PERSONNEL
PSYCHOLOGY
THE STUDY OF
PEOPLE AT WORK

PERSONNEL PSYCHOLOGY
2011, 64, 899–935

TOWARD BETTER META-ANALYTIC MATRICES: HOW INPUT VALUES CAN AFFECT RESEARCH CONCLUSIONS IN HUMAN RESOURCE MANAGEMENT SIMULATIONS

PHILIP L. ROTH
Department of Management
Clemson University

FRED S. SWITZER III
Department of Psychology
Clemson University

CHAD H. VAN IDDEKINGE
Department of Management
The Florida State University

IN-SUE OH
Department of Management
School of Business
Virginia Commonwealth University

Simulations and analyses based on meta-analytic matrices are fairly common in human resource management and organizational behavior research, particularly in staffing research. Unfortunately, the meta-analytic values estimates for validity and group differences (i.e., ρ and δ, respectively) used in such matrices often vary in the extent to which they are affected by artifacts and how accurately the values capture the underlying constructs and the appropriate population. We investigate how such concerns might influence conclusions concerning key issues such as prediction of job performance and adverse impact of selection procedures, as well as noting wider applications of these issues. We also start the process of building a better matrix upon which to base many such simulations and analyses in staffing research. Finally, we offer guidelines to help researchers/practitioners better model human resources processes, and we suggest ways that researchers in a variety of areas can better assemble meta-analytic matrices.

Some of the most central issues in staffing and human resource management are the validity of selection systems (e.g., Hunter & Hunter, 1984), the adverse impact against protected groups that can result from those systems (e.g., Aguinis & Smith, 2007; McKay, 2010; McKay & McDaniel, 2006; Reilly & Warech, 1993; Schmitt & Quinn, 2010), and

Correspondence and requests for reprints should be addressed to Philip L. Roth, Department of Management, Clemson University, Clemson, SC 29634-1305; rothp@clemson.edu.

withdrawal from selection and organizations in general (e.g., Scullen, Bergey, & Aiman-Smith, 2005). Researchers are increasingly turning to simulations to help understand how these and other HR practices and processes influence such outcomes and potential "trade-offs" (e.g., Harrison, Carroll, & Carley, 2007; Ployhart & Holtz, 2008; Sackett & Lievens, 2008; Sackett, Schmitt, Ellingson, & Kabin, 2001; see also Aguinis, Culpepper, & Pierce, 2010). HR simulations often use values from previous meta-analytic studies to form a matrix to generate simulated data (e.g., Hattrup, Rock, & Scalia, 1997; see also Finch, Edwards, & Wallace, 2009; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997).

Matrices derived from meta-analyses have also been widely used as a basis for path analysis and structural equation modeling (SEM; Viswesvaran & Ones, 1995). This is natural because meta-analysis has been characterized as revolutionizing the field of management and other related fields (Schmidt, Oh, & Hayes, 2009, see also McDaniel, Rothstein, & Whetzel, 2006), meta-analyses are cited roughly three times as often as primary empirical articles (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2010), and publication of meta-analyses have been increasing at a geometric rate since the late 1980s (Wanous, Sullivan, & Malinak, 1989, see also Aguinis, Dalton, et al., 2010). Furthermore, about half of meta-analytics articles use matrices to perform analyses such as path analysis or SEM within the same article (Aguinis, Dalton, et al., 2010). For example, Colquitt, LePine, and Noe (2000), in an article cited more than 500 times by researchers according to Google Scholar, used meta-analytic path analyses to analyze antecedents and consequences of training motivation, and Judge, Jackson, Shaw, Scott, and Rich (2007) used this approach to analyze correlates of self-efficacy.

A key set of issues in analyzing a meta-analytic matrix is the theoretical and methodological accuracy of the individual values comprising the matrix, which serve as the "input" to the subsequent analysis. In terms of simulations, researchers have noted that "the accurate compilation of such statistics is *the key* (emphasis added) to the accurate simulation of outcomes of the future selection process" (Doverspike, Winter, Healy, & Barrett, 1996, p. 263). Likewise, others have suggested that "input data values must be chosen carefully because the results of the calculation depend on them" (DeCorte, Lievens, & Sackett, 2006, p. 525). Yet, these concerns apply beyond just simulations and analyses. Meta-analysts have expressed concern over the accuracy of meta-analytic practices and values in research on organizational change (e.g., Bullock & Svyantek, 1985), strategic management (e.g., Geyskens, Krishnan, Steenkamp, & Cunha, 2009), and OB/HR (e.g., Aguinis, Dalton, et al., 2010).

Unfortunately, there appear to be relatively widespread potential problems in the use of meta-analytic matrices. Many simulation studies seek to model how unscreened applicants progress through a personnel system (e.g., Finch et al., 2009) or how faking behavior might influence who is selected (e.g., Berry & Sackett, 2009). Yet, the matrices used to generate simulated data are typically based on covariances/correlations that are restricted (e.g., based on incumbents and not job applicants). In other studies, researchers have failed to match the constructs to the values of the matrix (e.g., confusing task performance and overall job performance; Decorte et al., 2006, 2007). Thus, the subsequent analyses can provide problematic guidance to decision makers in designing personnel systems or understanding organizational phenomenon.

The purposes of this manuscript are to (a) highlight the importance of using accurate and theoretically appropriate population estimates in meta-analytic matrices and (b) investigate the importance of using accurate values by analyzing a pair of meta-analytic matrices. A key issue is the choice of the correct population for any meta-analytic matrix. That is, we do not just argue that "what goes in influences what comes out." Rather, we note that choosing the incorrect population values substantially and systematically distorts various results. Such practices can misdirect decision makers in their choices (e.g., which selection procedures to use) or potentially bias testing of various theories.

In order to accomplish these goals, we first discuss the importance of determining the appropriate population and then illustrate such thoughts with a relatively straightforward simulation. Later, we update values to a frequently used matrix to facilitate future simulation research and to demonstrate how the updated values systematically change results. Although we focus on simulation studies that rely on meta-analytic matrices to model staffing processes, we also discuss the implications of these issues for other research based on such matrices.

The Importance of Determining the Appropriate Population

As noted above, a typical goal in staffing simulations is to model a process in which simulated applicants initially apply for a given job and progress through a personnel system (e.g., Schmitt et al., 1997). Alternatively, some simulation studies incorporate selection and its subsequent relationship with turnover (e.g., Scullen et al., 2005). Thus, the appropriate population of data involves job applicants, and thus, covariances corrected back to the level of initial job applicants would be most appropriate for most such studies.

The logical approach of correcting back to the level of job applicants as the population of interest is also consistent with the routine use of such corrections in most bivariate meta-analyses of validity (e.g., Barrick & Mount, 1991; Huffcutt & Arthur, 1994; Hunter & Hunter, 1984;

Hurtz & Donovan, 2000; McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; Ones, Viswesvaran, & Schmidt, 1993), adverse impact potential (e.g., Roth, BeVier, Bobko, Switzer, & Tyler, 2001), and some nonsimulation based analyses of meta-analytic matrices (e.g., Schmidt, Shaffer, & Oh, 2008). In fact, the use of such corrections has been characterized as "expected" in selection research (Aguinis et al., 2010, p 7; see also Viswesvaran & Ones, 1995), and the lack of such corrections is seen as an important research limitation in most meta-analyses (note the cautions in Hunter & Schmidt, 2004; Whetzel, McDaniel, & Nguyen, 2008).

In other cases, the appropriate population may be job incumbents. For example, analysis of training motivation might focus on individuals already on the job (e.g., Colquitt et al., 2000). We illustrate the importance of choosing the correct population for staffing simulation research, as confusion over this concept seems particularly acute in this area (but we also note similar examples in other areas).

Current Input Matrices in Simulation Studies

To assess how widespread the above methodological concerns might be, we investigated the nature of the input values used in the meta-analytic matrices underlying staffing simulations. We searched for personnel simulation studies that were designed to help understand underlying personnel processes and outcomes in several leading personnel journals (i.e., *Journal of Applied Psychology*, *Personnel Psychology*, *Human Performance*, and *International Journal of Selection and Assessment*).

The largest group of studies in Table 1 focuses on validity and/or adverse impact. Early studies did not tend to tie their input matrix values to the literature as clearly as current studies. As time passed, the amount of description of the matrices has increased. For more recent studies aimed primarily at selection validity and/or adverse impact, 9 of the last 10 published studies in this table were conducted on uncorrected matrices. Further, three of the more recent studies appeared in the *Journal of Applied Psychology*. Thus, it appears that a large majority of such simulation studies are based on largely uncorrected matrices (and such studies appear in one of the premier journals). The trends are more mixed for other areas such as applicant response distortion and turnover/withdrawal studies (see Table 1).

Such trends are not necessarily unique to selection-based matrices and simulations. For example, Colquitt et al. (2000) analyzed training motivation, but they mix corrections such that some appear corrected back to the applicant level (e.g., the relationship between Conscientiousness and job performance from Barrick and Mount, 1991) whereas others are

TABLE 1
Use of Corrected and Uncorrected Matrices in Personnel Selection Simulations

Article	Use of corrected or uncorrected matrices
Selection Validity and/or Adverse Impact	
Sackett & Roth (1991)	Unclear
Doverspike et al. (1996)	Unclear
Sackett & Roth (1996)	Validities corrected for range restriction
Murphy & Shiarella (1997)*	Uncorrected
Sackett & Ellingson (1997)*[1]	Possibly not influenced by range restriction
Schmitt et al. (1997)*	Uncorrected
Hattrup et al. (1997)	Corrected and uncorrected matrices available
Bobko et al. (1999)*	Uncorrected
DeCorte (1999)	Uncorrected
Hattrup & Rock (2002)	Uncorrected
DeCorte & Lievens (2003)	Uncorrected
Potosky et al. (2005)	Corrected
DeCorte et al. (2006)	Uncorrected
DeCorte et al. (2007)	Uncorrected
Whetzel et al. (2008)	Uncorrected
Dunleavy et al. (2008)	Uncorrected
Finch et al. (2009)	Uncorrected
Personality Faking Studies	
Schmitt & Oswald (2006)	Unclear
Komar et al. (2008)	Corrected
Berry & Sackett (2009)	Uncorrected
Converse et al. (2009)	Some corrected & some not corrected
Turnover/Withdrawal Related Studies	
Tam et al.'s (2004) Study 1	Uncorrected[2]
Tam et al.'s (2004) Study 2	Unclear
Scullen et al. (2005)	Corrected

Notes. *Studies identified by Hattrup et al. (1997, p. 30) as using uncorrected values in simulations. We did not include Roth, Bobko, Switzer, and Dean (2001) in our analysis because it did not focus on the influence of personnel practices on outcome variables. Rather, it focused on biases in estimating *d* values.

