Industrial and Organizational Psychology (2023), **16**, 371–377
doi:10.1017/iop.2023.47

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

A reply to commentaries on "Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors"

Paul R. Sackett[1], Christopher M. Berry[2], Filip Lievens[3], and Charlene Zhang[4]

[1]Department of Psychology, University of Minnesota, Minneapolis, MN, USA, [2]Department of Management and Entrepreneurship, Indiana University, Bloomington, IN, USA, [3]Singapore Management University, Lee Kong Chian School of Business, Singapore and [4]Amazon, Alexandria, VA, USA
Corresponding author: Paul R. Sackett; Email: psackett@umn.edu

We greatly appreciate the thoughtful and thought-provoking comments that we have received in reaction to our focal article (Sackett et al., 2023). Generally, the commentaries could be divided into two groups. The largest group of commentaries suggested ways to expand or refine our assertions. This group thus saw our paper as a starting point and offered various new insights. Another smaller group of commentaries was more critical. A few even argued strongly that the original paper was in error and its conclusions were wrong. With very limited page space, this is not the place for a point-by-point response to each of the commentaries. Thus, we first briefly review some of the suggestions of the first group and then focus more on three critical commentaries, given readers likely want to know whether we accept the criticisms offered or have alternative perspectives. We want to emphasize that although we offer our rationales for why we at times take positions that differ from those offered by commentators, we do not claim to be the last word on the issues at hand. We present our rationales to permit the reader to compare the perspectives taken by us and by the commentators. And issues raised by these differing perspectives will, we hope, prompt research to help resolve areas of disagreement.

Suggestions to expand or refine on our focal article

Regarding the first group of commentaries, we agree with Murphy (2023) that discussions of the validity of selection predictors should not be a "horse race." That is why we mentioned that composite validity and incremental validity are also important and that diversity (Olenick & Somaraju, 2023) and other criteria such as candidate experience (Jones & Cunningham, 2023) and cost/time (Harms et al., 2023) should be factored in. Similarly, we concur about the "validity for what" question (i.e., predictor–criterion construct congruence, Hough & Oswald, 2023) and the importance of careful selection design to increase the likelihood that the validity of a given predictor obtained in a specific situation (DeSimone & Fezzey, 2023) is at least as high as the meta-analytic average of Sackett et al. (2022). Jones and Cunningham (2023) make a good point that our paper concentrated on the "usual suspects" in selection and that in the future similar large-scale reviews are needed to detail the effects of recent advances such as gamification, "flash" assessments, digital interviewing, and so on. Huffcutt and Murphy (2023) highlight what we view as an important issue, namely, the need to look not only at meta-analytic mean validity but also at the variance. They call attention to the large variability that accompanies the high mean validity estimate for structured interviews and offer useful insights into ways to shed light on the features driving this variability. Finally, we acknowledge that our paper did not tackle how to best

communicate the validities to a lay audience. Like Highhouse and Brooks (2023), we are in favor of the adoption of more realistic effect sizes and of specific approaches in the form of probabilities and/or frequencies to present the validities more persuasively.

Barrett and Doverspike (2023) are critical of meta-analytic findings and offer thoughts on their applicability to practice. They note that we make use of simulations that combine meta-analytic data across six predictors and express doubt that any applied selection system would use the full set of predictors we examine in the simulation. Although we used all six predictors as an illustration, we noted that we also examined all possible combinations of two through six predictors. Barrett and Doverspike conclude with five useful recommendations for practice, including cautioning against reliance on meta-analytic findings in any specific situation due to the variance across studies in validity and the difficulties in making sound range restriction corrections. This is good advice; we advocate meta-analytic findings as a road map for identifying predictors more or less likely to prove useful rather than as the source of a point estimate of validity for any specific setting.

Schoen's (2023) commentary is addressed at the field at large, not just at our paper. He challenges the fundamental notion of the use of corrections for range restriction. It is useful to see an integrated collection of the various concerns about corrections that have appeared in the literature over the years. Schoen argues because range restriction is present in the real world, corrections for it depart from the real world. We do think that in principle it is sound to address range restriction, as the question we are asking in designing selection systems is "how well will this selection system predict the outcome of interest in an applicant population?" If a selected sample is restricted relative to an applicant sample, a validity estimate in the selected sample will be an underestimate. If the information needed to estimate the degree of restriction is available, making a correction permits an estimate of the relationship of interest, namely, validity in the applicant population.

