Industrial and Organizational Psychology (2023), **16**, 283–300
doi:10.1017/iop.2023.24

CAMBRIDGE UNIVERSITY PRESS

FOCAL ARTICLE

Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors

Paul R. Sackett[1]*, Charlene Zhang[1], Christopher M. Berry[2], and Filip Lievens[3]

[1]University of Minnesota, Minneapolis, MN, USA, [2]Indiana University, Bloomington, IN, USA and [3]Singapore Management University, Singapore
*Corresponding author: Email: psackett@umn.edu

(Received 14 September 2022; revised 12 January 2023; accepted 13 January 2023; first published online 09 May 2023)

Abstract

Sackett et al. (2022) identified previously unnoticed flaws in the way range restriction corrections have been applied in prior meta-analyses of personnel selection tools. They offered revised estimates of operational validity, which are often quite different from the prior estimates. The present paper attempts to draw out the applied implications of that work. We aim to (a) present a conceptual overview of the critique of prior approaches to correction, (b) outline the implications of this new perspective for the relative validity of different predictors and for the tradeoff between validity and diversity in selection system design, (c) highlight the need to attend to variability in meta-analytic validity estimates, rather than just the mean, (d) summarize reactions encountered to date to Sackett et al., and (e) offer a series of recommendations regarding how to go about correcting validity estimates for unreliability in the criterion and for range restriction in applied work.

Keywords: selection; meta-analysis; validity

In a recent paper, we (Sackett et al., 2022) identified previously unnoticed flaws in the way range restriction corrections have been applied in prior meta-analyses of personnel selection tools. We offered revised estimates of operational validity (i.e., the level of validity expected after correcting for unreliability in the criterion and for range restriction, where appropriate), which are often quite different from the prior estimates. It is a quite technical article, and we believe that there are implications for practice that space considerations prevented us to fully develop in that article. The present article is therefore our attempt to more comprehensively draw out the applied implications of that work. We intend this as free-standing: it does not assume that readers have read the initial article, although interested readers are encouraged to do so.

In this article, we aim to do five things. First, we present a conceptual overview of our critique of prior approaches to correction. Second, we outline the implications of this new perspective for the relative validity of different predictors and for the tradeoff between validity and diversity in selection system design. Third, we highlight the need to attend to variability in meta-analytic validity estimates, rather than just the mean. Fourth, we summarize reactions encountered to date to our work. Finally, we offer a series of recommendations regarding how to go about correcting validity estimates for unreliability in the criterion and for range restriction in applied work.

Meta-analytic work on the validity of selection procedures: Past and present

In 1998, Schmidt and Hunter published a summary of meta-analytic findings for the relationship between a wide variety of predictors used for personnel selection and overall job performance.

This has been widely cited—over 6,500 citations as of this writing. One important feature was the focus on cognitive ability as a central predictor; for each predictor, they tabled its incremental validity over cognitive ability. Another important feature is that Schmidt and Hunter presented only validity values corrected for measurement error in the criterion and for restriction of range.

In Sackett et al. (2022), we updated this summary. Our work differed in a number of ways. First, we presented both observed and corrected validity values, thus making clear the effect of corrections on validity. Second, we added a variety of additional predictors that were not included in Schmidt and Hunter (1998), such as the Big Five personality traits, emotional intelligence, and situational judgment tests (SJTs). Third, we reported meta-analytic standard deviations (i.e., between-study variability in corrected validities) as well as means, thereby emphasizing that considerable variability is found across settings for each predictor. Fourth, we also reported White–Black mean differences for each predictor, reflecting the commonly noted joint concern for job performance and diversity as desirable outcomes of selection system use.

We first show the key finding of Sackett et al. (2022), and then outline how and why our findings differed considerably from Schmidt and Hunter's (1998) results. Table 1 contains the key information from Table 3 of Sackett et al. (2022), with some updates that are discussed below, presenting the Schmidt and Hunter and Sackett et al. estimates side by side. Figure 1 is a reproduction of Figure 2 from Sackett et al., showing the validity findings graphically in conjunction with information about White-Black mean differences. One overall conclusion is that our widely used predictors, on average, exhibit lower operational criterion-related validity with regard to overall performance than has previously been reported. For some predictors, the difference was quite substantial (e.g., .21 lower for work sample tests, .20 lower for cognitive ability, .19 lower for unstructured interviews). A second conclusion is that the relative predictive power for various predictors changed considerably. Cognitive ability is no longer the stand-out predictor that it was in the prior work. Structured interviews emerged as the predictor with the highest mean validity.

So how did this happen? Several factors contributed. In some cases, newer data for a predictor led to new conclusions. In other cases, the change in conclusions is a result of us challenging the range restriction corrections that had a large effect on the estimates of operational validity. Range restriction is present when the sample used to compute a validity coefficient is narrower in range (technically, has a smaller standard deviation) than in the applicant pool for which one proposes to use the predictor. Validity is lower in restricted than in unrestricted samples, and corrections for range restriction are commonly made to estimate validity in unrestricted samples. Sackett et al.'s (2022) key contribution was the documentation that the procedures long and widely used for range restriction corrections in meta-analysis were often inappropriate. Sackett et al. is a detailed and technical paper; here we present only a conceptual overview of the issues. To set the stage, the rarely attained ideal in meta-analysis is to individually correct each effect size estimate (here, an observed validity coefficient) using a sample-specific estimate of the degree of criterion unreliability and range restriction present in the sample. Typically, estimates of criterion unreliability and range restriction are available only for a small subset of the samples collected for meta-analysis. The general solution has been to compute the mean level of criterion unreliability in samples reporting it, the mean level of range restriction in samples reporting it, and then apply the mean criterion unreliability and range restriction as correction factors to the mean observed validity. Sackett et al. argued that this makes good sense if the set of studies providing criterion unreliability and range restriction estimates are a random or representative sample of the full set of studies in the meta-analysis. However, Sackett et al. made the point that in reality the samples providing range restriction information usually are not random or representative, and we built on that fact to challenge typical range restriction correction practice; and, in particular, to show that the typical practice has generally resulted in substantial overcorrections for range restriction.