[1] Standardized ethnic differences values (e.g., cognitive ability, dominance) were more hypothetical in nature than other such papers and appeared consistent with a concern to minimize range restriction (e.g., gender dominance $d = .5$).

[2] The authors defined their population of interest as being those individuals who passed a cognitive test. Thus, one could defend their choice as appropriate given the definitions in the article.

not corrected for range restriction (e.g., the relationship between general mental ability and job performance from Hartigan & Wigdor, 1989).[1]

[1] It is easily understandable how researchers might have trouble deciphering what the value of .30 from Hartigan and Wigdor (1989) might mean given the long and arduous discussion preceding this value in Chapter 8 (leading up to the value on p. 170).

Similarly, goal orientation studies have mixed the operational validity (i.e., no correction for predictor unreliability) of measures of general mental ability with true score correlations (i.e., correction for predictor unreliability) for personality measures and predictor intercorrelations in the same matrix (e.g., Payne, Youngcourt, & Beaubien, 2007).

Reasons for Use of Incumbent-Based Values in Selection Simulations

There are a variety of reasons that researchers give for focusing on uncorrected values in simulations. One group suggested, "we used uncorrected estimates, because we were focused on the operational validity of the predictors" (Finch et al., 2009, p. 323). A second group provided a mixed message by writing "we hope that other researchers will avail themselves of the values presented in this updated matrix in their simulations" (Bobko, Roth, & Potosky, 1999, p. 568), but the authors also stated "it is important to note that the estimates in these matrices are 'uncorrected' correlation estimates. . . .as such they are likely to be negatively biased estimates of latent correlations" (p. 563).[2] It is also interesting to note the trend for using uncorrected values within the validity and adverse impact studies in Table 1 appeared to solidify following publication of Bobko et al. A careful reading of their study suggests that methodological rigor in the compilation of meta-analytic matrices is very important, and we continue this theme below.

Study 1: Cognitive Ability and Work Samples

Input Values and Procedure

An organization might be contemplating the use of a cognitive ability test and a work sample to select employees. We investigate how three ways of assembling the "input" matrix can impact the results and provide recommendations for decision makers.

First, we investigated calculations based on the logic of existing simulation matrices using values from these matrices and augmenting them with work sample information (see the top panel of Table 2). For effect

[2] These authors also wrote (following the quoted material in the text above), "However, we based our initial thinking on Schmitt et al. (1997), who indicated that they worked with uncorrected correlations to consider the 'operational use' (p. 721) of potential sets of predictors." "It is important to point out that estimates of the operational validity of selection procedures would be corrected for unreliability in the job performance measure, as well as for range restriction . . . Nonetheless, we stay with Schmitt et al.'s convention for comparability purposes" (Bobko et al., [1999], p. 563). Thus, these authors appear to favor use of corrected matrices in simulation studies.

TABLE 2
Incumbent and Applicant Values for Work Sample Tests and Cognitive Ability Tests

Current or typical approach	Cognitive ability	Work samples	*d*
Cognitive ability			1.00 (Sackett et al., 2001)
Work samples	.32 ($k = 43, N = 17{,}563$) (Roth et al., 2005)		.38 ($k = 37, N = 15{,}738$) (Schmitt et al., 1996)
Job performance	.30 ($k = 515$, N not known) (Bobko et al., 1999)	.26 ($k = 54, N = 10{,}469$) (Roth et al., 2005)	.45 ($k = 40, N = 8{,}417$) (Bobko et al., 1999)
Composite validity	.34	Composite *d*	.85
Restricted (incumbent) values	Cognitive ability	Work samples	*d*
Cognitive ability			.41 ($k = 11, N = 3{,}315$) (Roth et al., 2001)
Work samples	.32 ($k = 43, N = 17{,}563$) (Roth et al., 2005)		.53 ($k = 12, N = 3{,}742$) (Roth et al., 2008)
Job performance	.30 ($k = 515$, N not known) (Bobko et al., 1999)	.26 ($k = 54, N = 10{,}469$) (Roth et al., 2005)	.27 ($k = 572, N = 109{,}974$) (McKay & McDaniel, 2006)
Composite validity	.34	Composite *d*	.65
Unrestricted (applicant) values	Cognitive ability	Work samples	*d*
Cognitive ability			.72 ($k = 18, N = 31{,}990$) (Roth et al., 2001)
Work samples	.42[1] ($k = 4, N = 1{,}156$) (Roth et al., 2005)		.73 ($k = 21, N = 2{,}476$) (Roth et al., 2008)
Job performance	.52 ($k = 151, N = 12{,}933$) (Hunter, 1986) ($k = 43, N = 4{,}744$) (Salgado et al., 2003)	.33 ($k = 54, N = 10{,}469$) (Roth et al., 2005)	.38 ($k = 572, N = 109{,}974$) (McKay & McDaniel, 2006)
Composite validity	.50	Composite *d*	.98

Note. [1]This value also converges with military studies that were not subject to range restriction (mean r of .48). d = standardized ethnic group difference.

sizes (which represent White–Black subgroup differences), we used the value of $d = 1.0$ for cognitive ability tests used by virtually all selection simulations (e.g., DeCorte et al., 2007; Finch et al., 2009), $d = .38$ for work samples based on incumbents (as per the frequently cited work of Schmitt, Clause, & Pulakos, 1996), and $d = .45$ for job performance (as per Bobko et al., 1999). For criterion-related validities, we used the uncorrected value of .30 for cognitive ability (e.g., DeCorte et al., 2006, 2007) and the uncorrected value of .26 for work samples (Roth, Bobko, & McFarland, 2005). We also used an uncorrected predictor intercorrelation of .32 (Roth et al., 2005). We labeled this the "typical approach."

Second, we investigated how using solely restricted/incumbent values to assemble all elements of a matrix might influence results. We set $d = .41$ for cognitive ability in order to have all values based on incumbents (as per Roth et al., 2001). We also found an updated *d* for work samples that was explicitly and solely based on job incumbents (.53, Roth, Bobko, McFarland, & Buster, 2008). We updated the *d* for job performance (.27) from a larger meta-analysis by McKay and McDaniel (2006; see the second panel of Table 2). We labeled this the "restricted approach." The consistency of this approach (e.g., all *d*s come from incumbents) might highlight the inconsistency of typical approaches as they mix an unrestricted value of *d* for cognitive ability and restricted values for other "cells" in the matrix.

Third, we assembled the matrix of values for unrestricted values, or what we called the "unrestricted approach," in the bottom panel of Table 2 (and we also use the term "applicants" to refer to the individuals within this approach, see Berry, Sackett, & Landers, 2007). We chose to focus on medium complexity jobs in this example, so we use the value of $d = .72$ for cognitive ability (Roth et al., 2001) and a *d* value of .73 for work samples because both illustrate unrestricted values based on job applicants (Roth et al., 2008). We also used the corrected validity of .52 for cognitive ability for medium complexity jobs averaged from two large scale meta-analyses (Hunter, 1986; Salgado, Anderson, Moscoso, Bertuna, Fruyt, & Rolland, 2003). The value of .52 is based on the average of two studies noted above. Specifically, we obtained the data from Hunter, which was corrected for direct range restriction (we continue to discuss the logic of this choice in more below). Next, we obtained the data from Salgado et al. and corrected the observed values for direct range restriction (and then we averaged the two values together). We also used the corrected intercorrelation of work samples and cognitive ability (.42).

To compare results across three different approaches discussed above, we initially examined the unit-weighted composite validity and adverse impact potential of the cognitive ability test and work sample. We used the program by DeCorte et al. (2006) to continue our analysis to an elementary sequential two-stage selection system in which the first hurdle was a measure of cognitive ability and the second hurdle was a work sample. For these multiple hurdle analyses we assume three scenarios such that Scenario 1 entails a selection ratio of .20 at both hurdles, Scenario 2 entails a selection ratio of .40 at both hurdles, and Scenario 3 entails a selection ratio of approximately .45 at both hurdles (as per scenarios from DeCorte et al., 2006). For both composite analyses and multiple hurdle analyses, we assume the sample comprises 20% Blacks and 80% Whites (e.g., see Bobko et al., 1999; DeCorte et al., 2006; Schmitt et al., 1997).

Results

Composite validity. Results varied markedly depending upon which values one chose to use in the "input matrix." In terms of validity, a unit-weighted composite was associated with a value of .34 for both the "typical" approach to matrix construction and for a consistently "restricted" matrix. In contrast, the estimated composite validity was .50 for the "unrestricted" (i.e., applicant) matrix (see Table 2).

Two things are important about these numbers. First, validity is misestimated for the typical approach and the restricted approach by .16. As a result, decision makers would not be informed of the actual level of validity, the average level of performance, or utility (e.g., Schmidt, Mack, & Hunter, 1984). This might lead to reluctance to invest in a staffing system (e.g., cognitive ability and work samples). Second, the composite validity based on the unrestricted matrix is roughly the same as it would be for cognitive ability alone, but decision makers might be willing to accept the additional expense in order to reduce adverse impact.

Composite adverse impact. Results in Table 2 also varied markedly for adverse impact potential. A unit-weighted composite based on the *typical* approach suggests that adverse impact potential (i.e., *d*) dropped from 1.00 for cognitive ability alone to .85 for the composite (when the work sample value of *d* was .38). We report composite *ds* as an illustration of focusing on effect sizes (Aguinis et al., 2010).