Response to Ones and Viswesvaran

Ones and Viswesvaran (2023) offer a particularly strong critique. They, along with Oh et al. (2023), challenge our starting observation, namely, that range restriction correction factors derived from applicant-based predictive studies cannot be assumed to also apply to incumbent-based concurrent studies. Our position, which is based on the standard range restriction correction formulas, is that at a given selection ratio at point of hire, restriction on the predictor x will be greater if x is used in selection (e.g., in a predictive study) than if selection is on the basis of another variable, z, with x measured later in an incumbent sample. The degree of restriction is driven by the correlation between x and z. The smaller the correlation, the smaller the amount of restriction on x in the incumbent sample. In contrast, as r_{zx} approaches 1.0, the two types of studies become more similar in the amount of restriction on x.

Thus, we state that the likelihood of sizable restriction is less in concurrent studies than predictive studies and is often negligible. Therefore, we argue against assessing the degree of restriction present in predictive studies and assuming that the same degree of restriction is present in concurrent studies. Other critics of our work accepted this starting premise but took positions different from us on the likelihood of large amounts of restriction in concurrent studies (Oh et al., in press).

In contrast, Ones and Viswesvaran (2023) assert that empirical work shows comparable observed validity for cognitive tests in predictive and current studies, from which one can infer that comparable degrees of range restriction are present in both. There are some issues with this conclusion. A first issue is looseness in terminology, to which we have, we acknowledge, contributed. Predictive studies usually, but not always, are done on applicants. But a study many might label as concurrent because incumbents are tested, might be labeled predictive if the

https://doi.org/10.1017/iop.2023.47 Published online by Cambridge University Press

criterion is obtained later rather than at the same time that the predictor is administered. In addition, applicant studies are meaningfully differentiated into studies where the predictor of interest was used operationally and where it was not. We would have been better served if we consistently said "predictive studies using applicants, *with the predictor used in selection*," rather than using "predictive" as shorthand. As we note in our focal article, the critical feature is whether the predictor was or was not used in selection, as large amounts of range restriction are much more readily attainable in the former than the latter.

Ones and Viswesvaran (2023) cite Hartigan and Wigdor (1989), among others, in support of their position of predictive and concurrent studies producing comparable observed validity. Hartigan and Wigdor analyzed GATB data, and in that dataset all studies did not use the GATB for selection. Some studies tested newly hired employees but did not use the GATB scores in selection; others tested incumbents. That both sets of studies produce comparable validity estimates speaks to the issue of whether the time interval between predictor and criterion matters but not to the central issue facing us here, namely, comparability of estimates from applicants where the predictor of interest *was used for selection* and incumbents where the predictor of interest was *not* used.

A second key issue is that Ones and Viswesvaran argue for continuing to use Hunter's (1983) analysis of GATB data as the basis for the field's conclusions about the validity of cognitive ability tests. In Sackett et al. (2022), we outlined why there is reason to question the resulting validity estimate. In a nutshell, Hunter did not have the applicant SD needed for a range restriction correction and instead pooled incumbent data across jobs and viewed this as an estimate of the applicant SD. Hunter found this pooled incumbent SD was on average about 50% larger than the incumbent SDs within the individual validity studies and thus obtained a large correction factor. However, Hunter focused on an older set of GATB studies; importantly, we have the raw data from a newer set, with old and new comparable in size (about 38,000 individuals in each). When we mimic Hunter's analysis with the new data, and apply Sackett and Ostgaard's (1994) correction factor for the use of pooled cross-job data to estimate applicant pool SDs for specific jobs (i.e., reduce the pooled incumbent SD by 10%), we get an estimated applicant pool SD that is only 6% larger than the average incumbent SD, thus signaling very limited restriction. Hunter's value of an estimated applicant pool SD 50% larger than the average incumbent SD using the older GATB studies is at odds with both the newer studies (which are procedurally equivalent; Hartigan & Wigdor, 1989) and with the analytic framework developed in Sackett et al. (2022), which shows that this degree of restriction would be hard to obtain in concurrent studies. Finally, we have only recently become aware of an article by Salgado and Moscoso (2019) in which they extracted information from hundreds of GATB technical reports that make up the data underlying Hunter's analysis and applied the same analyses reported by Hunter. They report a smaller estimate of restriction, namely a *u*-ratio of .77 where Hunter reported .67. When we apply the Sackett and Ostgaard (1994) correction factor described above we get a *u*-ratio of .86 Thus, there are reasons to question conclusions based solely on Hunter's analysis of the older set of GATB studies.

