

# Revisiting Meta-Analytic Estimates of Validity in Personnel Selection: Addressing Systematic Overcorrection for Restriction of Range

Paul R. Sackett<sup>1</sup>, Charlene Zhang<sup>1</sup>, Christopher M. Berry<sup>2</sup>, and Filip Lievens<sup>3</sup>

<sup>1</sup>Department of Psychology, University of Minnesota

<sup>2</sup>Department of Management and Entrepreneurship, Kelley School of Business, Indiana University

<sup>3</sup>Lee Kong Chian School of Business, Singapore Management University

This paper systematically revisits prior meta-analytic conclusions about the criterion-related validity of personnel selection procedures, and particularly the effect of range restriction corrections on those validity estimates. Corrections for range restriction in meta-analyses of predictor–criterion relationships in personnel selection contexts typically involve the use of an artifact distribution. After outlining and critiquing five approaches that have commonly been used to create and apply range restriction artifact distributions, we conclude that each has significant issues that often result in substantial overcorrection and that therefore the validity of many selection procedures for predicting job performance has been substantially overestimated. Revisiting prior meta-analytic conclusions produces revised validity estimates. Key findings are that most of the same selection procedures that ranked high in prior summaries remain high in rank, but with mean validity estimates reduced by .10–.20 points. Structured interviews emerged as the top-ranked selection procedure. We also pair validity estimates with information about mean Black–White subgroup differences per selection procedure, providing information about validity–diversity tradeoffs. We conclude that our selection procedures remain useful, but selection predictor–criterion relationships are considerably lower than previously thought.

*Keywords:* selection procedures, validity, meta-analysis, range restriction, artifact distribution

It is well understood that the goal of validation of selection procedures used as predictors of job performance is the estimation of operational validity in an applicant sample, using a criterion measure free of measurement error. Observed validity estimates are underestimates in the presence of range restriction (a narrower range of predictor scores in the validation sample than in the applicant pool) and measurement error in the criterion. Corrections for range restriction (Pearson, 1903) and measurement error (Spearman, 1904) were developed in the early days of psychometrics, and these are used to obtain estimates of operational validity.

In recent decades, meta-analyses of a wide range of selection procedures have been conducted; Schmidt and Hunter (1998) offer a summary. An ideal approach, rarely implemented in the personnel selection literature, would be to obtain an estimate of operational validity for each study and cumulate these findings. This would require that the amount of criterion measurement error and range restriction in each study were known. As this is rarely the case (see Rothstein et al., 1990, for an exception), the routinely used alternative is to make use of an artifact distribution. For example, estimates of the reliability of the criterion measure may be available for a subset of studies. The mean and variance of this artifact distribution

of reliability estimates are obtained, and the mean and variance of the full set of observed validity estimates are corrected using this artifact distribution. The mathematics and logic of this are sound if one assumes that the subset of studies used to build the artifact distribution are effectively a random draw from the full set of studies included in the meta-analysis. An additional assumption, in the case of multiple artifact distributions (e.g., for both range restriction and measurement error in the criterion) is independence of the artifacts (James et al., 1992).

Our focus in this paper is on the approaches used to build artifact distributions for range restriction on the predictor. We make the case that the approaches typically used to build range restriction artifact distributions have significant flaws that have generally led meta-analysts to substantially overcorrect for range restriction. We acknowledge that we ourselves have errantly drawn upon these approaches in our research, cited them in past studies, and taught them to our students. We now posit that this has caused our field to greatly overestimate the criterion-related validity of personnel selection predictors. We will show that a recalibration of criterion-related validity with appropriate corrections for range restriction (or lack of corrections, in some cases) suggests that our field's selection procedures are still valid, just not as valid as we thought they were.

## Approaches for Building Artifact Distributions for Range Restriction

### Background

Every selection predictor meta-analysis faces the problem of commonly only having available one of the two pieces of information needed to quantify the amount of predictor range restriction affecting any given study. What is typically available is the standard

Paul R. Sackett  <https://orcid.org/0000-0001-7633-4160>

Charlene Zhang  <https://orcid.org/0000-0001-6975-5653>

Filip Lievens  <https://orcid.org/0000-0002-9487-5187>

Correspondence concerning this article should be addressed to Paul R. Sackett, Department of Psychology, University of Minnesota, Elliott Hall, 75 East River Road, Minneapolis, MN 55455, United States. Email: psackett@umn.edu

deviation of scores on the selection procedure of interest  $x$  among those selected (i.e., among those for whom a validity coefficient is computed), which is referred to as the “restricted  $SD_x$ .” But what is also needed for calculating a correction factor is  $SD_x$  in the full applicant pool for the job in question, which is referred to as the “unrestricted  $SD_x$ .” This is often unavailable for a variety of reasons. One is that the original researcher had the information, but did not include it in the validation report. Another is that the study used a concurrent design in which the test was administered to job incumbents, and thus information about an unrestricted  $SD_x$  in an applicant pool was never available. Faced with this lack of information, meta-analysts adopt differing approaches for building artifact distributions, which we review in the coming sections.

Before reviewing approaches to generating artifact distributions, there is a critical observation we need to make and elaborate, namely, that meta-analyses of selection procedure validity to date have assumed that the artifact distribution applies to all studies used in the meta-analysis. In the context of analyzing intercorrelations among predictors (as opposed to selection method validation, which focuses on predictor–criterion relationships), Sackett et al. (2007) and Berry et al. (2007) noted that the application of the same correction factor (or artifact distribution correction factor) to all studies can be seriously misguided. Berry et al. (2007) focused on the relationship between cognitive ability and employment interviews. Some studies administered the two measures to all applicants; in this setting there was no range restriction whatsoever. Others screened initially on ability, and only interviewed a subset; in this case there was direct restriction on ability and indirect restriction on the interview. Others administered both predictors to current employees; in this case there was indirect restriction if the selection method used to select current employees was correlated with the interview, with ability, or with both. Berry et al. detailed additional scenarios beyond these three, but for our purposes the point is simply that applying a uniform correction across all studies makes no sense. Berry et al. separated the available research studies into subsets based on information about range restriction mechanisms in each subset, and applied appropriate corrections within each subset. Conceptually, one could apply appropriate corrections to subsets, and combine the subsets for an estimate of the parameter of interest (e.g., mean operational validity).

The implications of this notion of correcting different subsets of studies using different correction factors has had little effect on the examination of the validity of selection procedures. A meta-analysis of employment interview validity by Huffcutt et al. (2014) building explicitly on Berry et al., and applying differing correction factors to predictive and concurrent studies, is the only example we have been able to locate after reviewing meta-analyses of a wide range of selection procedures.

With this as the backdrop, we examine how meta-analysts have obtained the mean and variance of range restriction for artifact distributions. We discuss five approaches that have been used in the selection literature.

### Approach 1: Derive a Distribution of Unrestricted $SD_x$ From a Subset of Studies

This is the “textbook solution” (Hunter & Schmidt, 2004). A subset of studies in the meta-analysis that include both unrestricted and restricted  $SD_x$  are identified, and the mean and variance of the

ratio of unrestricted  $SD_x$  to restricted  $SD_x$ , commonly labeled “ $U$ ” are computed. Conceptually, the mean observed validity  $r_{xy}$  is corrected based on this  $U$  ratio correction factor, and the sampling error variance is adjusted (increased) to reflect the variance in  $U$  ratios; see Hunter and Schmidt (2004) for computational details. If a correction for unreliability in the criterion is to be made, that correction is first made to mean observed validity; that reliability-corrected mean validity is then corrected for range restriction.

This approach makes conceptual and computational sense if it is reasonable to assume that the subset of studies for which  $U$  ratios are available is a random draw from the full set of studies used in the meta-analysis. One must also assume that the set of studies for which  $U$  ratios are available is large enough that we have confidence that the mean  $U$  ratio in the subset reasonably estimates the unknown “true” mean  $U$  ratio that would be obtained if  $U$  were available for all studies.

However, we believe that the subset of studies providing  $U$  ratios for artifact distributions is typically not representative of all studies in selection predictor meta-analyses. The typical meta-analysis contains a mixture of applicant-based predictive validity studies and incumbent-based concurrent validity studies. In applicant-based predictive validity studies, range restriction could be direct (i.e., applicants were selected based solely on their scores on  $x$ ) or indirect (e.g.,  $x$  was not used in selecting applicants or  $x$  was just one of a number of predictors used to select applicants). In incumbent-based concurrent validity studies, range restriction can only be indirect because incumbents had already been hired prior to the study using some method  $z$  other than the focal predictor  $x$  (Hunter et al., 2006). Further, in typical settings, only with applicant-based predictive validity studies is it possible to directly obtain a  $U$  ratio: an unrestricted  $SD_x$  is by definition obtained from an applicant pool. (We will later address the scenario in which other data, such as normative data from a test publisher, is used as a proxy for an unrestricted applicant pool  $SD_x$ .) Thus, as the typical meta-analysis contains a mixture of predictive and concurrent studies, and the studies containing the needed information to compute a  $U$  ratio come solely from predictive studies, it follows that this violates the assumption that the studies providing the artifact distribution constitute effectively a random draw of studies. The next sections of the paper show that this assumption violation has serious consequences for accurate estimation of operational validity.

### Propositions

We offer here three propositions. The first is that if a selection predictor is administered to current employees for a validation study, we can be virtually certain that they were not selected at point of hire on that predictor. This is the essence of a concurrent validation strategy: the predictor of interest is administered to current employees, and thus we know only the predictor  $SD$  among current employees. Thus, in concurrent validity studies, any range restriction will only be indirect: employees were selected on some other basis than the predictor of interest, and the predictor  $SD$  in the current employee sample will be restricted only to the degree that the predictor of interest is correlated with the other predictor(s) actually used for selection. We document below that when range restriction is indirect, there must be a very strong correlation between the selection method by which employees were actually selected ( $z$ ) and the predictor of interest ( $x$ ) for there to be substantial restriction of range on  $x$ .

The second proposition is that we can be reasonably confident that with rare exceptions  $z$  is not highly correlated with  $x$ . We see two scenarios. In the first, which we view as common, incumbents were selected using a different type of method than the one under investigation in the validation study. For example, incumbents may have been selected via an interview, and an ability test is then administered to incumbents for validation purposes. We examined various sources to estimate the range of possible intercorrelations between selection predictors. Roth et al. (2011) compiled a meta-analytic matrix of unrestricted intercorrelations among commonly used predictors, which include intercorrelations among cognitive ability tests, structured interviews, conscientiousness measures, biodata, and integrity tests (e.g., Roth et al., 2011). The highest range restriction-corrected correlation among these measures is .37. Schmidt and Hunter's (1998) review of the validity of a variety of selection measures reported the correlation between each and cognitive ability. Correlations between ability and various measures include work samples (.38), structured interviews (.30), unstructured interviews (.38), job knowledge (.48), biodata (.50), and assessment centers (.50). Lower correlations between ability and assessment centers can be found in Meriac et al. (2008), where a composite of assessment dimensions correlated .45 with ability, and in Hoffman et al. (2015) where a composite of assessment exercises correlated .27 with ability. Meriac et al. also reported correlations between ability and personality, with average correlations in the .10 range. A clear message is that measures with a higher cognitive loading will correlate higher with one another (e.g., the cognitive ability—job knowledge correlation of .48); measures in the cognitive domain show small correlations with those in the noncognitive domain. Importantly, correlations in the .50 range are the highest we have been able to locate.

In the second scenario, which we view as possible but relatively uncommon, the method under investigation in the validity study is a variant of the method previously used for selection. For example, researchers may investigate a computer-adaptive cognitive ability test with incumbents having been initially selected on a paper-and-pencil ability test. Or a new structured interview may be under investigation, with incumbents selected on a prior interview. Additional examples include investigating a video situational judgment test (SJT) with incumbents having been initially selected on a paper-and-pencil SJT or replacing generic personality scales with contextualized ones. In these examples, it is likely that the focal predictor is indeed highly correlated with the predictor used in prior selection. While these examples are possible, we do not view them as prototypic: it does not seem plausible that the meta-analytic database for a given predictor is populated predominantly with studies where candidates were initially selected on a variant of the method under current investigation. We suggest that such studies may become more common in current and future research as technology offers new approaches to measurement (e.g., comparing a new game-based assessment of cognitive ability with a currently used traditional cognitive ability test). However, the meta-analytic databases examined in this paper predate this technology revolution.

To be clear: we are not arguing that a high correlation between the new predictor of interest and the predictor(s) used for initial selection is not possible. Rather, the argument is that such studies likely make up a relatively small portion of a meta-analytic data base, and that far more concurrent validity studies fall into our first category, namely investigating a new type of predictor rather than validating a close variant of a predictor type already in use.

The third proposition is that indirect range restriction restricts variance on the selection predictor of interest to only a very modest degree under virtually all realistic circumstances. We anticipate that this third proposition might be seen as provocative and startling, given that there are studies suggesting that correcting for indirect range restriction substantially increases criterion-related validity estimates (e.g., Hunter et al., 2006). However, it follows quite directly from the mechanics of range restriction. Building on the work of Sackett, Zhang, et al. (2021), in Panel 1 of Table 1 we show the effect on  $SD_x$  of direct range restriction on  $x$  and then in Panel 2 of Table 1 we show the effect on  $SD_x$  of indirect range restriction due to selection on a third variable  $z$  as two things vary (Panel 2 focuses on the scenario in which the reliability of  $x$  is 1.0; we later turn to scenarios in which the reliability is less than 1.0). The first is the unrestricted correlation between  $z$  and  $x$ , with no measurement error in either variable (the paragraph above argues this correlation between  $z$  and  $x$  is commonly modest). The second is the selection ratio on  $z$ ; the smaller the selection ratio, the greater the range restriction. The tabled values are easily obtained by rearranging the terms in the direct range restriction formula to solve for the restricted correlation between  $z$  and  $x$ , and then inserting this value in the definitional formula for the relationship between restricted and unrestricted validity to solve for the restricted  $SD_x$ . Unrestricted  $SD_x$  is set at 1.0. Using the direct correction formula is appropriate here, as  $z$  is the variable used for selection and restriction on  $z$  is direct by definition. Table 1 also includes information about the effects of direct range restriction on restricted  $SD_x$ . Footnote 1 gives technical details of the computations underlying Panel 2 of Table 1.<sup>1</sup> While Table 1 was generated using the equations presented in Footnote 1, at the suggestion of a reviewer we did a large-scale simulation with one million cases, and were able to reproduce the table, thus assuaging any concerns as to the appropriateness of our equations. The R-code is available upon request.

Panel 1 of Table 1 shows that the  $SD_x$  is restricted quite substantially under *direct* range restriction. For example, a selection ratio of .50 produces a restricted  $SD_x$  of .60 (relative to 1.0 as the unrestricted value). Dividing unrestricted  $SD_x$  by restricted  $SD_x$  gives the correction factor ( $U$  ratio) of 1.67: observed validity would be increased 67% when one corrects for this direct range restriction. Demonstrations of the effects of direct restriction, such as the above, are common in didactic treatments of range restriction, and so we are used to the notion that range restriction commonly has a large effect.

<sup>1</sup> Conceptually, we use the standard direct range restriction formula to examine the effects of selection on the unknown variable  $z$  on the correlation between  $z$  and the test  $x$ ; we then compare unrestricted and restricted  $r_{zx}$  to identify restricted  $SD_x$ . Thus, we use a two-step process. The first step is to use the direct range restriction formula to solve for restricted  $r_{zx}$  (e.g., correlations between the predictor of interest and the actual selection variable  $z$ ). Thus, inserting an unrestricted  $r_{zx}$  and the restricted  $SD_z$  corresponding to the selection ratio of interest, one solves for the restricted  $r_{zx}$ -restricted  $r_{zx} = (\text{unrestricted } r_{zx} \times \text{restricted } SD_z) / \text{SQRT}(1 + (\text{restricted } SD_z^2 - 1) \times r_{zx}^2)$ . The second step is to solve for restricted  $SD_x$  using the  $r_{zx}$  and restricted  $SD_z$  values input into the above equation and the  $r_{zx}$  value solved for in the above equation. The equation used is simply the definitional relationship between restricted and unrestricted validity (restricted  $r_{zx} = \text{unrestricted } r_{zx} (\text{restricted } SD_z / \text{restricted } SD_x)$ ), algebraically rearranged to solve for the restricted  $SD_x$ : Restricted  $SD_x = (\text{unrestricted } r_{zx} / \text{restricted } r_{zx}) \times \text{restricted } SD_z$ .