This analysis also allowed us to conduct sensitivity analyses to see if results changed based on changing a particular value (e.g., Stillwell, Seaver, & Edwards, 1981, see also Rich & Boudreau, 1987). Such a sensitivity analysis also can be thought of as one way researchers might use multiple values for meta-analytic covariances (*r* or *d*) so as not to be overly tied to only one value for a given relationship. One reason for using multiple values might be relatively wider credibility intervals from random effects meta-analytic models (Schmidt, Oh, & Hayes, 2009) or researchers might wish to consider additional levels of psychometric corrections (with larger or smaller *u* values to help explore the influence of range restriction; McDaniel, Whetzel, Schmidt, & Maurer, 1994).

An example of sensitivity analysis might be substituting the *d* of .53 for .38 because the value of .53 is based on a more comprehensive meta-analysis of job incumbents on work sample tests (Roth et al., 2008). In this case, the composite *d* was .95. In both cases, it appeared that adverse impact potential was *reduced* (from the level of 1.0 when using a cognitive ability test alone) to .85 or to .95 when the work sample was added.

Simulation results based on the consistent use of *restricted* values were associated with a composite *d* of .65. In this case, adverse impact potential *increased* from .41 to .65 as one moved from using a test of cognitive ability to a composite that included both ability and a work sample. If one used the frequently cited value of .38 for the work sample *d* (rather than .53), then the composite value is .55. In both cases, adverse impact was moderate in magnitude (and conclusions are not greatly sensitive to a change in values).

Results based on the *unrestricted* or applicant approach were associated with a composite *d* of .98. This value was .13 higher than the typical approach. Likewise, one could use sensitivity analysis and the value of $d = 1.00$ for cognitive ability (instead of the value of .72) and the composite *d* would be 1.17. Regardless of *d* value used for cognitive ability, the apparent decrease in adverse impact potential from the typical approach was *illusory* and the restricted approach provided much too small a point estimate of *d*. All told, decision makers and researchers were misinformed about the levels of both validity and adverse impact they may expect in operational settings with job applicants.

Multiple hurdles. We also conducted analyses examining sequential two-stage multiple hurdle selection systems in Table 3. Our comparisons involved projected levels of job performance and adverse impact (conceptually similar to Aguinis and Smith's 2007, p. 175, "expected" levels of various selection characteristics, e.g., expected hires).

For the restricted values (based on incumbents), Scenario 1 (with selection ratios of .20 for both hurdles) was associated with a value of .71 for the average level of job performance. This suggested that the average level of job performance of hires was .71 standard deviations above what would have occurred over random hiring. Further, the restricted approach was associated with a somewhat "optimistic" adverse impact ratio of .27 (as compared to the unrestricted value below). However, these values are logically flawed as they are based upon a restricted matrix (of incumbents). That is, although one had already hired the incumbents based on some system, now we were making projections based on hiring them *again* for the job they already hold.

Use of unrestricted values (based on applicants) showed higher levels of job performance at 1.00, which was .29 (or 41%) higher than the estimate based on the restricted values. The adverse impact ratio is .14, which suggested hiring approximately *half* the proportion of Blacks as the restricted/incumbent figure (.27) suggests. Similar results occur for Scenario 2 and Scenario 3 (as per the simulations from DeCorte et al., 2006).

We again illustrated the use of sensitivity analysis in such matrices with the validity for work sample exams. We substituted the values of

TABLE 3
Results of Multiple Hurdle Analyses of a Cognitive Ability Test and a Work Sample Exam

Scenario/approach	Average criterion score ("quality")[1]	AI ratio[2]
Scenario 1		
Applicants (unrestricted)	1.00	.14
Incumbents (restricted)	.71	.27
Typical	.74	.15
Scenario 2		
Applicants (unrestricted)	.71	.24
Incumbents (restricted)	.50	.40
Typical	.52	.26
Scenario 3		
Applicants (unrestricted)	.65	.27
Incumbents (restricted)	.46	.43
Typical	.48	.28
Sensitivity analysis for only the unrestricted/applicant matrix with .40 work sample validity		
Scenario 1		
Applicants	1.10	.14
Scenario 2		
Applicants	.78	.24
Scenario 3		
Applicants	.72	.27
Sensitivity analysis for only the unrestricted/applicant matrix with .45 work sample validity		
Scenario 1		
Applicants	1.17	.14
Scenario 2		
Applicants	.83	.24
Scenario 3		
Applicants	.76	.27

Notes. Scenario 1 entails a selection ratio of .20 at both hurdles, Scenario 2 entails a selection ratio of .40 at both hurdles, and Scenario 3 entails a selection ratio of approximately .45 at both hurdles (as per scenarios from DeCorte et al., 2006). For all analyses, we assume the sample comprises 20% Blacks and 80% Whites.
[1] Expected performance improvement in standard deviation form over what would have occurred hiring randomly.
[2] AI ratio = Adverse impact ratio: the selection ratio of the minority group divided by the selection ratio of the majority group.

.40 and .45 into our matrix in place of .33 (as decision makers might believe they have a particularly comprehensive work sample or wish to investigate different levels of values of *u*). Results for the unrestricted matrix are available in the middle and bottom panels of Table 3, and they suggest increases in the average level of job performance from 1.00 to

1.10 for a validity of .40 and further to 1.17 for a validity of .45 (and no changes in adverse impact as one might expect in such a two hurdle selection system).

To summarize, across both unit-weighted composites and two-hurdle selection systems, the use of the typical approach (using incumbent based values for all variables except the *d* for cognitive ability) and restricted approach appeared to present a systematically inaccurate picture of what will happen relative to values based on job applicants. Typical values were often too optimistic about adverse impact/minority hiring ratio in hiring and nearly universally too pessimistic about validity/expected job performance level for the simulated two-hurdle selection system. As such, we argue that correction for range restriction is particularly impactful for summarizing staffing research.

The results of this simulation are practically important in the realm of high stakes assessment (e.g., Sackett et al., 2001, see also Aguinis, Werner, et al., 2010). One reason for this is that validity/job performance and adverse impact/minority hiring ratio can have cumulative consequences for decision makers and organizations (Cortina & Landis, 2009). In terms of adverse impact, typical matrices suggest hiring a relatively larger number of minorities, but this may not occur in implementation of selection systems because actual implementation of selection systems will occur on job applicants. For example, use of applicants may result in an adverse impact ratio that is half of what is predicted (denoting fewer minority hires). This effect can be cumulated across factors such as the tenure of hires, the number of cohorts hired with an organization (e.g., Boudreau, & Berger, 1985), and the number of organizations using a type of selection system. Likewise, misrepresenting the validity of selection decisions or the efficacy of other HR systems (e.g., training, rewards) can deprive organizations of better performing employees, and, again, this consequence is multiplicatively cumulative across employees and tenure (e.g., Schmidt et al., 1984). It is also possible that underestimation of validity may cause decision makers to overestimate other aspects of personnel systems such as number of false negative hires (Arvey & Faley, 1988). We discuss the influence of misestimation on theory as well as other issues below in Study 2.

Study 2: Updating a Widely Used Matrix

Bobko et al. (1999) wrote perhaps the most highly cited article[3] in the area of HR simulations, and almost all the validity/adverse impact studies

[3] Bobko et al. (1999) is cited 107 times within the database of PsycINFO and 156 times in Google Scholar.

published thereafter have incorporated Bobko et al.'s values in their own input matrices. This matrix includes several popular predictors: cognitive ability tests, structured interviews, tests of Conscientiousness, and biodata instruments. The latest simulation study in personnel selection relies heavily on the work of Bobko et al. (1999). Finch et al. (2009) used this matrix and added the variable of integrity tests. We take this opportunity to update the Finch et al. matrix based on job applicant information. We use several inclusion criteria as we assemble this matrix. First, we look for unrestricted correlations and effect sizes/standardized group differences (e.g., estimates without prescreening as per Berry et al., 2007; Roth et al., 2001). Second, we look for validities that were corrected for criterion unreliability in order to model how using selection measures will influence underlying job performance (Hunter & Schmidt, 2004). That is, we look for "operational validities" (not corrected for predictor unreliability). Finally, we use data from studies designed to select employees rather than students whenever possible.

Values for the Input Matrices

Illustrations of the rationale behind input matrix value choices. We present the values for both the matrix used by Finch et al. (2009; and many previous researchers) and our updated matrix. Our inclusion criteria informed our choice of two key validities (that are illustrative of our process of choosing values). We believe the best point estimate of validity for biodata is .32 because it is corrected for range restriction and criterion unreliability (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990). This value is not greatly different than the value of .28 used by Finch et al. (2009) and others.

We spent substantial time examining the validity for cognitive ability tests. First, we combined the validities of Hunter (1986) and Salgado et al. (2003). This is advantageous because Hunter and Hunter based their results on the General Aptitude Test Battery in the U.S., and Salgado et al. used a number of cognitive ability tests given in Europe, so our results are cumulated across both U.S. and European samples across various tests. Second, we considered the criticism that the u values used to correct for range restriction by Hunter were potentially too small (and may have overcorrected mean correlations; Hartigan & Wigdor, 1989). However, there was substantial convergence between the mean validities across these two meta-analyses, which is interesting because Salgado et al. used u values derived from job applicants for specific jobs in their database.

We choose to use the validities corrected for direct range restriction in our analyses below for two reasons. First, there is substantial concern that a relatively small number of u values underlie corrections relative to the number of validity coefficients in many meta-analyses and that meta-analytic estimates are vulnerable to high levels of correction based on potentially unstable distributions of *u* (Schmitt, 2007; another side of this debate is presented by Schmidt, Le, Oh, & Shaffer, 2007). Thus, we remain conservative by using direct corrections in this case.[4] Second, we believe there is a need to avoid potentially large downward biases of range restriction. As such, the criterion-related validities for medium and low complexity jobs of .52. and .41 are higher than the value of .30 used by Finch et al. (2009) as well as many others (see Table 1). These two levels of job complexity represent the majority of jobs in the U.S. and Europe.

A key in examining validities for both biodata and cognitive ability is that the emphasis on corrected, applicant-level validities avoids/minimizes a pattern where prior validities were *differentially* influenced by range restriction (Hunter & Schmidt, 2004; Law, Schmidt, & Hunter, 1994). For example, the prior validity for biodata was only slightly downwardly biased, whereas the prior validity for cognitive ability tests was substantially downwardly biased. Avoiding this problem is critical to ensuring that the differential presence of artifacts does not "drive"/confound the results of the simulation as it can do in the "typical" approach to assembling matrices.