Four key insights come together to drive the overcorrection for range restriction: (a) correction factors generally come from predictive validity studies (i.e., studies wherein the predictor of

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

Table 1. Comparison of Schmidt and Hunter's (1998) and Sackett et al.'s (2022) validity estimates with updates

Predictor	Schmidt & Hunter (1998) validity estimate	Sackett et al. (2022)				Updated validity estimates since Sackett et al. (2022)
		Validity estimate (ρ)	*SD* of ρ	Lower 80% credibility value	B-W d	
Employment interviews (structured)	0.51	0.42	0.19	0.18	0.23	
Job knowledge tests	0.48	0.40	0.13	0.23	*0.54*	
Empirically keyed biodata	0.35	0.38	0.09	0.26	*0.33*	
Work sample tests	0.54	0.33	0.09	0.21	0.67	
GMA tests	0.51	0.31	0.14	0.13	0.79	0.23
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	0.33
SJT (knowledge)		0.26	0.10	0.13	*0.39*	
SJT (behavioral tendency)		0.26	0.12	0.11	*0.34*	
Conscientiousness – contextualized		0.25	0.00	0.25	*−0.07*	
Interests	0.10	0.24	0.25	−0.08	*0.33*	
Emotional Stability – contextualized		0.23	0.10	0.10	*0.09*	
Ability-based EI		0.22	0.05	0.16		
Rationally keyed biodata		0.22	0.06	0.14	*0.33*	
Extraversion – contextualized		0.21	0.08	0.11	*0.16*	
Conscientiousness- overall	0.31	0.21	0.15	0.02	*−0.07*	
Employment interviews (unstructured)	0.38	0.19	0.16	−0.01	0.32	
Agreeableness – contextualized		0.19	0.13	0.02	*0.03*	
Openness to Experience – contextualized		0.12	0.00	0.12	*0.01*	
Extraversion – overall		0.11	0.13	−0.06	*0.16*	
Agreeableness – overall		0.1	0.14	−0.08	*0.03*	
Emotional Stability – overall		0.09	0.08	−0.01	*0.09*	
Job experience (years)	0.18	0.07	0.11	−0.07	*0.49*	
Openness to Experience – overall		0.06	0.07	−0.03	*0.10*	

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.

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

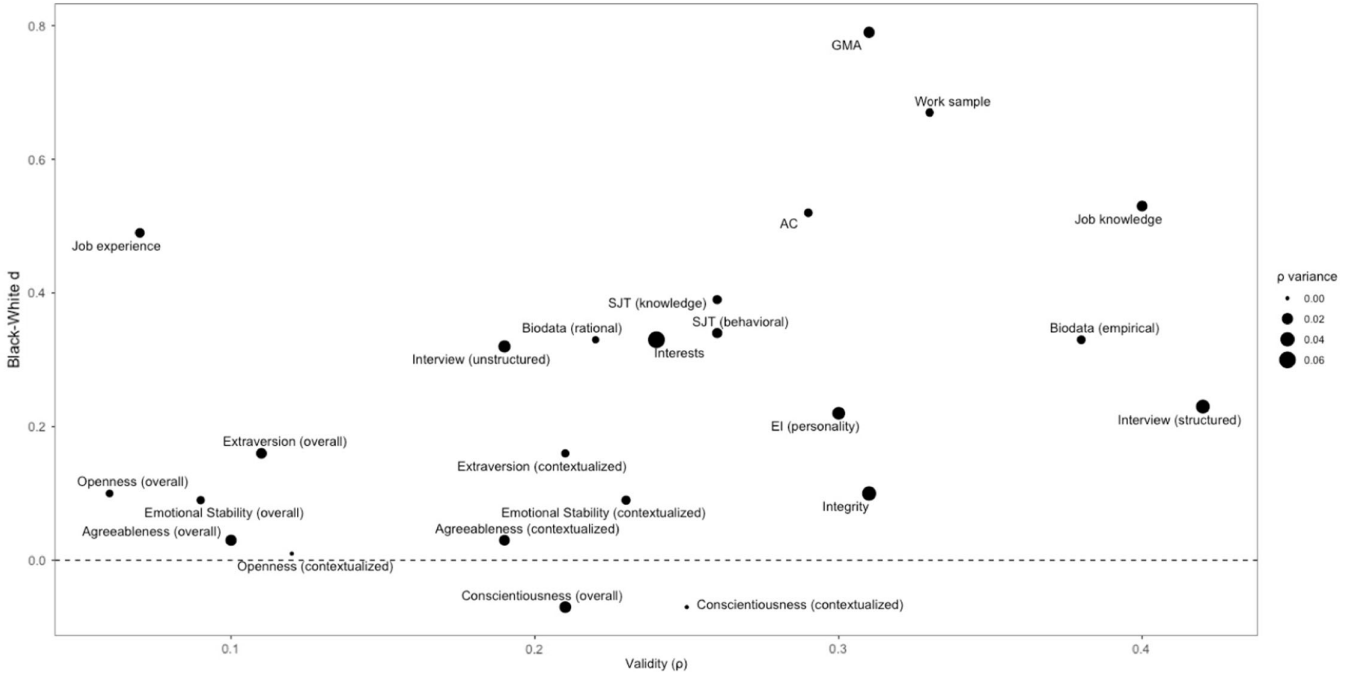

Figure 1. *A Visual Summary of Common Selection Procedures' Validity, Validity Variance, and Black–White d.*
Note. GMA = General Mental Ability, AC = Assessment Center, SJT = Situational Judgment Test, EI = Emotional Intelligence.

interest is administered atone point in time, usually to applicant samples, with criterion data collected at a later time), as the correction factors are based on comparing applicant and incumbent predictor standard deviations, with applicant data unavailable in concurrent validity studies (i.e., studies wherein the predictor of interest is administered to current employees); (b) the degree of range restriction can be quite substantial in predictive studies; however, Sackett et al. demonstrated that restriction will commonly (though not always) be quite small in concurrent studies; (c) applying a large correction factor derived from predictive studies to concurrent studies thus results in an overcorrection, which is often quite large; and (d) across meta-analyses about 80% of studies are concurrent. Thus, importantly, we do not question range restriction formulas; rather, we note that they have systematically been applied in ways which are not appropriate.