**Table 1***Restricted Test Standard Deviation as a Function of Selection Method (with Unrestricted Standard Deviation = 1.0)*

$r_{zx}$	Selection ratio						
	0.90	0.70	0.50	0.30	0.10	0.05	0.01
	Panel 1: Select on $x$						
	0.85	0.70	0.60	0.52	0.41	0.37	0.31
	Panel 2: Select on $z$ ( $r_{xx} = 1.0$ )						
0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	0.99	0.99	0.99	0.99	0.98	0.98	0.98
0.3	0.99	0.98	0.97	0.97	0.96	0.96	0.96
0.4	0.98	0.96	0.95	0.94	0.93	0.93	0.92
0.5	0.96	0.93	0.92	0.90	0.89	0.89	0.88
0.6	0.95	0.90	0.88	0.86	0.83	0.83	0.82
0.7	0.93	0.87	0.83	0.80	0.76	0.76	0.75
0.8	0.91	0.82	0.77	0.73	0.68	0.67	0.65
0.9	0.88	0.77	0.69	0.63	0.56	0.55	0.52
	Panel 3: Select on $z$ ( $r_{xx} = .90$ )						
0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	0.99	0.99	0.99	0.99	0.98	0.98	0.98
0.3	0.99	0.98	0.97	0.97	0.97	0.96	0.96
0.4	0.98	0.96	0.95	0.95	0.94	0.94	0.93
0.5	0.97	0.94	0.93	0.91	0.90	0.90	0.89
0.6	0.95	0.91	0.89	0.87	0.85	0.85	0.84
0.7	0.94	0.88	0.85	0.82	0.79	0.79	0.78
0.8	0.92	0.84	0.79	0.76	0.72	0.71	0.69
0.9	0.89	0.79	0.73	0.68	0.62	0.61	0.58
	Panel 4: Select on $z$ ( $r_{xx} = .80$ )						
0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	1.00	0.99	0.99	0.99	0.99	0.99	0.99
0.3	0.99	0.98	0.98	0.97	0.97	0.97	0.97
0.4	0.98	0.97	0.96	0.95	0.94	0.94	0.94
0.5	0.97	0.95	0.93	0.92	0.91	0.91	0.91
0.6	0.96	0.92	0.90	0.89	0.87	0.87	0.86
0.7	0.94	0.89	0.87	0.84	0.82	0.81	0.80
0.8	0.93	0.86	0.82	0.79	0.75	0.75	0.73
0.9	0.91	0.82	0.77	0.72	0.67	0.66	0.64
	Panel 5: Select on $z$ ( $r_{xx} = .70$ )						
0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	1.00	0.99	0.99	0.99	0.99	0.99	0.99
0.3	0.99	0.98	0.98	0.98	0.97	0.97	0.97
0.4	0.98	0.97	0.96	0.96	0.95	0.95	0.95
0.5	0.98	0.95	0.94	0.93	0.92	0.92	0.92
0.6	0.96	0.93	0.92	0.90	0.89	0.88	0.88
0.7	0.95	0.91	0.88	0.86	0.84	0.84	0.83
0.8	0.94	0.88	0.84	0.82	0.79	0.78	0.77
0.9	0.92	0.84	0.80	0.76	0.72	0.71	0.70

But the table shows that the *indirect* effects on  $SD_x$  of selecting on a third variable  $z$  are quite small. With a correlation between  $z$  and  $x$  of .50 or smaller and anything less than a very extreme selection ratio (i.e., any ratio larger than .1), restricted  $SD_x$  is .90 or larger. In other words, under indirect range restriction, the correction factor will generally be less than 10%, commonly much less than 10%. This is a major “takeaway” from this table.

We now amplify the argument and show that in concurrent studies (or in predictive studies in which the predictor under investigation is not used for selection) the already-modest restricted  $SD_x$  values in Panel 2 of Table 1 are overestimates of restricted  $SD_x$  under operational conditions where the reliability of  $x$  ( $r_{xx}$ ) is less than 1.0: the effects of indirect restriction on  $SD_x$  are even smaller than Panel 2 of Table 1 shows. The basis for this argument is an important insight offered by Hunter et al. (2006). They noted

that under indirect restriction where the focal predictor  $x$  is not part of the actual selection variable  $z$ , selection on  $z$  has a direct effect on the true score of  $x$ ; the effects of restriction on  $z$  on observed  $SD_x$  will be reduced as a function of the reliability of  $x$ . The values we show in Panel 2 of Table 1 are the effects of selection on  $z$  on the true  $SD_x$  (i.e.,  $x$  measured without measurement error). Hunter et al. offered an equation showing the relationship between true  $SD_x$  and observed  $SD_x$  when  $x$  is measured with less than perfect reliability.<sup>2</sup> In Panels 3, 4, and 5 of Table 1, we apply this formula to the values in Panel 2 for predictor reliability values of .90, .80,

<sup>2</sup> Step 5 in Table 2 of Hunter et al. (2006) gives the formula to solve for true  $SD_x$  if observed  $SD_x$  and the reliability at the applicant level are known. Re-arranging terms to solve for observed  $SD_x$  if true  $SD_x$  is known gives: observed  $SD_x = \text{SQRT}((\text{true } SD_x^2 \times r_{xx}) + (1 - r_{xx}))$ .

and .70, respectively. As one illustration, consider the  $SD_x$  value in Panel 1 of Table 1 for the situation in which  $r_{zx}$  is .5 and the selection ratio is .3, which results in a restricted  $SD_x$  of .90. Adding progressively more measurement error increases restricted  $SD_x$  systematically, producing values of .91, .92, and .93 when  $r_{zx}$  equals .90, .80, and .70, respectively. Taking the realistic step of adding measurement error shows that it is even harder to produce substantial reductions in observed  $SD_x$  than the initial discussion of Panel 1 outlined.

Having set the stage for the importance of differentiating between (applicant-based) predictive studies, where direct restriction is possible and where restriction can be sizeable, and (incumbent-based) concurrent studies, where restriction, if present, is indirect and small in magnitude, we return to our treatment of the way in which artifact distributions are used in meta-analyses of the validity of selection procedures. As we noted earlier, it is common to assemble an artifact distribution of  $U$  ratios and then correct the mean observed validity by the mean of this artifact distribution. This is conceptually inappropriate in the setting in which the meta-analytic data base comprises a combination of predictive studies (the subset from which the distribution of  $U$  ratios is extracted) and concurrent studies (for which the calculation of a  $U$  ratio is not possible because the applicant pool  $SD_x$  is unknown), which would have a very different—and much smaller—mean  $U$  ratio than would the predictive studies. The routine practice of correcting mean observed validity by the  $U$  ratio obtained in the subset of predictive studies thus produces an overcorrection. Studies with a concurrent design and little to no restriction (because it is indirect range restriction) are, in effect, corrected as if they were predictive studies with more—often much more—restriction (because  $x$  was used in selection). If a meta-analysis reports results separately for predictive and concurrent studies, the range restriction artifact distribution developed on predictive studies can be applied to predictive studies. Absent credible  $U$  ratios for concurrent studies, and in light of the above demonstration that  $U$  ratios are likely to be close to 1.0 in concurrent studies, we recommend no range restriction correction for concurrent studies. An  $N$ -weighted average of the separate corrected mean estimates for predictive and concurrent studies can then be computed to estimate mean operational validity.

### Examples

We offer several concrete examples. For our first example, we turn to meta-analyses of the validity of cognitive ability tests. There have been a number of such analyses; the best-known is Hunter (1983), which we treat later because a different approach to correction for restriction was used in that analysis. Here, we focus on two meta-analyses: Salgado et al.'s (2003) analysis of validity in the European community and Bertua et al.'s (2005) analysis in the United Kingdom. Salgado et al. located 120 validity studies, and obtained  $U$  estimates for 20 of them. The mean was  $U = 1.61$ , and correcting for criterion unreliability and range restriction increased the validity estimate from the mean observed value of .29 to a corrected value of .62. However, Salgado et al. mentioned that their sample included a mix of applicant and current employee studies; while the paper did not report the number of studies in each subset, Salgado (personal communication) reported that 78% of studies were current employee studies. Bertua et al. located 60 studies, and obtained a mean  $U$  estimate of 1.67 from an unreported number of studies. Like Salgado et al., they included both predictive and

concurrent studies, but did not report the number of studies in each subset; Salgado (personal communication) estimated that 80–85% of studies in that meta-analysis were concurrent. Correction increased the validity estimate from an observed .22 to a corrected .48. So, for both studies, a large  $U$  estimate was obtained from predictive studies and applied to both predictive and concurrent studies. We can conclude that the corrected estimate is an overestimate of cognitive ability test validity.

The second example we offer is the meta-analysis of integrity test validity by Ones et al. (1993). They located 655 validity studies. They obtained a distribution of  $U$  ratios from 79 predictive studies, finding a mean  $U$  ratio of 1.23, and applied a correction using this  $U$  ratio to all 655 studies. This increased the overall mean correlations between integrity tests and job performance from .29 to .34. However, 76% of the studies were concurrent, with an unknown, but necessarily small, amount of restriction. Thus, it is certainly the case that the corrected results reported by Ones et al. are overestimates of integrity test validity: the vast majority of studies came from concurrent studies with minimal restriction, but were treated as though they came from predictive studies with nontrivial restriction. We note that the correction factor here is relatively small compared with that found in other domains. We also note that Ones et al. subsequently focused on a subset of predictive studies using job applicants; that analysis is not subject to the concerns noted here.

The takeaway message is that applying a correction factor derived from predictive studies to concurrent studies, where restriction is generally minimal, results in an overestimate of validity. The degree of overestimation will vary as the relative proportion of predictive and concurrent studies varies. If the proportion of concurrent studies is large, then the overcorrection can be dramatic. We offer the proposition that the vast majority of validation research is concurrent. The percent of studies that are concurrent from various meta-analyses include 98% in Roth et al. (2005) work sample meta-analysis, 95% in McDaniel et al.'s (2007) SJT meta-analysis, 76% in Ones et al.'s (1993) integrity test meta-analysis, 74% in Huffcutt et al.'s (2014) interview meta-analysis, 78% in Salgado et al.'s (2003) cognitive ability meta-analysis in Europe, 80–85% in Bertua et al.'s (2005) cognitive ability meta-analysis in the United Kingdom, and 80% in the General Aptitude Test Battery (GATB) data base used in Hunter (1983) cognitive ability meta-analysis (reported in Bemis, 1968). Of course, there may be settings with very different percentages; what is needed is careful attention to the issue. So, while the  $U$  distribution may rightly apply to the predictive studies, it is common that this is a small piece of the data set, and thus overcorrection is large.

### Approach 2: Estimate a $U$ Ratio Based on Hiring Rate

If no actual  $U$  ratio is available, some researchers estimate one based on company-reported hiring rates. If a firm states “we typically hire half our applicants,” one can make strong assumptions and convert a selection ratio to a  $U$  ratio using a formula from Schmidt et al. (1976). For example, a selection ratio of .5 translates to a restricted  $SD_x$  of .6 and a  $U$  ratio of 1.67. Wiesner and Cronshaw's (1988) meta-analysis of interview validity illustrates this: they report obtaining hiring rates from a subset of studies, and converted these to an artifact distribution of  $U$  ratios.

This approach requires two strong assumptions. The first is that the sole basis for selection was the predictor of interest. That is

virtually never true, in our experience. We believe this wrongly gives an overinflated  $U$  ratio. We offer an empirical illustration. Sackett, Sharpe, et al. (2021) examined range restriction in SAT scores across 174 colleges and universities. They obtained both school-specific applicant pool data, and enrolled student data, and thus were able to compute SAT  $U$  ratios for each school; the mean  $U$  was 1.2. They also obtained a school-reported offer rate: that averages 62%. The assumption that all selectivity is “spent” on the predictor of interest (in this case, the SAT) would translate a 62% selection ratio to a  $U$  ratio of 1.54, suggesting much greater restriction than is actually present.

The second assumption is that all job offers are accepted. If say, 20% of job offers are rejected, a firm would need to extend offers to 70% of applicants in order to obtain a hiring rate of 50% of candidates. In addition, the implication for range restriction differs if rejection is random versus systematically related to standing on the predictor of interest (e.g., top candidates are more likely to have multiple offers; Murphy, 1986).

Additional examples of this problem are found in the cognitive ability meta-analyses by Salgado et al. (2003) and Bertua et al. (2005) discussed earlier. To generate a range restriction artifact distribution they used a combination of (a) actual  $U$  values from predictive studies where available, and (b)  $U$  ratios inferred from selection ratios, as critiqued here. They did not report how often they used each of the two approaches: their distribution of  $U$  ratios is a mixture of computed and inferred  $U$  values. However, Salgado (personal communication) reports that 53% of the studies used to create the artifact distribution in the Salgado et al. study inferred  $U$  ratios from selection ratios.

In short, obtaining hiring rates from a subset of studies, converting these to  $U$  ratios, and applying the result to the full set of validity studies gathered for the meta-analysis will overestimate the amount of range restriction in the meta-analytic data set. The magnitude of overestimation is, in our view, likely to be large, as the approach compounds two problems: (a) the assumption that the predictor of interest is the sole basis for selection, and (b) the application of hiring rate-based  $U$  ratios to concurrent studies.

### Approach 3: Use Incumbent Data Pooled Across Jobs to Estimate Unrestricted $SD_x$

The most prominent example of this is Hunter’s (1983) analysis of a large set of validity studies examining the GATB. Most studies were concurrent; 80% according to Bemis (1968); some studies used a version of predictive validity in which GATB scores were obtained from applicants but were not available to the hiring organization. Thus, in both cases any restriction would be indirect. Importantly, Hunter’s estimate of corrected validity (.51) for this data set is the value taken to represent the validity of cognitive ability tests in Schmidt and Hunter (1998) often cited summary article on the validity of selection procedures. As no applicant data was available in the GATB database, Hunter pooled incumbent data across studies and offered the resulting  $SD_x$  as a proxy for the applicant pool  $SD_x$ . The National Academy of Sciences Committee on the GATB (Hartigan & Wigdor, 1989) was skeptical of this, viewing it as an estimate of  $SD_x$  for the national workforce, rather than the needed  $SD_x$  for the job-specific applicant pool. Sackett and Ostgaard (1994) obtained applicant pool  $SD_x$  values for a large number of jobs and then pooled the data across jobs as an estimate of workforce  $SD_x$ ;

they reported that the applicant pool  $SD_x$  values were on average 10% smaller than the workforce  $SD_x$  estimate. So, based on Sackett and Ostgaard’s finding, it is at least hypothetically possible that it could be reasonable to pool incumbent data across jobs to estimate the unrestricted  $SD_x$ , and then reduce that  $SD_x$  by 10%.

However, we offer an argument for skepticism about this approach, at least in terms of the  $U$  ratio estimate it produced in Hunter (1983). Hunter reported a restricted  $SD_x$  of .67 (relative to the unrestricted value of 1.0), corresponding to a  $U$  ratio of 1.5. But as we discussed above and showed in Table 1, in studies where scores on the predictor of interest were not used in the hiring decision (e.g., concurrent studies) only in extreme settings will one see a restricted  $SD_x$  lower than .90. Table 1 shows that an  $SD_x$  value as extreme as Hunter’s .67 is only possible with an  $r_{zx}$  of .90 or larger, paired with an extreme selection ratio. We do not find it plausible that virtually all firms choosing to participate in a validation study of the GATB used another highly correlated cognitive ability test in initial selection, and used it with an extreme selection ratio. We note that the GATB validation research program continued beyond the set of studies analyzed by Hunter. A later set of procedurally comparable 264 studies produced a restricted  $SD_x$  of .94, and thus a  $U$  ratio of 1.06 (Sackett, Zhang, et al., 2021). In short, we cannot see a reasonable basis for Hunter’s .67 value for the restricted  $SD_x$ .

### Approach 4: Use Test Publisher Norm Data to Estimate Unrestricted $SD_x$

For settings in which a published instrument or set of instruments is used, norm group information from a test manual has been used as an estimate of unrestricted  $SD_x$ . For example, Shaffer and Postlethwaite’s (2012) personality meta-analysis obtained norm group information for each test in their database, and reported a mean  $U$  ratio of 1.10. In using norm-based unrestricted  $SD_x$  values, care is needed to determine the most appropriate value. For some measures, only overall workplace norms are reported; for others, applicant norms for specific occupations are reported. Shaffer and Postlethwaite did not specify anything other than that published norms were used. In the personality domain, Ones and Viswesvaran (2003) reported that job-specific applicant pools average 4% smaller than broad norm data, suggesting minimal self-selection in the personality domain. Thus, the issues of what norms are appropriate is less crucial for Shaffer and Postlethwaite’s analysis than it might be in other domains. We suggest that a meta-analyst proposing to rely on norms as an estimate of  $SD_x$  could examine a subset of predictive studies in which unrestricted  $SD_x$  is available, which would permit a verification of the correspondence between applicant pool information and norms. Close correspondence would support the use of norms as the unrestricted  $SD_x$  estimate for concurrent studies; lack of correspondence would indicate that the approach is not useful in that particular setting.

### Approach 5: Use a Made-up Assumed Distribution to Estimate $U$

Dye et al.’s (1993) meta-analysis of job knowledge tests offers an example of this. They had no  $U$  ratio data, so they borrowed an assumed distribution from Pearlman et al. (1980). However, Pearlman et al. offered no basis for their distribution: their purpose was to

illustrate the mechanics of how to apply an artifact distribution method. As a second example, Huffcutt et al. (2014) wanted to correct a set of concurrent validity studies of interviews for indirect range restriction. They had no information on actual prior selection procedures, so they made the following assumption: in all studies, candidates were first screened on an ability-based predictor, and the top 50% moved forward. These candidates were then interviewed, and the top 10% were selected. In other words, they assumed an overall selection ratio of 5%. They used this assumption to estimate unrestricted  $SD_x$  and reported large effects. Thus, their correction for indirect range restriction was substantial.

We do not find the assumption of a 5% selection ratio credible. Huffcutt et al. (2014) had two sets of studies; one set was predictive, while the other one was concurrent. For the predictive subset, they assumed direct range restriction and used an empirical  $U$  ratio of 1.63; under direct restriction, Table 1 shows that this  $U$  ratio corresponds to a selection ratio of about 50%. We would find reasonable an argument that the same selection ratio obtained in predictive studies would be expected as the original selection ratio in concurrent studies, and thus the use of 50% in estimating a  $U$  ratio for the concurrent studies. We struggle to understand an argument that the selection ratio is 50% in studies for which restriction information is available, but is assumed to be 5% in studies for which information is not available. Clearly, without a strong conceptual rationale, no trust should be placed in corrections based on (made up) assumed distributions.