We also made efforts to find intercorrelations among predictors that were consistent with the logic of the matrix. For example, it was challenging to find the best value for the relationship between integrity and Conscientiousness. Finch et al. (2009) reported a value of .39. This value most likely came from the true-score correlation of .39 between overt integrity tests and Conscientiousness from Sackett and Wanek (1996; see Finch et al., p. 323). If this is the case, the value appears to be inconsistent with the focus on observed correlations in most other cells of Finch's matrix (the observed or attenuated value would likely be .26).

To arrive at the best value available for our unrestricted, applicant-level analyses, we consulted the article by Sackett and Wanek (1996), which draws its data from the work of Ones (1993). We then read the dissertation by Ones (1993) and corresponded with Ones. The observed correlation was .28 for all integrity tests (see Table 17) and the fully corrected correlation was .42. However, we only wanted the intercorrelation of these measures corrected for range restriction (given this was a predictor intercorrelation). Ones reported the *u* value for integrity tests was .81. This leads to a correlation corrected for only range restriction of .34. This is the value we used in our analyses. We now turn to other values in our

[4] Validities corrected for indirect range restriction will likely increase the operational validities for cognitive ability and indirect correction values are available (Schmidt et al., 2008).

matrix, which were selected using the same inclusion rules as the values described above.

Predictor standardized ethnic group differences. Table 4 shows both the heavily incumbent-based estimates from past efforts (e.g., Finch et al., 2009) and our updated applicant-based estimates. Although we focus on the values used by Finch et al. because they are one of the most recent studies, we note that many of the values used by Finch et al. have been used by many of the simulation studies since Bobko et al. (1999). That is, Finch et al. followed what appeared to be a precedent-setting article in the literature and were acting consistently (and with good faith) with common practice in the field (again, see Table 1).

We use a *d* value of .57 for biodata predictors based on two studies (one in which a correction for range restriction was available and one in which there was no range restriction; Dean, 1999, and Kriska, 2001).

We use *d* values of .72 and .86 for a test of cognitive ability for job applicants for medium and low complexity jobs, respectively (Roth et al., 2001). These values are somewhat different than the value of 1.00 used by other simulation researchers (though the results are not atypical of other meta-analyses, e.g., Sackett & Shen, 2010).

We use a *d* value of .06 for Conscientiousness from Potosky, Bobko, and Roth (2005) because (a) it is based on unrestricted samples and (b) it was based only on measures of Conscientiousness. The value of .09 used by previous authors appears to have been based on multiple dimensions of personality (e.g., see the work of Schmitt et al., 1997 and Bobko et al., 1999). The value of .06 is also consistent with other minimal differences for ethnic groups on personality (e.g., Foldes, Duehr, & Ones, 2008; Ones & Viswesvaran, 1998).

We use a *d* value of .32 for the structured interview (Potosky et al., 2005). It is based on Huffcutt, Roth, and McDaniel's (1996) corrected for range restriction and, the unrestricted value from Roth et al. (2002). Potosky et al. weighted these two studies together by sample size. Although we note the importance of considering the role of constructs within various method based predictors (see Arthur & Villado, 2008), we continue to parallel previous analyses by using the same predictors in order to demonstrate the effects of changing the input values.

Finally, we use a *d* value of .04 for integrity tests based on the meta-analysis by Ones and Viswesvaran (1998). It is not greatly different than the value of .00 used by Finch (2009), but we use corrected values from meta-analyses when available.

Operational predictor validities. Table 5 presents validities for overall job performance. We focused on operational validities (corrections for criterion unreliability and range restriction) rather than true-score correlations (corrected for predictor and criterion reliability as well as range

TABLE 4

Standardized Ethnic Group Differences (d) for Predictors of Job Performance

Predictor	Uncorrected values[1]	Corrected values	Comments
Biodata	.33 (Bobko et al., 1999) ($k = 2, N = 6{,}115$)	.57 (Potosky et al., 2005) ($k = 2, N = 6{,}115$)	No data to look at constructs.
Cognitive ability	1.00 (Bobko et al., 1999)	.72/.86 (Roth et al. 2001) ($k = 18, N = 31{,}990$) ($k = 64, N = 125{,}654$)	Data medium (.72) and low (.86) complexity jobs from job applicants.
Conscientiousness	.09 (Bobko et al., 1999) ($k = 6, N = 801$)	.06 (Potosky et al., 2005) ($k = 3, N = 4{,}545$)	Potosky et al. values from data that were not restricted and only from tests of Conscientiousness.
Structured interview	.23 (Bobko et al., 1999) ($k = 21, N = 8{,}817$)	.32 (Potosky et al., 2005 & Roth et al., 2002) ($k = 22, N = 9{,}175$)	
Integrity	.00 (Finch et al., 2009) (Unknown)	.04 (Ones & Viswesvaran, 1998) ($k = 4, N = 481{,}523$)	Data from job applicants from over 100 organizations. Reported k is number of publishers.

Note. [1] Uncorrected values for biodata, cognitive ability, Conscientiousness, and structured interviews are from Bobko, Roth, and Potosky (1999). These four values were also used by DeCorte et al. (2006). The integrity value is from the work of Finch et al. (2009).

TABLE 5

Validity Estimates for Predictors of Job Performance

Predictor	Uncorrected values[1]	Corrected values	Comments
Biodata	.28 (Bobko et al., 1999) ($k = 5, N = 11{,}332$)	.32 (Rothstein et al., 1990) ($k = 5, N = 11{,}332$)	Rothstein et al. corrects for range restriction and unreliability
Cognitive ability	.30 (Bobko et al., 1999) ($k = 515$, no N given)	.52/.41 ($k = 194, N = 17{,}677$) ($k = 213, N = 15{,}267$)	Medium and low complexity jobs; data from Hunter (1986) and Salgado et al. (2003)
Conscientiousness	.18 (Barrick & Mount, 1991, 3 other studies)	.22 (Hurtz & Donovan, 2000) ($k = 42, N = 7{,}342$)	Hurtz and Donovan focus on measure based on the five factor model
Structured interview	.30 (Bobko et al., 1999) (Several meta-analyses)	.48 (Potosky et al., 2005) (Two meta-analyses)	Potosky et al. use only meta-analyses that focus on job performance
Integrity	.25[2] (Ones et al., 1993) ($k = 25, N = 7{,}831$)	.42 (Ones et al., 1993) ($k = 25, N = 7{,}831$)	Values for applicant populations (Table 6 in Ones et al., 1993)

Notes. [1] Uncorrected values for biodata, cognitive ability, Conscientiousness, and structured interviews are from Bobko, Roth, and Potosky (1999). These four values were also used by DeCorte et al. (2006). The integrity value is from the work of Finch et al. (2009).
[2] We found it difficult to trace the value from Finch et al. (2009) but it appears to be from Ones et al. (1993).

restriction). We made this choice because we were interested in how selection systems would influence organizations in terms of underlying performance (though some practitioners or researchers may choose to focus on observed performance and might forgo reliability corrections). As noted above, we believe the best point estimate of operational validity for biodata is .32. We also believe that the values of .52 (for medium complexity jobs) and .41 (for low complexity jobs) are good values for the validity of cognitive ability tests.

The use of two different values for validity for cognitive ability tests (as well as *d*s) is important to illustrate the use of multiple values in simulations and related meta-analytic analyses (Schmitt et al., 1997, see also Aguinis, Pierce, et al., 2011). The use of multiple values as inputs in simulations is particularly important when studying inputs to decisions to see if multiple levels of parameter estimates might influence results (Rich & Boudreau, 1987). For example, the variable of job complexity is theoretically important for understanding validity estimates and subgroup differences for cognitive ability tests (e.g., Hunter, 1986). As such, these analyses can keep researchers and decision makers from becoming too dependent or confident upon one given value per variable.

We believe the value of .22 from Hurtz and Donovan (2000) is a good estimate of the validity for Conscientiousness because it is corrected for range restriction and for criterion unreliability, and because it only includes measures designed to assess Conscientiousness (also, the meta-analysis by Hurtz and Donovan focused on personality measures designed to capture the five factor model of personality). Further, this value is virtually the same as the value reported in Schmidt et al. (2008) in which six relatively independent prior meta-analyses on Conscientiousness are aggregated.

Regarding interview validity, we believe that the value of .48 from Potosky et al. (2005) is a better estimate than the previously used value of .30 used by many previous simulations. Although the two meta-analyses used by Potosky et al. have some overlapping samples, the estimate of .48 addresses the problem of downward bias in individual validities due to corrections for range restriction and unreliability. Finally, we use a validity of .42 for integrity tests. Again, this is a corrected value corrected to the level of job applicants.

Two thoughts may be important for our validity estimates. First, we reiterate the importance of differential influence of artifacts and range restriction in particular (e.g., Law et al., 1994). In addition to differential influences on cognitive ability and biodata noted above, there was also differential influence of artifacts on predictor validities of Conscientiousness measures (with relatively low restriction) and structured interviews (with relatively high restriction). Second, the validity values for job performance from Finch et al. (2009) and Bobko et al. (1999) are for overall performance. In some previous studies, these values have been used as representing only task performance (e.g., DeCorte et al., 2006, 2007). Theoretically, it is likely that overall job performance is a function of both task and contextual performance (Johnson, 2001). So, equating overall job performance figures from a series of meta-analyses designed to look at that variable with task performance might not be conceptually appropriate. One possible implication of such a practice is that the influence of contextual performance is captured twice — once in the correlations between predictors and contextual performance and again in the correlations between predictors and overall performance (and the issue relates to judgments concerning the operational definition of variables in studies for potential inclusion in meta-analysis as per steps 2c and 4 as noted by Wanous et al., 1989). That is, we suggest careful consideration of operational definitions (and interpretations) of both predictors (independent variables) and criteria (dependent variables) later in the paper.

Intercorrelations. Modeling multiple-hurdle or composite selection systems requires one to estimate the predictor intercorrelations. Of course, such intercorrelations can also be influenced by range restriction (see Sackett, Lievens, Berry, & Landers, 2007). In Table 6, we suggest the use of several values that are not restricted in range, but we caution readers that sample sizes tend to be small. In some cases, there appear to be no available estimates of unrestricted predictor intercorrelations. The use of restricted values and small sample sizes is somewhat troubling given the importance of predictor intercorrelations in the results of such work (DeCorte et al., 2006; Sackett & Ellingson, 1997 for the importance of intercorrelations).