Thus, we argued against applying a *uniform* correction to all studies in a meta-analysis. We suggested dividing the set of studies available for meta-analysis into groups, based on differing range restriction mechanisms. For example, one may have both applicant and incumbent predictor standard deviations for the subset of predictive studies, and can therefore apply an appropriate correction to that subset of studies. However, for concurrent studies, one would usually not have applicant standard deviations because these are current employees who would almost never have been hired using the predictor of interest in the concurrent validity study. So, any attempt at correction would have to be based on knowledge of the correlation between the predictor of interest in the concurrent validity study and the predictor or predictor composite originally used to select the employees (i.e., there would only be concerns about substantial range restriction if the predictor of interest is substantially correlated with the predictor or predictor composite originally used to select the employees). We note that this correlation is generally not known. Even if it is known that, say, a particular test was used as part of the original selection process, in most cases we believe that selection processes involve multiple sources of information, many of them unquantified (e.g., hiring manager subjective judgement), and thus the correlation between the predictor of interest and the entire original basis for selection will be unknown. Further, we demonstrated that unless the current employees were hired solely based on the predictor of interest in the concurrent validity study (which would virtually never be the case), the amount of range restriction will be small in most concurrent validity studies. Therefore, in such cases, we proposed what we call the "principle of conservative estimation": absent a sound basis for estimating the degree of range restriction, it is better to not attempt a correction, especially when the amount of range restriction is likely to be small anyway and the potential for substantial overcorrection (due to reliance on range restriction estimates from predictive studies) is large. Thus, we did not make range restriction corrections in settings where we did not see a clear basis for estimating the degree of restriction (e.g., in meta-analyses that do not differentiate predictive and concurrent studies).

In the present paper, we discuss implications for I-O practice of this revised view of the personnel selection space. An overly simplistic takeaway is that Sackett et al. (2022) showed that validity in general is lower than we had thought, which may be viewed as an unwelcome finding. We believe the implications are far broader. That is, as noted below, they change our view of the relative value of different predictors, of managing the long-discussed validity–diversity tradeoff, and of the role of cognitive ability in selection systems.

What selection procedures work best?

Our findings indicate that the predictors at the top of our list in terms of criterion-related validity are those specific to individual jobs, such as structured interviews, job knowledge tests, work sample tests, and empirically keyed biodata. These tend to fare better than more general measures of psychological constructs, such as measures in the ability and personality domain. This is important for a variety of reasons.

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

First, it supports the "sample" end of the sign-sample dichotomy (Wernimont & Campbell, 1968) as a strategy for building effective selection instruments. Conducting a careful job analysis and designing measures to sample job behaviors—either directly, via a work sample, or less directly via strategies such as using "tell me about a time when . . . " approaches to interviewing—emerges as more effective, on average, than a "sign" strategy of identifying psychological constructs judged as relevant to the job in question.

Second, and following from the first, it shows the potential value of an investment in a custom-designed selection system, as opposed to an off-the shelf solution—at least in terms of validity. At the same time, other issues such as development costs and time to implementation merit consideration. Our findings should not be viewed as a mandate for preferring one predictor over another regardless of circumstances. Integration of multiple outcomes, such as validity, cost, time constraints, testing volume, subgroup differences, and applicant reactions, is needed for an informed decision.

Third, a number of the predictors at the top of our list in terms of validity are not generally applicable for entry-level hiring in situations where knowledge, skills, and abilities are developed after hire via training or on the job. Work samples and job knowledge tests clearly fall into this category. For others, we see some uncertainty. Empirical biodata and structured interviews can conceptually be used with both experienced and inexperienced applicants, but item content would be quite different, namely, limited to content not specific to the job for selection tools designed for use with inexperienced candidates. Our read of the current literature is that we cannot readily tease out whether comparable levels of validity are found for biodata scales and structured interviews designed for use with inexperienced versus experienced candidates.

Fourth, many, but not all, of the predictors at the top of our list reflect domains arguably more changeable with investment of time and effort than others. Measures such as work samples and knowledge tests lend themselves more readily to skill development via study and practice than measures of attributes generally viewed as more stable, at least in the short run, such as cognitive ability and personality. Although there are general norms calling for permitting candidates to retest and improve, it is advantageous to be able to offer candidates advice as to what they might do to improve their standing on the construct of interest.

How have the new findings been received?

Of course, we look forward to the responses to our article in the current journal issue and this will give one sense of how the findings have been received. However, we have also already received a lot of feedback on our original article, and have compiled a listing of reactions as gathered via email, conference conversations, blogs, LinkedIn posts, social media, and so on; along with comments from us in response to the reactions.

An overall reaction is a sense of "once you point it out, it's obvious." The key insight regarding the error in applying large range restriction corrections in settings where restriction can be expected to be small is readily grasped. Our favorite reaction is from one of the *Journal of Applied Psychology* reviewers, who opened the review with "Why didn't you publish this 40 years ago and save the field a lot of trouble?" The issue has been hiding in plain sight for a long time; we didn't see it, nor did the rest of the field.

Another reaction that was common is concern that our work focuses on what might be termed "legacy predictors," rather than on novel predictors emerging in the era of machine learning and artificial intelligence. Readers want to see a comparison between the validity of legacy predictors and gamified assessments, asynchronous video interviews, natural language processing-based scoring of essays, and the like. We share that interest and are also eager to see such comparisons. The scope of our project, however, was limited to a revisiting of existing meta-analyses; we look forward to a point that we hope is coming soon where a body of validity information will

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

accumulate about these novel predictors. That said, the set of predictors reviewed in our paper remain widely used, and we argue for as clear a picture as possible regarding how they function.

Third, we note concerns about our finding that validity has been overstated for quite a number of predictors. Those making such comments are not questioning our findings, but rather expressing disappointment with them. A version of this concern is that many nonpsychologists (e.g., computer scientists) are now competing in the selection space, with questions about the relative value of traditional versus new predictors or selection methods not yet widely addressed. Revised lower estimates of the validity of traditional predictors do not help these predictors compete for success in the marketplace. In a similar vein, there are concerns about how to present these new findings to clients: the argument is that we have been overstating the value of selection systems, and it is difficult to admit this. One response is to note that our field is constantly updating its knowledge base, and new insights replace prior beliefs. A second is that the findings apply to single predictors used in isolation. The use of multiple predictors is widespread, and by using composites of predictors we can obtain overall levels of criterion-related validity quite similar to what prior work would indicate. To illustrate, in ongoing work we make use of a revised version of a widely used matrix of meta-analytic validity coefficients, interpredictor correlations, and White-Black standardized mean differences among a set of widely used predictors: cognitive ability, conscientiousness, biodata, structured interviews, and integrity tests (Roth et al., 2011). We replace the validity values and White-Black mean differences used by Roth et al. with those from Sackett et al. (2022), with one exception: there is a new meta-analysis of the White-Black difference on biodata measures (Tenbrink et al., 2022), and we use that value. We make modest adjustments to the interpredictor correlations based on new data and/or the insights into range restriction provided by Sackett et al.; our updated matrix is available upon request. We compared mean validity for all possible combinations of 1, 2, 3, 4, and 5 predictors; the mean across these composites is .51 based on the prior validity values from Roth et al. and .47 based on our revised values. Thus, it would be incorrect to view our work as challenging the value of I-O selection practice.