## Summary

We have described reasons for skepticism about the approaches used to estimate the degree of range restriction operating in meta-analytic databases. Several points are critical. First, studies using a design in which selection predictor scores were not used in the selection decision (e.g., concurrent studies) will not exhibit substantial range restriction except for in rare and extreme cases (i.e., when the predictor being examined is a close proxy for the predictor used in prior selection and thus correlates highly with it). In concurrent designs, selection was done on the basis of a predictor other than the focal predictor under examination in the meta-analysis. Given what we know about correlations among selection procedures (Roth et al., 2011), and given the mathematics of indirect range restriction, the effects of indirect restriction in concurrent studies can be expected to be small: a correction factor typically ranging from zero to roughly 10%.

Second, as our examples in the literature illustrated, in many meta-analyses the vast majority of studies are current employee studies (concurrent designs). Relatedly and importantly, it is typical to obtain an artifact distribution of  $U$  ratios from the small subset of predictive validity studies available, and then apply them to the full set of studies in the meta-analytic database—both predictive and concurrent. The result is overcorrection, and the overcorrection can be severe.

Third, in correcting for range restriction we endorse the principle of conservative estimation. If one realizes that an existing correction approach is incorrect and results in an overcorrection one has three options: (a) do additional work to obtain and apply an appropriate correction; (b) use the correction anyway; or (c) apply no correction, perhaps noting that that the result is an underestimate should there be some restriction. Clearly, the first is ideal. However, it may be time

consuming, taking years to assemble a large meta-analytic database with new information. In the interim, one's choices are the second and the third: overcorrect or apply no correction. Given what we reviewed above suggesting that most validation studies are concurrent and that there is little range restriction in concurrent validation studies, we argue for no correction. We cannot see how one can knowingly report an overestimate of validity when that overestimate could be so substantial. Reporting a conservative underestimate is prudent when that underestimate is likely to be small in comparison to the size of the overestimate.

## Revisiting Meta-Analytic Estimates of the Validity of Selection Procedures: Scrutinizing the Correction Factors Used

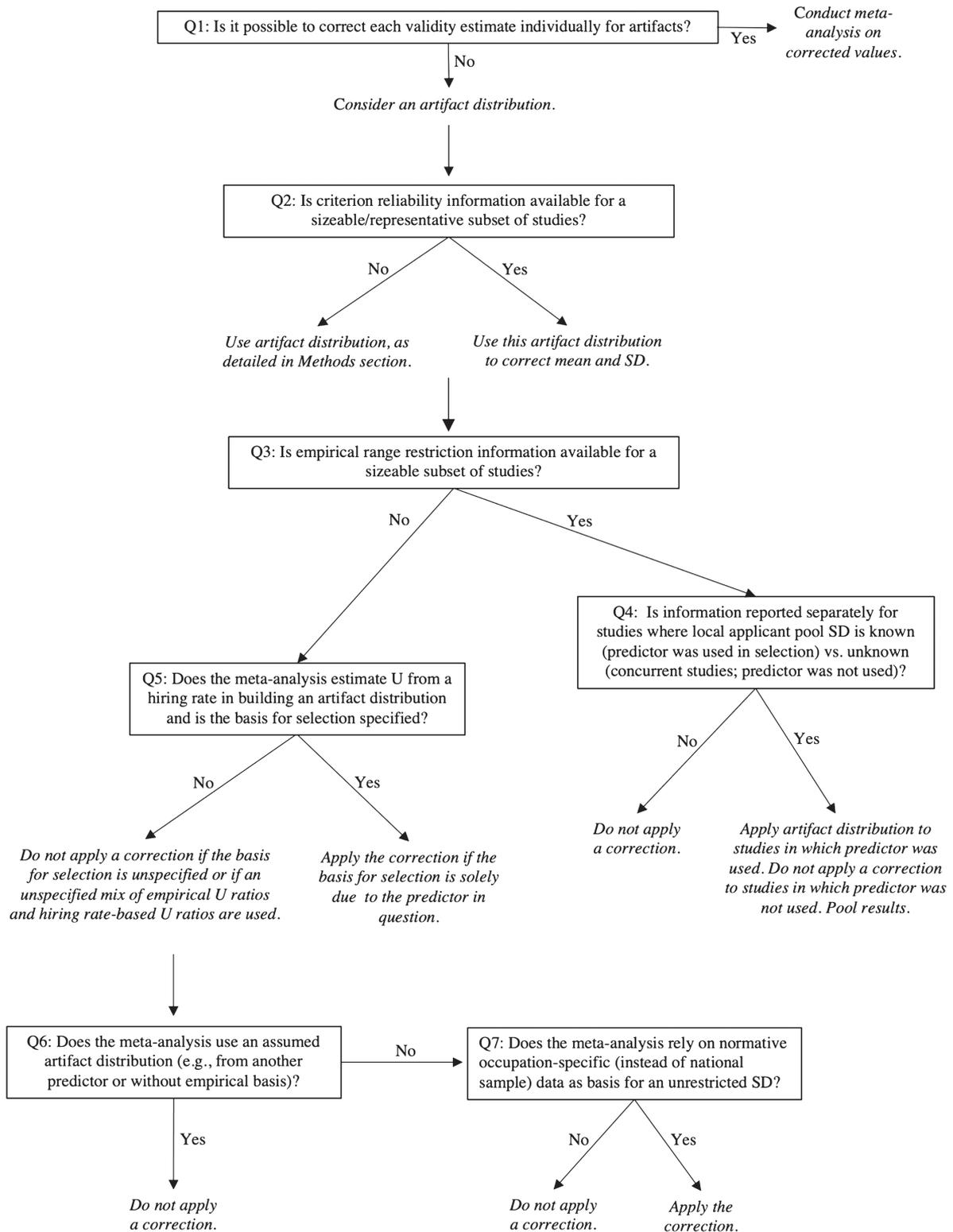
In light of our critique of range restriction correction methods, we now turn to revisiting prior summaries of meta-analytic estimates of the validity of a broad range of predictors used in selection. Our starting point is Schmidt and Hunter's (1998) summary of meta-analytic evidence for the validity of 19 types of measures used as predictors of job performance. Our primary focus is on a critical examination of the range restriction corrections used in the various meta-analyses. In some cases, we rely on the same meta-analyses as Schmidt and Hunter; in other cases we incorporate more recent meta-analyses. We also add a number of additional predictors not included in the Schmidt and Hunter summary.

## Method/Analytical Strategy

Schmidt and Hunter (1998) summarized meta-analytic estimates of operational validity for the prediction of job performance for 19 selection procedures. We offer revised estimates, at times based on the addition of meta-analyses not available at the time of the Schmidt and Hunter summary. We evaluate the appropriateness of range restriction corrections in light of the set of issues outlined in the prior section of this paper. Where appropriate we offer alternative corrections; at times we conclude that no correction should be made as the information needed for a trustworthy correction is not available. Figure 1 presents a decision tree designed to capture our process of deciding whether a correction for range restriction was or was not appropriate. We also examine a number of additional predictors not included in Schmidt and Hunter, namely, the Big Five personality traits, SJTs, and emotional intelligence. The selection of additional predictors reflects a judgment call regarding predictors that have emerged in recent years as the subject of substantial interest and inquiry.

Schmidt and Hunter presented only mean operational validity estimates corrected as fully as possible (e.g., for measurement error in the criterion and for range restriction). We present mean observed validity estimates, estimates corrected only for measurement error in the criterion, and estimates corrected for both criterion measurement error and range restriction from the meta-analyses we are summarizing. Presenting these is useful, as it makes clear the degree to which each correction affects mean validity estimates. After critiquing the range restriction correction made, we offer our best estimate of mean operational validity, which often reflects either a revised range restriction correction or no correction at all.

For a number of selection procedure, we review multiple meta-analyses, offer our best estimate of operational validity for each, and

**Figure 1***Decision Tree for When a Correction for Range Restriction Was Appropriate*

compute a sample-size-weighted average across meta-analyses for a final estimate. We acknowledge that at times there is a lack of independence of the meta-analyses (e.g., in the case of a more recent analysis also including studies included in a prior meta-analysis). If a more recent meta-analysis fully incorporates all studies from a prior meta-analysis, we use only the more recent one.

In deciding whether to incorporate new meta-analyses, we made a number of decisions. First, like Schmidt and Hunter (1998), we focused on meta-analyses of the relationship between selection procedures and overall job performance. Therefore, we excluded meta-analyses of the relationships between predictors and narrower job performance facets such as organizational citizenship behavior. Second, we focused on meta-analyses that attempted an overall evaluation of validity across the spectrum of jobs for which the predictor is intended. Thus, an analysis focused on a single job family (e.g., clerical jobs) would be excluded. Third, we excluded meta-analyses that were narrowly focused. For example, Sackett et al. (2017) meta-analyzed assessment center validity in studies in which the same participants completed a cognitive ability test in addition to the assessment center. Similarly, Van Iddekinge et al. (2018) meta-analyzed cognitive ability validity in studies in which the same participants also completed a measure of state or trait motivation. Fourth, we focused on overall validity for a predictor, or on widely established variants that have come to be treated as essentially separate predictors (e.g., structured vs. unstructured interviews, or ability-based vs. personality-based emotional intelligence). For two predictors we present findings for moderators where the difference in validity is so substantial that we conclude that what had been a single predictor is now best viewed as two predictors. These two are biodata and personality, where we now treat empirically keyed versus rationally keyed biodata separately, and contextualized versus noncontextualized personality (i.e., personality at work vs. personality in general) separately; we detail our examination of these features below. We acknowledge that in many domains a wide variety of moderators have been usefully examined. For example, one can examine the validity of Big Five personality traits among measures using a single stimulus versus forced choice format (Salgado et al., 2014), or examine the validity of situational versus past behavior description interviews (e.g., Taylor & Small, 2002). A detailed examination of moderators within each predictor would be useful, but is beyond the scope of this paper. Fifth, in settings in which meta-analyses present findings for more than one possible mode of use, we attempted to focus on the validity of predictors as they would typically be used in operational settings. For example, in the vocational interest domain, we see the question of interest to be one of correspondence between the interest profile of a person and that of a job, and thus focus on studies examining person-job congruence (e.g., does having realistic interests predict performance in a job categorized as realistic), rather than main effects for an interest dimension (e.g., does a realistic interest score predict performance, regardless of job). Sixth, we focus on variables that have a history of being used in selection systems. There are a wide variety of variables that have been examined as correlates of job performance, yet have not become part of the psychologist's arsenal in designing selection systems (e.g., job satisfaction, family obligations). Seventh, we focused on predictors as operationally measured, rather than at the construct or latent variable level, as our interest is in the operational value of various predictors.

We also note that Schmidt and Hunter (1998) focused exclusively on mean validity. We believe that it is important to pair such means with estimates of variability. Virtually all meta-analyses do this, typically reporting an observed variance or standard deviation, an estimate of the amount of variability expected due to artifacts, and a residual estimate that subtracts expected variance from observed variance. Including estimates of the *SD* of validity in a summary table comparing predictors is useful for differentiating predictors with little validity variance from those with greater variance. Thus, we included such estimates. The values we present are residual *SDs*, in which the observed *SD* has been adjusted for sampling error variance, variance due to the variability in criterion measurement error across studies, and (if range restriction corrections are made) variance due to variability in range restriction across studies. In cases where we pool results across multiple meta-analyses, we computed a pooled *SD* estimate that incorporates variance across the meta-analyses in mean validity estimates as well as variance within each meta-analysis (i.e., total variance equals the variance of the means plus the mean of the variances).

Apart from the different approaches for building range restriction artifact distributions (which is the main focus of this paper), we also want to highlight an important issue regarding the correction for measurement error in the criterion. All of the meta-analyses we review make use of interrater reliability as a reliability estimate. We note that while there is a minority position opposing the use of interrater reliability as a reliability estimate in correcting validity coefficients (Murphy & De Shon, 2000), its use is as close to a consensus practice as one can find in our field. The meta-analyses we review and summarize also differ in how they obtain estimates of interrater reliability; we find four different approaches taken. The first creates an artifact distribution of interrater reliability estimates from the subset of studies reporting such estimates. For example, Salgado et al.'s (2003) cognitive ability meta-analysis contained 19 studies in which interrater reliability was reported, and the mean of this distribution was used. The mean was .52; rising to .59 if two outliers were removed. The second does not use reliability values from the set of studies gathered for the meta-analysis, but rather relies on a distribution with a mean of .52 obtained by Viswesvaran et al. (1996) in meta-analytic work focusing explicitly on interrater reliability of job performance ratings. Using this value as a correction factor increases mean validity by 39% (i.e., dividing mean observed validity by the square root of .52). The third is conceptually identical to the second, but instead of the .52 value from the Viswesvaran et al. (1996) meta-analysis, it uses a value of .60, citing various sources, including Pearlman et al. (1980) and Bobko et al. (1999). Using this value as a correction factor increases mean validity by 29%. The fourth is a hybrid, used when performance measures are a mix of supervisory ratings and objective performance measures. Here, a different value is used for each subset. For example, Roth et al. (2005) meta-analysis of work sample test validity uses a mean reliability of .60 for ratings criteria and .80 for objective measures.

We note that some selection procedures are more likely to be used in some settings than in others. Assessment centers, for example, are most commonly used for higher-level positions, while interviews are used broadly across job types. If all predictors are being correlated with overall job performance, would we expect that the true mean interrater reliability of performance ratings should be the same for all predictors? Or might average interrater reliability

be expected to differ across predictors, as the jobs for which validity is examined are likely noncomparable across predictors? Conway and Huffcutt (1997) reported that the interrater reliability of performance ratings varies with job complexity, with means of .60 and .48 for low and high complexity jobs, respectively. Thus, if meta-analyses used empirical artifact distributions from studies included in the meta-analysis to estimate interrater reliability, we are comfortable with different corrections being applied for different predictors. However, we are less comfortable with settings in which meta-analysts relied on distributions borrowed from other settings. In these settings, some meta-analysts chose .60 as their mean value, while others choose .52. Choosing the lower value means a 10% larger increase in estimated operational validity than choosing the higher value. When estimating validity across predictors, it would be troubling to conclude that, for two predictors with equal mean observed validity, predictor A has higher operational validity than predictor B due to nothing more than differing assumptions about criterion reliability. Thus, we propose to use a consistent value in settings in which an artifact distribution external to the set of studies in the meta-analysis is used. We will use .60 rather than .52 for two reasons. First, research indicates higher reliability for lower complexity jobs, which quite naturally make up a larger proportion of the body of validation literature due to larger samples being available for validation. Second, our principle of conservative estimation leads us to prefer a conservative estimate in the case of uncertainty. While we use .60 as our estimate of mean criterion reliability, we will also present findings using the .52 value, permitting the reader to compare results. We also need a value for the *SD* of the reliability artifact distribution. We use an *SD* value of .095 from a meta-analysis of criterion reliability from Viswesvaran et al. (1996).

Finally, for as many selection predictors as information is available, we also present the mean Black–White difference for the predictor. Given the longstanding concerns about considering both validity and any adverse impact that results from group differences when designing a selection system (e.g., Sackett & Wilk, 1994; Sackett et al., 2001) we view it as useful to gather in one place information about each of these two features. We acknowledge that presenting information about only a single subgroup comparison reflects less than complete information (see Roth et al., 2017); however, the Black–White mean difference has been examined for more predictors than for other groups. For example, Dahlke and Sackett (2017) located meta-analytic or nationally representative sample estimates for 38 predictors for the Black–White comparison and for 18 predictors for the Hispanic–White comparison. We rely heavily on Dahlke and Sackett as the source of Black–White predictor differences. We note that the subgroup difference values are either from applicant samples, where no range restriction correction is needed, or from mixtures of applicant and incumbent samples, to which no range restriction correction has been applied due to lack of the information needed for appropriate correction. The values used reflect the best data available to the field at present, but the inclusion of incumbent samples adds some uncertainty. This remains an important issue for the field that requires further clarification, as subgroup difference estimates can differ depending on whether one examines national samples, job-specific applicant pools, job incumbents, or mixtures of these. We argue that the Black–White data we present are useful in the sense of locating various predictors in terms of having large, modest, or minimal group differences, even as more precise point estimation will require

additional data. We also emphasize that the focus on only the Black–White subgroup comparison is a function of data availability rather than a lack of interest in other subgroups. We believe though, that presenting at least some subgroup difference information is a useful counterpart to validity information, as concerns about both validity and diversity are widespread. In our table of findings, we will note which subgroup difference estimates are based on applicant samples versus other samples.

## Results

We now turn to a predictor-by-predictor review of the meta-analytic evidence. Table 2 presents detailed information about each meta-analysis we draw on, including number of studies, total sample size, mean observed validity, mean validity corrected for measurement error in the criterion, mean validity also corrected for range restriction, the artifact distribution mean criterion reliability value used in the analysis, an evaluation of the appropriateness of the range restriction correction factor used in the study, and our revised estimate of operational validity. Table 3 presents Schmidt and Hunter's (1998) estimates of operational validity alongside our revised validity estimates, and also presents estimates of the residual *SD*, the lower bound of the 80% credibility interval, and the Black–White subgroup standardized mean difference (*d*). Figure 2 presents the key information from Table 3 in graphic form, plotting each selection predictor in two-dimensional space, with validity on one axis and the Black–White mean difference on the other. In addition, the size of the dot representing a predictor shows the residual *SD* for the predictor.