We first discuss biodata-related correlations and suggest in Table 6 that a better estimate of the biodata–cognitive ability correlation is .37 because it is not range restricted (Potosky et al., 2006). Unfortunately, there appear to be no unrestricted estimates for the relationship between biodata and Conscientiousness or biodata and structured interviews. Finch et al. used the values of .51 and .16 (see Bobko et al., 1999 for more discussion of these values).

Correlations between cognitive ability and other predictors are also reported in Table 6. We used .03 for the relationship between cognitive ability and Conscientiousness (based on a small meta-analysis focusing on unrestricted samples from Potosky et al., 2005). We suggest that a better estimate for the cognitive ability–structured interview relationship is .31 because this value is corrected for range restriction (again, we used results from Potosky et al., 2005). We did not use the values from Berry et al. (2007) as their study included data from interviews designed to select students (e.g., for graduate school) as well as employees, and our interest was in employee selection. We also suggest that a better value for

TABLE 6

Intercorrelations Among Predictors of Job Performance

Predictor pair	Uncorrected values[1]	Corrected values	Comments
Biodata— cognitive ability	.19 (Bobko et al., 1999) ($k = 2, N = 1{,}363$)	.37 (Potosky et al., 2005) ($k = 2, N = 5{,}475$)	
Biodata— Conscientiousness	.51 (Bobko et al., 1999) ($k = 2, N = 1{,}363$)	None available	
Biodata— structured interview	.16 (Bobko et al., 1999) ($k = 2, N = 1{,}046$)	None available	
Cognitive ability— Conscientiousness	.00 (Bobko et al., 1999) ($k = 2, N = 4{,}504$)	.03 (Potosky et al., 2005) ($k = 7, N = 6{,}759$)	Potosky et al.'s data is from unrestricted samples
Cognitive ability— structured interview	.24 (Bobko et al., 1999) ($k = 41, N = 8{,}890$)	.31 (Potosky et al., 2005) ($k = 21, N = 8{,}817$)	Potosky et al.'s data corrects the value of .24 for range restriction

Continued

TABLE 6 (continued)

Predictor pair	Uncorrected values[1]	Corrected values	Comments
Structured interview— Conscientiousness	.12 (Bobko et al., 1999) ($k = 1, N = 465$)	.13 (Salgado & Moscoso, 2002) ($k = 13, N = 1{,}497$)	Salgado & Moscoso (2002) data is corrected for range restriction
Integrity— cognitive ability	.00[2] (Finch et al., 2009)	.02 (Ones, 1993) ($k = 106, N = 23{,}306$)	New value is the exact value from Ones (1993)
Integrity— structured interview	.00[3] (Van Iddekinge et al., 2004) ($k = 1, N = 427$)	-.02 (Van Iddekinge et al., 2004) ($k = 1, N = 427$)	Correlation from one sample of job applicants (Van Iddekinge et al., 2004)
Integrity— Conscientiousness	.39 (Finch et al., 2009) (Unclear)	.34 (Ones, 1993) ($k = 423, N = 91{,}360$)	Value likely from Sackett and Wanek (1996)
Integrity— biodata	.25 (McFarland & Ryan, 2000) ($k = 1, N = 192$)	.25 (McFarland & Ryan, 2000) ($k = 1, N = 192$)	Value from a lab study of undergraduates in an honest condition

Notes. [1]Uncorrected values for biodata, cognitive ability, Conscientiousness, and structured interviews are from Bobko, Roth, & Potosky (1999).
[2]The integrity values are from the work of Finch et al. (2009).
[3]It appears that the value of .00 by Finch et al. was based on the work of Van Iddekinge et al. (2004).

the structured interview–Conscientiousness relationship is .13 because this estimate is corrected for range restriction and is based on substantial sample size (Salgado & Moscoso, 2002).

There are also a number of correlations with integrity tests, though sometimes the data are sparse. For the integrity–Cognitive ability correlation, we use the value of .02 from Ones (1993, p. 158). This is not greatly different than the value of .00 from Finch et al. (2009), but we use the most accurate meta-analytic data available.

For integrity and structured interviews, it appears that Finch et al. (2009) referred to Van Iddekinge, Raymark, Eidson, and Attenweiler (2004). We averaged the two values for this correlation (one from each form of the interview by Van Iddekinge et al.) from this primary study and use a value of -.02. For integrity and Conscientiousness, we reiterate the value of .34 noted above. For integrity and biodata, it appears that the only unrestricted value is .25 based on a primary study of 192 undergraduates in an experimental condition in which they were to honestly report their integrity scores (McFarland & Ryan, 2000).

Criterion ethnic group differences on performance. We suggest a better value for the Black–White *d* for overall job performance is the value of .38 from McKay and McDaniel (2006; and it is corrected for unreliability). We again focus on only two ethnic groups and assume these groups made up 20% and 80% of the population, respectively.

Results

We use our updated matrix to reexamine two sets of practical and important issues from meta-analytic matrices that underlie HR simulations (but the principles likely also apply more broadly). We first apply our new matrix to understanding the combined validity of the predictors. Second, we examine previous research on avoiding adverse impact in employee selection.

Multiple regression analyses. We use multiple regression analysis to compare uncorrected and corrected estimates of the standardized regression (beta) weights for our five predictors of job performance in Table 7. We assume a relatively straightforward model in which the five predictors are immediate precursors to job performance for medium complexity jobs as an illustration. The multiple *R* for the entire model increases substantially from .48 to .75 as one moves from uncorrected inputs to corrected validity estimates in the matrix. Perhaps of more interest are the changes in the beta weights. Several beta weights for the more valid predictors such as cognitive ability tests increased markedly (from roughly .20 to roughly .40). Perhaps of even more interest is that the beta weight for biodata decreases from .14 to .00 as other, more valid, predictors

TABLE 7

Regression Analysis of Job Performance on the Predictors of Biodata, Cognitive Ability, Conscientiousness, Structured Interviews, and Integrity Tests

Input matrix	Uncorrected matrix	Corrected matrix
Predictor	β	β
Biodata	.14	.00 (.07)
Cognitive ability	.22	.40 (.25)
Conscientiousness	.03	.02 (−.01)
Interview	.22	.36 (.40)
Integrity	.21	.41 (.41)
Multiple *R*	.48	.75 (.70)

Note. β = standardized regression weight; medium job complexity results are presented first and low complexity results are presented in parentheses.

receive greater weight in the corrected matrix. Results are similar, though not quite as dramatic for low complexity jobs (also in Table 7). These findings suggest that some variables might be included in inductive theory building based on uncorrected results but not included when using corrected results. Further, the size of the effects for other variables depends substantially on the application of corrections (again, moving from roughly .20 to almost .40). Relatedly, theory testing results might be somewhat different if variables such as biodata were associated with beta weights of near zero. Of course, such results could vary depending upon what constructs are targeted by biodata (or potentially interviews). Our point is that given current practices in input matrix development, it is possible that use of corrections can influence theory development and testing such that certain variables are more impacted than others in terms of inclusion or exclusion and relative standing in theoretical models.

Selected results from Finch et al. (2009). One of the interesting findings by Finch et al. (2009) was that there were a number of selection systems that were associated with no adverse impact according to the 4/5ths rule (see Finch et al., table 4).[5] We focus on the 4/5ths rule and medium complexity jobs in order to keep our illustrations parsimonious. We also focus on this particular table given the importance of the issue of adverse impact and to see how results of simulations might change based on the use of unrestricted inputs. We used the program developed by DeCorte et al. (2006) to compute estimates of minority hiring ratio.

[5] Finch et al. state "in contrast to previous research, the current simulation demonstrated numerous strategies that produced no adverse impact" (p. 326).

TABLE 8

Results for Scenarios Previously Shown to Have No Adverse Impact According to the 4/5ths Rule With Results for Applicant Level Analyses (n = 500, 20% Minority Representation)

Predictor combination	Selection ratios	Applicant AIR	Finch et al. AIR	Applicant job performance	Finch et al. job performance
Biodata, Consc.	.60/.50	.72	.80 (.80)	.32	.27
	.75/.40	.82	.85 (.84)	.29	.24
Integrity, (Biodata + Consc. + SI)	.45/.67	.73	.83 (.80)	.60	.40
Integrity, (Biodata + Consc.)	.15/.67[1]	.75	.86 (.88)	.77	.50
	.45/.67	.77	.86 (.84)	.50	.34
Integrity, (Biodata + SI)	.45/.67	.66	.80 (.80)	.64	.42
Integrity, biodata	.15/.67[1]	.65	.82 (.82)	.79	.52
	.45/.67	.67	.82 (.80)	.52	.36
Integrity, (Consc.+ SI)	.15/.67[1]	.81	.88 (.88)	.87	.54
	.45/.67	.83	.88 (.88)	.60	.38
	.60/.50	.79	.83 (.80)	.61	.40
	.75/.40	.76	.80 (.80)	.60	.40
Integrity, Consc.	.15/.67[2]	.91	.95 (.98)	.71	.44
	.25/.40	.90	.91 (.88)	.65	.43
	.45/.67	.94	.95 (.96)	.44	.28
	.60/.50	.93	.93 (.92)	.38	.27
	.75/.40	.93	.91 (.92)	.33	.24
Integrity, SI	.15/.67[1]	.78	.88 (.88)	.92	.55
	.45/.67	.80	.88 (.88)	.64	.39
	.60/.50	.74	.82 (.84)	.66	.41
	.75/.40	.70	.79 (.80)	.65	.40
Biodata, integrity	.60/.50	.68	.82 (.84)	.52	.36
	.75/.40	.77	.89 (.88)	.52	.33

Notes. Consc. = Conscientiousness; SI = structured interview. Estimates of the Finch et al. (2009) results using DeCorte's (2006) programs are reported with the values reported by Finch et al. in parentheses.

[1] These selection scenarios involve a total or net selection ratio of .10. All other involve a total or net selection ratio of .30. We use these selection ratios because they are the ratios used by Finch et al. (2009).