Another common reaction is a sense of relief: our findings better match what is seen in practice. A number of practitioners express frustration that they have been unable to match the large values suggested in prior meta-analyses in their own local validation work, and express skepticism about the credibility of the prior values. The message is that our findings correspond more closely to what they see in applied practice, thus eliminating self-doubts as to what they are doing wrong in their validation work.

Finally, a quite common question is "did Frank Schmidt have a chance to see and react to your findings before his unexpected death in 2021?" The immediate answer is no, he did not. We acknowledge that he was on our mind as we worked on the paper, knowing that we were challenging common range restriction correction practices, and thus his conclusions about the relative validity of various predictors. We looked forward to discussing our findings with him, and were dismayed to learn of his death shortly before we finished the manuscript. We have the utmost respect for his work; the senior author is on record as viewing the original Schmidt and Hunter (1977) paper on the theory of validity generalization as the most influential paper published in the course of his 40+ years in the profession (Sackett, 2021). Our hope is that he would view our paper as an attempt to move us closer to a fuller understanding of the validity of selection tools.

We have yet to hear a serious challenge to the central message of our paper, namely, that applying a uniform correction to all validity studies in a domain, regardless of whether they are concurrent or predictive, is inappropriate. Oh, Le, and Roth (in press) offer a critique of our paper that focuses on identifying conditions under which range restriction can have meaningful effects in concurrent studies. We do not disagree with Oh et al. that it is *possible* for range restriction to have meaningful effects in concurrent studies. For example, we documented in Sackett et al. (2022) that range restriction can have meaningful effects on validity in concurrent studies when the predictor of interest is very strongly correlated (i.e., in excess of $r = .70$) with the

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

predictor or predictor composite that was used to hire employees combined with a small selection ratio (i.e., .10 or lower) having been used to hire those employees. Our key point and response to Oh et al. is that we view it as implausible that the existing meta-analyses that we reviewed in Sackett et al. are populated with primary studies that *on average* included samples of employees that were hired using a predictor or predictor composite that is so strongly correlated with the predictor of interest combined with such small selection ratios. Instead, it is much more plausible that the primary studies in those meta-analyses include, on average, more modest correlations and selection ratios, which Sackett et al. demonstrated would result in range restriction having little meaningful effect on predictor validity. With this said, that we choose to present this paper in a forum (IOP) where reactions are solicited signals our interest in opportunities for other views to be heard. As we noted, considerable uncertainty remains about many of our predictors, and we expect new data to help resolve these uncertainties.

Do new developments alter the Sackett et al. (2022) findings?

Sackett et al. (2022) emphasized that their findings reflected what was known at the time, and are subject to updating as new evidence emerges. We highlight here two new developments in the year or so since Sackett et al. was published.

First, in the domain of cognitive ability, Griebe et al. (2022) noted that the data used as the basis for the Schmidt and Hunter (1998) estimate of validity for general cognitive ability was based exclusively on studies at least 50 years old. They conducted a meta-analysis of 113 validity studies of the general cognitive ability—overall job performance relationship conducted in the 21st century, and find lower mean validity than the .31 estimate reported by Sackett et al. (2022), namely a mean observed validity of .16, and a mean corrected for unreliability in the criterion and for range restriction of .23. Using this value drops cognitive ability's rank among the set of predictors examined from 5th to 12th. Griebe et al. hypothesize that the lower validity reflects the reduced role of manufacturing jobs (which dominated the Schmidt and Hunter data) in the 21st century economy and the growing role of team structures in work. These changes result in a broader conceptualization of job performance than the quantity and quality of task performance measures used in the past for manufacturing jobs, incorporating less cognitively loaded interpersonal aspects of work, such as citizenship and teamwork, thus resulting in cognitive ability accounting for a smaller portion of this broader performance space.

Second, we also have revisited findings regarding assessment centers. Zhou et al. (2022) report meta-analytic findings of lower mean criterion reliability for measures of overall job performance for managerial jobs (.46) than nonmanagerial jobs (.61). Sackett et al. (2022) used a mean reliability value of .60 in the corrections they applied. Assessment centers stand alone among the set of predictors examined in terms of being used predominantly for managerial positions, and thus it is reasonable to use the reliability value for managerial jobs. This increases the mean corrected operational validity estimate from .29 to .33, which raises assessment centers from 8th on the validity ranking list to a tie with work samples for 4th. As assessment centers are a specific form of work sample tests, we find symmetry in these findings of comparable validity. Also, the move up of assessment centers in our ranking of predictors solidifies our finding of higher validity for job-specific predictors, with assessment centers joining structured interviews, work samples, job knowledge tests, and empirically keyed biodata as the top five predictors.

Rethinking validity–diversity tradeoffs

As our Table 1 and Figure 1 show, cognitive ability is the predictor with the largest White-Black mean difference. This has created a dilemma for selection system designers who value diversity, as the received wisdom for decades has been that cognitive ability is the most valid predictor of

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

Table 2. Standardized Regression Weights for Predictors Using Old and New Validity Values

	Weights based on Roth et al. (2011) validity values	Weights based on Sackett et al. (2022) validity values
Cognitive ability	.40	.23
Biodata	.06	.26
Conscientiousness	.09	−.04
Structured interview	.38	.39
Integrity	.16	.28
Situational judgment test	−.09	−.13

overall job performance. A system relying solely on cognitive ability will generally produce substantial adverse impact. Adding additional predictors and forming a composite can reduce this to some degree, but a key driver of the degree to which adding additional predictors reduces group differences is the weight given to each predictor (see Sackett and Ellingson, 1997, for useful tables showing the effect of adding predictors). In the past, several strategies have been used for addressing the validity–diversity tradeoff. One traditional strategy has been predictor weighting. Given the pattern of findings reported by Schmidt and Hunter (1998), the use of regression weighting to maximize the validity of the resulting predictor composite would result in giving very high weight to the cognitive ability measure, which substantially reduces the degree to which adding predictors decreases adverse impact. Equal weighting of predictors then becomes appealing as a way to reduce group differences, but this comes at the expense of validity. An alternate traditional strategy has been to use tests with separate cutoffs, rather than forming a composite, and setting a very low cutoff on the cognitive measure. This also mitigates the group difference problem, but dramatically reduces the performance gain that could potentially be attained via the use of the ability measure.