Third, Ones and Viswesvaran (2023) note potential bias in a new meta-analysis of cognitive ability validity studies done in the 21st century (Griebe et al., 2022). They acknowledge that it is a SIOP paper of limited length and thus without full detail. Ones and Viswesvaran posit that the new data are biased by studies designed to show the incremental validity of novel predictors over cognitive ability. One of us (Sackett) is an author of the 21st century meta-analysis, and this concern about potential bias was voiced by a number of people when we presented at SIOP. We have subsequently added additional studies to the database, solicited additional studies from practitioners, and compared studies that (a) examined and found incremental validity, (b) examined and did not find incremental validity, and (c) did not address incremental validity. Critically, we did not find validity differences across this partitioning of studies based on incremental validity, and studies provided to us by consultants did not differ significantly from

https://doi.org/10.1017/iop.2023.47 Published online by Cambridge University Press

studies in the searchable literature. Again, all this is preliminary, but it is responsive to the concerns voiced by Ones and Viswesvaran and shows that the bias is not present.

Response to Oh, Mendoza, and Le

Many of the ideas in Oh et al. (2023) are also included in Oh et al.'s (in press) commentary on Sackett et al. (2022). We have been invited by *Journal of Applied Psychology* to respond to that commentary (Sackett et al., in press), so we will leave our detailed responses to those ideas for our *JAP* response. In brief, Oh et al. (in press) questioned a number of choices made in Sackett et al. (2022). They interpreted our paper as recommending against correcting for range restriction in general in concurrent validation studies; yet, we emphasize that we endorse correction when one has access to the information needed to do so. Our focus was on making range restriction corrections when conducting *meta-analyses*, where it is common for primary studies to be silent as to the prior basis for selection of the employees later participating in the concurrent validation study. As such, the applicant pool information needed for correction is typically not available. Sackett et al. highlighted that in many situations range restriction will be small; so, the inability to correct for it results in only a modest underestimate of validity. Oh, Mendoza, and Le mentioned settings that would result in substantial range restriction. Indeed, there are settings that can produce substantial range restriction (e.g., when r_{zx} is very high or the selection ratio is very low). We view such settings as uncommon, and it seems implausible that they made up the bulk of the studies contributing to the meta-analyses reviewed by Sackett et al.

In the following, we respond to a handful of new ideas presented by Oh et al. (2023), structured around some of their subsections. First, Oh, Mendoza, and Le wrote that they are "shocked" by our dichotomous thinking. Let us clarify that our thinking is only dichotomous in that, when considering correcting for range restriction, one must ask themselves whether one has the information to do so. That is, does one have key information needed for a range restriction correction such as u_x (restricted SD divided by unrestricted SD), or a pairing of u_z and r_{zx}? If the answer is "yes," then one should correct for range restriction. If the answer is "no," (which is very often the case, especially in meta-analyses) then we believe one should not correct for range restriction. In the latter situation, Sackett et al. (2022) demonstrated that in most circumstances the lack of a correction will result in only a small underestimate.