We report results comparing Finch et al.'s more typical, mostly uncorrected matrix to our updated, corrected applicant matrix in Table 8. Table 8 contains information on the predictor combination (e.g., "Biodata. Consc." represents a two-hurdle selection system with a biodata screen followed by a measure of Conscientiousness) and the selection ratios for each predictor (e.g., ".60, .50" indicates the biodata selection ratio was .60 and the Conscientiousness ratio was .50, as per Finch et al.). We also report the applicant adverse impact ratio from our matrix to augment Finch et al.'s analyses. Given we did not have Finch et al.'s algorithms, we used the programs by DeCorte et. al (2006) to estimate the adverse impact ratio (AIR), and these are placed under the header of "Finch et al. AIR" (we follow our estimation of the Finch et al. results using DeCorte's programs with the actual values reported by Finch et al. in parentheses so that readers can see results both ways). Finally, we report the results for job performance for our updated applicant matrix compared to Finch et al.'s matrix (and positive values represent the number of standard deviations mean that performance is above the mean level of performance if employees were selected randomly).

We note three trends to illustrate the changes in results in Table 8. First, the lack of adverse impact disappears in 14 of the 23 selection systems when unrestricted applicant values are used to model selection. Thus, advice to researchers and decision makers can change when applicant-based values are used in the input matrix.

Second, the remaining scenarios with no adverse impact appear to share two characteristics. Many of the scenarios involve only personality-based measures for both hurdles. For example, the integrity and then Conscientiousness selection systems account for five of the remaining no-adverse-impact scenarios. Other scenarios involve selection ratios of .60 or greater on the moderate *d* predictor (see Sackett & Ellingson, 1997 for a discussion of such a trend). For example, the integrity then structured interview scenario involves an interview selection ratio of .67. A possible interpretation of these trends is that only personality-based prediction systems, or systems with high selection rates for predictors with moderate *d* values (e.g., biodata and structured interviews), allow organizations to escape violations of the 4/5ths rule.

Third, use of incumbent-based estimates underestimates the predicted level of job performance, often by 33–50%. For example, the integrity then biodata, Conscientiousness, and structured interview system predicts the average level of job performance to be .40 standard deviations above the mean associated with random selection, whereas the applicant data suggest the value will be .60.

Overall, results changed markedly when we used applicant-level inputs into the correlations matrix, and such a majority of selection systems thought to avoid adverse impact actually result in adverse impact (there are few easy solutions to subgroup differences and adverse impact; Schmitt, Sackett, & Ellingson, 2002). Further, there is a downward bias in results aimed at understanding levels of job performance.

Discussion

Changed Results

Our purpose was to reconsider the choice of values in meta-analytic input matrices underlying personnel simulations and analyses (and SEM models). We believe that use of unrestricted (or applicant) values can change results. In our matrix of cognitive ability and work samples, use of the typical approach of primarily restricted values suggested that adverse impact potential (i.e., *d*) would drop when adding a work sample to a test of cognitive ability. Use of applicant values shows the composite *d* would *increase* substantially over and above the use of a cognitive ability test. Further, validities were changed substantially from .34 using typical or restricted approaches to .50 using unrestricted/applicant data. The change of .16 (.34 to .50) is roughly three times the change due to corrections for range restriction in previous research in general and research in selection in particular (Aguinis et al., 2010). The large change may be due to the markedly higher levels of range restriction for cognitive ability variables (Schmidt, Oh, & Le, 2006; Schmidt et al., 2008).

We also updated perhaps the most influential matrix in HR simulation research (i.e., Bobko et al., 1999). In the first application of this matrix, we demonstrated that multiple regression validity results changed in terms of the magnitude of beta weights and which predictors were related to the job performance (e.g., biodata's beta weights changed markedly). Second, we focused on previous results that suggested 23 selection systems that might avoid adverse impact, and we found the majority of such systems resulted in adverse impact (and underestimated levels of job performance).

Implications and Suggestions for Future Simulation Studies and Other Research

This manuscript appears to have implications for researchers using simulations and researchers assembling meta-analytic matrices to underlie SEM analyses. The overall guiding principle is that meta-analyses and meta-analytic matrices should be compiled with care (Aguinis, Pierce, et al., 2011; Bobko & Stone-Romero, 1998; see Table 9). We order these suggestions such that those with the most applicability to HR simulations (and staffing) are set forth first and our suggestions increase in importance towards other uses of meta-analytic matrices in organizational research (e.g., use of meta-analyses as inputs for SEM) as we progress toward our latter suggestions.

TABLE 9

Suggestions for Assembling a Meta-Analytic Matrix for Future Simulation or Analytic Solution Research

1	Correct correlations and standardized group differences for range restriction, as appropriate	Researchers who wish to model applicants should use corrected values in their matrix such that the values are not influenced by range restriction. Alternatively, researchers might find results from samples that are not influenced by range restriction due to prior organizational selection (e.g., Berry et al., 2007; Roth et al., 2001).
2	Carefully consider the role of constructs in the criteria	Researchers should clearly delineate what criteria are of interest. Criteria might include overall job performance, training success, withdrawal, or deviance. Similarly, researchers should consider if they are primarily interested in task performance, contextual performance, or overall job performance. Data from primary studies or meta-analyses that for the input matrix should be carefully screened to match the population(s), goals, or objectives of the study.
3	Carefully consider the role of constructs in the predictors	Again, data from primary or meta-analytic studies should match the predictors as conceptualized by the researchers or decision makers. For example, care should be taken to use data for a particular personality dimension rather than aggregating across personality dimensions.
4	Researchers and practitioners should consider using ranges of values	Ranges of values can be useful if decision makers want to systematically vary an aspect of the situation (e.g., percentage of minorities), believe there is uncertainty in some of the meta-analytic values (SD_ρ is greater than zero), or if there needs to be investigation into how sensitive the solution is to variance in the value of a key variable (e.g., lack/paucity of data available for statistical corrections).
5	Report the source for all values in the meta-analytic matrix	Readers should be able to readily tell where each value in the meta-analytic matrix originates. Researchers should report the source and sample sizes.
6	Researchers might consider the nature of the target job, including level of job complexity	Meta-analytic matrices designed for use as decision aids should consider the nature of the job targeted for decisions. For example, certain jobs may involve a great deal more customer contact and others more technical problem solving. Researchers may wish to use as many coefficients from studies that are similar to the target job as possible. One key variable could be job complexity where validity or standardized group differences could vary by complexity (as per Hunter, 1986; Hunter & Hunter, 1984).
7	Carefully consider the role of corrections for criterion unreliability	Researchers should determine if they are interested in studying the actual influence of HR systems on operational job performance (or other constructs) or if they are interested in studying the influence of HR systems on observed job performance. We generally lean towards the former, but the objectives of the simulation may vary.
8	Consider the consistency of corrections throughout the matrix	HR and OB researchers should determine if they are interested in correlations corrected for various factors (e.g., predictor unreliability) such that all values in a matrix are logically consistent with each other (e.g., all true correlations or operational validities). Likewise, mixing corrected and uncorrected correlations without considerable justification may be problematic.

First, researchers and practitioners should carefully deal with the artifact of range restriction consistent with study goals (see also Geyskens et al., 2009, in strategic management research). The issue of range restriction is particularly impactful in HR simulations where the population being modeled is job applicants, and input statistics should be consistent with that population. Further, there does not appear to have been wide application of this principle in HR simulations, which is somewhat surprising given the acceptance of such corrections in many selection meta-analyses. Specific manifestations of this principle for personnel selection involve examining validity or adverse impact of applicants from initial application through simulated selection systems by initially choosing statistics corrected for range restriction (or not subject to range restriction as per Berry et al., 2007; Roth et al., 2001) in their input matrices because such artifacts are typical in selection studies (Aguinis, Culpepper, & Pierce, 2010). Alternatively, studies focusing on downsizing in organizations might consider using incumbent based (restricted values that are not corrected for range restriction), as their purpose is to measure the influence of downsizing on various work outcomes based on individuals currently performing the job in question.

Second, researchers should carefully define their predictor or independent variables (Wanous et al., 1989). For example, one might need to be careful in defining work samples in order to be clear whether situational judgment tests are included or not in the database (and results). Likewise, researchers should try to consider the influence of various constructs (e.g., cognitive ability, social skills) within such methods of measurement (Arthur & Villado, 2008). One way to construct such matrices would be to find data from the literature (or company files) that considers how various constructs are measured using certain methods (e.g., what is the *d* for social skills when using situational judgment tests to measure such skills). Again, this issue may be particularly acute in HR simulations as many predictors in HR are methods (e.g., interviews, work samples).

Third, researchers and practitioners should carefully define and model job performance and other dependent variables (Viswesvaran & Ones, 1995). Within HR simulations (and as noted above), it is easy to potentially confuse overall job performance with task performance, perhaps due to the lack of data in this area. We suggest that prior matrices (e.g., Schmitt et al., 1997; Schmidt et al., 1998) have focused primarily on overall job performance as the dependent variable. Our recommendation is generally to focus on overall job performance as there are data sufficient to make stable meta-analytic estimates at the present time. Alternatively, researchers may wish to explicitly and carefully measure both task performance and contextual performance (or even counter productive work behaviors) as they assemble their matrices.

Fourth, researchers and practitioners should consider the use of a range of values in their input matrices (Aguinis, Pierce, et al., 2011; Schmitt et al., 1997). Although the primary purpose of this manuscript was to present better starting point-estimates, we also suggest there are a variety of reasons that researchers and practitioners may wish to use a range of values in their simulations. The range of values might involve a desire to look at different levels of a variable in terms of "experimental design." In HR simulations, decision makers might want to see if job complexity or the percentage of minorities influences decisions (Schmitt et al., 1997). Other researchers might suggest that there is often variability in meta-analytic estimates after variance attributable to artifacts is removed from analyses and such variance should be modeled (e.g., DeCorte et al., 2006; McDaniel, Hartman, Whetzel, & Grubb, 2007; Whetzel et al., 2008) or that one should consider typical, best, and worst types of scenarios. One might consider using the observed value of *r* or *d* as a lower value in a range of values as one approach. Such sensitivity analyses could be used more widely in OB/HR or strategy meta-analytic matrices used as a basis for path analysis or SEM. Although such use may not be typical, the logic of approaches such as sensitivity analysis is similar.