Our Figure 1 shows that the revised validity findings lead to a very different picture of the validity–diversity tradeoff. As cognitive ability no longer emerges as a top predictor in terms of validity, we do not see a basis for the prior argument that cognitive ability merits a place as the centerpiece of many/most selection systems. With the new validity findings, cognitive ability would merit considerably smaller weight in a regression model involving multiple predictors. We return here to some findings from our ongoing work, which we described above; i.e., making use of an updated version of a widely used matrix of meta-analytic validity coefficients, interpredictor correlations, and White-Black standardized mean differences among a set of widely used predictors: cognitive ability, conscientiousness, biodata, structured interviews, and integrity tests (Roth et al., 2011). We added a sixth widely used predictor, namely, SJTs, to this matrix, as the needed information to do so is now available. In Table 2 we show the standardized regression weights that are found using the prior validity estimates for these six predictors versus using the updated values.

We offer several observations. First, note that although cognitive ability had the highest weight among the six predictors using the prior validity estimates, it has only the fourth highest weight using the new estimates. Second, and critically, we examined the consequences of giving ability a weight of zero in a composite of all six predictors. Using the prior estimates, doing so reduces validity by .20 (from .66 to .46); it is easy to see the argument for the importance of ability based on the prior validity estimates. However, using the new validity estimates, giving ability a weight of zero reduces validity by .05 (from .61 to .56). Although the findings we show above make use of a composite of all six predictors, we also examined all possible predictor subsets (i.e., 1, 2, 3, 4, and 5 predictors). Holding constant the number of predictor methods, the mean validity for those not

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

including cognitive ability is virtually identical to those including ability. Third, the new validity estimates give a different picture regarding effects on diversity. The finding just noted that composites with and without cognitive ability can achieve comparable validity means that equally valid composites with smaller group differences can be created, thus aiding in reducing the validity–diversity tradeoff.

Importance of variability in validity, not just the mean

There is an unfortunate tendency in the use of meta-analysis to focus on mean validity values, with little to no attention given to the variance or standard deviation across studies. As noted earlier, for example, the Schmidt and Hunter (1998) summary of the validity of various predictors presents only mean values. This can wrongly contribute to a sense that the meta-analytic mean is a value that can be expected in subsequent applied settings. This expectation may be reasonable if the standard deviation around the mean value is small, but if it is large this means that, although the average study may find validity close to the mean value, many will find validity is considerably higher or lower than that mean value. Therefore, Sackett et al. (2022) presented standard deviations of operational validity estimates for each predictor to index how much between-study variability there is around the mean estimates.

A first observation is that the standard deviations do vary considerably. As Table 1 shows, for some predictors, the standard deviation is quite small: near zero for several contextualized personality measures (i.e., measures of personality at work, rather than in general). However, for others, the standard deviation is substantial: .25 for interests, .20 for integrity, .19 for structured interviews, .17 for personality-based emotional intelligence, .16 for unstructured interviews, and .15 for decontextualized conscientiousness. Meta-analytic convention is to report an 80% credibility interval (i.e., identifying the 10^{th} and 90^{th} percentiles of the distribution of operational validity values). Structured interviews, which top our list in terms of a mean validity of .42 have an 80% credibility interval ranging from .18 to .66. Thus, the validity of structured interviews should really be viewed as ".42, plus or minus .24", which conveys a different flavor than "validity is .42."

Questions that surface immediately are "why the variability? And "what is a practitioner to do given this variability and thus this uncertainty?" For the "why" question, there are multiple possible contributors. The first is that the predictor label encompasses too broad a set of measures which should in fact be viewed as distinct. The field has been making progress in exploring this. For example, the Schmidt and Hunter (1998) review separated interviews into structured versus unstructured. Meta-analyses since Schmidt and Hunter have produced findings which led Sackett et al. (2022) to differentiate between empirically keyed versus rationally keyed biodata, knowledge-based situational judgment tests versus behavioral tendency-based situational judgment tests, ability-based versus personality-based emotional intelligence measures, and contextualized versus decontextualized personality measures. Each of these distinctions has helped reduce the validity standard deviation that would have been observed had the distinction not been made. We encourage future research to make use of Lievens and Sackett's (2017) modular framework of predictor method factors to explore the effects of other distinctions (e.g., multiple choice versus constructed response formats) on validity. We thus suspect that deeper inquiry into current validity databases and new validity data will likely lead the field to examine such useful, novel distinctions, and further reduce validity standard deviations.

A second, related reason is that not all selection procedures are designed to meet comparable quality standards. This is especially the case for measures that cannot be readily taken off the shelf. Take assessment centers as an example: Some organizations might carefully follow all steps to design assessment center exercises, whereas others might take shortcuts and invest less in assessor training or use less experienced assessors. Along these lines, Gaugler et al.'s (1987) meta-analysis showed that the criterion-related validity of assessment center ratings is higher when the assessor

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

team consists of *both* psychologists and managers (as compared to when the assessor team consists either solely of psychologists or solely of managers). So, the large variability we reported can also be seen as a call to invest in high-quality design of selection procedures so that the validity obtained is more likely to be closer to the higher end of the credibility interval.