Second, we agree with Oh, Mendoza et al., and Morris (2023) that, even in predictive validity studies, direct range restriction is relatively uncommon. Thus, even in predictive studies, range restriction will likely be indirect. As we mentioned in our response to Ones and Viswesvaran, how much indirect restriction affects the validity of x depends in large part on the correlation between x and the selection method, z. In concurrent studies, r_{zx} will generally not reach the level needed for indirect restriction to have a large effect (i.e., about $r_{zx} = .70$ or higher) because x was not used in selecting the job incumbents, and the correlation between different selection procedures (e.g., x and z) is generally .50 or lower, as we reviewed in Sackett et al. (2022). However, in predictive studies, x often will be used to hire applicants. For example, x may be part of a battery of other predictors used to select applicants. In this case, r_{zx} is in large part a function of the number of predictors included in the battery. For example, per the theory of composites (Ghiselli et al., 1981), if the average intercorrelation among the predictors is .25, then x will correlate with z (here z is the composite of predictors, of which x is one part) at .79 for a two-predictor battery, .71 for a three-predictor battery, and .66 for a four-predictor battery (if the average intercorrelation of predictors is .00, then these r_{zx} values become .71, .58, and .50, respectively; and if the average intercorrelation is .50, they become .87, .82, and .79, respectively). So, even with indirect range restriction, it is at least possible for range restriction to be substantial due to high r_{zx} values in predictive validity studies, whereas such high r_{zx} values would be much less common in concurrent studies. That is

https://doi.org/10.1017/iop.2023.47 Published online by Cambridge University Press

why we suggested that one should not apply a value of u_x derived from predictive studies to concurrent studies.

To make the case that amounts of range restriction are similar in predictive and concurrent designs, Oh, Mendoz, and Le cited Schmitt et al. (1984) as writing "concurrent validation designs produce (observed) validity coefficients roughly equivalent to those obtained in predictive validation designs." That was not the entire sentence from Schmitt et al., The rest of the sentence reads "and . . . both of these designs produce higher validity coefficients than does a predictive design which includes use of the selection instrument" (Schmitt et al., p. 407). Schmitt et al. referred to the latter research design as "a very common situation in validation research" (p. 408). Schmitt et al. found that in concurrent designs the observed validity of cognitive ability tests was .34, in pure predictive designs (test not used for selection) validity was .30, and in predictive designs in which the test was used for selection the validity was .26. Thus, Schmitt et al. actually supports our idea that range restriction will generally be greater in predictive designs than in concurrent designs, especially if the test was used for selection. Oh, Mendoza, and Le also cited Pearlman et al. (1980) for the idea that observed validity of cognitive ability tests is similar for concurrent and predictive designs. Although this is what Pearlman et al. reported, they did not distinguish between predictive designs in which the test was or was not used for selection, as did Schmitt et al. which we view as a crucial distinction. They also did not distinguish between predictive designs that used job incumbents versus applicants, which is also a key distinction (see our response to Ones and Viswesvaran, 2023).

Third, Oh, Mendoz, and Le note our skeptical tone in a quote from Sackett et al. (2022). Let us clarify that our skepticism was more about the value of u_x that Hunter (1983) obtained than the method that was used to obtain it. Specifically, we voiced skepticism about the value of $u_x = .67$ that Hunter arrived at from using an approach wherein many incumbent samples were pooled together to compute an estimate of the applicant pool SD of cognitive ability. That skepticism has two bases. First, $u_x = .67$ seems too low for the *average* amount of range restriction in the set of concurrent studies contributing to Hunter's meta-analysis; we showed in Sackett et al. (2022) that one can only get u-ratios that low in concurrent studies when a high value of r_{zx} is paired with a low selection ratio. Based on Salgado and Moscoso's (2019) reanalysis of the GATB studies examined by Hunter, we obtain $u_x = .86$, as we noted in our response to Ones and Viswesvaran (2023). Second, as we also mentioned in our response to Ones and Viswesvaran (2023), using this same approach of pooling across incumbent samples to estimate the applicant pool SD in the newer, procedurally equivalent GATB studies, we get an applicant pool SD estimate that is only slightly larger than the average incumbent SD, signaling very limited range restriction.

Response to Cucina and Hayes

We agree with a lot that Cucina and Hayes (2023) put forward in their commentary. So, let us clarify upfront that we do not argue for the demise of cognitive ability testing (see also Kulikowski, 2023). Their commentary focuses on a very useful documentation of the range of other criteria that cognitive ability predicts. This echoes a section in our focal article: We acknowledged that our conclusions are limited to relationships with supervisor ratings of overall job performance, and cognitive ability is likely an important predictor of other key criteria such as performance in learning/training situations, or when predicting specific technical aspects of performance.

