CHAPTER 12

RETHINKING THE VALIDITY OF INTERVIEWS FOR EMPLOYMENT DECISION MAKING

Implications of Recent Developments in Meta-analysis

In-Sue Oh, Bennett E. Postlethwaite, and Frank L. Schmidt

Employment interviews are one of the most widely used selection tools across organizations, industries, and countries (Dipboye, 1992, 1997; Dipboye & Jackson, 1999; Ryan, McFarland, Baron, & Page, 1999; Salgado, Viswesvaran, & Ones, 2001; Wilk & Cappelli, 2003, Table 1). Interviews also play an important role in government employment decisions, particularly at the Federal level (U.S. Merit Systems Protection Board, 2003). Likewise, employment interviews have long been a focus of both laboratory (e.g., Highhouse & Bottrill, 1995; Motowidlo & Burnett, 1995; Paunonen, Jackson, & Oberman, 1987; Purkiss, Perrewe, Gillespie, Mayes, & Ferris, 2006) and field (e.g., Chapman & Zweig, 2005; Maurer & Solamon, 2006; van der Zee, Bakker, & Bakker, 2002) research. Although the use of employment interviews is widespread, a wealth of research indicates that not all interviews are equally valid predictors of future job performance. In particular,

Received Wisdom, Kernels of Truth, and Boundary Conditions in Organizational Studies, pp. 297–329

discrepancies in validity have been frequently observed in regard to interview structure. Both narrative (e.g., Arvey & Campion, 1982; Campion, Palmer, & Campion, 1997) and meta-analytic reviews (Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994) have consistently demonstrated that structured interviews are superior in validity to unstructured interviews to varying extents, leading Campion et al. (1997) to conclude that "in the 80-year history of published research on employment interviewing, … few conclusions have been more widely supported than the idea that structuring the interview enhances reliability and validity" (p. 655). Chapman and Zweig (2005) stated, "what is evident from Campion et al.'s seminal article, and those on which it was based, is that one could easily replace the term *structured interview* with *good interview* or a *valid interview*" (p. 675).

Au: Is emphasis in original quote?

No, please add, "italic added" after p. 675

In their comprehensive narrative review, Campion et al. (1997) broadly define structure as "any enhancement of the interview that is intended to increase psychometric properties by increasing standardization or otherwise assisting the interviewer in determining what questions to ask or how to evaluate responses" (p. 656). Campion et al. identified 15 elements of interview structure and provided predictions regarding how applicants and interviewers should react to and utilize those elements. However, Chapman and Zweig (2005), based upon a very large-scale two-sample field survey, found that interview structure was best described by only four elements out of Campion et al.'s 15 elements: (a) questioning consistency, (b) evaluation standardization, (c) question sophistication, and (d) rapport building. Structured employment interviews can take many forms, with behavioral description interviews and situational interviews among the most common (Dipboye, 1997).

Three major meta-analyses (Huffcutt & Arthur, 1994; McDaniel et al., 1994; Wiesner & Cronshaw, 1988) have addressed the issue of interview structure, with each concluding that structured interviews are more valid than their unstructured counterparts, albeit to varying degrees, in predicting job performance. In particular, in the most comprehensive meta-analysis currently available, McDaniel et al. (1994) estimated an overall mean operational validity[1] of .44 (k = 106, N = 12,847) for structured interviews compared with .33 (k = 39, N = 9,330) for unstructured interviews in predicting job performance. Although not well known to many researchers and practitioners, McDaniel et al. (1994) also reported that the operational validity of unstructured interviewers (ρ = .36, k = 30, N = 45,576) is slightly higher than that of structured interviews (ρ = .34, k = 26, N = 3,576) in predicting training performance.

Although the superior operational validity of structured interviews may appear well established, several researchers interested in research synthesis methodologies have recently begun to challenge this widely held belief

(e.g., Duval, 2005; Le & Schmidt, 2006). In the present paper, we introduce new evidence based on more accurate calibrations of range restriction (Hunter, Schmidt, & Le, 2006) and publication bias (Duval, 2005; Duval & Tweedie, 2000a, 2000b), both of which are known to attenuate and/or distort validity (Rosenthal, 1979; Rothstein, Sutton, & Borenstein, 2005a; Thorndike, 1949). The methods described and illustrated in this chapter are relatively new. To date there has been no published research that has simultaneously incorporated both of these methodological advancements when estimating the operational validity of employment interviews. Accordingly, in the current study we correct/adjust for indirect range restriction, publication bias, and criterion unreliability in order to reestimate and rethink the operational validities of structured and unstructured interviews for predicting both job and training performance using the data set used in McDaniel et al.'s (1994) meta-analysis.[2] We begin by reviewing two recent methodological advancements in meta-analysis, indirect range restriction and publication bias.

Range Restriction

Even after corrections for criterion unreliability have been made, the validity of some selection tools often remains an underestimate due to restriction of range (Thorndike, 1949). Range restriction is a very common problem in validation studies, and the phenomenon has been widely investigated in personnel selection research (Hunter et al., 2006; Raju & Brand, 2003; Ree, Caretta, Earles, & Albert, 1994; Sackett & Yang, 2000; Thorndike, 1949). Validities estimated in range restricted samples are biased downward, and in this sense range restriction is a form of biased sampling (Sackett & Yang, 2000). Researchers and practitioners often want to estimate the validity of a selection tool (e.g., the Graduate Management Admissions Test [GMAT]) for an unrestricted population (the applicant population), but have data only for the restricted population (the incumbent or admitted student population in the case of the GMAT). Typically, the incumbent population has a higher mean score and a smaller standard deviation than the applicant population, thereby creating a biased sample (Thorndike, 1949). Validity estimates in such biased samples are attenuated, so it is necessary to correct (disattenuate) the validity estimates for these biasing effects due to range restriction. However, in order to do so, it is first necessary to differentiate between two types of range restriction, direct and indirect (Thorndike, 1949), as each type requires different correction methods (Hunter et al., 2006).

It is useful to examine range restriction using an illustrative example. Suppose that all applicants with a GMAT scores above 650 are admitted

to a MBA program, and all applicants scoring below 650 are rejected without exception. This scenario represents *direct* or *explicit* range restriction (DRR). In direct range restriction situations, it is assumed that applicants are selected both solely and directly on their selection procedure scores in a "top-down" manner. Most validation studies to date have corrected for DRR (Hunter & Schmidt, 2004). In reality, however, MBA admission decisions, like most employment decisions, are not solely based on a single test score, but rather are made using multiple (quantified and un-quantified) variables correlated with GMAT scores (such as a composite score of work experience, statement of purpose, undergraduate GPA, and letters of recommendation) (Oh, Schmidt, Shaffer, & Le, 2008). This latter scenario represents a case of *indirect* or *implicit* range restriction (for a figurative distinction between DRR and IRR, see Oh et al., 2008, Figure 1). Stated another way, we rarely use only one predictor in a complete top-down manner, but rather decisions are made using other sources of information in virtually all selection situations (Hunter et al, 2006). Furthermore, because applicants typically apply simultaneously for multiple MBA programs and jobs, some applicants will obtain multiple offers of admission and employment. This violates an underlying assumption of DRR. In such situations, MBA programs or hiring organizations may not obtain acceptances from those applicants at the high end of the score distribution thereby prohibiting true top-down admission or selection decisions. That is, in reality, range restriction is almost always indirect, as Thorndike (1949) noted more than half the century ago. Nevertheless, most researchers have believed that correcting for direct range restriction, even when the restriction is known to be indirect, provides corrections that are of satisfactory accuracy. This erroneous belief is still prevalent (Hunter et al., 2006; Oh et al., 2008).

Correction methods for both direct and indirect range restriction have been available since the late 1940s (Thorndike, 1949). However, Thorndike's Case III correction method for indirect range restriction (IRR) requires information and assumptions unavailable and unrealistic in almost all studies (Hunter et al., 2006; Linn, Harnisch, & Dunbar, 1981; Thorndike, 1949). Due to this lack of information, researchers have instead used the correction method for DRR (Thorndike's Case II), which is simple and easy to apply. Researchers were aware that the Case II correction leads to underestimation; however, they believed that this underestimation was not substantial (e.g., Berry, Sackett, & Landers, 2007; Le & Schmidt, 2006; Schmidt, Oh, & Le, 2006). However, this traditional belief was discovered to be erroneous when a new correction method for IRR (Case IV) recently developed by Schmidt and his colleagues was applied to previous large-scale databases (e.g., Hunter et al., 2006; Schmidt et al., 2006).