A third potential contributor to variability across studies is the broad category of design features for validity studies. One example is the criterion used. The meta-analyses summarized by Sackett et al. (2022) were limited to those using measures of overall job performance as the criterion and it is often unclear how overall performance is conceptualized and operationalized in a given study. Modern conceptualizations of performance are multifaceted and can include task performance, citizenship, and counterproductive work behavior, among others. It is often unclear which of these components are included in overall performance measures, as the items making up multi-item performance measures are commonly not documented in validity study write-ups (e.g., no detail beyond "performance was a composite across 8 items"). And in other studies, a single-item global performance rating is used, with no information provided as to how this task is framed for raters (e.g., limit to task performance versus go beyond task performance). As a field we need more clarity here to help understand the role of criterion dimensionality in contributing to variability in findings across studies. Another example is the use of predictive versus concurrent designs. These differ on a variety of dimensions, including potential differences in levels of test-taking motivation among applicants versus incumbents and differences in incentives to fake. Importantly, Sackett et al. noted that many meta-analyses do not differentiate between predictive and concurrent studies. Again, clarity is needed.

A fourth potential contributor to variability is differences between jobs or job types. For some predictors, we see no clear conceptual reason why job type should moderate validity (e.g., why would we expect a job-specific knowledge test for accountants to be more predictive than a job-specific test for engineers?). For others, one can readily hypothesize about differences (e.g., the possibility that agreeableness matters more for jobs requiring considerable team work versus solitary jobs).

This leads to the question of "what are practitioners to do in light of this uncertainty?" A first recommendation is that practitioners look to the literature for guidance to go beyond the meta-analytic grand means reported in Sackett et al. (2022) and examine the original meta-analyses for predictors under consideration for use and carefully examine moderator analyses to determine if there are any subcategorization of studies particularly relevant to the job and situation under consideration (e.g., findings broken down by job type or the see the above example of assessor type in Gaugler et al., 1987). A second recommendation is to compare predictors not only in terms of meta-analytic mean validity but also in terms of the low end of the 80% credibility interval. Taking one example from Sackett et al. (2022), consider structured interviews and empirically keyed biodata. The interview has higher mean validity (.42 versus .38), but biodata has a higher value for the lower end of a credibility interval (.26 versus .18). A risk-averse employer might prefer the predictor with less downside risk, and thus focus on the lower-end credibility value in identifying potential predictors. Others may use both the mean and the credibility value. A third recommendation is to search the published literature and available test publisher archives for individual validity studies in situations similar to that under current consideration. Meta-analyses are a useful summary of broad bodies of literature, but their existence does not negate the evidentiary value of a local validity study of the predictor of interest (see Newman et al. (2007) for a consideration of a Bayesian approach to integrating meta-analytic and local evidence).

Is there still a place for cognitive ability?

The finding of a considerably smaller corrected correlation between cognitive ability and overall performance (.31) in older data than previously believed to be true (.51), paired with finding

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

smaller ability-performance correlations in 21st century samples (.23), merits some elaboration. The finding certainly leads to a substantial revision of our view of the role of ability in predicting overall job performance, relative to other predictors. Overall, we believe the findings lead to a reduced role for cognitive ability in selection in the prototypical setting in which one's interest is in predicting overall job performance. In fact, given our analyses above demonstrating that validity of a predictor composite is about the same regardless of whether cognitive ability tests are included or not, paired with the adverse impact that cognitive ability tests can cause, some organizations might make the value judgment that inclusion of cognitive ability does more harm to diversity than it does to improving validity for predicting overall job performance. However, it would be incorrect to interpret these findings as a broad refutation of the value of cognitive ability measures in all settings. We offer a series of observations.

First, although current data produce smaller correlations (e.g., the corrected mean of .23 reported by Griebe et al., 2022), the relationship can still be useful in a variety of settings. Many of the predictors producing larger correlations are only applicable in settings in which applicants have prior experience and/or training in job-specific skills (e.g., work sample tests, job knowledge tests). Others are not readily amenable to use in high-volume screening settings (e.g., structured interviews). Still others require substantial development time and the availability of large validation samples (e.g., empirically keyed biodata). This highlights the potential value in research focused on finding ways to adapt such predictors for use in a wider variety of settings or for high-volume selection, while retaining their validity. For example, there is research on adapting interviews to high-volume selection using automated scoring of asynchronous video interviews (Hickman et al., 2022). Still, in terms of off-the-shelf measures deployable quickly, cognitive ability measures merit consideration, along with measures of personality, integrity, and emotional intelligence.

Second, cognitive ability has long been found to be a very strong predictor of performance in learning settings (e.g., higher education, training in civilian and military work settings, knowledge acquisition on the job). In jobs where lengthy and expensive training investments are made, cognitive ability remains the best available predictor, based on current knowledge. We note, however, that the relationship between ability and training performance has not been revisited in light of the issues with range restriction corrections that led to revised estimates of the ability-performance relationship.

Third, much depends on the criterion of interest. The historic view is that ability is a better predictor of task performance than of other aspects of performance, such as citizenship and counterproductive work behavior (Gonzalez-Mulé et al., 2014), though that finding merits revisiting in light of the range restriction correction issues highlighted by Sackett et al. (2022). Also, within the task performance domain we draw a distinction between typical and maximum performance. Maximum performance reflects the level of performance attainable under maximum effort conditions. A good example of this is the large-scale military research in the 1980's, including the Army's Project A (Campbell, 1990). In this research, a set of "hands-on performance" criteria were developed, in which incumbents performed carefully designed tasks (e.g., fix a broken radio) while observed by a trained observer who used a checklist to record whether needed task steps were or were not performed correctly. Although it is possible that some individuals may intentionally perform badly, we see this as a setting where maximum performance is likely: while under observation for a short period of time, we expect most soldiers to give their best effort. Average performance over an extended period of time reflects typical performance. As typical performance is influenced by effort in addition to procedural and declarative knowledge we would expect a smaller ability-performance relationship than would be the case for maximum performance. Whether to design a selection system targeting typical or maximum performance should be a strategic choice for an organization. Sackett (2007) argued that an organization might select for maximum performance if they believed they had strong structures in place (e.g., training, leadership, culture) designed to produce levels of typical performance that come close to maximum

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

performance. They might select for typical performance at lower levels of these structures. In short, ability merits stronger consideration in the design of a selection system targeting maximum performance than for one targeting typical performance.

How should practitioners and researchers estimate operational validity?