Table 4 represents a sensitivity analysis wherein we report several alternate estimates of operational validity. This table demonstrates how much our validity estimates would change if different plausible criterion reliabilities and/or amounts of range restriction are assumed. First, for criterion reliability, in addition to our main validity estimates drawn from Table 3 and presented in the first column of Table 4, which are based on a criterion reliability value of .60, we also present estimates based on a criterion reliability value of .52. We have presented above our rationale for the use of .60, but show both corrections here for the interested reader. The table then presents separate validities using the .60 and .52 reliabilities corrected for one illustrative level of indirect range restriction which we view as the maximum plausible average amount of range restriction in studies; namely, with  $r_{zx}$  set at .50, with the selection ratio at the relatively extreme value of .05, and the reliability of the predictor of interest set at .80. We used the Case IV method for correction. We have argued against making a range restriction correction unless one has a sound empirical basis for doing so. However, we recognize that there is often likely to at least be some range restriction affecting validity estimates, even if that amount of range restriction is typically small. Thus, we offer these range-restriction-corrected validity estimates to approximate the highest corrected values for a meta-analytic operational validity estimate that could plausibly be obtained in an indirect restriction scenario (i.e., .50 is the high end of the range of plausible values for  $r_{zx}$ ; it is by no means the expected value).

### Cognitive ability

As already noted, Schmidt and Hunter's (1998) validity estimate for cognitive ability is based on an analysis by Hunter (1983) of validity evidence for the general ability score from the GATB. The

**Table 2**  
*Meta-Analytic Validity Estimates by Selection Procedure*

Selection procedure	<i>K</i>	<i>N</i>	Mean observed $r_{xy}$	$r_{xy}$ used by meta-analysis	Validity corrected for $r_{yy}$	<i>U</i>	RR corrected validity	Basis for RR correction	Is RR correction credible?	Current estimate of operational validity ( $\rho$ ) with best available correction
<b>Cognitive ability</b>										
Hartigan and Wigdor (1989); GATB: early studies	515	38,620	0.25	0.8	0.28	NA	NA	No correction	NA	0.32
Hartigan and Wigdor (1989); GATB: later studies	264	38,251	0.21	0.8	0.23	NA	NA	No correction	NA	0.27
Salgado et al. (2003)	93	9,554	0.29	0.52	0.4	1.61	0.62	Mix of empirical <i>U</i> values and inferred values from hiring rate	No: (a) applies correction to concurrent studies; (b) uses hiring rate as <i>u</i>	0.4
Bertua et al. (2005)	12	2,469	0.22	0.52	0.31	1.67	0.48	Unclear	No: (a) applies correction to concurrent studies; (b) uses hiring rate as <i>u</i>	0.28
Overall			0.24		0.27		0.59			0.31
<b>Employment interviews</b>										
McDaniel et al. (1994): structured	36	3,069	0.28	0.6	0.37	1.47	0.51	<i>u</i> of subset of studies	No: applies empirical <i>u</i> to concurrent studies	0.37
McDaniel et al. (1994): unstructured	9	531	0.21	0.6	0.27	1.47	0.38	<i>u</i> of subset of studies	No: applies empirical <i>u</i> to concurrent studies	0.27
Huffcutt et al. (2014): structured	69	4,795	0.35	0.52	0.48	1.64	0.70	Assumed (SR = 5% for concurrent and 50% for predictive)	No: assumed distribution with no empirical basis	0.45
Huffcutt et al. (2014): unstructured	23	2,594	0.12	0.52	0.16	1.64	0.40	Assumed (SR = 5% for concurrent and 50% for predictive)	No: assumed distribution with no empirical basis	0.18
Overall			0.32		0.44		0.63			0.42
Structured			0.13		0.18		0.40			0.19
<b>Work sample tests</b>										
Roth et al. (2005)	54	10,469	0.26	.60 for supervisory ratings; .80 for objective measures	0.33	NA	NA	NA	Rightly makes no correction	<b>0.33</b>
<b>Job knowledge</b>										
Dye et al. (1993)	59	3,965	0.31	0.6	0.4	1.69	0.62	Assumed (SR = .5)	No: assumed distribution with no empirical basis	<b>0.40</b>
<b>Situational judgement tests</b>										
McDaniel et al. (2007): knowledge-based	96	22,050	0.2	0.6	0.26	NA	NA	NA	Makes no correction—almost all are concurrent	<b>0.26</b>
McDaniel et al. (2007): behavioral tendency	22	2,706	0.2	0.6	0.26	NA	NA	NA	Makes no correction—almost all are concurrent	<b>0.26</b>

(table continues)



This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

**Table 2** (continued)

Selection procedure	K	N	Mean observed $r_{xy}$	$r_{yy}$ used by meta-analysis	Validity corrected for $r_{yy}$	U	RR corrected validity	Basis for RR correction	Is RR correction credible?	Current estimate of operational validity ( $\rho$ ) with best available correction
Salgado et al. (2003)—FFM framework	48	7,562	0.05	0.52	0.07	1.18	0.08	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.06
Salgado et al. (2003)—non-FFM framework	29	4,364	0.05	0.52	0.07	1.18	0.08	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.06
Judge et al. (2013)	47	16,068	0.06	0.82	0.07	NA	NA	NA	NA	0.08
Shaffer and Postlethwaite (2012)—contextualized	14	2,178	0.09	0.52	0.12	1.08	0.14	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.12
Agreeableness	206	36,210	0.06	NA	NA	NA	NA	NA	NA	0.08
Barrick et al. (2001)	68	10,716	0.08	0.52	0.11	1.22	0.13	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.1
Salgado et al. (2003)—FFM framework	31	4,573	0.08	0.52	0.11	1.22	0.13	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.1
Judge et al. (2013)	40	14,321	0.13	0.82	0.14	NA	NA	NA	NA	0.17
Shaffer and Postlethwaite (2012)—contextualized	21	3,357	0.15	0.52	0.21	1.10	0.24	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.19
Extraversion	222	39,432	0.06	NA	NA	NA	NA	NA	NA	0.08
Barrick et al. (2001)	75	11,940	0.04	0.52	0.06	1.16	0.07	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.05
Salgado et al. (2003)—FFM framework	26	4,338	0.05	0.52	0.07	1.16	0.08	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.06
Judge et al. (2013)	63	19,868	0.16	0.82	0.18	NA	NA	NA	NA	0.21
Shaffer and Postlethwaite (2012)—contextualized	18	2,692	0.16	0.52	0.22	1.09	0.25	Combination of $u$ and manual national norms	No: applies empirical $u$ to concurrent studies	0.21
Overall										
Conscientiousness			0.16		0.23		0.26			0.19
Emotional stability			0.07		0.09		0.13			0.09
Openness			0.04		0.07		0.08			0.05
Agreeableness			0.08		0.12		0.13			0.10
Extraversion			0.08		0.13		0.07			0.10
Contextualized										
Conscientiousness			0.19		0.26		0.30			0.25
Emotional stability			0.18		0.25		0.27			0.23
Openness			0.09		0.12		0.14			0.12
Agreeableness			0.15		0.21		0.24			0.19
Extraversion			0.16		0.22		0.25			0.21

(table continues)

Table 2 (continued)

Selection procedure	<i>K</i>	<i>N</i>	Mean observed $r_{xy}$	$r_{yy}$ used by meta-analysis	Validity corrected for $r_{yy}$	<i>U</i>	RR corrected validity	Basis for RR correction	Is RR correction credible?	Current estimate of operational validity ( $\rho$ ) with best available correction
Interest										
Nye et al. (2017)	80	14,522	0.17	0.60	0.22	[1.08, 1.16]	0.24	Combination of <i>u</i> and manual national norms	Somewhat: applied <i>u</i> to all studies but 73.5% were predictive	<b>0.24</b>
Emotional intelligence										
Joseph et al. (2015)—Ability EI	13	1,287	0.17	0.88	0.18	1.01	0.18	Not provided	No: does not distinguish concurrent from predictive	<b>0.22</b>
Joseph et al. (2015)—Mixed EI	16	2,343	0.23	0.79	0.26	1.05	0.27	Not provided	No: does not distinguish concurrent from predictive	<b>0.30</b>
Work experience										
Van Iddekinge et al. (2019)	32	8,463	0.06	.6 for single-rater; .79 for k-rater; .72 for productivity; .80 for work samples; .76 for CWB	0.07	.95	NA	NA	Rightly makes no correction	<b>0.07</b>
Biographical data										
Rothstein et al. (1990)	79	11,288	0.285	0.64 for duties ratings; .69 for ability ratings	0.35	1.10	0.35	Complete local artifact	Yes: correction made to individual studies	0.35
Speer et al. (2021)—Empirically keyed	49	20,564	0.31	0.52	0.44	IRR: 1.01; DRR: 1.09	0.44	Combined local and Rothstein et al. (1990) for distribution	Minimal restriction—only three studies had DRR	0.4
Speer et al. (2021)—Rationally keyed	22	16,279	0.17	0.52	0.24	IRR: 1.01; DRR: 1.09	0.24	Combined local and Rothstein et al. (1990) for distribution	Minimal restriction—only three studies had DRR	0.22
Overall										
Empirically keyed			0.30		0.41		0.41			0.38
Rationally keyed			0.17		0.24		0.24			0.22

Note. *K* = number of studies. *N* = total sample size. RR = range restriction. DRR = direct range restriction. IRR = indirect range restriction. SR = selection ratio.

<sup>a</sup> A second-order meta-analysis.

**Table 3***Comparison of Schmidt and Hunter's (1998) Validity Estimates With Present Study's Validity Estimates and Subgroup Differences*

Selection procedure	Schmidt and Hunter (1998) Validity estimate	Current meta-analysis			
		Validity estimate ( $\rho$ )	<i>SD</i> of $\rho$	Lower 80% credibility value	B-W <i>d</i>
Employment interviews—structured	0.51	0.42	0.19	0.18	0.23
Job knowledge tests	0.48	0.40	0.13	0.23	<i>0.54</i>
Empirically keyed biodata	0.35	0.38	0.09	0.26	<i>0.33</i>
Work sample tests	0.54	0.33	0.09	0.21	0.67
Cognitive ability tests	0.51	0.31	0.14	0.13	0.79
Integrity tests	0.41	0.31	0.20	0.05	0.10
Personality-based EI		0.30	0.17	0.08	0.22
Assessment centers	0.37	0.29	0.09	0.17	0.52
SJT—knowledge		0.26	0.10	0.13	<i>0.39</i>
SJT—behavioral tendency		0.26	0.12	0.11	<i>0.34</i>
Conscientiousness—contextualized		0.25	0.00	0.25	<i>-0.07</i>
Interests	0.1	0.24	0.25	-0.08	<i>0.33</i>
Emotional stability—contextualized		0.23	0.10	0.10	<i>0.09</i>
Ability-based EI		0.22	0.05	0.16	
Rationally keyed biodata		0.22	0.06	0.14	<i>0.33</i>
Extraversion—contextualized		0.21	0.08	0.11	<i>0.16</i>
Conscientiousness—overall	0.31	0.19	0.15	0.02	<i>-0.07</i>
Employment interviews—unstructured	0.38	0.19	0.16	-0.01	0.32
Agreeableness—contextualized		0.19	0.13	0.02	<i>0.03</i>
Openness to experience—contextualized		0.12	0.00	0.12	<i>0.01</i>
Extraversion—overall		0.10	0.12	-0.06	<i>0.16</i>
Agreeableness—overall		0.10	0.13	-0.08	<i>0.03</i>
Emotional stability—overall		0.09	0.08	-0.01	<i>0.09</i>
Job experience (years)	0.18	0.07	0.11	-0.07	<i>0.49</i>
Openness to experience—overall		0.05	0.07	-0.03	<i>0.10</i>
Selection procedures excluded for insufficient information					
Years of education	0.10				
Peer ratings	0.49				
T&E behavioral consistency method	0.45				
Job tryout procedure	0.44				
Reference checks	0.26				
Graphology	0.02				
Age	-0.01				
T&E point method	0.11				

*Note.* EI = emotional intelligence. SJT = situational judgment test. B-W *d* = Cohen's *d* of predictors between Black and White from Dahlke and Sackett (2017) except GMA from Roth et al. (2011) and Job Knowledge from Roth et al. (2003). Job Knowledge estimate is from concurrent samples, as there is no meta-analysis of applicant data. Italicized B-W *d* values are from nonapplicant samples, mixed samples, or did not provide sufficient information to classify type of sample; unitalicized B-W *d* values are from applicant samples.

U.S. Employment Service conducted over 700 validity studies of the GATB over many decades. The intent was to develop specific test batteries for each of a large number of jobs. Studies followed a common template. For most studies, current employees were tested; for other studies, applicants were tested at Employment Service offices, but test scores were not made available to the hiring employer. Thus, both approaches have the same key common characteristic: test scores were not part of the selection process, and thus only indirect range restriction is possible. The vast majority of studies used supervisory ratings as criteria; studies contributing to meta-analyses considered here were those using rating criteria. A standard rating form was used across jobs, with the standard set of rated dimensions sometimes augmented if an employer had a specific interest in additional dimensions. See a National Academy of Sciences volume reviewing the GATB program for details (Hartigan & Wigdor, 1989). The validity database contains two sets of studies: 515 earlier studies and 264 later studies. Hunter's (1983) analysis relied on the earlier set of studies; Hunter sorted jobs into five levels of complexity, and Schmidt and Hunter (1998) used

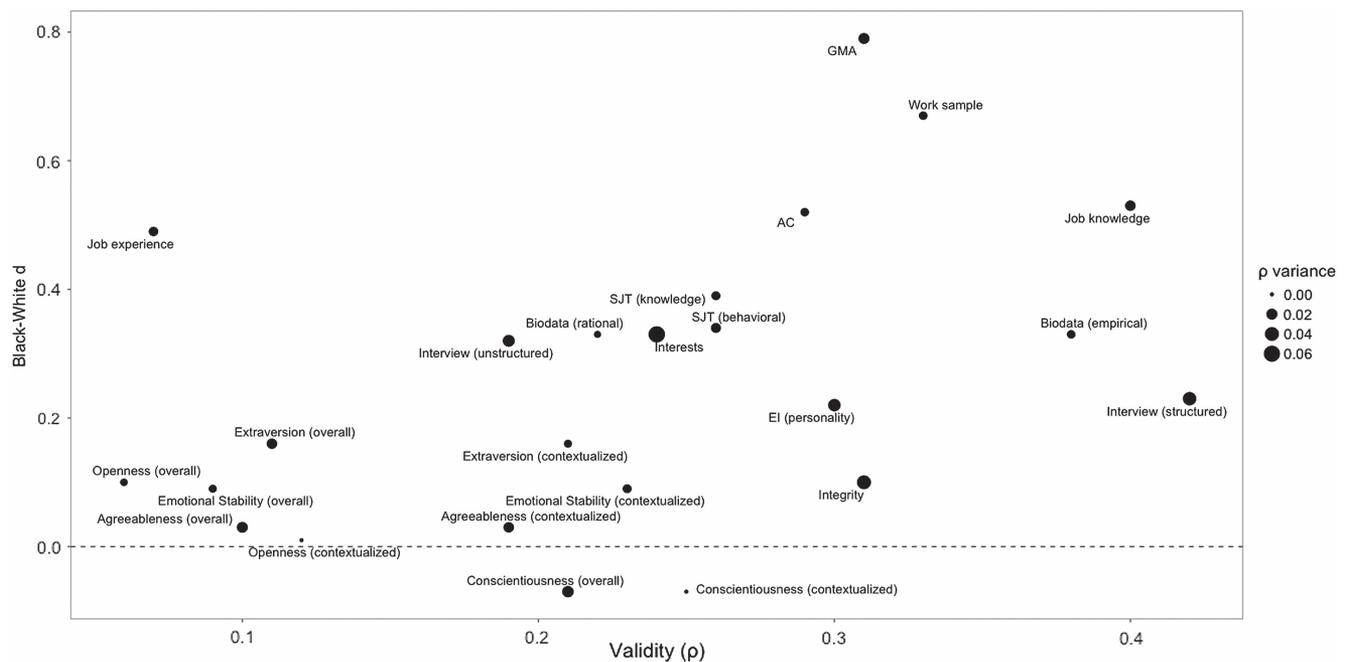
data from studies of jobs of medium complexity as the basis for the validity value of .51 that they report for cognitive ability.

The mean observed validity for these studies was .28; Hunter (1983) used a value of .60 as the basis for a correction for unreliability in the criterion, which raises the mean validity to .36. Hunter's correction for range restriction increases this value by 42%, to .51. In the introduction above we critiqued Hunter's range restriction correction: Hunter estimated a *U* value of 1.5 by pooling incumbent data across jobs and using this as the estimates of applicant *SD* for each job. Above we showed that a *U* value of this magnitude is not possible in concurrent studies unless the prior basis for selection correlates .80 or higher with the GATB ability score. We view that as implausible, and thus conclude that there is no trustworthy basis for further correction beyond the correction for measurement error.

The National Academy of Sciences (NAS) evaluation of the GATB contains meta-analyses of both the newer and older sets of GATB studies. For the older studies, we prefer the NAS estimate to Hunter's, as it includes all jobs, rather than limiting the estimate to

**Figure 2**

*A Visual Summary of Common Selection Procedures' Validity, Validity Variance, and, Black-White  $d$*



Note. GMA = General Mental Ability. AC = Assessment Center. SJT = Situational Judgment Test. EI = Emotional Intelligence.

jobs of medium complexity. After all, meta-analyses of other predictors are also not restrictive as to job complexity level. Thus, not limiting the estimate for cognitive ability to jobs of medium complexity makes the comparison across predictors more apt. Mean observed validity was .25 for the older studies and .21 for the newer. The NAS endorsed a conservative correction for measurement error, using a value of .80, which raises these to .35 and .29, respectively. We recorrected the mean observed validities for measurement error in the criterion using .60, resulting in .32 for the older studies and .27 for the newer studies. As GATB scores were not used in selection, range restriction is expected to be minimal, and we apply no correction. The NAS report did not find differences between the older and newer studies on variables such as study design and sample demographics; the one notable difference was the use of much larger sample sizes in the new set of studies, with average  $N$  rising from 75 to 146.