This new IRR method was analytically derived (see Hunter et al., 2006 for statistical derivations) and shown to be accurate via extensive Monte-Carlo simulation studies (Le & Schmidt, 2006). This method is incorporated in the Hunter-Schmidt meta-analysis programs (Schmidt & Le, 2004), the most technically advanced psychometric meta-analysis software programs available (Roth, 2008). The new IRR methods have also been applied to several existing large databases and the reanalysis results have been published in several major journals (e.g., *Journal of Applied Psychology*, *Personnel Psychology*, *Psychological Methods*, and *Academy of Management Learning & Education*). First, when Hunter et al. (2006) applied this new correction method to a well-known large-sample database (United States Employment Service's General Aptitude Test Battery), they found that previous meta-analyses have underestimated the operational validity of the general mental ability (GMA) test in predicting job performance by approximately 25% (.51 for DRR corrections versus .68 for IRR corrections for medium-complexity jobs). Subsequent meta-analyses (Schmidt et al., 2006) have also confirmed this finding for various cognitive measures across diverse job families for both job and training performance. This is a significant finding because GMA is the single best predictor of job performance and its dominance compared to other predictors in terms of validity has been a central foundation in theories of job performance (Borman, White, Pulakos, & Oppler, 1991; Campbell, McCloy, Oppler, & Sager, 1993; Ree, Earles, & Teachout, 1994; Schmidt & Hunter, 1998; Schmidt & Hunter, 1992). Further, operational validity estimates of the two most valid personality predictors important across most selection situations (conscientiousness and emotional stability; Barrick, Mount, & Judge, 2001) were also found to be underestimates by about 5% (Schmidt, Shaffer, & Oh, 2008) due to the previous application of suboptimal range restriction correction methods. Oh et al. (2008) also found that the operational validity estimate of the GMAT has been underestimated by 7% due to the application of suboptimal range restriction corrections.

Relative to our focus in the current chapter on interviews, Le and Schmidt (2006) further applied this new range restriction correction method (called Case IV) to McDaniel et al.'s (1994) interview data set on job-related structured and unstructured interviews. In McDaniel et al.'s original meta-analysis, which used DRR correction methods, the difference between the operational validities for structured and unstructured interviews was .11 (.44 vs. .33) or a 25% difference, while it was .03 (.44 vs. 41) or a 8% difference in the reanalysis conducted using the new IRR correction method (Le & Schmidt, 2006, see Table 4 for details). Le and Schmidt (2006) concluded that "the use of inappropriate meta-analysis methods for range restriction led to erroneous conclusions

about the relative magnitude of the validities of structured and unstructured interviews" (p. 433). In sum, counter to widely held beliefs, it appears that the operational validity estimate of structured interviews is not much greater than that of unstructured interviews.

Publication Bias

According to Rothstein, Sutton, and Borenstein (2005b), "publication bias is the term for what occurs whenever the research that appears in the published literature is systematically unrepresentative of the population of completed studies" (p. 1). This bias is what Rosenthal (1979) refers to as the "file-drawer problem," or the inability to detect studies that were conducted but never actually reported (thus "hidden" in researchers' filing cabinets). Bias may take the form of a disproportionate number of published studies that are either statistically significant or have a large, positive effect size (Greenwald, 1975; Lipsey & Wilson, 1993), and "the concern here is that studies revealing smaller effect sizes are systematically censured from the published literature" (Cooper, 2003, p. 3). However, publication bias may not be the sole source of bias: Study results may also be suppressed for a variety of other reasons apart from the inability to gain acceptance for publication (Duval & Tweedie, 2000a, 2000b; Hunter & Schmidt, 2004). Hunter and Schmidt (2004) maintain that "even among unpublished studies, those that are retrievable may not be representative of all unpublished studies" (p. 492). For example, privately funded research might only be released to the public if the results are determined to be beneficial to the sponsor. Given this, some researchers (e.g. Hunter & Schmidt, 2004; Lipsey & Wilson, 1993) prefer to use the broader term "availability bias."

To address the file-drawer problem, Rosenthal (1979) developed a test (which subsequently became known as the failsafe *N*) to determine the number of studies necessary to reduce a statistically significant research effect to a level of nonsignificance (Becker, 2005; McDaniel, Rothstein, & Whetzel, 2006). Orwin (1983) extended Rosenthal's methodology to cover various effect sizes in meta-analysis. His method estimates the number of missed (or "file-drawer") studies required to reduce an effect size to some critical value which is "the smallest mean value that we would consider theoretically or practically significant" (Hunter & Schmidt, 2004, p. 500). However, research has suggested that failsafe *N* methods lack the sensitivity to detect publication bias in meta-analysis in some cases where it is known to exist (Becker, 1994, 2005; McDaniel et al., 2006).

Au: The prefix *non* does not require hyphen in APA, unless base word is capitalized, an abbreviation, a number, or more than one word.

Okay, thanks!

Given the low statistical power of the failsafe *N* methods, a number of other methods have been developed to assess publication bias in meta-

analysis. One of the most widely used of these is the funnel plot (Light & Pillemer, 1984). Funnel plots are constructed by plotting individual study effect sizes on the horizontal graph (X) axis against measures of individual study sampling error variance, standard error or precision (= 1 / standard error) on the vertical graph (Y) axis. If there is no publication bias, the plot is symmetrical (like an inverted funnel) with the true mean effect size in the center. However, when studies are missing, the plot is asymmetrical, suggesting the presence of publication bias. According to Duval and Tweedie (2000a), the funnel shape is based not completely on statistical modeling but rather on two empirical observations. First, sampling error variances in a meta-analysis tend to be distributed in a manner such that there are fewer precise studies (larger sample sizes) and more imprecise studies (smaller sample sizes). Second, at any fixed level of sampling error variance, studies are distributed around the true mean effect size. Figure 1 presents a funnel plot of the structured interview validity estimates for predicting job performance using McDaniel et al.'s (1994) interview data set. The open circles in the plot represent individual study effect sizes plotted against their respective standard errors. While funnel plots can be useful for illustrating publication bias, a significant limitation is that their interpretation is subjective and prone to interpretive errors (Duval & Tweedie, 2000a, p. 456).

More recently, researchers have developed more advanced quantitative methods to detect possible publication bias [e.g., Begg & Muzumdar's (1994) rank correlation test, Egger, Smith, Schneider, and Minder's (1997) regression-based test]. However, these tests suffer from low statistical power and are often not sensitive enough to detect bias (Duval & Tweedie, 2000a). Furthermore, neither method provides an estimate of the number of studies potentially missing from the study population or a validity estimate that is adjusted for the effect of the missing studies (Duval & Tweedie, 2000b). In response to these limitations, Duval and Tweedie (2000a, 2000b) developed the "trim and fill" method for detecting publication bias.

The trim and fill method is based on statistically sophisticated iterative adjustments of the effect size distribution based on its departure from an assumed symmetrical shape. Specifically, the trim and fill method "trims" extreme effect sizes from the skewed side of the funnel plot causing asymmetry. This trimming procedure is repeated until the distribution of effect sizes become symmetrical. Next, the previously trimmed effect sizes are added back to the funnel plot with the addition of imputed (i.e., "filled") effect sizes mirroring filled observed effect sizes on the opposite side in the funnel plot distribution to maintain symmetry. That is, the trim and fill method identifies missing studies and then imputes their values in an iterative manner (McDaniel et al., 2006). These imputed values

(presented as solid circles in Figure 12.1) are meta-analytically combined (or "filled") with the original data to estimate the impact of publication bias on the overall effect size. Thus, a particular advantage of the trim and fill method is its ability to produce a validity estimate that has been adjusted for the effects of publication bias. Duval (2005) notes that "how well trim and fill works depends largely on the somewhat untestable assumption that the most extreme results in one direction are those which have been suppressed" (p. 135), which is fundamentally consistent with all aforementioned methods for publication bias. Given this, Duval proposes that the trim and fill method is most appropriately used for sensitivity analysis. That is, trim and fill is an effective method for *adjustment* for publication bias rather than *correction* per se. Accordingly, we adopt adjustment terminology when discussing our subsequent application of the trim and fill method. The trim and fill method has been applied across a diversity of disciplines including clinical medicine (Nelson, Humphrey, Nygren, Teutsch, & Allan, 2002), ecology (Jennions & Moller, 2002), education (Robbins et al., 2004), developmental psychology

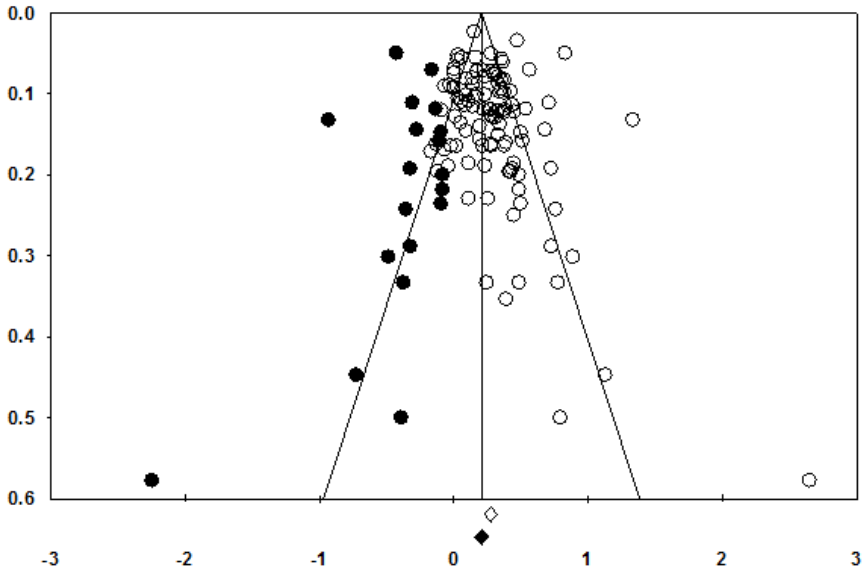

Figure 12.1. *The funnel plot combined with the trim-and-fill test for the structured interview validity data for job performance;* individual study effect sizes on the horizontal graph (X) axis against standard error on the vertical graph (Y) axis; values imputed by the trim and fill test are presented as solid circles, while reported values are presented as open circles.