Operational validity refers to the criterion-related validity of a selection procedure in the applicant pool (so, free from range restriction introduced via selection) for predicting a criterion free of measurement error. Well-known psychometric formulas are available for estimating operational validity by correcting for criterion measurement error and range restriction, when there is range restriction. Sackett et al. (2022) provided a number of insights that should guide when and how these correction formulas are used. Here we distill some of these insights and offer 18 pieces of advice on how and when to correct for unreliability in the criterion and for range restriction. Note that here we are discussing making corrections in an individual validity study; Sackett et al. (2022) focused on correction in meta-analysis. More clarity about correcting individual studies will aid future meta-analyses.

1. *Correct for reliability first, then for range restriction*

If correcting for both, in the prototypic case in which criterion reliability is estimated on incumbents (i.e., on a restricted sample), then the reliability estimate is also range restricted. So, correct for reliability first, and then apply the range restriction correction to the reliability-corrected correlation.

2. *There is considerable measurement error in virtually all our criteria, and correcting for unreliability will be important for virtually all validity studies.*

Although there are many settings in which range restriction will be minimal, and no correction is warranted, measurement error is an issue for all of the widely used criterion measures in our field (e.g., rated performance, objective measures).

3. *Use an estimate of interrater reliability, not of internal consistency*

In the prototypic setting in which ratings are used as the criterion, the source of error of greatest interest and influence is the idiosyncratic perspective of a single rater. So, our interest is in an estimate of interrater reliability, which focuses on variance that is shared among raters. Internal consistency is based on the intercorrelations among rating dimensions, which is typically quite high, resulting in a high internal consistency coefficient. So internal consistency is not a major issue for validation research.

4. *Consider a local interrater reliability estimate if available*

If there are two qualified raters available for each participant in the validity study, use this to obtain an interrater reliability estimate.

5. *If a local interrater reliability estimate is not available, consider reliability estimates from highly similar settings with highly similar measures, if available*

For example, a consulting firm may use a common performance rating tool in multiple validity studies with multiple clients, and a measure-specific reliability estimate may be available.

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

6. *If the above two interrater reliability estimates are not available, consider a relevant meta-analytic reliability estimate*

For example, Zhou et al. (2022) offer current meta-analytic mean estimates of .46 for managerial jobs and .61 for non-managerial jobs. Similarly, Conway and Huffcutt (1997) offered meta-analytic mean estimates of .60 for low complexity jobs and .48 for high complexity jobs.

7. *Triangulate among relevant reliability estimates (e.g., local and meta-analytic) if multiple estimates are available*

Any value, whether local, extracted from a similar study, or meta-analytic, is a fallible estimate, affected by sampling error and a variety of potentially biasing factors. If multiple equally credible estimates are available, using a weighted or unweighted average is a good option.

8. *Realize that lower reliability estimates produce larger corrections*

We advocate the principle of conservative estimation in making corrections (i.e., better to underestimate validity than overclaim). The reliability correction divides a validity coefficient by the square root of the reliability estimate, which means that the lower the reliability the higher the correction. Point 7 above suggests averaging if multiple reliability estimates are available; here we are noting the error of instead thinking "given two reliability estimates I'll be conservative and use the lower one"; in fact, using the higher one would be conservative.

9. *If objective performance measures are used, consistency over time is the basis for a reliability estimate*

Measures reflect a fixed time period, and the question is consistency over time. For example, Sackett et al. (1988) assessed items processed per minute for supermarket cashiers for four different one-week periods; a composite of these four measures had a coefficient alpha value of .95, and thus measurement error is minimal. But reliability of a single one-week estimate was only .80 (i.e., the average intercorrelation between the four different one-week periods was .80); had a one-week measure of performance been used it would be important to correct for measurement error.

10. *Correcting for range restriction requires a credible estimate of the predictor standard deviation in the applicant pool and the standard deviation among selected employees*

Different types of restriction require different correction formulas. But at the heart of all of them is the ratio of unrestricted standard deviation (i.e., standard deviation in the applicant pool) to restricted standard deviation (i.e., standard deviation among selected employees); often referred to as a U-ratio. If one does not have a credible estimate of the U-ratio, then correction for range restriction should be avoided, with the acknowledgment that the uncorrected validity is likely an underestimate. If one does have a credible estimate of the U-ratio, then well-known formulas can be used to correct for range restriction. When correcting, it is important to remember the distinction between direct versus indirect range restriction. Thorndike's (1949) Case II correction can be used when there is direct range restriction. However, direct range restriction is exceedingly rare in personnel selection because it only occurs when incumbents were selected using *only* the selection method in question (e.g., incumbents in the validity study were hired using only the cognitive ability test being validated in the study). Indirect range restriction, which encompasses any scenario in which incumbents were selected using some method in addition to or other than the selection method in question (e.g., incumbents were selected using some method other than the validation study predictor, or were selected using a battery of methods in addition to the validation study predictor), is much more common. In this scenario, if information about the

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

intercorrelations between the actual method of selection (e.g., the battery of tests actually used to hire incumbents), the validation study predictor (e.g., a cognitive ability test administered to job incumbents for the purpose of validation), and the criterion are known (e.g., supervisor ratings of job performance), then Thorndike's Case III correction formula can be used. Otherwise, the relatively new Case IV correction method (Hunter et al., 2006) can be used, as it only requires the ratio of incumbent to applicant predictor standard deviations. See Sackett and Yang (2000) for treatment of additional range restriction scenarios.

11. *Range restriction is a particularly important issue if the predictor in question was used in selecting the validation sample*

Addressing restriction is generally important if a predictor was used operationally, either alone or in conjunction with other predictors. It should not be assumed that "range restriction corrections are only important for direct restriction, but not for indirect restriction": the key is whether the predictor was used.

Technically, Sackett et al. (2022) showed that even indirect range restriction has a meaningful effect if the correlation between the predictor of interest (x) and the actual basis for selection (z) is large. In concurrent studies, we showed that $r_{zx} = .5$ is roughly the high end for this in cases where selection took place on another variable (and there are some exceptions with higher values), and that this is not a large enough correlation between z and x for indirect range restriction to have a meaningful effect on validity. But consider an indirect restriction scenario in which a unit weighted composite of the predictor of interest to us (e.g., a cognitive ability test) and another predictor (e.g., a personality test) was used for selection. The correlation between our predictor of interest and the two-predictor composite is a part-whole correlation. Its magnitude depends on the correlation between the two predictors, and is lowest when the two predictors are uncorrelated, where it takes a value of .71. Sackett et al. showed that at or above $r_{zx} = .70$ indirect range restriction can have a meaningful effect on validity, and thus careful correction may be needed.