Fifth, researchers and practitioners should clearly articulate the source of their estimates (see Schmitt et al., 1997 and Schmidt et al., 2008 for good examples). The current simulation literature generally notes that meta-analytic matrices were used but seldom lists where estimates came from on a "cell by cell" basis. Having authors explicitly note where each estimate originated would help replication and extension of these studies. Although we have attempted to model such an approach, we are not the first researchers to do so (e.g., see Judge et al., 2007 and Schmitt et al., 1997).

Sixth, researchers might consider the nature of the target job. We believe that such concerns are important for both HR simulations and wider applications of meta-analytic matrices. Within HR simulations, job complexity is seldom modeled (though we made efforts to begin this process). Likewise, the nature of some jobs could be markedly different than other jobs. For example, certain jobs may involve a great deal more customer contact and others more technical problem solving (e.g., managerial jobs).

Seventh, researchers may wish to correct their validities for criterion unreliability if they wish to model how various organizational practices influence changes in "true" or underlying job performance (e.g., Scullen et al., 2005). It is also possible that practitioners may prefer not to correct for criterion unreliability. The choice of whether to correct for this artifact appears to be heavily influenced by the research question. If the question is focused on how organization interventions (e.g., training, selection) influence observed performance, corrections for criterion unreliability could be bypassed. On the other hand, corrections may be important if the researcher is interested in construct-level relationships or in how interventions influence underlying performance.

Eighth, researchers and practitioners should strive for logical consistency within matrices. We have noted how simulations from the "typical approach" can be inconsistent given that some values apply to applicants and some to incumbents. We have noted in some cases that input matrices mix operational validities (criterion unreliability and predictor range restriction corrections) with true correlations (corrections for predictor unreliability as well). Recall this concern in research on goal orientation (e.g., Payne et al., 2007). This problem seems particularly acute when assembling meta-analytic matrices from multiple sources as opposed to when a group of researchers is doing all their own analyses.

Limitations

We note several limitations of our efforts. First, as with many previous studies, the studies within the meta-analyses that made up our matrices came from a variety of sources. Thus, it is possible that the values in the meta-analytic matrix were not statistically consistent with each other, which might be more likely when some meta-analytic correlations are based on small sample sizes. In a sense, this is a similar problem to the use of pairwise missing data because each correlation is potentially based on a different sample. There is little guidance in the meta-analytic literature on what boundary conditions make this issue problematic.

Second, the validity estimates in our matrices were generally corrected for direct range restriction (given the conventions at the time when they were conducted). Yet, recent research on indirect range restriction should also be considered (e.g., Aguinis, & Whitehead, 1997; Schmidt et al. 2006; 2008, see also Schmitt, 2007). We chose to use validity estimates that were corrected for direct range restriction such that all validities were corrected for the same type of range restriction. However, some of the validity estimates we report may be conservative, and we encourage more investigation into indirect range restriction in simulation research.

Third, some of our estimates of validity, standardized group differences, and intercorrelations will need continual updating as the research literature expands. For example, a new meta-analysis of biodata validity might appear that updates and expands previous work. Similar work might occur for group differences or intercorrelations.

Future Research

Staffing issues. There is a surprising lack of research on Hispanic–White differences in the selection literature. This is especially surprising given that Hispanics have overtaken Blacks in terms of absolute numbers in the workforce (Whetzel et al., 2008). Similarly, there is little work on unrestricted Black–White group differences (and constructs) for important predictors such as assessment centers, biodata, and situational judgment tests (Ployhart & Holtz, 2008). As a result, sample sizes in meta-analytic efforts can be small (and more primary studies are needed). Another interesting issue is how occupational interests and personality traits might influence range restriction both in the process of attraction and selection, and the process of individuals leaving jobs and organizations after employment has begun.

Further, much work is needed to examine constructs that are measured by various methods (e.g., biodata). For example, it is possible that biodata items targeting scholastic achievement may differ substantially from biodata items targeting social tendencies or skills. Unfortunately, there is very little information on biodata standardized ethnic group differences for job applicants. Construct-specific information would greatly help both the understanding of psychometric properties of biodata items (interested readers might consult Schmitt et al., 2009). Similar thoughts might also apply to situational judgment tests, and there are needs for more primary studies in both areas (and some integrity measures may "map" onto multiple personality constructs). Finally, personality (e.g., Conscientiousness) measured using self-reports may differ substantially from personality measured using observer ratings (e.g., recommendation forms) in operational validity (Oh, Wang, & Mount, 2011). Accordingly, more multitrait multimethod research is needed in the field of staffing and other research fields.

There is also a great need for research on the intercorrelations of predictors/independent variables (Hunter & Hunter, 1984; Schmidt & Hunter, 1998). Primary or meta-analytic work that targets both methods (work samples, situational judgment tests) and constructs (perhaps in a methods-by-construct matrix) deserves substantial attention, though there has been lamentably little such work. This concern is not limited to staffing because many cells in a variety of HR and OB matrices tend to have intercorrelations based on small numbers of studies.

Researchers may also wish to examine corrections for artifacts such as criterion unreliability. On one hand, not correcting for this artifact leads to downward biases in validities or standardized group differences, and such biases are often nontrivial (e.g., Hunter & Schmidt, 2004). On the other hand, primary studies often do not report estimates of reliability

(Geyskens et al., 2009). Empirical research might examine the proportion of studies that report relevant artifact information (including range restriction information), and Monte Carlo studies might examine how stable artifact distributions appear to be and how changes in artifact distributions influence covariance parameter estimates. Similarly, it would be useful to continue research into *u* distributions to increase sample size and assess stability of such distributions (Schmitt, 2007).

Decision aid research. We conducted some of our research using DeCorte et al.'s (2006) programs, which are quite useful for helping model the effects of HR practices. The design of more user-friendly decision support systems may be an interesting issue. Research regarding user reactions to such systems might examine models such as the technology acceptance model (e.g., Lee, Kozar, & Larsen, 2003) and the management information systems literature as a whole. Such work is certainly nontrivial as past efforts of personnel researchers in the area of utility analysis have not been met with high levels of managerial approval (e.g., Carson, Becker, & Henderson, 1998), though this might improve using state of the art Internet technology as per Aguinis and Smith (2007).

In summary, we have suggested that the increasingly common approach of using meta-analytic matrices as a basis for simulation research (or SEM research) could benefit from careful consideration of the nature of the matrices underlying these simulations. We suggest consideration of the population of interest and note the importance of correcting covariances for range restriction and criterion unreliability in many situations. Further, we urge researchers to consider the operationalization of both independent/predictor and dependent/criterion constructs in their matrices (in a variety of areas of organizational research). Careful consideration of these factors should enable simulation research to accurately inform researchers, aid decision makers, and move forward to bring even more of the promise of this methodology.

REFERENCES

Aguinis H, Culpepper SA, Pierce CA. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology*, *95*, 648–680.

Aguinis H, Dalton DR, Bosco FA, Pierce CA, Dalton CM. (2010). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management*, doi:10.1177/0149206310377113.

Aguinis H, Pierce CA, Bosco FA, Dalton DR, Dalton CM. (2011). Debunking myths and urban legends about meta-analysis. *Organizational Research Methods, 14*, 306–331.

Aguinis H, Smith MA. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact. PERSONNEL PSYCHOLOGY, *60*, 165–199.

Aguinis H, Werner S, Abbot JL, Angert C, Park JH, Kohlhausen D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, *13*, 515–539.

Aguinis H, Whitehead R. (1997). Sampling variance in the correlations coefficient under indirect range restriction: Implication for validity generalization. *Journal of Applied Psychology*, *82*, 528–538.

Arthur W, Jr., Villado AJ. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection and research in practice. *Journal of Applied Psychology*, *93*, 435–442.

Arvey RD, Faley RH. (1988). *Fairness in selecting employees* (2nd Ed). Reading, MA: Addison Wesley.

Barrick MR, Mount MK. (1991). The Big Five personality dimensions and job performance: A meta-analysis. PERSONNEL PSYCHOLOGY, *44*, 1–26.

Berry CM, Sackett PR. (2009). Faking in personnel selection: Tradeoffs in performance versus fairness resulting from two cut-score strategies. PERSONNEL PSYCHOLOGY, *62*, 835–863.

Berry CM, Sackett PR, Landers RN. (2007). Revisiting interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. PERSONNEL PSYCHOLOGY, 60, 837–874.

Bobko P, Roth PL, Potosky D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. PERSONNEL PSYCHOLOGY, *62*, 561–589.

Bobko P, Stone-Romero E. (1998). Meta-analysis is another useful research tool, but it is not a panacea. In Ferris GR (Ed.), *Research in personnel and human resources management*, (Vol. 16, pp. 359–397). Greenwich, CT: JAI Press.

Boudreau JW, Berger CJ. (1985). Decision-theoretic utility analysis applied to employee separations and acquisitions. *Journal of Applied Psychology*, *70*, 581–612.

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

Carson KP, Becker JS, Henderson JA. (1998). Is utility really futile? A failure to replicate and an extension. *Journal of Applied Psychology*, *83*, 84–96.

Colquitt JA, LePine JA, Noe RA. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, *85*, 678–707.

Converse PD, Peterson MH, Griffith RL. (2009). Faking on personality measures: Implications for selection involving multiple predictors. *International Journal of Selection and Assessment*, *17*, 47–60.

Cortina JM, Landis RS. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In Lance CE, Vandenberg RJ (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 287–308). New York, NY: Routledge/Taylor & Francis Group.

Dean M. (1999). *On biodata construct validity, criterion validity and adverse impact.* Unpublished doctoral dissertation, Louisiana State University.

DeCorte W. (1999). Weighing job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology*, *84*, 695–702.

De Corte W, Lievens F. (2003). A practical procedure to estimate the quality and the adverse impact of single-stage selection decisions. *International Journal of Selection and Assessment*, *11*, 89–97.

DeCorte W, Lievens F, Sackett PR. (2006). Predicting adverse impact and mean criterion performance in multi-stage selection. *Journal of Applied Psychology*, *91*, 523–537.