(Goldberg, Prause, Lucas-Thompson, & Himsel, 2008), and health psychology (Vitaliano, Zhang, & Scanlan, 2003). More recently, researchers in I/O psychology have begun utilizing the trim and fill technique (McDaniel et al., 2006; McDaniel, Hartman, Whetzel, & Grubb, 2007).

Since publication bias can have significant negative consequences, Sutton (2005) maintains that "researchers should always check for the presence of publication bias and perform a sensitivity analysis to assess the potential impact of missing studies" (p. 175). Similarly, the editor of *Psychological Bulletin*, the field's premier review journal, recently required all meta-analyses submitted to the journal to address concerns over publication bias: "Synthesists submitting manuscripts to *Psychological Bulletin* who have the goal of making summary claims about a particular relationship, hypothesis, or treatment will be expected to conduct thorough searches to locate both published and unpublished research. Deviations from the exhaustive search strategy will require convincing justification" (Cooper, 2003, p. 6). While the formal assessment of publication bias using a robust method is becoming increasingly common among meta-analysts in the biomedical sciences (Sutton, 2005), it is far rarer among researchers in the organizational sciences (see McDaniel et al., 2006; McDaniel et al., 2007 for notable exceptions).

Relative to our focus on interviews in the current paper, Duval (2005) analyzed the interview validity data in McDaniel et al. (1994) using the trim and fill method. The results of this analysis indicate that the validity data for structured interviews suffers from publication bias, thereby resulting in an overestimation of structured interview validity. No publication bias was observed in the validity data for unstructured interviews (see Duval, 2005, Table 8.3, p. 140). McDaniel et al. (2006) maintain that the bias identified by Duval raises two significant issues. First, many practitioners may have relied on the biased findings and invested substantial effort in constructing (highly) structured interviews, believing that they were notably more valid than unstructured interviews. Second, "the number of research studies comparing the two types of interviews decreased after the meta-analysis was published, reducing the potential for contradictory findings" (McDaniel et al., 2006, p. 928).

STUDY PURPOSE

The purpose of the current study is to compare more accurate estimates of operational validity for structured interviews with those of unstructured interviews based on recent developments in meta-analysis: indirect range restriction and the trim and fill method. Accordingly, to illustrate the impact of these recent developments in meta-analysis on employment

decisions, we reestimate the operational validities of structured and unstructured interviews for job and training performance. We do so by correcting for error of measurement in the criterion, correcting for range restriction (using both DRR and IRR correction methods to compare the difference due to range restriction type), and adjusting for publication bias using the trim and fill method. We test one known moderator for the predictor—performance relationship: the source of performance ratings (Jawahar & Williams, 1997; Schmidt & Zimmerman, 2004). Finally, we provide 95% confidence intervals to enrich comparisons between structured and unstructured interview validities. To the best of our knowledge, these simultaneous analyses have not been previously conducted.

METHOD

Database

For reanalysis, we sought large data sets used for meta-analysis in order to avoid the compounding effect of sampling error on our reanalysis results. We identified several published meta-analyses on the validity of interviews (e.g., Huffcutt & Arthur, 1994; McDaniel et al., 1994; Wiesner & Cronshaw, 1988). Interestingly, we found that a number of studies (e.g., Duval, 2005; Le & Schmidt, 2006; Schmidt & Zimmerman, 2004) have used the McDaniel et al. (1994) data set for reanalysis. We examined why the McDaniel et al. (1994) data set was reanalyzed instead of other interview data sets and found that (cf. Schmidt & Zimmerman, 2004): (a) all the psychometric information needed for reanalysis are available; (b) it is the only data set which differentiates between administrative and research ratings for job performance. As discussed in Schmidt and Zimmerman (2004), "administrative ratings suffer from more leniency and exhibit more halo than do ratings collected solely for research purposes" (p. 554; see also Jawahar & Williams, 1997). Moreover, "the interview studies in the research-only category represent a broad range of job complexity levels" (Schmidt & Zimmerman, 2004, p. 555), from which more credible conclusions can be drawn; (c) it is the only data set which includes the training performance criterion (Although interviews are less frequently used in predicting training performance, we reasoned that it would provide a more robust conclusion if our reanalysis results generalize across both job and training performance); (d) McDaniel's et al.'s (1994) meta-analytic results have been widely cited in a variety of HR handbooks, textbooks, book chapters, and interview guides for practitioners; (e) McDaniel et al. (1994) also showed high levels of coding reliability (p. 604); and (f) this data set the largest *meta-analytic* data set

currently available in the published literature. It is noted that the purpose of our study is to illustrate the impact of indirect range restriction and publication bias on interview validities, not to update McDaniel et al.'s (1994) meta-analysis. Holding the data set constant between McDaniel et al. (1994) and the current study allows the effects of indirect range restriction corrections and publication bias adjustments to be isolated. That is, changes in validities can be solely attributed to methodological factors.

Last, a careful examination of this data set reveals that the mechanism by which range restriction operates on the predictor side is mainly indirect (see Berry et al., 2007 for an in-depth report): across the meta-analyses shown in Tables 12.1 and 12.2, an average of 73% (61% ~ 100%) of validities analyzed were based either on concurrent validation design or on studies where other test scores are available to interviewers. In the remainder of validities analyzed, we do not have explicit information on the nature of the mechanism by which range restriction operates on the predictor side. We believe that few if any companies hire their employees solely based on interview scores in a complete top-down manner, where the mechanism of range restriction on the predictor side is direct (Hunter et al., 2006). Given these reasons, we selected McDaniel et al.'s (1994) data set for reanalysis.

Artifact Data

We further decided to use the same the psychometric information reported in McDaniel et al. (1994), with the exception of interview reliability. First, with regard to range restriction corrections, we used the $\bar{u}_x$ value of .68 reported in McDaniel et al. (1994), which was empirically determined for both structured and unstructured interviews. By way of comparison, Huffcutt and Arthur (1994) reported a $\bar{u}_x$ value of .74 and Salgado and Moscoso (2002) reported a $\bar{u}_x$ value of .61 ($k = 31$) based on the largest number of independent observations. Second, we used the mean criterion reliability ($\bar{r}_{yy_i}$) estimates of .60 and .80 for job and training performance measures as reported in McDaniel et al. (1994), which were also empirically determined. Last, as Schmidt and Zimmerman (2004) noted, Conway, Jako, and Goodman's (1995) meta-analysis is the only one reporting unbiased estimates of interinterviewer reliability. That is, individual studies included in Conway et al.'s meta-analysis assess the reliability of interviewers' overall evaluation ratings by correlating scores of two different interviewers who interview the same sample of interviewees on two different occasions. This is the appropriate procedure for calibrating interview reliability because this procedure

Decrease the font size of "mean r_yy_i" to be consistent with other text

Au: The prefix *inter* does not require hyphen in APA.

Oaky, thanks!

Au: The font size of Tables 12.1 and 12.2 has been decreased to accommodate table on one page.

Okay, thanks!

Table 12.1. Reanalysis of Validity of Employment Interview Data

	Structured Interview					*Unstructured Interview*				
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*
Criterion = Job Performance	*k*	$\overline{r}$	$\overline{\rho}$ *DRR*[c]	$\overline{\rho}$ *IRR*[d]	% *under*	*k*	$\overline{r}$	$\overline{\rho}$ *DRR*[c]	$\overline{\rho}$ *IRR*[d]	% *under*
Validity not adj. for pub bias	106	.266	.473	.562	16%	39	.187[a]	.343	.556	38%
Lower Limit of the 95% CI		.223	.404	.489	17%		.138	.257	.445	42%
Upper Limit of the 95% CI		.308	.537	.626	14%		.235	.424	.640	34%
Validity adj. for pub bias	125	.202[b]	.368	.450	18%	39	.187[b]	.343	.556	38%
Lower Limit of the 95% CI		.151	.281	.349	20%		.138	.257	.445	42%
Upper Limit of the 95% CI		.251	.449	.537	16%		.235	.424	.640	34%
Criterion = Training Performance										
Validity not adj. for pub bias	32	.206[a]	.329	.405	19%	42	.202[a]	.322	.530	39%
Lower Limit of the 95% CI		.162	.261	.325	20%		.157	.253	.439	42%
Upper Limit of the 95% CI		.250	.394	.478	18%		.246	.387	.604	36%
Validity adj. for pub bias	34	.199[b]	.318	.392	19%	42	.202[b]	.322	.530	39%
Lower Limit of the 95% CI		.153	.247	.309	20%		.157	.253	.439	42%
Upper Limit of the 95% CI		.244	.384	.467	18%		.246	.387	.604	36%

Note. Original data was obtained from Michael McDaniel; (1) Number of validity coefficients; (2) Sample size weighted mean observed validity; (3) Mean operational (true) validity corrected for direct range restriction; (4) Mean operational (true) validity corrected for indirect range restriction; (5) Percent underestimation of operational validity by direct correction method; CI = Confidence Interval.