We also draw on two additional meta-analyses: Salgado et al. (2003) reported a mean observed validity of .29 in 93 ability test validity studies in the European community, and Bertua et al. (2005) reported a mean observed validity of .22 in 12 ability test validity studies in the United Kingdom. They corrected for measurement error in the criterion using reliability estimates of .52, which raised these mean values to .40 and .31, respectively. They further corrected for range restriction, producing fully corrected mean values of .62 and .48, respectively. Both obtained a  $U$  ratio artifact distribution from a subset of studies. Salgado et al. reported that information was available for 17% of studies; Bertua et al. did not report the percentage. Salgado (personal communication) reported that 78% of studies were concurrent, whereas Bertua et al. did not report the percentage. Both reported that their artifact distribution

was a mix of empirical  $U$  values and  $U$  values estimated by converting a reported overall hiring rate into a  $U$  value. As we noted earlier in the paper, the use of hiring rates requires an assumption that we view as untenable, namely, that the predictor under examination was the sole basis for selection. That we cannot separate studies with an empirical  $U$  value from studies estimating  $U$  based on hiring rate, paired with the fact that the meta-analyses applied correction factors derived from predictive studies to concurrent studies as well, leads us to the conclusion that the corrections used in these meta-analyses are substantial overestimates. Without a trustworthy estimate of range restriction, and with the fact that the vast majority of studies were concurrent and were thus affected only minimally by range restriction, we conclude that our best estimates at present are the mean validity values corrected only for measurement error in the criterion. We retained Salgado et al.'s reliability-corrected validity estimate of .40 given that their criterion reliability estimate of .52 was computed locally. However, Bertua et al. used the .52 criterion reliability value from Viswesvaran et al. (1996), and we therefore recorrected with the .60 value, resulting in a validity estimate of .28.

Taken together, the  $N$ -weighted mean reliability-corrected value across the GATB datasets, the Salgado et al.'s (2003) study, and the Bertua et al.'s (2005) study is .31. This is our estimate of cognitive ability test validity, in contrast with the .51 value presented by Schmidt and Hunter (1998).

### *Structured and Unstructured Interviews*

The structured and unstructured interview validity estimates of .51 and .38 tabled in Schmidt and Hunter (1998) were drawn from

**Table 4***Comparison of Current Validity Estimates With Alternate Estimates Based on Different Reliability and Range Restriction Corrections*

Selection procedure	Current validity estimate ( $\rho$ ) corrected with $r_{yy} = .60^a$	Current validity estimate ( $\rho$ ) corrected for range restriction <sup>b</sup>	Alternative validity estimate ( $\rho$ ) corrected with $r_{yy} = .52^c$	Alternative validity estimate ( $\rho$ ) corrected for range restriction <sup>b</sup>
Employment interviews—structured	0.42	0.45	0.45	0.48
Job knowledge tests	0.40	0.44	0.43	0.47
Empirically keyed biodata	0.38	0.42	0.40	0.44
Work sample tests	0.33	0.36	0.33	0.36
Cognitive ability tests	0.31	0.34	0.33	0.36
Personality-based EI	0.30	0.33	0.32	0.35
Assessment centers	0.28	0.29	0.29	0.30
SJT—knowledge	0.26	0.29	0.28	0.31
SJT—behavioral tendency	0.26	0.29	0.28	0.31
Conscientiousness—contextualized	0.25	0.27	0.26	0.29
Interests	0.24	0.24	0.24	0.26
Emotional stability—contextualized	0.23	0.26	0.25	0.28
Ability-based EI	0.22	0.24	0.24	0.26
Rationally keyed biodata	0.22	0.24	0.24	0.26
Extraversion—contextualized	0.21	0.23	0.22	0.25
Conscientiousness- overall	0.19	0.22	0.20	0.22
Employment interviews—unstructured	0.19	0.21	0.19	0.22
Agreeableness—contextualized	0.19	0.22	0.21	0.23
Openness to experience—contextualized	0.12	0.13	0.12	0.14
Extraversion—overall	0.10	0.11	0.10	0.11
Agreeableness—overall	0.10	0.11	0.10	0.12
Emotional stability—overall	0.09	0.09	0.09	0.10
Job experience (years)	0.07	0.08	0.07	0.08
Openness to experience—overall	0.05	0.06	0.06	0.06

*Note.* EI = emotional intelligence. SJT = situational judgment test.  $r_{yy}$  = reliability of job performance. Integrity test is not included in this table as both meta-analytic studies used local artifact for reliability and made an appropriate range restriction correction.

<sup>a</sup>Current estimate is the same as in Table 4 and thus based on .60 as the reliability estimate except for meta-analytic studies that applied a local reliability distribution. <sup>b</sup>Case IV indirect range restriction was applied. <sup>c</sup>Alternative validity was corrected using .52 as the reliability estimate of job performance except for meta-analytic studies that applied a local reliability artifact.

McDaniel et al. (1994). The structured interview estimate reflects a mean observed validity of .28, increasing to .37 when correcting for measurement error in the criterion, and then increasing to .51 when correcting for range restriction. Thus, the restriction correction increased the validity estimate by 38%. The unstructured interview estimate reflects a mean observed validity of .21, increasing to .27 when correcting for measurement error in the criterion, and then increasing to .38 when correcting for range restriction. Thus, the restriction correction increased the validity estimate by 41%. We note that Schmidt and Hunter limited their analysis to studies using performance ratings obtained for research purposes, which reflects 36 of the 106 structured interview validity estimates and 9 of the 39 unstructured validity estimates. In the full sample, the corresponding fully corrected values are .44 and .33, respectively.

McDaniel et al. (1994) did not differentiate between predictive and concurrent studies, or report how many studies were of each type. They obtained empirical  $U$  estimates from 14 of 245 studies, and used this set of  $U$  values as their artifact distribution. The mean  $U$  value was 1.47. This artifact distribution was used to correct all validity coefficients, both predictive and concurrent, and thus we expect it to produce a substantial overestimate of operational validity. Absent information about the proportion of studies that are predictive versus concurrent, we cannot apply an alternate correction (i.e., apply the artifact distribution only to the predictive studies), and thus we use the reliability-corrected .37 and .27 as our validity estimates for structured and unstructured validity, respectively.

Huffcutt et al. (2014) presented a new meta-analysis of interview validity. Only 10 of the 91 studies in their central analyses overlapped with McDaniel et al. (1994). They used four categories of interview structure, noting that their categories 1 and 2 corresponded to McDaniel et al.'s unstructured label, and categories 3 and 4 to the structured label; we thus collapsed Huffcutt et al.'s findings as structured versus unstructured accordingly. They reported a mean observed validity of .36 and .13 for structured and unstructured interviews; correcting for measurement error using .52 (Rothstein et al., 1990) produced values of .49 and .18, respectively. Their further correction for range restriction produces values of .70 and .36.

We first recorrected mean observed validities for criterion unreliability using .60, obtaining .46 for structured interviews and .17 for unstructured interviews. We also offered a critique of Huffcutt et al.'s (2014) range restriction correction earlier in the paper. For concurrent studies, they assumed an extreme prior selection ratio of 5%, while assuming approximately a 50% selection ratio for predictive studies. We note that 74% of their studies were concurrent, and we apply no correction to the concurrent studies. They used prior work by Salgado and Moscoso (2002) which produced an empirical  $U$  estimate of 1.63 for predictive interview studies. We apply that correction to the predictive studies and then produce an  $N$ -weighted overall estimate, combining predictive and concurrent studies, of .45 for structured interview validity and .18 for unstructured interviews.

We computed sample size-weighted means to combine our estimates from the McDaniel et al. (1994) and Huffcutt et al. (2014) meta-analyses, producing mean validity estimates of .42 for structured interviews and .19 for unstructured interviews.

### **Work Samples**

We rely on a meta-analysis of work samples by Roth et al. (2005), which provided much more current information than the narrative review by Asher and Sciarino (1974) used by Schmidt and Hunter (1998). Roth et al. located 54 studies, 53 of them concurrent and one predictive. With a mean observed validity of .26, correcting for measurement error in the criterion using .60 for supervisor ratings and .80 for objective criteria produced a mean corrected validity of .33. Importantly, with 98% concurrent studies, we, like Roth et al., make no correction for range restriction.

### **Job Knowledge Tests**

We rely on a meta-analysis of job knowledge tests by Dye et al. (1993). Schmidt and Hunter's (1998) estimate is based on validity coefficients from a technical report by Dunnette (1972) on job knowledge tests for jobs relevant to the petroleum industry. The narrow scope and age of the studies led us to focus on the Dye et al.'s meta-analysis. Dye et al. made the important differentiation between knowledge tests developed for the job in question versus settings in which a knowledge test that is conceptually irrelevant to the job is used; they use the example of a mechanical aptitude test being used for stenographers. While they located 164 studies, with a mean observed validity of .22, 59 of the studies used job-specific knowledge tests, with a mean observed validity of .31. Although we report both sets of findings, we focus on the job-specific tests. While it is of some academic interest to examine how well irrelevant knowledge predicts performance, the argument for considering the use of job knowledge tests is surely that there is a body of knowledge relevant to a given job, and candidates who possess more of that knowledge will be more effective on the job. The mean observed validity of .31 increases to .40 once corrected for measurement error using a reliability value of .60.

Dye et al. (1993) did correct their findings for range restriction. However, they based their correction on an assumed range restriction distribution. They assumed direct range restriction, with a selection ratio of .50. They did not report whether studies were predictive or concurrent. As discussed earlier, with assumed distributions and with no information about the distribution of predictive and concurrent studies, we cannot feel confident of any range restriction correction. Thus, we offer the unreliability-corrected value of .40 as a conservative validity estimate, subject to future research that might address range restriction concretely and credibly.

### **Situational Judgment Tests**

We rely on a meta-analysis by McDaniel et al. (2007). They differentiated between SJTs using what they term knowledge instructions (e.g., pick the best/worst response instructions: asking the respondent to identify what one "should" do) and SJTs using what they term behavioral instruction (e.g., pick the response you would engage in: asking the respondent to identify what one "would" do).

They located 96 studies using knowledge instructions and 22 using behavioral instructions. Mean observed validity was .20 for both, increasing to .26 after correcting for error of measurement using a reliability of .60. No correction for range restriction was made, as virtually all studies were concurrent. Thus, we offer .26 as our estimate of operational validity.

### **Assessment Centers**

We summarize and integrate the findings from three meta-analyses. Gaugler et al. (1987) was the basis for Schmidt and Hunter's mean value of .37 for assessment centers. We also include Gaugler et al. However, while Schmidt and Hunter relied on a grand mean validity estimate that combined findings across different criteria, including job performance, training performance, and advancement; we limit our examination to studies using job performance as a criterion. An observed validity of .25 across 44 samples increased to .32 after correcting for measurement error in the criterion. Gaugler et al. obtained a mean criterion reliability value of .61 from a subset of studies, which closely matches the .60 value that we use across selection procedures for consistency. Gaugler et al. further corrected the validity estimate for range restriction, coding each study on a dichotomous variable: no restriction (90% of studies) versus restriction (10% of studies). They used the mean of this dichotomous variable (i.e., .90) as their estimate of  $u$ , corresponding to a  $U$  value of 1.1, which leads to a range-corrected value of .36. We do not find this a viable approach to correction. A  $u = .9$  across all studies would only make sense if the  $u = 0$  for the 10% of studies with restriction (i.e., 90% of studies have  $u = 1.0$  because they have no restriction, so you can only have a mean  $u = .9$  if the remaining 10% have  $u = 0$ ). Mathematically,  $u$  cannot be zero unless the numerator of  $u$  (the restricted  $SD$ ) were zero, meaning all selected individuals had identical predictors scores. In any case, with 90% of studies having no restriction, correcting for range restriction would have little effect on the correlation, and we make no correction. Thus, we use the reliability-corrected value of .32 as our mean operational validity estimate.

Hermelin et al. (2007) reported a meta-analysis of 27 studies conducted since Gaugler et al. (1987). They found a mean observed validity of .17 using supervisor ratings of performance criterion measures. They corrected each study individually for range restriction, as 12 studies used the assessment center operationally and study-specific  $U$  ratios could be calculated; no correction was made for the other studies. They corrected for unreliability in the criterion using a mean value of .52. As we are using a value of .60 for consistency across predictors, we obtain an operational validity estimate of .26, slightly smaller than their estimate of .28.

Hardison and Sackett (2006) also conducted a meta-analysis of more recent studies than Gaugler et al. (1987). They located 49 studies with performance ratings criteria. About one-third of studies (16) were also included in Hermelin et al. (2007). We include validity estimates from both meta-analyses given the considerably different, though not completely unique, set of studies. Hardison et al. reported a mean validity of .20, increasing to .26 after correction for error of measurement in the criterion using the reliability estimate of .60. They reported that too few studies provided the information needed for range restriction corrections, and did not attempt a correction.

In short, an  $N$ -weighted average across the three meta-analyses yields a value of .29. We use this as our operational validity estimate.

### **Integrity Tests**

We summarize and integrate the results of two meta-analyses. Ones et al. (1993) reported an extensive meta-analysis of the relationship between integrity test scores and job performance. An overall analysis was based on 222 studies. However, the analysis used as the basis for their conclusions was based on a much smaller subset: 23 studies that used a predictive validity design with job applicants. They made the case that this subset best represents the operational value of the tests in selection settings. They reported a mean observed validity of .25, increasing to .35 correcting for a criterion reliability value of .52 (which they obtained from a subset of studies in their meta-analysis), and further increasing to .41 correcting for range restriction. They located 79 studies reporting both restricted and unrestricted test  $SD$ s, and thus had a sizeable amount of data to create an artifact distribution. As their artifact distribution was drawn from predictive validity studies and the subset of 23 studies that were their focus were also predictive validity studies, we are comfortable with using their artifact distribution for correction. This value of .41 was used by Schmidt and Hunter (1998) as their operational validity estimate. However, we have reason to believe that .41 is an underestimate because they corrected for direct range restriction when range restriction was likely indirect in their studies (i.e., it is unlikely employees were selected into their jobs using only the integrity test scores). Correcting for direct range restriction when range restriction is actually indirect can (depending on the specific details of how the direct range restriction correction was performed) result in an underestimate of validity (Hunter et al., 2006; Linn et al., 1981; Sackett et al., 2007). We can estimate what the indirect-range-restriction-corrected operational validity would be using Hunter et al.'s (2006) Case IV method of correcting for indirect range restriction. This method is based on the observation that in the case of indirect range restriction, restriction takes place on latent predictor scores, not observed scores; see Hunter et al. (2006) for procedural details. Hunter et al.'s method requires as input the observed validity (.25), the restricted criterion reliability (.52), the restricted predictor reliability (.81, reported in Ones et al.'s Table 3), and the  $u$  ratio (.81, which corresponds to  $U = 1.23$ ; reported in Ones et al.); using those values results in an operational validity estimate of .44. We therefore use this value.

We conducted another meta-analysis, with about half the studies also included in Ones et al. (1993) and half more recent. They imposed a more restrictive set of study inclusion criteria. For example, they required that detail about study design and data analysis be available, thus excluding secondary reports of studies from qualitative reviews of the literature. They obtained findings markedly different from Ones et al. Like Ones et al., they focused on predictive studies with applicant samples. They located 24 studies, where Ones et al. located 23; total  $N$  for the two studies was roughly 7,000 for each. Observed mean validity was .11. They created a criterion reliability artifact distribution with a mean of .56 based on a subset of the studies in their meta-analysis; thus, we are comfortable using .56 to correct, rather than .60. The observed validity of .11 increased to .15 after correcting for error of

measurement in the criterion. Further correction for range restriction using an artifact distribution with a mean  $u$  ratio of .90 ( $U = 1.11$ ) from a subset of these applicant studies increased the mean validity to .18, which is our estimate of operational validity.

Thus there is quite a discrepancy between the two meta-analyses, with estimates of .44 and .18. Sackett and Schmitt (2012) carefully examined the two in an attempt to reconcile the findings. They were unsuccessful: they concluded that exclusion criteria, correction for artifacts, and second-order sampling error are not likely explanations for the differences. They noted that the two meta-analyses did not contain enough information for an independent reviewer to re-examine the data. Modern meta-analytic reporting standards call for a more complete reporting (e.g., a listing of all studies, with the effect sizes extracted from each); Ones et al. (1993) was published well before those reporting norms emerged. Ones et al. admirably searched for and located a great many unpublished studies and unpublished data sets from test publishers; however, those data are not available to those trying to reconcile the findings. Sackett and Schmitt concluded that both meta-analyses were of high quality, yet yield different results for reasons not yet understood. Thus, we report an  $N$ -weighted average of the two validity estimates, namely, .31, as our estimate of operational validity.

### **The Big Five Personality Traits**

We review three meta-analyses that examined the validities of the Big Five factors for predicting job performance: Barrick et al. (2001), Salgado (2003), and Judge et al. (2013). These meta-analyses had largely nonoverlapping studies. We will then turn to an additional meta-analysis that examined the effects of contextualization (e.g., specifying the work context either in items or in response instructions; Shaffer & Postlethwaite, 2012).

Barrick et al. (2001) conducted a second-order meta-analysis using 11 meta-analyses of the personality–performance relationships conducted in the 1990s. They reported mean observed validities for extraversion, emotional stability, agreeableness, conscientiousness, and openness to experience to be .06, .06, .06, .12, and .03, respectively. They also reported correlation estimates that were corrected for measurement error in the predictor and the criterion as well as range restriction by aggregating those of prior meta-analyses. We apply a criterion unreliability correction to the sample size-weighted observed validities using the reliability estimate of .60, resulting in corrected validities of .08, .08, .08, .15, and .04, respectively. We do not correct for range restriction because the  $U$  values used by Barrick et al. reflect key problems we have identified above, namely, applying  $U$  values derived from predictive studies to concurrent studies.

