΣN_i	$\bar{r}$	s^2_r	s^2_e	s^2_r	r
Hunter & Hunter (1984)							
Supervisor Ratings	10	2,694	-	-	-	.0121	.140
Promotion	5	1,744	-	-	-	.0000	.080
Training Success	9	3,544	-	-	-	.0049	.100
Tenure	3	1,925	-	-	-	.0000	.030
McDaniel et al. (1994)[1]							
Job Performance Criterion Measures							
Structured	106	12,847	.240	.0324	-	-	.440
Test information available	9	1,031	.090	.0121	-	-	.160
Test information unavailable	36	4,865	.220	.0400	-	-	.400
Unstructured	39	9,330	.180	.0121	-	-	.330
Test information available	5	433	.180	.0036	-	-	.340

Table 1 continued

Study	K	ΣN_i	$\bar{r}$	s^2_r	s^2_e	s^2_r	r
Test information unavailable	9	1,854	.32	.0144	-	-	.570
Training Performance Criterion Measures							
Structured	26	3,576	.210	.0144	-	-	.340
Unstructured	30	47,576	.230	.0064	-	-	.360
Reilly & Chao (1982)[2]	12	987	.190	-	-	-	-

Note, r = estimate of population criterion-related validity corrected for statistical artifacts (usually measurement reliability and range restriction), $\bar{r}$ = estimate of population criterion-related validity uncorrected for statistical artifacts.

1. All of McDaniel et al.?s (1994) results are not summarized here, as they also compiled meta-analytic results by type of criteria and interview content, structure, and purpose.

2. Based on Reilly and Chao?s (1982) Table 3, most studies used structured or semi-structured interviews.