[a] Observed mean validity was estimated based on random-effects (RE) methods using the Comprehensive Meta-Analysis (CMA) software Version 2 (Borenstein, Hedges, Higgins, & Rothstein, 2005).

[b] Publication bias was adjusted for by using RE methods.

[c] The mean observed validity corrected for mean criterion unreliability ($\overline{r}_{yyi}$) and direct range restriction (using Case II correction methods).

[d] The mean observed validity corrected for mean criterion unreliability ($\overline{r}_{yyi}$) and indirect range restriction (using Case IV correction methods).

Make the font size for both the circles be the same

Table 12.2. Reanalysis of Validity of Employment Interview Data for Job Performance by Rating Type

	Structured Interview					*Unstructured Interview*				
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*
Criterion = Research-purpose Ratings of Job Performance	*k*	$\overline{r}$	$\overline{\rho}$ *DRR*[c]	$\overline{\rho}$ *IRR*[d]	% *under*	*k*	$\overline{r}$	$\overline{\rho}$ *DRR*[c]	$\overline{\rho}$ *IRR*[d]	% *under*
Validity not adj. for pub bias	44	.283[a]	.501	.590	15%	9	.218[a]	.396	.613	35%
Lower Limit of the 95% CI		.207	.378	.460	18%		.112	.211	.376	44%
Upper Limit of the 95% CI		.357	.606	.691	12%		.319	.553	.742	25%
Validity adj. for pub bias	46	.269[b]	.478	.567	16%	9	.218[b]	.396	.613	35%
Lower Limit of the 95% CI		.189	.347	.425	18%		.112	.211	.376	44%
Upper Limit of the 95% CI		.345	.590	.676	13%		.319	.553	.742	25%
Criterion = Administrative Ratings of Job Performance										
Validity not adj. for pub bias	58	.231[a]	.418	.504	17%	24	.189[a]	.348	.561	38%
Lower Limit of the 95% CI		.182	.336	.413	19%		.120	.225	.398	43%
Upper Limit of the 95% CI		.279	.493	.582	15%		.257	.459	.672	32%
Validity adj. for pub bias	66	.193[b]	.353	.432	18%	24	.189[b]	.348	.561	38%
Lower Limit of the 95% CI		.142	.264	.329	20%		.120	.225	.398	43%
Upper Limit of the 95% CI		.242	.436	.523	17%		.257	.459	.672	32%

Note. Original data was obtained from Michael McDaniel; (1) Number of validity coefficients; (2) Sample size weighted mean observed validity; (3) Mean operational (true) validity corrected for direct range restriction; (4) Mean operational (true) validity corrected for indirect range restriction; (5) Percent underestimation of operational validity by direct correction method; CI = Confidence Interval.

[a] Observed mean validity was estimated based on random-effects (RE) methods using the Comprehensive Meta-Analysis (CMA) software Version 2 (Borenstein, Hedges, Higgins, & Rothstein, 2005).

[b] Publication bias was adjusted for by using RE methods.

[c] The mean observed validity corrected for mean criterion unreliability ($\overline{r}_{yyi}$) and direct range restriction (using Case II correction methods).

[d] The mean observed validity corrected for mean criterion unreliability ($\overline{r}_{yyi}$) and indirect range restriction (using Case IV correction methods).

> controls for all sources of measurement error in the interview scores: random response error on the part of the interviewees, transient error in the responses of the interviewees, and conspect error (disagreement between the raters in how they score or rate the interviewee's responses). (Schmidt & Zimmerman, 2004, p. 554)

Conway et al. (1995) reported mean interinterviewer reliability estimates for five levels of interview structure. Following Schmidt and Zimmerman (2004, p. 555), we reasoned that Conway et al.'s Level 1 corresponds to the unstructured interview case and Level 4 to the structured interview case. Level 5 is too highly structured to be the case for real-life structured interviews. Conway et al.'s estimated mean reliability estimates for Levels 1 and 4 were used in the current study to estimate more accurate validities of unstructured and structured interviews. However, it should be noted that these reliabilities ($\bar{r}_{xx_i}$) are derived from the highly restricted (incumbent) groups. When corrected for range restriction using the formula presented in Hunter et al. (2006, p. 106, Equation 3.17b), the unrestricted reliabilities ($\bar{r}_{xx_a}$) are .71 and .84 for unstructured and structured interviews, respectively. Further, the impact of unreliability is better evaluated based upon the square root of the reliability (a.k.a., the reliability index), because this is the actual number used in the unreliability correction in meta-analysis (Gaugler, Rosenthal, Thornton, & Bentson, 1987, p. 494). Reliability indices are .84 and .92 for unstructured and structured interviews, respectively.

Some readers might wonder why we emphasize interview reliability when our interest is in operational validity rather than true-score correlation. Interinterviewer reliability plays an important role in terms of accurately estimating interview validity in several ways. First, the reliability of a measure sets the upper limit of its validity. That is, the reliability index (the square root of reliability, which is the correlation between the measure and its own true score) is the maximum correlation the measure can have with any other variable (Nunnally & Bernstein, 1994). Thus, one way to increase the validity of interview is to increase its reliability. In a reanalysis of McDaniel et al.'s (1994) interview data, Schmidt and Zimmerman (2004) presented evidence indicating that the higher operational validity estimated for structured interviews may result from higher interinterviewer reliability compared to that of unstructured interviews. They concluded that if unstructured interviews could be made as reliable as structured interviews, validities would likely be equal. One way to increase the reliability (and therefore the validity) of unstructured interviews is to average across several unstructured interviews. They showed that it takes approximately three to four unstructured interviews (interviewers) to have validity equal to that of one structured interview (inter-

viewer). That is, "by increasing the number of interviews (or interviewers), one can raise the validity of unstructured interviews to that of structured interviews" (Schmidt & Zimmerman, 2004, p. 558).

Second, our concern with interview validity stems from the fact that the IRR correction procedure requires that the observed validity first be corrected for measurement error in both the predictor and criterion (Hunter et al., 2006). It should be noted that the major difference between the classical method (Case II) and the new IRR correction method (Case IV) is that the former is based on range restriction in observed scores (u_x) whereas the latter uses range restriction in true scores (u_t) to correct for the biasing effect of range restriction. That is, $u_x = s_x/s_x$, the ratio of the restricted to the unrestricted observed *SD*s of *x*, whereas $u_t = s_t/s_t$ is the degree of range restriction on the true scores (*t*) underlying the observed scores (*x*). Specifically, under DRR, applicants are selected solely on their observed test scores. Thus, they are selected partly on their true scores and partly on the measurement errors in their observed scores. However, under IRR, scores on the test of interest are not used in selection and thus measurement errors in the test scores have no effect on the selection process. That is, the impact of IRR is only on true scores underlying observed scores. It is important to note that u_t is a function of u_t and r_{xx_a} as shown:

change to u_x; sub t should be changed to sub x.

$$u_t = \sqrt{\frac{u_x^2 - (1 - r_{xx_a})}{r_{xx_a}}}$$

where u_t and u_x are as defined above, and r_{xx_a} is the reliability of the predictor in the unrestricted group (i.e., applicant group) which in a function of the reliability of the predictor in the incumbent group and u_x (Hunter & Schmidt, 2004, Equation 3.17b, p. 106). Using this formula, we obtained u_t values of .49 and .60 for unstructured and structured interviews, respectively. Since u_t is usually smaller than u_x (except when the predictor measure is perfectly reliable or there is no range restriction), the new IRR method usually yields larger estimates of true score correlations (and operational validity) than those provided by the classical DRR correction method (see Hunter et al. 2006 for more details; see also their Tables 1 and 2 for correction steps and formulae for DRR and IRR, respectively).

change "in" to "is"

In sum, the new IRR method requires that an accurate estimate of interview reliability be obtained before the IRR procedure can be applied. The IRR correction method initially yields true score correlations that have been corrected for measurement error in both the predictor and cri-

terion. Thus, in order to estimate operational validity it is necessary to attenuate true score correlations for measurement error in the predictor in the population (r_{xx_a}). Accordingly, we reintroduced measurement error in the predictor by multiplying the true score validities by the square root of the mean interview reliabilities in the unrestricted group.

Data Analysis

The observed mean validities of structured and unstructured interviews, and those adjusted for publication bias, were estimated using Comprehensive Meta-Analysis (CMA) software Version 2 (Borenstein, Hedges, Higgins, & Rothstein, 2005). CMA is the only commercially available meta-analysis program able to (a) adjust for publication bias using Duval and Tweedie's (2000a, 2000b) trim and fill method and (b) produce funnel plots with missing effect size estimates (see Figure 12.1). In general, meta-analyses based on the random-effects (RE) model provide more accurate and less biased estimates than meta-analyses based on the fixed-effects (FE) model, so we selected the RE model in the current study (Field, 2001, 2003, 2005; Schmidt, Oh, & Hayes, 2009). In order to minimize second-order sampling error (cf. Hunter & Schmidt, 2004, Ch. 9), we set the minimum number of primary studies subject to reanalysis to $k \geq 7$.