DeCorte W, Lievens F, Sackett PR. (2007). Combining predictors to achieve optimal tradeoffs between selection quality and adverse impact. *Journal of Applied Psychology*, *92*, 1380–1393.

Doverspike D, Winter JL, Healy MC, Barrett GV. (1996). Simulations as a method of illustrating the impact of differential weights on personnel selection outcomes. *Human Performance*, *9*, 259–273.

Dunleavy EM, Stuebing KK, Campion JE, Glenn DM. (2008). Using the 4/5ths rule as an outcome in regression analysis: A demonstrative simulation. *Journal of Business and Psychology*, *23*, 103–114.

Finch DM, Edwards BD, Wallace CJ. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology*, *94*, 318–340.

Foldes H, Duehr E, Ones D. (2008). Group differences in personality: Meta-analyses comparing five U.S. racial groups. PERSONNEL PSYCHOLOGY, *61*, 579–616.

Geyskens I, Krishnan R, Steenkamp JEM, Cunha PV. (2009). A review and evaluation of meta-analysis practices in management research. *Journal of Management*, *35*, 393–419.

Hartigan JA, Wigdor AK. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, D.C.: National Academies Press.

Harrison JR, Carroll GR, Carley KM. (2007). Simulation modeling in organizational and management research. *Academy of Management Review*, *32*, 1229–1245.

Hattrup K, Rock J. (2002). A comparison of predictor-based and criterion-based methods of weighting predictors to reduce adverse impact. *Applied H.R.M. Research*, *7*, 22–38.

Hattrup K, Rock J, Scalia C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology*, *82*, 656–664.

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

Huffcutt AI, Roth PL, McDaniel MA. (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 JE. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, *29*, 340–362.

Hunter JE, Hunter RF. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98.

Hunter JE, Schmidt FL. (2004). *Methods of meta-analysis: Correcting for error and bias in research findings* (2nd ed). Newbury Park, CA: Sage.

Hurtz GM, Donovan JJ. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, *85*, 869–879.

Johnson JW. (2001). The relative importance of task and contextual performance dimensions to supervisor judgments of overall performance. *Journal of Applied Psychology*, *86*, 984–996.

Judge TA, Jackson CL, Shaw JC, Scott BA, Rich BL. (2007). Self-efficacy and work-related performance: The integral role of individual differences. *Journal of Applied Psychology*, *92*, 107–127.

Komar S, Brown DJ, Komar JA, Robie C. (2008). The impact of faking on the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93*, 140–154.

Kriska SD. (2001, April). *The validity-adverse impact trade-off: Real data and mathematical model estimates*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Law KS, Schmidt FL, Hunter JE. (1994). A test of two refinements in procedures for meta-analysis. *Journal of Applied Psychology*, *79*, 978–986.

Lee Y, Kozar KA, Larsen KRT. (2003). The technology acceptance model: Past, present, and future. *Communications of AIS*, *12*, 752–780.

McDaniel MA, Hartman NS, Whetzel DL, Grubb L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. PERSONNEL PSYCHOLOGY, *60*, 63–91.

McDaniel MA, Morgeson FP, Finnegan EB, Campion MA, Braverman EP. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

McDaniel MA, Rothstein HR, Whetzel DL. (2006). Publication bias: A case study of four test vendors. PERSONNEL PSYCHOLOGY, *59*, 927–953.

McDaniel MA, Whetzel DL, Schmidt FL, Maurer S. (1994). The validity of the employment interview: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*, 599–616.

McFarland LA, Ryan AM. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*, 812–821.

McKay PF. (2010). Perspectives on adverse impact in work performance: What we know and what we could learn more about. In Outtz J. (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 249–270). New York, NY: Routledge.

McKay PF, McDaniel MA. (2006). A reexamination of Black-White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology*, *91*, 538–554.

Murphy KR, Shiarella AH. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. PERSONNEL PSYCHOLOGY, *50*, 823–850.

Oh I-S, Wang G, Mount MK. (2011). Validity of observer ratings of the five-factor model of personality: A meta-analysis. *Journal of Applied Psychology, 96*, 762–773.

Ones DS.(1993). The construct validity of integrity? tests. Unpublished doctoral dissertation, University of Iowa.

Ones DS, Viswesvaran C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large-scale applicant data sets. *Journal of Applied Psychology*, *83*, 35–42.

Ones DS, Viswesvaran C, Schmidt FL. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*, 679–703.

Payne SC, Youngcourt SS, Beaubien JM. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology*, 92, 128–150.

Ployhart RE, Holtz BC. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. PERSONNEL PSYCHOLOGY, *61*, 153–172.

Potosky DP, Bobko P, Roth PL. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Selection and Assessment*, *13*, 304–315.

Reilly RR, Warech MA. (1993). The validity and fairness of alternatives to cognitive ability tests. In Wing L, Gifford B (Eds.), *Policy issues in employment testing*. Boston, MA: Kluwer.

Rich JR, Boudreau JW. (1987). The effects of variability and risk in selection utility analysis: An empirical comparison. PERSONNEL PSYCHOLOGY, *40*, 55–84.

Roth PL, BeVier CA, Bobko P, Switzer FS III, Tyler P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. PERSONNEL PSYCHOLOGY, *54*, 297–330.

Roth PL, Bobko P, McFarland LA. (2005). A meta-analysis of work sample test validity: Updating some classic literature. PERSONNEL PSYCHOLOGY, *58*, 1009–1037.

Roth P, Bobko P, McFarland L, Buster M. (2008). Work sample tests in personnel selection: A meta-analysis of Black-White differences in overall and exercise scores. PERSONNEL PSYCHOLOGY*, 61*, 637–662.

Roth PL, Bobko P, Switzer FS III, Dean MA. (2001). Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. PERSONNEL PSYCHOLOGY, *54*, 591–617.

Roth PL, Van Iddekinge CH, Huffcutt AI, Eidson CE, Jr., Bobko P. (2002). Correcting for range restriction in structured interview ethnic group differences: The values may be larger than we thought. *Journal of Applied Psychology*, *87*, 369–376.

Rothstein H, Schmidt FL, Erwin F, Owens W, Sparks CP. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology*, *75*, 174–184.

Sackett PR, Ellingson J. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. PERSONNEL PSYCHOLOGY, *50*, 707–721.

Sackett PR, Lievens F. (2008). Personnel selection. *Annual Review of Psychology, 59*, 1–32

Sackett PR, Lievens F, Berry CM, Landers RN. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology*, *92*, 538–544.

Sackett PR, Roth L. (1991). A Monte Carlo examination of banding and rank order methods of test score use in personnel selection. *Human Performance*, *4*, 279–295.

Sackett PR, Roth L. (1996). Multistage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. PERSONNEL PSYCHOLOGY, *49*, 1–18.

Sackett PR, Schmitt N, Ellingson JE, Kabin ME. (2001). High stakes testing in employment, credentialing, and higher education: Prospects in a post affirmative action world. *American Psychologist*, *56*, 302–318.

Sackett PR, Shen W. (2010). Subgroup differences on cognitive tests in contexts other than personnel selection. In Outtz J (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 323–346). New York, NY: Routledge.

Sackett PR, Wanek JE. (1996). New developments in the use of measures of honest, integrity, conscientiousness, dependability, trustworthiness, and reliability. PERSONNEL PSYCHOLOGY, *49*, 797–829.

Salgado JF, Anderson N, Moscoso S, Bertuna C, Fruyt F, Rolland JP. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology*, *88*, 1068–1081.

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

Schmidt FL, Hunter JE. (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 FL, Le H, Oh I-S, Shaffer JA. (2007). General mental ability, job performance, and red herrings – Responses to Osterman, Hauser, and Schmitt. *Academy of Management Perspectives*, 21, 64–76.

Schmidt FL, Mack MJ, Hunter JE. (1984). Selection utility in the occupation of US Park Ranger for three modes of test use. *Journal of Applied Psychology*, *69*, 490–497.

Schmidt, FL, Oh I-S, Hayes TL. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–128.

Schmidt FL, 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 FL, Shaffer JA, 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.

Schmitt N. (2007). The value of personnel selection: Reflections on some remarkable claims. *The Academy of Management Perspectives*, 21, 19–23.

Schmitt N, Clause CS, Pulakos ED. (1996). Subgroup differences associated with different measures of some common job relevant constructs. In Cooper CL, Robertson IT (Eds.), *International review of industrial and organizational psychology* (pp. 115–139). New York, NY: Wiley.

Schmitt N, Keeney J, Oswald FL, Pleskac TJ, Billington AQ, Sinha R, Zorzie M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, *94*, 1479–1497.

Schmitt N, Oswald FL. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology*, *91*, 613–621.

Schmitt N, Quinn A. (2010). Reductions in measured subgroup mean differences: What is possible? In Outtz J (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 425–452). New York, NY: Routledge.

Schmitt N, Rogers W, Chan D, Sheppard L, Jennings D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology*, *82*, 719–730.

Schmitt N, Sackett PR, Ellingson JE. (2002). No easy solution to subgroup differences in high stakes testing. *American Psychologist, 57*, 305–306.

Scullen SE, Bergey PK, Aiman-Smith L. (2005). Forced choice distribution rating systems and the improvement of workforce potential: A baseline simulation. PERSONNEL PSYCHOLOGY, *58*, 1–32.

Stillwell WG, Seaver DA, Edwards W. (1981). A comparison of weight approximation techniques in multiattribute utility decision making. *Organizational Behavior & Human Performance*, *28*, 62–77.

Tam AP, Murphy KR, Lyall JT. (2004). Can changes in differential dropout rates reduce adverse impact? A computer simulation study of a multi-wave selection system. PERSONNEL PSYCHOLOGY, *57*, 905–934.

Van Iddekinge CH, Raymark PH, Eidson CE, Jr., Attenweiler B. (2004). What do structured interviews really measure? The construct validity of behavior description interviews. *Human Performance, 17*, 71–93

Viswesvaran C, Ones DS. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. PERSONNEL PSYCHOLOGY, *48*, 865–885.

Whetzel D, McDaniel M, Nguyen N. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*, 291–308.

Wanous JP, Sullivan SE, Malinak J. (1989). The role of judgment call in meta-analysis. *Journal of Applied Psychology*, *74*, 259–264.