Salgado (2003) examined two newer sets of studies that examined the validities of the Big Five factors measured by inventories developed within the five-factor model (FFM) framework, and those measured by non-FFM inventories. Based on the former, they found mean observed validities of .04, .09, .08, .17, and .05 for extraversion, emotional stability, agreeableness, conscientiousness, and openness to experience, respectively. Based on the latter, they found mean observed validities of .05, .03, .08, .11, and .05, respectively. For both sets of observed validities, they corrected for criterion unreliability using .52 found by Viswesvaran et al. (1996). In keeping with our decision to correct with a consistent value when an external criterion reliability artifact distribution is

used, we computed unreliability-corrected validities using .60, obtaining .05, .12, .10, .22, and .06 for FFM inventories, and .06, .04, .10, .14, and .06 for non-FFM inventories. Their range restriction artifact distribution was computed based on restricted and unrestricted *SDs* when reported, and unrestricted *SDs* reported in instrument manuals for the rest. The resulting *U* ratios were 1.16, 1.23, 1.22, 1.20, and 1.18 for the Big Five, respectively. However, no information was reported regarding the number of studies that were predictive and we could not determine the appropriate range restriction correction to be applied and thus applied no correction.

Judge et al. (2013) reviewed meta-analytic relationships between the Big Five at the facet level and job performance, which were then aggregated to the trait level. They reported mean observed validities of .16, .08, .13, .21, and .06 for extraversion, emotional stability, agreeableness, conscientiousness, and openness to experience, respectively. They corrected for measurement error in both predictors and criteria using internal consistency reliability artifact distributions compiled locally, resulting in corrected validities of .20, .10, .17, .26, and .08, respectively. To obtain estimates that are corrected for unreliability in the criterion only, we recorrected the mean observed validities. However, for the sake of comparability with the other meta-analyses corrected using interrater reliabilities, we did not use Judge et al.'s local internal consistency reliability artifact distributions. Instead, we used an interrater reliability of .60, resulting in criterion unreliability-corrected validities of .21, .10, .17, .27, and .08. Judge et al. did not make range restriction corrections, and did not report information that would permit us to do so.

In sum, *N*-weighting these four sets of unreliability-corrected validities for the Big Five factors across Barrick et al. (2001), Salgado (2003), and Judge et al. (2013) results in our estimates: .10 for extraversion, .09 for emotional stability, .10 for agreeableness, .19 for conscientiousness, and .05 for openness to experience.

As noted earlier, we also considered a meta-analysis by Shaffer and Postlethwaite (2012) that examined validity in studies of contextualized personality, that is, personality in the workplace. Contextualization can be built into items (e.g., add "at work" to items), or imposed via instructions (e.g., "respond in terms of how you behave at work"). They found substantial differences in validity for contextualized personality versus personality in general, such that we choose to view contextualized personality as a separate predictor category, and thus present findings for contextualized personality in addition to the findings reported above, which come from studies that do not differentiate contextualized and noncontextualized personality. Shaffer and Postlethwaite used a mean criterion reliability value of .52 from Viswesvaran et al. (1996); we used .60 to be consistent with our practice across other predictors. They assembled a range restriction artifact distribution using local estimates where available as well as estimates based on applicant *SDs* obtained from test manuals. As they applied their mean *U* ratio to both predictive and concurrent studies we do not view the range restriction correction as credible, and therefore do not correct for restriction. Our estimates of operational validity for contextualized personality measures are .21 for extraversion, .23 for emotional stability, .19 for agreeableness, .25 for conscientiousness, and .12 for openness to experience.

## Interests

Three meta-analyses investigated the validity of vocational interests. However, Nye et al. (2017) included all unique studies

analyzed by both Nye et al. (2012) and Van Iddekinge et al. (2011). Therefore, we focus our review on Nye et al. In addition, we specifically review the validity estimate of interest congruence as we believe that evaluating the importance of the match between one's interests and the interest profile of a given job is more meaningful than the aggregate validity of heterogeneous interest scales for heterogeneous jobs.

Nye et al. (2017) used a regression approach to examine possible moderators of the validity of interests and reported an overall operational validity (corrected for criterion measurement error and indirect range restriction) of interest congruence for all criteria of .16 (p. 142). We estimated the operational validity of interest congruence for task performance from their regression model as .25 (i.e., an intercept of .16, minus an average effect across 5 types of measures of .02, minus an effect for the task performance criterion type of .05, plus an effect for use of a congruence measure of .16). Although they did not report the mean observed validity, Nye (personal communication) provided it to us: the value is .17. We note that the criterion is reported as task performance, rather than overall performance; we include it with that caveat, and also note that most meta-analyses examined here do not present enough information for us to judge whether the performance criterion includes performance facets beyond task performance.

For the indirect range restriction correction, they used restricted *SDs* from the primary studies and unrestricted *SDs* from the technical manuals of the interest measures used when available. Each study was corrected individually, with the average *U* ratio used for studies where study-specific information was unavailable. Furthermore, they reasoned that due to the process of attraction-select-attrition (Schneider, 1987), the degree of range restriction in interests will depend on the level of fit between interest and job characteristics. They used *U* ratios of 1.16, 1.08, 1.10, and 1.15 for matching, adjacent, alternate, and opposite fit between interest and occupation for indirect range restriction corrections, respectively. However, the authors did not report the proportion of correlations that fall into each level of fit. Therefore, we use the *U* ratios of 1.16 as the lower bound and 1.08 as the upper bound of range restriction.

Finally, their operational validity estimate of .25 was not corrected for unreliability in the predictors but a predictor unreliability value was needed for their indirect range restriction correction (Hunter et al., 2006). Reliability estimates from the primary studies were used when available. Technical manuals of the interest measures were used as a secondary source. For studies that provided neither, the average reliability across all measures was used. Although the authors did not provide the mean interest reliability estimate used, Nye (personal communication) informed us that the mean reliability was .87.

Thus, mean observed validity is .18; using .60 as a criterion reliability estimate, as did Nye et al., we obtain a corrected validity estimate of .23. While Nye et al. (2017) did not report the number of studies included that were concurrent versus predictive, Nye (personal communication) indicated that 73.5% of the studies were predictive. Applying range restriction corrections to the reliability-corrected estimate with a weight of .735 and applying no range correction to the same estimate with a weight of .265 produced a weighted average of .24, which we use as our operational estimate of validity.

### **Emotional Intelligence**

We review the results of a meta-analysis of the validity of emotional intelligence (EI) for job performance by Joseph et al. (2015), which included studies from prior meta-analyses on the topic by Van Rooy and Viswesvaran (2004), Joseph and Newman (2010), and O'Boyle et al. (2011), as well as additional newer primary studies that measured supervisory ratings of job performance.

Joseph et al. (2015) reported mean observed validities of ability EI to be .17 and mixed EI to be .23 for supervisor ratings of job performance. They corrected for measurement error in both the predictor and criterion, using empirical distributions of reliabilities found in the primary studies. They also corrected for range restriction using artifact distributions of .99 for ability EI and .95 for mixed EI, obtaining validity estimates of .20 and .29, respectively. The authors did not specify the number of studies that were predictive versus concurrent, thus while applying range restriction corrections to all studies likely produced an overestimate, the magnitude of the overestimation is unclear. The authors also did not report validities without correcting for measurement error in the predictors. However, they did provide a table that lists the primary studies included, their effect sizes, sample sizes, and reliability values. Although the study did not specify what type of criterion reliabilities were used for correction, Joseph (personal communication) reported to us that they used internal consistency reliabilities. In order to be consistent with the other meta-analyses we reviewed that used interrater reliabilities, we re-computed criterion-reliability-corrected validities of ability and mixed EI, correcting for measurement error using an interrater reliability value of .60, resulting in .22 and .30, respectively.

### **Work Experience**

Schmidt and Hunter relied on values from Hunter and Hunter (1984), reporting an operational validity estimate of .18. Hunter and Hunter reported the results of original analyses of the validity of these variables in the GATB data base. As noted earlier, this data base is made up of predominantly concurrent studies, and thus addresses the question of whether workers with a longer tenure in their current job have higher performance. However, when considering experience as a potential selection method, the focus must be on prior experience at the point of job application. Recently, Van Iddekinge et al. (2019) offered a meta-analysis reporting a mean correlation corrected for measurement error in the criterion of .07 for relevant pre-hire experience. Only three studies reported the needed information for a range restriction correction, with a mean  $U$  ratio of 1.05, and thus Van Iddekinge et al. did not correct for range restriction in their estimate of operational validity. We use their value of .07 as our estimate of operational validity.

### **Biodata**

For biodata, Schmidt and Hunter (1998) relied on a meta-analysis by Rothstein et al. (1990). This involved multiple studies using a single empirically keyed biodata instrument, and the authors had full information about applicant and incumbent  $SD$ s for all studies. Thus, this study does not reflect any of the concerns we addressed in this paper about the use of artifact distributions, and thus we do not need to revisit the meta-analysis. From that meta-analysis we accept their value of .35 as the estimate of operational validity.

Speer et al. (2021) provided a new meta-analysis of biodata-job performance relationships. They obtained a mean observed validity of .27 for a biodata composite score, increasing to .37 correcting for measurement error in the criterion using the mean value of .52 drawn from Viswesvaran et al. (1996). They reported a mean  $U$  value of 1.01 for studies with indirect range restriction, and 1.09 for studies with direct restriction. Speer (personal communication) reported to us that only three of the studies involved direct restriction. Thus, restriction is minimal, and applying a correction does not change the mean reliability-corrected value. We revise their estimate using the more conservative .60 value for criterion reliability, producing a value of .35. Thus, this new meta-analysis produces the same operational validity estimate as the prior estimate from Rothstein et al. (1990).

We do note a moderator with a substantial effect on validity. We estimate operational validity at .40 for empirically keyed biodata and .22 for rationally keyed biodata in Speer et al.'s (2021) data. Importantly, Speer et al. examined a set of 18 studies in which both approaches were applied to the same item set; the difference noted above is retained in this subset of studies. This suggests that the two approaches be treated separately in summaries of the validity of selection procedures (as is done with structured vs. unstructured interviews). Thus, we use a weighted average of .38 across Rothstein et al. (1990) and Speer et al.'s empirically keyed findings as our estimate of operational validity for empirically keyed biodata, and Speer et al.'s value of .22 as our estimate of operational validity for rationally scored biodata.

### **Seven Selection Procedures With No New Analyses**

For peer ratings, job tryout, graphology, two types of training and experience evaluations, age, and educational level, Schmidt and Hunter relied on values from Hunter and Hunter (1984). For these we do not have sufficient information to re-evaluate the reported validity estimates. Hunter and Hunter often reported very limited information beyond the mean validity estimate. The  $SD$  of the validity distribution was never reported; in a number of cases the number of studies ( $k$ ) and total  $N$  were not reported. Thus, while we report the Schmidt and Hunter values in Table 2 for completeness, we will not include these values when we turn to discussing relative validity across predictors.

We do note that three of these selection procedures were reported as among the most predictive in the predictor set examined: peer ratings, the behavioral consistency approach to training and experience evaluation, and job tryout. Given the magnitude of reported validity we offer more detail here as to why we do not view the values provided for those three as informative.

For peer rating, the mean validity offered (.49) would be the highest in our list of updated validity estimates. Hunter and Hunter (1984) computed this value from a set of validity coefficients offered in Table 1 of Kane and Lawler's (1978) narrative review of the peer rating literature. Hunter and Hunter reported 31 studies in which peer ratings are correlated with supervisor ratings. However, these were commonly not supervisor ratings of job performance, but rather of a specific trait (e.g., ratings of emotional adjustment, ratings of industriousness). Thus, we conclude that these data simply do not address the question of interest. For the behavioral consistency approach to training and experience evaluation, Hunter and Hunter offered a mean validity of .45, drawn from five studies, with

unspecified  $N$ , included in a technical report by Schmidt et al. (1979). That report is not available on any searchable platform, and thus we are unable to evaluate it (e.g., we do not know observed validity; we do not know the approach to corrections taken in the study). Regarding job tryout, Hunter and Hunter reported a mean validity of .44, from a technical report by Dunnette (1972) on the validity of selection procedures relevant to the petroleum industry. Again, we have no procedural information other than a mean validity estimate.

## Discussion

### Main Conclusions

We presented a detailed critique of the current use of range restriction corrections in meta-analysis. We do not take issue with the mechanics of range restriction correction, but rather with the approaches used to obtain a range restriction correction factor to apply in a given setting. As detailed in our critique, these approaches have led to substantial overcorrections for range restriction in many existing meta-analyses of selection predictor validity. In light of these issues we revisited the estimates of validity for a wide range of predictors that were previously summarized by Schmidt and Hunter (1998).

We made revised range restriction corrections where the needed information was available. When the needed information was not available and a credible correction factor could not be obtained, we argued against making any correction, especially given how small the effects of range restriction are in typical concurrent validity studies. We tabulated the various decisions in the meta-analyses summarized in Table 2. The dominant categorization is labeling the correction made by the meta-analyst as not appropriate, and the data needed for an appropriate correction are not available (25 meta-analyses). For 13, no correction was made by the original authors. For seven, we deem the correction made by the authors as appropriate. For three, we made our own correction. For the studies where we judged the corrections made as inappropriate, failure to differentiate between predictive and concurrent studies, and applying a common correction to all studies was the dominant reason (21 meta-analyses). Four meta-analyses used a selection ratio derived from a hiring rate. Three meta-analyses used an assumed artifact distribution. Thus, differentiating predictive and concurrent studies is the greatest need for the field in moving forward.

The result of this process is markedly lower estimates of operational validity for a number of selection predictors. Our predictors are useful; but the predictive relationships are considerably weaker than previously thought. We acknowledge that these are not findings that will be eagerly embraced by the field. We wish we could report better news. But we believe that a clear-eyed understanding of operational validity is needed.

### Which Selection Procedures Were Most Affected by Our New Calibration?

For many predictors, our new estimates of mean validity are substantially smaller than the values presented by Schmidt and Hunter (1998). The following predictors had validity estimates that changed by .05 or more, ordered from the largest change to the smallest change:

**Work Samples.** The work sample validity estimate decreased by .21, from .54 to .33. This is the result of a new meta-analysis, with the earlier estimate coming from a study pre-dating the development of meta-analysis.

**Cognitive ability.** The cognitive ability validity estimate decreased by .20, from .51 to .31. While this estimate is based on a number of additional meta-analyses, the prime driver of the decrease is a revised estimate of the degree to which the set of studies in question were affected by range restriction.

**Unstructured Interviews.** The unstructured interview validity estimate decreased by .19, from .38 to .19. This estimate combines a prior meta-analysis with a new one. The validity estimates presented in the two meta-analyses are very similar. The driver of the decrease is a revised estimate of the degree to which the studies were affected by range restriction.

**Interests.** The vocational interest validity estimate increased by .14, from .10 to .24. This estimate is based on a new meta-analysis which re-conceptualizes how interests are evaluated. The new estimate is based on the validity of interest congruence indices, which assess fit between a person's interests and the job in question, rather than a main effect across jobs for a particular interest dimension.

**Conscientiousness.** The conscientiousness validity estimate decreased by .12, from .31 to .19. This results from multiple new meta-analyses in this domain. Range restriction is not a significant factor in this domain.

**Experience.** The work experience validity estimate decreased by .11, from .18 to .07. This is the result of a new meta-analysis which focuses on prior work experience at point of hire. Prior work had addressed a different issue, namely, the relationship between current tenure and performance among incumbents in a given job.

**Integrity Tests.** The integrity test validity estimate decreased by .10, from .41 to .31. This is due to a new meta-analysis, producing a markedly lower estimate than a prior meta-analysis; our estimate is an  $N$ -weighted average of the two. Range restriction is not a significant factor in the integrity test domain.

**Structured Interviews.** The structured interviews validity estimate is reduced by .09, from .51 to .42. The estimate incorporates a new meta-analysis in addition to one used in the prior estimate. Our estimate is markedly smaller than that offered in the published meta-analysis, as we did not make use of a range restriction correction that relied on an assumed distribution.

**Job Knowledge.** The job knowledge test validity estimate decreased by .08, from .48 to .40. This estimate relies on a more current meta-analysis. The estimate from that meta-analysis is adjusted downward due to the use of a range restriction correction that we do not find trustworthy, namely, reliance on an assumed distribution.

**Assessment Centers.** The assessment center validity estimate is reduced by .08, from .37 to .29. The estimate incorporates new meta-analyses, and, in the case of one meta-analysis, sets aside a range restriction correction that we do not find trustworthy.

As the above review makes clear, there are three different factors contributing to the changes in the validity estimates. First, in a number of cases more recent meta-analyses bring new findings to bear. Second, in a number of cases we did not adapt the range restriction correction offered in the published meta-analyses, based on the arguments we have developed in this paper (e.g., applying corrections developed on applicant samples to incumbent data, or

using assumed distributions). Third, in two cases new work has led to rethinking how a predictor is conceptualized and evaluated, namely, taking a person-job congruence perspective in evaluating interests, and focusing on prior experience at point of hire.

### ***What Are Now Our Strongest Predictors of Job Performance?***

Table 3 lists predictors in decreasing order of validity based on our revised validity estimates. The Schmidt and Hunter (1998) findings identified three predictors with mean validity above .50: work samples, general cognitive ability, and structured interviews. Two additional predictors that relied on reasonably contemporary meta-analytic findings had mean validity above .40: job knowledge tests and integrity tests. Three others had mean validity above .40, but we set these aside as either inappropriate (i.e., the peer ratings validity estimate included correlations based on ratings of personality traits, while all other predictors are evaluated against job performance) or based on so little information that we could not evaluate the estimates (i.e., data for the behavioral consistency approach to training and experience evaluation and data for job tryouts each came from Hunter and Hunter (1984) with insufficient information reported, for example, unsure of number of studies, of total sample; no information on variability across studies).