Duval (2005) also reanalyzed the McDaniel et al. (1994) data set using CMA software. However, our reanalysis methods, and our results that follow, differ from hers in several ways. First, Duval (2005) estimated mean observed validities adjusted only for publication bias and not measurement error in the criterion or range restriction. In the current study, we corrected the mean observed validities adjusted for publication bias for both criterion unreliability and range restriction to estimate operational validity. We applied both DRR and IRR to illustrate the extent to which traditionally used DRR correction methods underestimate the actual validity of interviews for job and training performance when range restriction is actually indirect. Second, while Duval focused on job performance, we include both job and training performance. Last, we also applied the same procedures described above to a subset of the data which includes information on rating type (i.e., research-only ratings vs. administrative ratings) to explore whether the difference in validity estimates of job-related interviews depends on the type of rating. We expected that the operational validity of job-related interviews conducted with research-purpose ratings would be higher than that of job-related interviews conducted with administrative ratings given that research-

purpose ratings are more construct valid (Schmidt & Hunter, 1998; Schmidt & Zimmerman, 2004).

RESULTS

Tables 12.1 and 12.2 show the results of our reanalysis of the McDaniel et al. (1994) data set. Below, we focus on the difference between structured and unstructured interview validities in three ways. First, we compare the two validities corrected for all three artifacts: criterion reliability, IRR, and publication bias. Second, we compare validities corrected only for criterion unreliability and IRR. Lastly, we compare the two validities in the traditional way. That is, we compare validities corrected for criterion unreliability and DRR. Further, we show additional information about how much the traditional way of correcting for range restriction (DRR) underestimates the operational validity of employment interviews (the % under column in Tables 12.1-12.2) and the 95% confidence intervals for each meta-analytic estimate.

Table 12.1 shows the overall validities of structured and unstructured interviews for job and training performance regardless of performance rating type. For job performance, when corrected/adjusted for IRR, publication bias, and measurement error in the criterion measure, the operational validity of unstructured interviews ($\bar{\rho}$ = .556) is about 24% greater than that of structured interviews ($\bar{\rho}$ = .450). This difference is noticeable, although there is some overlap between the 95% confidence intervals for the two estimates; the lower limit of the mean operational validity of unstructured interviews ($\bar{\rho}$ = .445) is practically the same as the mean operational validity of structured interviews ($\bar{\rho}$ = .450). This is mainly due to the fact that only structured interviews were found to suffer from publication bias, so the trim and fill method filled in 19 missing validity estimates (see Figure 12.1—the black solid circles indicate the 19 filled-in validity estimates), thereby lowering the structured interview validity. When no adjustment is made for publication bias, the operational validity of unstructured interviews ($\bar{\rho}$ = .556) is, in fact, virtually equal to that of structured interviews ($\bar{\rho}$ = .562). However, if traditional operational validity estimates (DRR) without adjustment for publication bias were relied upon, then our conclusion would be that structured interviews ($\bar{\rho}$ = .473) are more valid than unstructured interviews ($\bar{\rho}$ = .343) by 38%. Taken together, these findings suggest that suboptimal adjustment/correction methods for artifacts and publication bias changed the sign of advantage in favor of unstructured interviews over structured interviews, a finding that is likely to be very surprising and counterintuitive to many people. Further, as shown under the "% under" columns, operational validity estimates

change this to "optimal"

corrected only for DRR, when IRR is the case, are considerable underestimates (by 16% to 38%).

In the case of training performance (Table 12.1), when corrected/adjusted for IRR, publication bias, and measurement error in the criterion, the operational validity of unstructured interviews ($\bar{\rho}$ = .530) is about 35% greater than that of structured interviews ($\bar{\rho}$ = .392). This difference is substantial and credible given that the 95% confidence intervals for the two estimates hardly overlap; as shown in Table 1, the upper limit of the mean operational validity of structured interviews ($\bar{\rho}$ = .467) is only slightly greater than the lower limit of the mean operational validity of unstructured interviews ($\bar{\rho}$ = .439). This is partly due to the fact that only structured interviews were found to suffer from publication bias, though to a less severe extent than with job performance (the trim and fill method filled in only 2 missing validity estimates). Even when no adjustment is made for publication bias, the operational validity of unstructured interviews ($\bar{\rho}$ = .530) is still greater than that of structured interviews ($\bar{\rho}$ = .405) by about 31%. However, if traditional operational validity estimates (DRR) without adjustment for publication bias were relied upon, then our conclusion would be that structured interviews ($\bar{\rho}$ = .329) and unstructured interviews ($\bar{\rho}$ = .322) have equal validity. As was the case with job performance, operational validity corrected for DRR is a considerable underestimate compared with that corrected for IRR (19%–39%). In sum, when artifacts are corrected/adjusted for their biasing effects using recent advancements in meta-analysis, the operational validity estimate of unstructured interviews for predicting job performance is found to be about 24% greater than that of structured interviews. Likewise, the operational validity estimate of unstructured interviews for predicting training performance is found to be about 35% greater than that of structured interviews.

In Table 12.2, following Schmidt and Zimmerman (2004), in addition to interview structure, we also considered the types of job performance ratings (research-purpose vs. administrative) as a moderator which could affect the magnitude of operational validity estimates for job performance. That is, we examined the two moderators (interview structure and rating type) in a hierarchical manner, which can be a highly informative meta-analytic strategy (Hunter & Schmidt, 2004). To our knowledge, no single study has examined these two moderators in a hierarchical manner. For research-purpose job performance ratings, when corrected/adjusted for IRR, publication bias, and criterion unreliability, the operational validity of unstructured interviews ($\bar{\rho}$ = .613) is 8% greater than that of structured interviews ($\bar{\rho}$ = .567). This difference is not substantial and thus caution is needed given that the 95% confidence intervals for the two estimates completely overlap. As shown in Table 12.2, the 95% confidence interval for

the operational validity of structured interviews (.425 – .676) is nested within the 95% confidence interval for the operational validity of unstructured interviews (.376 – .742). Further, the operational validity of unstructured interviews is estimated based on a relatively small number of studies ($k = 9$). This is partly due to the fact that only structured interviews were found to suffer from publication bias (the trim and fill method filled in two missing validity estimates). Even when no adjustment is made for publication bias, the operational validity of unstructured interviews ($\bar{\rho} = .613$) is still slightly greater than that of structured interviews ($\bar{\rho} = .590$) by about 4%. However, if traditional operational validity estimates corrected for DRR and unadjusted for publication bias are relied upon, then our conclusion would be that structured interviews ($\bar{\rho} = .501$) are more valid than unstructured interviews ($\bar{\rho} = .396$) by around 27%.

For administrative job performance ratings, when corrected/adjusted for IRR, publication bias, and criterion unreliability, the operational validity estimate of unstructured interviews ($\bar{\rho} = .561$) is about 30% greater than that of structured interviews ($\bar{\rho} = .432$). This difference is not small, but caution is still needed given that the 95% confidence intervals for the two estimates overlap substantially. As shown in Table 12.2, the 95% confidence interval for the operational validity of structured interviews (.398 – .672) include a large portion of the 95% confidence interval for the operational validity of unstructured interviews (.329 – .523). This is mainly due to the fact that only structured interviews were found to suffer from publication bias (the trim-and-fill test filled in 8 missing validity estimates). Even when no adjustment is made for publication bias, the operational validity estimate of unstructured interviews ($\bar{\rho} = .561$) is still around 11% greater than that of structured interviews ($\bar{\rho} = .504$). However, if traditional operational validity estimates corrected for DRR and unadjusted for publication bias are relied upon, then our conclusion would be that structured interviews ($\bar{\rho} = .418$) are more valid than unstructured interviews ($\bar{\rho} = .348$) by about 20%. Taking research-purpose and administrative job performance ratings together, as found in Table 12.2, these findings suggest that optimal correction methods for artifacts changed the sign of advantage in favor of unstructured interviews over structured interviews. Further, as shown under the "% under" columns, operational validity corrected for DRR is a considerable underestimate (15%–38%).