12. *Range restriction generally does not have a sizeable influence if the predictor was not used in selecting the validation sample*

Sackett et al. (2022) showed that a correlation of about .70 or higher between the actual method of selection (z) and the predictor of interest (x) is needed for the effects of indirect range restriction to have the potential to be sizable, and that the correlation between different types of predictors is almost always less than .50; at these levels effects of indirect range restriction are minimal. An exception is when the "old" and "new" predictors are measures of the same construct (e.g., validation of a new alternate form of the prior predictor). One I-O colleague pointed out another scenario where this takes place: an organization replaces its test vendor with another one. Employees were selected using a tool from the prior vendor; the new vendor has a similar tool and wants to document its validity in a concurrent study. Here a concurrent study can have substantial range restriction.

13. *Obtain a local applicant and incumbent sample standard deviation if possible*

Textbooks would equate this with a predictive study. But not discussed anywhere we have seen is the reasonably plausible method of doing a concurrent study and at the same time administering the predictor of interest to an applicant sample (and not using the scores for selection) in order to obtain local applicant and incumbent standard deviations.

14. *Be cautious of the use of formulas that convert a selection ratio into the U-ratio needed for range restriction correction*

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

This only works if the predictor in question was the sole basis for selection, which we argue is virtually never the case. It is wrong to take information from a firm that they hire about half their applicants, and assume that this means the firm used a 50% selection ratio on the measure one is trying to validate.

15. *Be cautious about using publisher norms as an estimate of the applicant pool standard deviation*

Publisher norms may be feasible if the norms are occupation-specific; we are very skeptical of using general population norms.

16. *Do not use a mean range restriction correction factor from a meta-analysis as the basis for correction for concurrent studies*

This is a key message from Sackett et al. (2022). Such correction factors generally come from predictive studies and overcorrect substantially if applied to concurrent studies.

17. *Use a mean range restriction correction factor from a meta-analysis with extreme caution as the basis for correction for predictive studies*

Look closely at the range of U-ratios making up the artifact distribution. The range is often quite broad. Perhaps choose a value near the high end of the distribution as a conservative correction.

18. *Make no correction unless confident in the standard deviation information at hand*

The principle of conservative estimation suggests no correction is better than an incorrect overestimation. Tell your client that your validity value is likely an underestimate, but you are not in a good position for a more precise estimate.

Implications for research

Although our focus in this article is on implications for practice, we also note that Sackett et al. (2022) has many implications for research as well. As noted earlier, one is a call for those conducting meta-analyses to consider corrections separately tailored for different subsets of studies, rather than applying a common correction. Another is calling attention to reporting limitations in prior meta-analyses. Many do not differentiate between predictive and concurrent studies, thus precluding following the advice above regarding making differential correction for different subsets of studies. Various predictors could be revisited, now making these differentiations and permitting more nuanced corrections. Yet another is applying these ideas of differentiating between predictive and concurrent studies to additional domains. Our focus here was solely on the relationships between predictors and overall job performance. Relationships with other criteria can also be usefully considered. For example, one of us has a follow-up project in progress focusing on training performance as the criterion. We encourage similar efforts in other domains.

Conclusion

Our goal has been to further elaborate on issues emerging from our revisiting of prior meta-analyses of personnel selection tools in light of new insights into the correction factors

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

used to estimate validity (Sackett et al., 2022). We (a) presented a conceptual overview of our critique of prior approaches to correction, (b) outlined the implications of this new perspective for the relative validity of different predictors and for the tradeoff between validity and diversity in selection system design, (c) highlighted the need to attend to variability in meta-analytic validity estimates, rather than just the mean, (d) summarized reactions encountered to date to our work, and (e) offered a series of recommendations regarding how to go about correcting validity estimates for unreliability in the criterion and for range restriction in applied work.

References

Campbell, J. P. (1990). An overview of the army selection and classification project (Project A). *Personnel Psychology*, **43**(2), 231–239.

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**, 331–360.

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.

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.

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.

Griebie, A., Bazian, I., Demeke, S. Priest, R., Sackett, P. R., & Kuncel, N. R. (2022). A contemporary look at the relationship between cognitive ability and job performance [Poster]. Society for Industrial and Organizational Psychology Annual Conference, Seattle, WA.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, **107**, 1323–1351

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, **91**, 594–612. https://doi.org/10.1037/0021-9010.91.3.594

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**, 43–66.

Newman, D. A., Jacobs, R. R., & Bartram, D. (2007). Choosing the best method for local validity estimation: Relative accuracy of meta-analysis versus a local study versus Bayes-analysis. *Journal of Applied Psychology*, **92**(5), 1394–1413.

Oh, I., Le, H., & Roth, P. L (in press). Revisiting Sackett et al.'s (2022) recommendation against correcting for range restriction in concurrent validation studies. *Journal of Applied Psychology*.

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

Sackett, P. R. (2007). Revisiting the origins of the typical-maximum performance distinction. *Human Performance*, **20,** 179–185.

Sackett, P. R., & Ellingson, J. E. (1997). On the effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, **50**, 708–721.

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

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum performance. *Journal of Applied Psychology*, **73**, 482–486.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, **107,** 2040–2068.

Sackett, P.R. (2021). Reflections on a career studying individual differences in the workplace. *Annual Review of Organizational Psychology and Organizational Behavior*, **8**, 1–18.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, **62**(5), 529–540. https://doi.org/10.1037/0021-9010.62.5.529

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, **124**(2), 262.

https://doi.org/10.1017/iop.2023.24 Published online by Cambridge University Press

Tenbrink, A.P., Wegmeyer, L., Rowley, S.J., Sendra, C, & Speer, A. (2022, April). Group differences in biographical data: A meta-analysis. Presented at the annual conference of the Society for Industrial and Organizational Psychology.

Thorndike, R. L. (1949). *Personnel selection*. Wiley

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, **52**(5), 372–376.

Zhou, Y., Shen, W., Beatty, A. S., & Sackett, P. R. (2022, April). An updated meta-analysis of the interrater reliability of supervisory performance ratings. Presented at the Society for Industrial and Organizational Psychology conference, Seattle.

Cite this article: Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2023). Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors. *Industrial and Organizational Psychology* **16**, 283–300. https://doi.org/10.1017/iop.2023.24