Our current “top five” includes a tie between cognitive ability tests and integrity tests in the fifth position. Of the “top five” predictors from Schmidt and Hunter (1998), all five remain in our current “top five plus a tie.” Empirically keyed biodata is in our top five, but not in Schmidt and Hunter’s. So, at a very high level, there remains considerable similarity between prior estimates and our estimates in terms of what rises to the top of a list of the strongest predictors. However, the magnitudes of the validity estimates differ considerably. The mean across Schmidt and Hunter’s top five was .49, while the mean across our top five is .37.

While structured interviews fared well in prior work, they do emerge in the present work as the strongest predictor of job performance. This suggests a reframing: while Schmidt and Hunter (1998) positioned cognitive ability as the focal predictor, with others evaluated in terms of their incremental validity over cognitive ability, one might propose structured interviews as the focal predictor against which others are evaluated. We call attention to the fact that the strongest predictors in our re-analysis (structured interviews, job knowledge tests, empirically keyed biodata, and work samples) are all job-specific measures. Several more “psychological” constructs emerge next in our ranking: cognitive ability, integrity tests, and personality-based measures of emotional intelligence. This suggests a closer behavioral match between predictor and criterion (Schmitt & Ostroff, 1986) as a contributor to strong predictive relationships. Measures on the “sample” side of the classic sign versus sample dichotomy tend to fare well (Wernimont & Campbell, 1968).

We do note that more work needs to be done. For example, some predictors (e.g., work samples, knowledge tests) require a selection scenario in which candidates are expected to have prior training and/or experience, while others (e.g., cognitive ability tests) can be used with untrained individuals who will acquire job-specific KSAs on the job or in post-hire training. Conceptually, a structured interview can be used in either setting (e.g., situational interviews for candidates without domain-specific knowledge/experience versus

behavior description interviews for candidates with such knowledge/experience). But it is not readily evident that one can expect comparable levels of validity in the two scenarios, as the content of a structured interview is likely to be quite different across the two. Work to date does not clearly differentiate these scenarios and doing so would be fruitful. We also note that in their traditional in-person form, structured interviews are generally not a viable strategy for high-volume jobs, with interviews commonly used among a smaller subset screened via less time-intensive predictors. As administration and scoring of interviews move toward reliance on technology and automation (e.g., Hickman et al., 2021), they are more amenable to mass administration; at the same time validity needs to be evaluated under these changing circumstances.

### ***Would Different Choices About Reliability and Range Restriction Correction Affect Study Conclusions?***

We noted that meta-analysts differ in the criterion reliability estimate they choose when using a mean value reported in the broader literature, rather than using a local artifact distribution, with some using .52 and some using .60. We used .60, and earlier offered our conceptual rationale for doing so. In Table 4 we reported mean operational validity estimates using .52 in addition to our estimates based on .60. Estimates are on average .02 higher with the .52 correction; the largest increase is .03. So, the choice of a reliability estimate has a discernable, but modest, effect on operational validity estimates.

In Table 4 we also offer validity estimates corrected for range restriction, in contrast to our choice not to correct in the absence of the needed data for an appropriate correction. Earlier we reviewed literature on intercorrelations among predictors, and concluded that the highest values obtained were in the .50 range, though most were far smaller. To estimate the highest plausible corrected validity obtainable under indirect range restriction, we paired .50 as the value of the correlation between the predictor of interest and the variable  $z$  that had been used for selection with a relatively extreme selection ratio of .05. As the table shows, estimates average .02 higher with this range restriction correction; the largest increase is .04. So, this hypothetical level of range restriction would also have a discernable but modest effect on operational validity estimates.

Again, we note that we do not view it as plausible that the meta-analytic data base for a given selection procedure is populated exclusively by studies with high  $r_{zx}$  values and extreme selection ratios. While our Table 4 alternate estimates use an  $r_{zx}$  value of .50 and a selection ratio of .05, we think it more plausible that the meta-analytic data base for a given predictor will be made up of studies with a wide range of selection ratios and  $r_{zx}$  values. Our best approximation may be to assume sampling from the range of  $r_{zx}$  from .00 to .50, and the full range of selection ratios shown in Table 1. Table 5 facilitates this approximation and its effects on operational validity. In Table 1, we showed the effects of indirect range restriction on restricted  $SD_x$ . Table 5 uses the same format as Table 1, except that it tables validity (i.e.,  $r_{xy}$ ), corrected for Case IV indirect range restriction. It shows, in essence, the consequences for validity of the degree of restriction documented in Table 1. Table 5 is built on panel 4 of Table 1, in that it assumes a reliability of .80 for the predictor of interest. It also assumes a criterion reliability value of .60. The table requires a mean observed validity; as an illustration we use .236, which is the mean observed validity for cognitive

**Table 5**  
*Operational Validity for Cognitive Ability Tests at Varying Levels of Indirect Range Restriction*

$r_{zx}$	Selection ratio						
	.90	.70	.50	.30	.10	.05	.01
0.1	.31	.31	.31	.31	.31	.31	.31
0.2	.31	.31	.31	.31	.31	.31	.31
0.3	.31	.31	.31	.32	.32	.32	.32
0.4	.31	.32	.32	.33	.33	.33	.33
0.5	.32	.33	.33	.34	.34	.34	.34
0.6	.32	.34	.35	.35	.36	.36	.37
0.7	.33	.35	.36	.38	.39	.39	.40
0.8	.33	.37	.39	.40	.43	.43	.44
0.9	.34	.39	.42	.45	.48	.48	.50

*Note.* Case IV correction based on observed validity = .236, which is the mean observed validity estimate for cognitive ability tests; predictor reliability = .80; criterion reliability = .60.

ability tests in our analysis. Correcting for unreliability in the criterion using a reliability estimate of .60 increases the correlation to .31; the table shows the degree to which correcting for indirect range restriction increases this value. As the table shows, for plausible  $r_{zx}$  values (e.g., .50 or smaller), even extreme selection ratios do not raise the correlation higher than .34. If we sample from the range of  $r_{zx}$  from .00 to .50, and the full range of selection ratios shown in Table 5, the average increase in operational validity due to indirect range restriction would be .01. In other words, indirect range restriction has small effects on validity in the conditions under consideration here.

The bottom line is that alternate corrections would not alter our conclusion that validity of many selection procedures has been overestimated. For example, our estimate of operational validity for cognitive ability tests is .31. The highest alternate value in Table 4, pairing alternate reliability estimates and an indirect range restriction correction, is .36. This is still substantially smaller than the .51 value offered by Schmidt and Hunter (1998).

### **Moving Beyond Mean Validity: Incorporating Variability in Validity and Subgroup Differences**

The Schmidt and Hunter (1998) summary of the validity of various predictors focused on mean validity, and on the increment each predictor provided over cognitive ability. Table 4 presents three additional pieces of information in addition to mean validity. The first is a residual standard deviation, obtained by correcting the observed standard deviation across studies for sampling error and any artifacts used in corrections (i.e., measurement error in the criterion, and sometimes range restriction). This can be viewed as our best current estimate of the degree to which validity varies across settings. It is an essential reminder that a given employer cannot count on the mean value as applicable to their organization. We also see value in identifying predictors with above and below average levels of this standard deviation of operational validity estimates. All else equal, the larger the estimate, the greater the need for work to identify the underlying causes of the variability across samples. Ideally, moderators can be identified, leading to more specific estimates of mean and variability for either specific sets of

jobs, for more nuanced variants of the predictor (analogous to the differentiation of structured and unstructured interviews, or ability-based and personality-based emotional intelligence), or for specific predictor design considerations (Lievens & Sackett, 2017).

The second piece of additional information in Table 3 is related to the first, namely the lower end of the 80% credibility interval around each mean validity estimate (i.e., the value above which 90% of operational validity values are expected to fall). This is readily computed: mean operational validity minus 1.28 times the residual standard deviation. This shows the implications of combining the mean and residual standard deviation. For example, while the structured interview has the highest mean operational validity (.42), it also has a large residual *SD* of .19. Thus, while the mean is high, there is also a higher risk of obtaining a lower value. One could rank the selection procedures by the low end of the credibility interval, rather than by the mean, as a risk-averse employer might prefer a selection predictor with a high value of this credibility value over a predictor with a high mean. In terms of the low end of the credibility interval, the five highest-ranked predictors are, in order, empirically keyed biodata, contextualized conscientiousness, job knowledge, work samples, and structured interviews. Four of these five are also in the top five based on mean validity; the lower credibility value ranking drops integrity tests and cognitive ability tests and adds contextualized conscientiousness.

The third piece of additional information in Table 3 is the Black–White subgroup mean difference for each selection predictor. Detailed information for other groups does not exist for many predictors, so we focus on Black–White mean differences. With two exceptions, the Black–White values in Table 3 are drawn from the summary of meta-analytic evidence across predictors provided by Dahlke and Sackett (2017). The value for GMA was drawn from Roth et al. (2011), as the value provided by Dahlke and Sackett is based on population samples, rather than job-specific samples. The value for job knowledge tests was drawn from Roth et al. (2003). Their value is based on incumbent data, as we are unaware of meta-analytic estimates of the Black–White mean difference on job knowledge for job applicants. We note that this incumbent value is likely at least somewhat affected by range restriction relative to applicant pools, and thus job applicant *d*-values are likely at least this large, and perhaps somewhat larger. Future research is needed examining the Black–White mean job knowledge difference in applicant pools to test this conjecture. Finally, we do not include estimates of Black–White differences for emotional intelligence, as we could not locate a meta-analytic summary.

There is much written about the validity-diversity tradeoff, that is, that a number of highly valid selection procedures also have substantial subgroup mean differences (Sackett et al., 2001). But Table 3 makes clear that our “top five” predictors include three with substantial group mean differences (work samples, job knowledge tests, and cognitive ability tests) and three with much smaller mean differences (structured interviews, biodata, and integrity tests). Using combinations of predictors, rather than single predictors, has long been advocated as a mechanism for reducing group differences; our findings reinforce the value of thinking broadly about a selection system with multiple components.

We are not calling for setting aside predictors just because they have large group differences. The choice of predictors follows from the needs of the organization. For example, work samples make good sense when one needs to identify individuals capable of

stepping into a job requiring training and experience. Similarly, job knowledge tests can be particularly useful when applicants must possess specific knowledge in order to be qualified for a job (e.g., electrician, mechanic). However, organizations should be aware that the use of such predictors can create more adverse impact than predictors with smaller subgroup differences. They can use the information in Table 3 to evaluate the magnitude of this validity-diversity tradeoff. For example, organizations interested in a quick and relatively inexpensive way to screen large numbers of applicants might consider biodata, cognitive ability tests, integrity tests, or personality-based emotional intelligence (or perhaps a combination of these). Prior to this review, meta-analytic evidence would have suggested that cognitive ability tests held a substantial validity advantage over these other predictors (Schmidt & Hunter, 1998), which might legitimately offset cognitive ability tests' increased potential for adverse impact. However, our review demonstrates all of these predictors have similar validity, but that cognitive ability tests still have much greater adverse impact potential. Organizations should find these sorts of comparisons in Table 3 useful.

### ***Implications for Future Meta-Analysts***

A primary message of this work is the need to carefully identify range restriction scenarios and to apply different corrections to different subsets of studies as appropriate. We explained and showed here that the common practice of building an artifact distribution and applying it to all studies is flawed in most settings. Settings in which the predictor in question is used in the selection of applicants can readily produce sizable  $U$  ratios leading to substantial corrections. Settings in which the predictor in question is not used in selection (such as concurrent studies) can only produce sizeable  $U$  ratios in rare situations. Yet many meta-analyses do not categorize studies in this way (predictor used vs. predictor not used). Doing so is essential.

The above paragraph addressed how a meta-analyst should apply an artifact distribution. We also call out implications for determining when to include a  $U$  value in an artifact distribution (see Figure 1). First, we advocate an end to the practice of simply assuming a degree of restriction with no empirical basis. Second, we advocate extreme caution in the use of hiring rates as the basis for estimating a  $U$  ratio. An overall hiring rate likely reflects the use of multiple pieces of information, and assuming that overall selectivity reflects the degree of restriction on the predictor of interest to the researcher is rarely warranted. Absent information on the specific role of a predictor in the hiring process, we recommend against using an overall hiring rate as the basis for a  $U$  ratio. If the selection rate for the predictor of interest is known (e.g., in the case of multi-stage selection with an initial screening in of the top 50% on an ability test, followed by a subsequent screening of the top 50% of the remaining pool on an integrity test), the predictor-specific selection rate can be used to estimate a  $U$  ratio. Third, we advocate caution in relying on published norms as the estimate of the unrestricted predictor standard deviation. Doing so requires a thoughtful evaluation of the degree to which the published norms reflect the selection scenario at hand. For example, for some tests norms are presented for working adult samples; for others norms are presented for specific occupational groups. We are skeptical of the first, and more open to the second if there is a close match between the norm group and the job and applicant pool at hand.

We reiterate our principle of conservative estimation: if one is not confident in the basis for a range restriction correction, it is better to forego a correction than to use a value that results in an overestimate. We suggest presenting the value obtained without the correction, and noting that as some degree of restriction is likely, the presented value is a conservative estimate.

### **Cautionary Notes and Future Research Directions**

We turn to a treatment of a series of issues that we view as important in understanding our findings and putting them in perspective. These include a number of cautionary notes about issues that still remain unresolved in the field and therefore should prompt future research.

#### ***Is the Overall Job Performance Criterion Comparable Across Studies?***

An implicit assumption in the field is that the overall criterion of "supervisor ratings of job performance" reflects the same things across studies. Only with this assumption can one meaningfully compare findings across studies of a given selection predictor, or compare meta-analytic findings across predictors. But either by design or by happenstance raters can be directed to focus on different things. If raters perceive the request for an evaluation of performance to reflect primarily task performance, they will give a different evaluation than if asked to evaluate the citizenship or counterproductivity components of performance.

We offer two useful illustrations. First, Gonzalez-Mulé et al. (2014) report a meta-analytic summary of the relationship between general cognitive ability and task performance, organizational citizenship behavior, and counterproductive work behavior (CWB), showing strong relationships with task performance, weaker with citizenship, and near-zero with CWB. We note that rating measures can differ in the degree to which they emphasize each of these. Second, Sackett et al. (2017) compared the criterion measures used in meta-analyses of cognitive ability and assessment centers, finding a near-total criterion focus on task performance in the data base on cognitive ability, and a much broader criterion focus for assessment centers. They then identified a set of studies in which both a cognitive ability test and an assessment center were administered to the same individuals, with both measures correlated with a common performance measure. They termed these "head-to-head comparisons," and found that while the separate meta-analyses of the two domains produced higher range restriction-corrected validity for ability (mean validity = .51) than for assessment centers (mean validity = .37), the finding was reversed in the head-to-head comparisons (.44 for assessment centers vs. .22 for ability). These findings suggest caution in comparing findings across meta-analyses, as we do not have clear understanding of the specific components underlying performance ratings. This could be a fruitful avenue for future inquiry.

#### ***Do Predictive and Concurrent Designs Actually Estimate the Same Thing?***

Most of the meta-analyses reviewed here included both predictive and concurrent studies. Only a small number treat predictive versus concurrent as a moderator (i.e., report findings separately for the

two categories). Not differentiating between these study designs reflects an implicit assumption that both estimate a common true operational validity. In fact, it is possible that in different settings each of these designs may contribute systematic bias to the validity estimate.

For noncognitive predictors, a common concern is faking. The incentives to present oneself in a socially desirable manner are higher in an operational selection setting than in a research setting, and thus there may be systematic differences between predictive studies in operational settings and concurrent studies in research settings. While lower validity would be expected in settings where faking is more prevalent, the question of interest is the operational validity of a predictor, and thus the validity estimate in predictive settings would provide the better answer to the research question. The Ones et al. (1993) meta-analysis of integrity test validity reflects this concern. Although Ones et al. reported validity estimates separately for predictive and concurrent studies, they used only predictive studies as the basis for their estimate of operational validity.

For cognitive predictors, a common concern is level of effort on the part of the test-taker. An implicit assumption is that applicants are motivated to give their best effort in pursuit of a job of interest; in contrast, individuals asked to take a test for research purposes may exhibit less effort when confronting difficult test questions. The issue of study design has been disputed in the domain of cognitive ability testing, with Barrett et al. (1981) arguing that the designs produce comparable findings and Guion and Cranny (1982) countering that there are important design variants that affect the comparison. For example, predictive designs where the predictor of interest is used in selection decisions can be expected to be less comparable to concurrent designs than predictive designs where the predictor of interest is not used.

In addition to level of effort and social desirability issues, the use of a concurrent design carries the implicit assumption that experience on the job does not affect an individual's location in the score distribution. If experience varies across current employees and/or affects scores differentially, the validity obtained in the concurrent setting may not accurately estimate that obtained in the setting of real interest, namely, a predictive setting with job applicants. Thus, there is further reason to differentiate between predictive and concurrent studies in building meta-analytic data bases, namely, to shed light on whether findings differ across these settings.

Thus, we note here that with limited exceptions the literature summarized here does not carefully differentiate findings from predictive and concurrent studies. Our revisiting of the meta-analytic database does differentiate between the two in terms of differences in the degree of range restriction, but we do include both predictive and concurrent studies in our estimates of operational validity. Thus, our treatment also treats both types of studies as estimates of a common operational validity.