When looking at the results in a hierarchical manner, the highest operational validity estimate was found for unstructured interviews with research-purpose ratings ($\bar{\rho} = .613$), followed by structured interviews with research-purpose ratings ($\bar{\rho} = .567$), unstructured interviews with administrative ratings ($\bar{\rho} = .561$), and structured interviews with administrative ratings ($\bar{\rho} = .432$). When interviews are unstructured, the rating type does not make a large difference in validity (.052 = .613 – .561) and

the 95% confidences for both the estimates completely overlap, whereas when interviews are structured the rating type makes a larger difference (.135 = .567 – .432) in validity and the 95% confidence intervals do not overlap substantially (see Table 12.2 for more details). This interaction pattern is shown in Figure 12.2. In sum, we found that unstructured interviews with research-purpose ratings have the highest operational validity estimate while structured interviews with administrative ratings have lowest operational validity estimate. However, if operational validity is estimated using the traditional DRR method and no adjustment is made for publication bias, a different picture emerges: the highest operational validity estimate would be found for structured interviews with research-purpose ratings ($\bar{\rho}$ = .501) and lowest validity estimate for unstructured interview with administrative ratings ($\bar{\rho}$ = .348). Table 12.3 provides an overall summary of the interview types with the higher validity estimate categorized by correction method(s) applied and criterion measure/interview purpose.

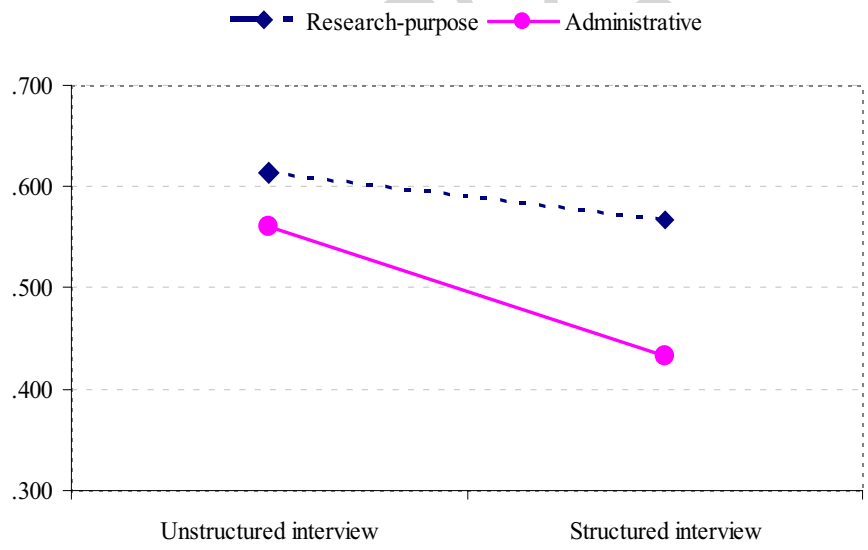

Figure 12.2. *The interaction between interview structure and rating type on interview validity for job performance;* the validity estimates used for this plot are those corrected for measurement error in the criterion measure, indirect range restriction, and publication bias (see the note for Table 12.2 for details) as reported in the intersections between the second row (validity adj. for pub bias) and the fourth column (IRR) in Table 12.2.

Table 12.3. Preferred Interview Types According to Correction Method Used

	Interview Type Displaying Higher Operational Validity			
	Excluding Adjustments for Publication Bias		*Including Adjustments for Publication Bias*	
Criterion/Purpose	*Traditional (DRR Only)*[d]	*IRR*[e]	*Traditional (DRR Only)*[d]	*IRR*[d]
Job Performance (all)[a]	Structured	~ Equal	Structured/Equal	Unstructured
Research ratings[b]	Structured/Equal	Unstructured/Equal	Structured/Equal	Unstructured/Equal
Administrative ratings[b]	Structured/Equal	Unstructured/Equal	~ Equal	Unstructured/Equal
Training Performance (all)[a]	~ Equal	Unstructured	~ Equal	Unstructured

Note. Original data was obtained from Michael McDaniel.
~ Equal represents a validity difference between structured and unstructured interviews of less than .01; Structured/Equal represents the superiority of structured interviews over unstructured interviews at the mean level, but the 95% confidence intervals overlap considerably; Unstructured/Equal represents the superiority of unstructured interviews over unstructured interviews at the mean level, but the 95% confidence intervals overlap considerably.

[a] See Table 1 for details

[b] See Table 2 for details

[c] Corrections for mean criterion unreliability ($\overline{r}_{yyi}$) and direct range restriction using Case II correction methods.

[d] Corrections for mean criterion unreliability ($\overline{r}_{yyi}$) and indirect range restriction using Case IV correction methods.

DISCUSSION

One of the primary goals of science is the accurate accumulation of generalizable knowledge (Toulmin, 1961; Schmidt et al., 2006). This goal is particularly important in making high-stakes employment decisions, as the selection tools chosen based on empirical evidence can have a significant impact on organizational performance (Schmidt & Hunter, 1998; Schmidt, Hunter, McKenzie, & Muldrow, 1979). Accordingly, it is imperative to obtain the most accurate estimates of operational validity (Schmidt et al., 2006). Meta-analysis has become an increasingly useful tool for achieving this objective; however, it is not a static technology (Schmidt, 2008). Rather, meta-analysis is a dynamic methodology that has become increasingly accurate with systematic refinements and advancements that continue to be developed. As the results of this study indicate, applying recent developments in meta-analytic methodology to an existing interview data set can result in rather dramatic changes in long-held assumptions and knowledge about the operational validity estimates of structured and unstructured employment interviews.

It is useful to briefly consider in turn how each of the recent meta-analytic advancements applied in this study affect the difference in validity between structured and unstructured interviews (a summary is presented in Table 12.3). When interview validities were estimated in a traditional manner (using DRR correction methods), structured interviews displayed a higher operational validity estimate than unstructured interviews in all cases except for overall training performance (Table 12.1). In this exception, structured and unstructured interviews were estimated to have approximately equal validities. However, this situation changes rather dramatically when more accurate (that is, IRR) correction methods are applied (Hunter et al., 2006; Le & Schmidt, 2006; Schmidt et al., 2006). When interview operational validities were estimated using IRR corrections (unadjusted for publication bias), the structured and unstructured interviews displayed approximately the same operational validity estimate for predicting overall job performance (Table 12.1). Correcting for IRR ~~[illegible]~~ with publication bias adjustment, ~~[illegible]~~ reversed the sign advantage for structured interviews in favor of unstructured interviews for overall training performance (Table 12.1) and job performance measured using research-purpose ratings or administrative ratings (Table 12.2). These findings illustrate the consequences that may result from correcting for direct range restriction when range restriction is known to be indirect. The assumption that this practice does not make a practical difference in the ultimate conclusions reached is erroneous. As such, it may have retarded both the science and practice of personnel selection.

Should read "Correcting for IRR with the publication bias adjustment reversed the sign advantage for …"

It should again be noted that the information in Table 12.3 is based on comparisons *at the mean level* alone and thus should be carefully interpreted with reference to the relevant 95% confidence intervals shown in Tables 12.1 and 12.2 because the 95% confidence intervals overlap considerably in some cases. However, it can be safely concluded that unstructured interviews may be as valid as structured interviews in most cases. *Even before adjusting for publication bias*, it was found that structured interviews are not superior to unstructured interviews at the mean level (remember the caution given in text). This reversal results from the fact that u_t, the degree of indirect range restriction on the true score (t) underlying observed interview scores (x), is lower for unstructured, rather than structured, interviews. Thus, larger corrections are made for unstructured, rather than structured, interviews. Because we used the same value of u_x for both interview types, this reversal is actually due to the difference in measurement error between structured and unstructured interviews. Accordingly, this underscores the importance of using appropriate measurement error corrections when estimating operational validities (Hunter et al., 2006). Again, it should be noted that we used interview reliabilities only to accurately correct for indirect range restriction. We did not correct for predictor unreliability. In sum, many readers (particularly, researchers) will find these conclusions and results surprising and implausible. However, readers should note that our analyses are based on the most advanced correction methods for range restriction that have been published in premier I/O psychology (Berry et al., 2007; Hunter et al., 2006; Schmidt et al., 2006) and psychological methodology journals (Le & Schmidt, 2006), and on methods of adjusting for publication bias that appear in the best statistics journals (Duval & Tweedie, 2000a, 2000b).

Adjusting for publication bias using the trim and fill method also reduced the advantage of structured interviews by lowering their estimated validities (right panel in Table 12.3). Publication bias was found only in structured interview data. This finding suggests that researchers may have been reluctant to report results that indicate lower or nonsignificant operational validities for structured interviews. Relevant to this, Cooper (2003) argued as follows: "In particular, research that fails to achieve standard levels of statistical significance is frequently left in researchers' file drawers.... Published estimates of effect may make relationships appear stronger than if all estimates were retrieved by the synthesist" (p. 6). In our reanalysis, publication bias was greatest in the overall sample of structured interviews used to predict job performance, with the trim and fill method indicating 19 missing studies in addition to the original 106 (Table 12.1 and Figure 12.1). This adjustment resulted in a 20% reduction of the operational validity estimate for structured interviews used to predict job performance, which broadly confirms Cooper's (2003) argument.

However, we believe that since it is impossible to test whether structured interview studies reporting negative or lower results have *actually* been suppressed (Duval, 2005), the results we present which have been adjusted for publication bias should be considered as sensitivity tests rather than an actual corrected estimates. However, it would be hard to account for the pattern of observed validities for the structured interview shown in Figure 12.1 without recourse to the hypothesis of publication or availability bias. As noted previously, as a matter of practice, most organizational researchers do not systematically test their data for the presence of publication bias (McDaniel et al., 2006; Sutton, 2005). We support Cooper's (2003) and Sutton's (2005) recommendation that sensitivity tests for publication bias be conducted for each meta-analysis unless there are reasons to believe that publication bias cannot occur in a particular research situation (cf. Hunter & Schmidt, 2004, pp. 496-498). Given this, the trim and fill method appears to be a promising method for accomplishing this goal, and our results support previous findings that have shown it to be more sensitive than other methods such as the failsafe *N* test. (In fact, when applied to the data in the current study, the failsafe *N* test did not detect any publication bias). Nevertheless, we suggest that additional investigation of the trim and fill method using real and simulated data would be informative.

It was not our initial intention to consider why unstructured interviews might work as well as (or even better than) structured interviews. However, given our seemingly surprising results, we sought to understand why this might be the case. Accordingly, we offer a few suggestions based on a review of the literature.

First, several studies on the construct validity of structured vs. unstructured interviews suggest that unstructured, rather than highly structured, interviews are more highly related to general mental ability (GMA; which is the single best predictor of job and training performance; Schmidt & Hunter, 1998) as well as Conscientiousness and Emotional Stability, two of the Big Five personality factors that are most predictive of job performance across situations (Barrick et al., 2001). Huffcutt, Roth, and McDaniel's (1996) meta-analysis found that interviews ratings and GMA are moderately correlated at the true-score level ($\bar{\rho} = .40$; $k = 49$) and that the true-score correlation between interview scores and GMA tends to decrease as the level of structure increases [.52 ($k = 8$), .40 ($k = 19$), and .35 ($k = 22$) for low, medium, and high levels of interview structure, respectively]. Judge and Klinger's (2007) recent reanalysis of data from Cortina, Goldstein, Payne, Davison, and Gilliland (2000) found that unstructured interviews ($\bar{\rho} = .25$; $k = 72$) are slightly more related to GMA than structured interviews ($\bar{\rho} = .22$; $k = 137$). Salgado and Moscoso's (2002) meta-analysis also found that unstructured (conventional)

interviews ($\bar{\rho}$ = .41, k = 53) are more related to GMA than structured (behavioral) interviews ($\bar{\rho}$ = .28; k = 21). Using methods of improved accuracy and carefully updated data sets, Berry et al. (2007) found that the GMA and interview relationship at the true score level is moderated by level of interview structure: .22 (k = 27), .48 (k = 6), and .29 (k = 3) for high, medium, and low levels of interview structure, respectively. Taken together, these findings suggest that unstructured interviews may tap more GMA than (highly) structured interviews.

Using data from situational and behavior interviews, Roth and colleagues (2005) found that structured interview ratings have relatively low correlations with self-reports of personality. Likewise, meta-analytic results suggest that "there is relatively little relationship between structured interviews and self-reported personality factors" (Roth et al., 2005, p. 261). Specifically, unstructured interviews ($\bar{\rho}$ = .21; k = 23; Cortina et al., 2000) were found to be more related to Conscientiousness than structured interviews ($\bar{\rho}$ = .14; k = 39; Judge & Klinger, 2007). Likewise, unstructured interviews ($\bar{\rho}$ = .27; k = 24) were found to be more related to Emotional Stability than structured interviews ($\bar{\rho}$ = .02; k = 21; Judge & Klinger, 2007). Lastly, it was found that unstructured interviews ($\bar{\rho}$ = .14; k = 40) are slightly more related to grade point average, a proxy of GMA than structured interviews $\bar{\rho}$ = .11; k = 6; Judge & Klinger, 2007). Salgado and Moscoso's (2002) meta-analysis showed similar results: Unstructured, rather than structured, interviews are more related to ideal employee personality factors: Conscientiousness ($\bar{\rho}$ = .28 vs. .17), Extraversion ($\bar{\rho}$ = .34 vs. .21), and Emotional Stability ($\bar{\rho}$ = .38 vs. .08). In sum, compared with highly structured interviews, unstructured interviews appear to tap more strongly into basic individual difference variables that predict job and training performance. In sum, we echo the conclusion in Sitzmann, Kraiger, Stewart, & Wisher (2006) that training program effectiveness is determined more by a program's content than mode of delivery or media (Clark, 1994). It may also be the case that interview validity is determined more by content (the constructs being measured) rather than structure (the mode of measurement).

Second, Schmidt (1993, p. 506) noted that the validity of the unstructured interview equals that of assessment centers, integrity tests, and tests of single aptitudes (e.g., verbal, quantitative, or technical ability), a fact that may appear surprising to researchers as well as practitioners. One possibility is that just as it is possible to conduct poor structured interviews, it may also be possible to conduct good unstructured interviews. Particularly, we speculate that unstructured interviews designed and conducted for research proposes may be similar to semi-structured interviews. In support of this speculation, Schmidt hypothesized that the seemingly high validity of unstructured interviews

may reflect that "the unstructured interviews on which validity studies have been conducted are more carefully planned, more through, and perhaps longer than the shorter conversational unstructured interviews that are commonly used in business and industries" (p. 506). Nevertheless, they are still unstructured.

Likewise, unstructured interviews may be conducted by skilled human resource professionals or managers with significant interviewing experience and skill. Such interviewers may possess a repertoire of effective interview techniques (e.g., follow-up questions, interrogation or probing methods) that are applied rather consistently, though not identically, across candidates. Research by Dipboye and colleagues (e.g., Dipboye, Gaugler, Hayes, & Parker, 2001) supports this hypothesis. These researchers found that some interviewers' judgments are more valid than others even when unstructured interviews are used. Specifically, the validities of aggregated interview scores from five interviewers who interviewed and assessed 446 interviewees ranged from .07 to .12. However, when the validities of individual interview scores were examined, they found that two interviewers had considerably higher validities (.29 and .44). We speculate that the variation in validity across interviewers are partly, albeit not completely, attributable to their difference in interviewing experience and skill. This is consistent with Dipboye and Gaugler's (1993) notion that the interviewer differences "may result from differences among interviewers in their ability, experience, and other characteristics" (p. 155). Accordingly, we believe that future research on this issue is warranted and may yield promising insight.

Moreover, many structured interview formats necessitate that interviewers ask the same questions to each candidate, even in the same order. The rigid procedures characteristic of traditional structured interviews may increase the cognitive load on the part of interviewers, which in turn could reduce the accuracy of their judgments. In the context of assessment centers, Lievens and Klimoski (2001) argued that reducing assessors' cognitive load may lead to more accurate assessments and ultimately enhance criterion-related validity. In contrast, in an unstructured format, skilled interviewers may dedicate less attention to a topic or line of questioning after a candidate demonstrates competence, thereby reducing interviewers' cognitive load while also allowing a more in-depth assessment of areas where a candidate's proficiency is unproven or less evident. Therefore, when qualified interviewers are given an appropriate level of discretion to deviate, whenever needed, from the standardized interview, they may measure each applicant's ability, knowledge, and skills with increased accuracy. Taking together our first and second propositions as to why unstructured interviews may be at least as valid as structured interviews, we suggest that interviewers are more likely to make

accurate judgments when more relevant and valid cues about interviewees are available. More specifically, we believe, as Townsend, Bacigalupi, and Blackman (2007) suggested with regard to assessing integrity using simulated employment interviews, that "the free-flowing structure of the informal interview puts the target subject at ease and perhaps 'off guard,' and allows more diagnostic cues" (p. 555) relevant to construct(s) that predict job performance (e.g., GMA, integrity). In sum, we agree with Chapman and Zweig (2005, p. 673) who, argued, after evaluating the current interview and interviewer training literature (Campion et al., 1997; Palmer, Campion, & Green, 1999), that

> despite the widespread acceptance of the employment interview in personnel selection practice, we know surprisingly little about how interviews are typically conducted, how interviewers are trained, whether interviewer training influences the way interviews are conducted, or how interviewers and applicants react to the way that interviews are carried out. (p. 672)

Last, we again note that the purpose of the current study is not to update McDaniel et al. (1994) but to illustrate the potential influence of methodological advancements and refinements on meta-analytic results and conclusions using the data used in McDaniel et al. Given this, our reanalyses were conducted by focusing on a broad difference in validity between structured and unstructured interviews across conditions. Thus, we cannot rule out the possibility that publication bias might be more (or less) severe under some specific conditions (e.g., in predictive validation studies, among less experienced interviewers). In addition, our findings may not be generalizable beyond McDaniel et al.'s data set. It is rather obvious that many primary studies have been conducted since the publication of McDaniel et al. Accordingly, future meta-analytic updates of McDaniel et al. should employ the most updated range restriction correction and publication bias detection/adjustment methods and also examine all important moderators to be more informative.

CONCLUSION

Although the widely held belief is that structured, rather than unstructured, interviews are superior in validity, our results suggest that unstructured interviews may indeed possess greater validity than previously recognized. In other words, it may be the case that unstructured interviews are as valid as structured interviews. Overall, our results based on constructive, meta-analytic re-analyses of the McDaniel et al. (1994) data set methodologically *illustrate* that improvements in meta-analytic methods may produce important changes in validity estimates and may

Please add this sentence here in the first line. "This research is an example of how data, if not properly analyzed, lie to researchers (Schmidt, 2010)."

alter theoretical and practical conclusions. We hope that the current study encourages primary researchers to continue to explore issues of interview structure, content, construct, process, and quality, thereby further benefiting both the science and practice of personnel selection via future larger-scale meta-analytic efforts.