### ***What About Restriction on the Criterion?***

One potential objection to our argument that range restriction will be minimal in most concurrent validity settings is that we focus solely on restriction on the predictor. Should one not also consider restriction on the criterion? While this is a potential issue in all studies, it is arguably of greater concern in concurrent studies. In such studies, incumbents tend to have been on the job for longer

periods of time, and over time high performers may have been promoted out of the job (or left the organization for a higher-level job), and low performers may have been dismissed. The result would be restriction on the criterion, and thus reduced validity estimates (cf. Sackett et al., 2002).

Restriction on the criterion has generally not been addressed in the meta-analytic literature we review here. The issue was been raised in recent work by Huffcutt (2020), who modeled the effects of losing the top 5% of performers on the criterion measure to promotion and the bottom 5% to dismissal. However, we posit that while promotion and dismissal certainly take place, it is unrealistic to model this as taking place solely on the basis of the measured criterion variable. Rather, we suggest that a variety of different mechanisms contribute to the loss of individuals from the validation sample. Some new hires are let go because of persistent attendance problems; others have difficulty interacting with co-workers; others have difficulty mastering needed job tasks. Higher performers are also likely to be promoted on grounds other than the measured criterion in the validation study. Perceived leadership attributes, for example, are likely to influence promotion to a greater degree than they influence overall ratings of performance in an individual contributor role. The actual bases for promotion and dismissal decisions likely have nonzero correlations with the measured criterion used in the validation study, and the result will be indirect restriction on the criterion rather than the direct restriction modeled by Huffcutt. In a prior section of this paper, we have shown that indirect restriction on the predictor side will have small effects on validity unless the actual selection variable is highly correlated with the focal predictor; this same argument and the same mathematics apply on the criterion side. Recall also that we showed that indirect restriction effects are even smaller when the focal predictor is measured with less than perfect reliability. This phenomenon also applies on the criterion side, as the measured criterion typically has quite modest reliability; the corrections we make in this paper use an interrater reliability value of .60. Thus, while restriction on the criterion side is likely to be nonzero, we expect the effects to be very small.

### ***The Case IV Approach to Correcting for Restriction***

There is nothing in this paper that challenges the conceptual or mathematical underpinnings of the Case IV approach proposed by Hunter et al. (2006). As noted earlier, that approach makes use of the insight that when selection is on a third variable,  $z$ , the resulting indirect restriction on  $x$  takes place not on the measured variable  $x$ , but on the latent construct underlying measured  $x$ . By definition,  $z$  must be uncorrelated with the random measurement portion of measured  $x$  (i.e., nothing can correlate with a random variable). Thus, the relationship between  $z$  and latent  $x$  is larger than the relationship between  $z$  and measured  $x$ . The result is twofold: (a) the larger the amount of measurement error in  $x$ , the harder it is to get a small  $u_x$  ratio, and (b) with a given  $u_x$  ratio, range restriction-corrected  $r_{xy}$  will be larger when measurement error in  $x$  is larger. Thus Hunter et al.'s Case IV correction approach produces a larger corrected validity for a given  $u_x$  value than prior correction methods. There is a growing body of research revisiting prior meta-analyses and reporting higher validity for various predictors when applying the Case IV correction (e.g., Huffcutt et al., 2014; Le et al., 2016). Those re-analyses have the same problems that we highlighted in

this paper, namely, that the mean  $u_x$  ratios used as artifact distributions do not accurately reflect the actual distribution of restriction in the meta-analytic data base. Thus, just as this paper has argued that meta-analytic estimates of selection method validity have been overinflated, the same concerns apply to newer estimates using Case IV corrections. Again, the issue is not about the Case IV approach, but the  $u_x$  ratios used in applications of this approach.

### ***New Estimates of Validity Yield a Coherent Pattern***

A key question is whether the new validity estimates we report here make sense. To us, they do. Here is a reality check: Schmidt and Hunter's (1998) compilation identified general ability tests and work sample tests as among our best predictors. The field has internalized their findings as "general ability is conceptually appropriate when hiring untrained workers for entry jobs; work samples are conceptually appropriate when hiring experienced workers, with mean corrected validity about the same." That made sense. But in an unpublished update to Schmidt and Hunter's (1998) paper, Schmidt et al. (2016) drop the work sample validity estimate from .54 to .33 based on a new meta-analysis by Roth et al. (2005). Roth et al.'s sample contained 98% concurrent studies, so they made no correction for range restriction. So, has the state of affairs now changed and work samples are now markedly inferior? We argue no: our new ability validity estimate of .31 fits with the .33 for work samples of Schmidt et al. (2016). So, there is a coherent pattern.

### ***Our Estimates Are Based on What We Know Now***

Some prior meta-analyses could be revisited in light of the ideas developed in this paper and more refined estimates provided (e.g., meta-analysts who did not differentiate between predictive and concurrent studies in the original published meta-analysis may have the information in their files to identify studies using each strategy and correct each as appropriate). Thus, we do not put a stake in the ground regarding the meta-analytic estimates we put forward here; rather, we eagerly await the availability of the data that will permit further refinement of our estimates.

## **References**

- Asher, J. J., & Sciarino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology*, 27(4), 519–533. <https://doi.org/10.1111/j.1744-6570.1974.tb01173.x>
- Barrett, G. V., Phillips, J. S., & Alexander, R. A. (1981). Concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, 66(1), 1–6. <https://doi.org/10.1037/0021-9010.66.1.1>
- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, 9(1–2), 9–30. <https://doi.org/10.1111/1468-2389.00160>
- Bemis, S. E. (1968). Occupational validity of the General Aptitude Test Battery. *Journal of Applied Psychology*, 52(3), 240–244. <https://doi.org/10.1037/h0025733>
- Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview–cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology*, 60(4), 837–874. <https://doi.org/10.1111/j.1744-6570.2007.00093.x>
- Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, 78(3), 387–409. <https://doi.org/10.1348/096317905X26994>
- Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, 52(3), 561–589. <https://doi.org/10.1111/j.1744-6570.1999.tb00172.x>
- Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, 10(4), 331–360. [https://doi.org/10.1207/s15327043hup1004\\_2](https://doi.org/10.1207/s15327043hup1004_2)
- Dahlke, J. A., & Sackett, P. R. (2017). The relationship between cognitive-ability saturation and subgroup mean differences across predictors of job performance. *Journal of Applied Psychology*, 102(10), 1403–1420. <https://doi.org/10.1037/apl0000234>
- Dunnette, M. D. (1972). *Validity study results for jobs relevant to the petroleum refining industry*. American Petroleum Institute.
- Dye, D. A., Reck, M., & McDaniel, M. A. (1993). The validity of job knowledge measures. *International Journal of Selection and Assessment*, 1(3), 153–157. <https://doi.org/10.1111/j.1468-2389.1993.tb00103.x>
- Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, 72(3), 493–511. <https://doi.org/10.1037/0021-9010.72.3.493>
- Gonzalez-Mulé, E., Mount, M. K., & Oh, I.-S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology*, 99(6), 1222–1243. <https://doi.org/10.1037/a0037547>
- Guion, R. M., & Cranny, C. J. (1982). A note on concurrent and predictive validity designs: A critical reanalysis. *Journal of Applied Psychology*, 67(2), 239–244. <https://doi.org/10.1037/0021-9010.67.2.239>
- Hardison, C. M., & Sackett, P. R. (2006). Kriteriumsbezogene validität des assessment centers: Lebendig und wohlauf? (Assessment center criterion-related validity: Alive and well?). In H. Schuler (Ed.), *Assessment Center als Methode der Personalentwicklung (Assessment center as a method of personnel development)* (pp. 192–202). Hogrefe.
- Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. National Research Council.
- Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centers for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment*, 15(4), 405–411. <https://doi.org/10.1111/j.1468-2389.2007.00399.x>
- Hickman, L., Bosc, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*. Advance online publication. <https://doi.org/10.1037/apl0000695>
- Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology*, 100, 1143–1168. <https://doi.org/10.1037/a0038707>
- Huffcutt, A. I. (2020). Range restriction in employment interviews. In E. F. Stone-Romero & P. J. Rosopa (Eds.), *Research methods in human research management: Toward valid research-based inferences* (pp. 173–196). Information Age Publishing.
- Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology: Employment interview validity. *International Journal of Selection and Assessment*, 22(3), 297–309. <https://doi.org/10.1111/ijsa.12078>
- Hunter, J. (1983). *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the GATB*. US Employment Service, US Department of Labor.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96(1), 72–98. <https://doi.org/10.1037/0033-2909.96.1.72>

- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage Publications.
- Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology, 91*(3), 594–612. <https://doi.org/10.1037/0021-9010.91.3.594>
- James, L. R., Demaree, R. G., Mulaik, S. A., & Ladd, R. T. (1992). Validity generalization in the context of situational models. *Journal of Applied Psychology, 77*, 3–14. <https://doi.org/10.1037/0021-9010.77.1.3>
- Joseph, D. L., Jin, J., Newman, D. A., & O'Boyle, E. H. (2015). Why does self-reported emotional intelligence predict job performance? A meta-analytic investigation of mixed EI. *Journal of Applied Psychology, 100*(2), 298–342. <https://doi.org/10.1037/a0037681>
- Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology, 95*(1), 54–78. <https://doi.org/10.1037/a0017286>
- Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*(6), 875–925. <https://doi.org/10.1037/a0033901>
- Kane, J. S., & Lawler, E. E. (1978). Methods of peer assessment. *Psychological Bulletin, 85*(3), 555–586. <https://doi.org/10.1037/0033-2909.85.3.555>
- Le, H., Oh, I.-S., Schmidt, F. L., & Wooldridge, C. D. (2016). Correction for range restriction in meta-analysis revisited: Improvements and implications for organizational research. *Personnel Psychology, 69*(4), 975–1008. <https://doi.org/10.1111/peps.12122>
- Lievens, F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology, 102*(1), 43–66. <https://doi.org/10.1037/apl0000160>
- Linn, R. L., Harnish, D. L., & Dunbar, S. B. (1981). Corrections for range restriction: An empirical investigation of conditions resulting in conservative corrections. *Journal of Applied Psychology, 66*, 655–663. <https://doi.org/10.1037/0021-9010.66.6.655>
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63–91. <https://doi.org/10.1111/j.1744-6570.2007.00065.x>
- McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*(4), 599–616. <https://doi.org/10.1037/0021-9010.79.4.599>
- Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042–1052. <https://doi.org/10.1037/0021-9010.93.5.1042>
- Murphy, K. R. (1986). When your top choice turns you down: Effect of rejected offers on the utility of selection tests. *Psychological Bulletin, 99*(1), 133–138. <https://doi.org/10.1037/0033-2909.99.1.133>
- Murphy, K. R., & De Shon, R. (2000). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology, 53*(4), 913–924. <https://doi.org/10.1111/j.1744-6570.2000.tb02423.x>
- Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). Vocational interests and performance: A quantitative summary of over 60 years of research. *Perspectives on Psychological Science, 7*(4), 384–403. <https://doi.org/10.1177/1745691612449021>
- Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2017). Interest congruence and performance: Revisiting recent meta-analytic findings. *Journal of Vocational Behavior, 98*, 138–151. <https://doi.org/10.1016/j.jvb.2016.11.002>
- O'Boyle, E. H., Jr., Humphrey, R. H., Pollack, J. M., Hawver, T. H., & Story, P. A. (2011). The relation between emotional intelligence and job performance: A meta-analysis. *Journal of Organizational Behavior, 32*(5), 788–818. <https://doi.org/10.1002/job.714>
- Ones, D. S., & Viswesvaran, C. (2003). Job-specific applicant pools and national norms for personality scales: Implications for range-restriction corrections in validation research. *Journal of Applied Psychology, 88*(3), 570–577. <https://doi.org/10.1037/0021-9010.88.3.570>
- Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78*(4), 679–703. <https://doi.org/10.1037/0021-9010.78.4.679>
- Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65*(4), 373–406. <https://doi.org/10.1037/0021-9010.65.4.373>
- Pearson, K. (1903). I. Mathematical contributions to the theory of evolution—XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 200*(321–330), 1–66. <https://doi.org/10.1098/rsta.1903.0001>
- Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*(4), 1009–1037. <https://doi.org/10.1111/j.1744-6570.2005.00714.x>
- Roth, P. L., Huffcutt, A. I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88*(4), 694–706. <https://doi.org/10.1037/0021-9010.88.4.694>
- Roth, P. L., Switzer, F. S., III, Van Iddekinge, C. H., & Oh, I.-S. (2011). Toward better meta-analytic matrices: How input values can affect research conclusions in human resource management simulations. *Personnel Psychology, 64*(4), 899–935. <https://doi.org/10.1111/j.1744-6570.2011.01231.x>
- Roth, P. L., Van Iddekinge, C. H., DeOrtentiis, P. S., Hackney, K. J., Zhang, L., & Buster, M. A. (2017). Hispanic and Asian performance on selection procedures: A narrative and meta-analytic review of 12 common predictors. *Journal of Applied Psychology, 102*(8), 1178–1202. <https://doi.org/10.1037/apl0000195>
- Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology, 75*(2), 175–184. <https://doi.org/10.1037/0021-9010.75.2.175>
- Sackett, P. R., Laczko, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology, 55*(4), 807–825. <https://doi.org/10.1111/j.1744-6570.2002.tb00130.x>
- Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology, 92*(2), 538–544. <https://doi.org/10.1037/0021-9010.92.2.538>
- Sackett, P. R., & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology, 79*(5), 680–684. <https://doi.org/10.1037/0021-9010.79.5.680>
- Sackett, P. R., & Schmitt, N. (2012). On reconciling conflicting meta-analytic findings regarding integrity test validity. *Journal of Applied Psychology, 97*(3), 550–556. <https://doi.org/10.1037/a0028167>
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-Stakes testing in employment, credentialing, and higher education. Prospects in a post-affirmative-action world. *American Psychologist, 56*(4), 302–318. <https://doi.org/10.1037/0003-066X.56.4.302>
- Sackett, P. R., Sharpe, M. S., & Kuncel, N. R. (2021). Relative reliance on tests versus Grades in college admissions. *Applied Measurement in*

- Education*. Advance online publication. <https://doi.org/10.1080/08957347.2021.1987903>
- Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology, 102*(10), 1435–1447. <https://doi.org/10.1037/apl0000236>
- Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49*(11), 929–954. <https://doi.org/10.1037/0003-066X.49.11.929>
- Sackett, P. R., Zhang, C., & Berry, C. M. (2021). Challenging conclusions about predictive bias against Hispanic test-takers in personnel selection. *Journal of Applied Psychology*. Advance on-line publication.
- Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*(3), 323–346. <https://doi.org/10.1348/096317903769647201>
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European Community meta-analysis. *Personnel Psychology, 56*(3), 573–605. <https://doi.org/10.1111/j.1744-6570.2003.tb00751.x>
- Salgado, J. F., Anderson, N., & Tauriz, G. (2014). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology, 88*(4), 797–834. <https://doi.org/10.1111/joop.12098>
- Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology, 11*(3), 299–324. <https://doi.org/10.1080/13594320244000184>
- Schmidt, F. L., Caplan, J. R., Bemis, S. E., Decuir, R., Dunn, L., & Antone, L. (1979). *The behavioral consistency method of unassembled examining*. US Office of Personnel Management, Personnel Resources and Development Center.
- Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Journal of Applied Psychology, 124*(2), 262–274. <https://doi.org/10.1037/0033-2909.124.2.262>
- Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61*(4), 473–485. <https://doi.org/10.1037/0021-9010.61.4.473>
- Schmidt, Frank L., In-Sue, O., & Shaffer, J. A. (2016). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years* [Working paper].
- Schmitt, N., & Ostroff, C. (1986). Operationalizing the “behavioral consistency” approach: Selection test development based on a content-oriented strategy. *Personnel Psychology, 39*(1), 91–108. <https://doi.org/10.1111/j.1744-6570.1986.tb00576.x>
- Schneider, B. (1987). The people make the place. *Personnel Psychology, 40*, 437–453. <https://doi.org/10.1111/j.1744-6570.1987.tb00609.x>
- Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*(3), 445–494. <https://doi.org/10.1111/j.1744-6570.2012.01250.x>
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. <https://doi.org/10.2307/1412159>
- Speer, A., Sendra, C., & Shihadeh, M. (2021). *Meta-analysis of biodata in employment settings: Providing clarity to criterion and construct-related validity estimates*. The 36th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA, United States.
- Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75*(3), 277–294. <https://doi.org/10.1348/096317902320369712>
- Van Iddekinge, C. H., Aguinis, H., Mackey, J. D., & DeOrtentiis, P. S. (2018). A meta-analysis of the interactive, additive, and relative effects of cognitive ability and motivation on performance. *Journal of Management, 44*(1), 249–279. <https://doi.org/10.1177/0149206317702220>
- Van Iddekinge, C. H., Arnold, J. D., Frieder, R. E., & Roth, P. L. (2019). A meta-analysis of the criterion-related validity of prehire work experience. *Personnel Psychology, 72*(4), 571–598. <https://doi.org/10.1111/peps.12335>
- Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology, 96*(6), 1167–1194. <https://doi.org/10.1037/a0024343>
- Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology, 97*(3), 499–530. <https://doi.org/10.1037/a0021196>
- Van Rooy, D. L., & Viswesvaran, C. (2004). Emotional intelligence: A meta-analytic investigation of predictive validity and nomological net. *Journal of Vocational Behavior, 65*(1), 71–95. [https://doi.org/10.1016/S0001-8791\(03\)00076-9](https://doi.org/10.1016/S0001-8791(03)00076-9)
- Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557–574. <https://doi.org/10.1037/0021-9010.81.5.557>
- Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372–376. <https://doi.org/10.1037/h0026244>
- Wiesner, W. H., & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*(4), 275–290. <https://doi.org/10.1111/j.2044-8325.1988.tb00467.x>

Received June 13, 2021

Revision received August 30, 2021

Accepted October 13, 2021 ■