ACKNOWLEDGMENTS

We thank Michael McDaniel and Deb Whetzel for providing us with the interview data set used in this study, Michael Borenstein for providing us with the Comprehensive Meta-Analysis software program, Huy Le for his useful comments on data analysis, and Ryan Klinger for his information on the new meta-analytic findings on the constructive validity of employment interviews.

NOTES

1. Operational validity is defined as an observed correlation between scores on the predictor and criterion measures corrected for predictor range restriction and criterion unreliability.
2. In fairness to McDaniel et al. (1994), it is noted that the new range restriction correction and publication bias adjustment methods to be employed in this study were not available to them.

REFERENCES

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

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*, 9-30.

Becker, B. J. (1994). Combining significance levels. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 215–230). New York, NY: Russell Sage.

Becker, B. J. (2005). The failsafe *N* or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111-126). Chichester, England: Wiley.

Begg, C. B., & Muzumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088-1101.

Berry, C. M., Sackett, P. R., & Landers, R. (2007). Revisiting interview-cognitive ability relationships: attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology*, 837-874.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. R. (2005). *Comprehensive meta-analysis. Version 2*. Englewood, NJ: Biostat.

Borman, W. C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76*, 863-872.

Campbell, J. P., McCloy, R. A. Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organization* (pp. 71-98). San Francisco, CA: Jossey-Bass.

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

Chapman, D. S., & Zweig, D. I. (2005). Developing a nomological network for interview structure: Antecedents and consequences of the structured selection interview. *Personnel Psychology, 58*, 673-702.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development, 42*, 21-29.

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.

Cooper, H. (2003). Editorial. *Psychological Bulletin, 129*, 3-9.

Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K., Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology, 53*, 325-351.

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

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

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

Dipboye, R. L., Gaugler, B. B., Hayes, T. L., & Parker, D. (2001). The validity of unstructured panel interviews: More than meets the eye? *Journal of Business and Psychology, 16*, 35-49.

Dipboye, R. L., & Jackson, S. L. (1999). Interviewer experience and expertise effects. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 259-278). Thousand Oaks, CA: SAGE.

Duval, S. (2005). The "Trim and Fill" method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 127-144). Chichester, England: Wiley.

Duval, S., & Tweedie, R. (2000a). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455-463.

Duval, S., & Tweedie, R. (2000b). A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89-98.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias is meta-analysis detected by a simple graphical test. *British Medical Journal, 315*, 629-634.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*, 161-180.

Field, A. P. (2003). The problem in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics, 2*, 105-124.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population effect sizes vary? *Psychological Methods, 10*, 444-467.

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

Goldberg, W. A., Prause, J., Lucas-Thompson, R., & Himsel, A. (2008). Maternal employment and children's achievement in context: A meta-analysis of four decades of research. *Psychological Bulletin, 134*, 77-108.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

Highhouse, S., & Bottrill, K. V. (1995). The influence of social (mis)information on memory for behavior in an employment interview. *Organizational Behavior and Human Decision Processes*, *62*, 220-229.

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

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology, 81*, 459-473.

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

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*, 594-612.

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

Jennions, M. D., & Moller, A. P. (2002). Publication bias in ecology and evolution: An empirical assessment using the "trim and fill" method. *Biological Reviews, 77*, 211-222.

Judge, T. A., & Klinger, R. L. (2007). *Distal characteristics and the employment interview: an investigation of their relative and unique contributions to job performance*. Manuscript submitted for publication.

Le, H., & Schmidt, F. L. (2006). Correcting for indirect range restriction in meta-analysis: Testing a new analytic procedure. *Psychological Methods, 11*, 416-438.

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? *International Review of Industrial and Organizational Psychology, 16*, 246-286.

Light, L., & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Linn, R. L., Harnish, D. L., & Dunbar, S. B. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology, 66*, 655-663.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209.

Maurer, T. J., & Solamon, J. M. (2006). The science and practice of a structured interview coaching program. *Personnel Psychology, 59*, 433-456.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. III (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*, 63-91.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case of four test vendors. *Personnel Psychology, 59*, 927-953.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 799*, 599-616.

Motowidlo, S. J., & Burnett, J. R. (1995). Aural and visual sources of validity in employment interviews. *Organizational Behavior and Human Decision Processes*, *61*, 238-249.

Nelson, H. D., Humphrey, L. L., Nygren, P., Teutsch, S. M., & Allan, J. D. (2002). Postmenopausal hormone replacement therapy. *Journal of the American Medical Association, 288*, 872-881.

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

Oh. I.-S., Schmidt, F. L., Shaffer, J. A., & Le, H. (2008). The graduate management admission test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. *Academy of Management Learning & Education. 17,* 563-570.

Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157-159.

Palmer, D. K., Campion, M. A., & Green P. C. (1999). Interviewing training for both applicant and interviewer. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 337–352), Thousand Oaks, CA: SAGE.

Paunonen, S. V., Jackson, D. N., & Oberman, S. M. (1987). Personnel selection decisions: Effect of applicant personality and the letter of reference. *Organizational Behavior and Human Decision Processes*, *40*, 96-114.

Purkiss, S. L. S., Perrewe, P. L., Gillespie, T. L., Mayes, B. T., & Ferris, G. R. (2006). Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes*, *101*, 152-167.

Raju, N. S., & Brand, P. A. (2003). Determinig the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement, 27*, 52-71.

Ree, M. J., Caretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79*, 298-301.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79*, 518-524.

Robbins, S.B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Roth, P. L. (2008). Software review: Hunter-Schmidt Meta-analysis program 1.1. *Organizational Research Methods, 11,* 192-196.

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., Jr., & Schmit, M. J. (2005). Personality saturation in structured interviews. *International Journal of Selection and Assessment, 13*, 261-273.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005a). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Chichester, England: Wiley.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005b). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 1-7). Chichester, England: Wiley.

Ryan, A. M., McFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology, 52*, 359-392.

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*, 112-118.

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology, 11*, 299-324.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology* (Vol. 1, pp. 165-199). London, England: SAGE.

Schmidt, F. L. (1993). Personnel psychology at the cutting edge. In N. Schmitt & W. C. Borman and Associates (Eds.), *Personnel selection in organizations* (pp. 497-515). San Francisco, CA: Jossey-Bass.

Schmidt, F. L. (2008). Meta-analysis: A constantly evolving research integration tool. *Organizational Research Methods, 11,* 96-113.

Schmidt, F. L., & Hunter, J. E. (1992). Development of causal models of processes determining job performance. *Current Directions is Psychological Science, 1*, 89-92.

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

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

Schmidt, F. L., & Le, H. (2004). *Software for the Hunter-Schmidt meta-analysis methods*. Department of Management and Organizations. University of Iowa, Iowa City, IA.

Schmidt, F. L., Oh, I.-S., & Hayes, T. (2009). Fixed vs. random effects models in meta-analysis: An empirical comparison of differences in results in the psychological literature. *British Journal of Mathematical and Statistical Psychology, 62*, 97-128.

Please add this reference: Schmidt, F. L. (2010). How to detect and correct the lies that data tell. Perspectives on Psychological Science, 5, 233 – 242.

Schmidt, F. L., Oh, I.-S., & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology, 59*, 281-305.

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology, 61*, 827-868.

Schmidt, F. L., & Zimmerman, R. D. (2004). A counterintuitive hypothesis about employment interview validity and some supporting evidence. *Journal of Applied Psychology, 89*, 553-561.

Sitzmann, T. M., Kraiger, K., Stewart. D. W., & Wisher, R. A. (2006). The comparative effectiveness of web-based and classroom instruction: A meta-analysis. *Personnel Psychology, 59*, 623-664.

Sutton, A. J. (2005). Evidence concerning the consequences of publication and related biases. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 175-192). Chichester, England: Wiley.

Thorndike, R. L. (1949). *Personnel selection*. New York, NY: Wiley.

Toulmin, S. S. (1961). *Foresight and understanding: An enquiry into the aims of science*. New York, NY: Harper.

Townsend, R. J., Bacigalupi, S. J., & Blackman M. C. (2007). The accuracy of lay integrity assessments in simulated employment interviews. *Journal of Research in Personality, 41*, 540-557.

U.S. Merit Systems Protection Board. (2003). *The federal selection interview: Unrealized potential*. Washington, DC: Government Printing Office.

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

Vitaliano, P. P., Zhang, J., & Scanlan, J. M. (2003). Is caregiving hazardous to one's physical health? A meta-analysis. *Psychological Bulletin, 129*, 946-972.

Wiesner, 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.

Wilk, S. L., & Cappelli, P. (2003). Understanding the determinants of employer use of selection methods. *Personnel Psychology, 56*, 103-124.