One of their key points is that supervisor ratings may be deficient as a criterion, and we know little about their construct validity. We acknowledge these potential problems. That said, correlations with supervisor ratings are at the heart of much to most of what we know about the relationships between various predictors and job performance. Rating criteria dominate the meta-analyses of predictors used in personnel selection, as well as the data making up the Schmidt and Hunter (1998) summary that has dominated the field for decades.

https://doi.org/10.1017/iop.2023.47 Published online by Cambridge University Press

Cucina and Hayes (2023) focus on a series of what they label paradoxes; however, we do not see them as such. The first is that ability predicts performance on a single task involving aiming and shooting a handgun better than it predicts supervisory ratings of overall performance, which is surprising if supervisor ratings are all that matter. As we noted above, we did not mean to suggest that supervisor ratings are all that matter; other criteria can certainly be of interest.

Second, Cucina and Hayes (2023) note that our finding that GATB data are minimally restricted is at odds with the result one gets when applying Schmidt and Hunter's (1998) big correction to the GATB training studies, which produced a validity (.56) comparable to other data in the training space. Two lines of evidence might help to explain this. First, we followed Hunter's (1983) method of pooling incumbent data to estimate an unrestricted SD for studies in the newer GATB data set that used training criteria and found a *u*-ratio of 1.01, indicating no range restriction. Second, a key difference between the military studies and the GATB studies was whether the cognitive ability measure was used in selection. This was not the case with the GATB studies. Trainees in the military studies had been selected in part based on their AFQT scores. So, large amounts of range restriction are at least plausible for those military studies but less so for the GATB studies.

Third, Cucina and Hayes (2023) note that other predictors, such as job knowledge, are substantially *g*-loaded, and thus selecting on one of these such measures inadvertently taps general ability. This is certainly true. That job knowledge indirectly taps general ability would be a problem if we were advocating the elimination of cognitive ability from selection systems. But this is not the case, as discussion above makes clear.

Fourth, they offer a critique also expressed by Ones and Viswesvaran (2023), namely, that if the cognitive demands of jobs are increasing, how can the validity of cognitive ability be decreasing? Apart from potential increases in the cognitive demands of jobs, jobs are also changing on dimensions other than their cognitive demands, with increased demands for interpersonal and teamwork skills as one key example. To the extent that (a) overall performance ratings include a larger interpersonal/teamwork component, and (b) these components are not well-predicted by cognitive ability, but rather by noncognitive factors, then (c) cognitive ability may be found to predict a smaller piece of a growing criterion space.

Finally, they observe that it is folly to select for contextual performance and yet expect technical task proficiency. We concur that selection systems should target the criteria of interest to the employer.

In sum, the commentaries have added a lot to our thinking about selection system design. We do believe that our central message remains intact in the face of the issues raised, as we have outlined in this response.

References

Barrett, G. V., & Doverspike, D. (2023). Rearranging deck chairs on the titanic: What are practitioners to do? *Industrial and Organizational Psychology*, **16**(3), 349–352.

Cucina, J. M., & Hayes, T. L. (2023). Rumors of general mental ability's demise are the next red herring. *Industrial and Organizational Psychology*, **16**(3), 301–306.

DeSimone, J., & Fezzey, T. (2023). Is it also time to revisit situational specificity. *Industrial and Organizational Psychology*, **16**.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. W. H. Freeman and Company.

Griebe, A., Bazian, I. M., Demeke, S., Priest, R., Sackett, P. R., & Kuncel, N.R. (2022). A contemporary look at the relationship between general cognitive ability and job performance. Presented at the 37th Annual Conference of Society for Industrial and Organizational Psychology, Seattle, WA.

Harms, P. D., Foster, J., & Brummel, B. (2023). Ideal solutions don't necessarily inform reality. *Industrial and Organizational Psychology*, **16**(3), 313–316.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. National Academies Press.

https://doi.org/10.1017/iop.2023.47 Published online by Cambridge University Press

Highhouse, S., & Brooks, M. E. (2023). Interpreting the magnitude of predictor effect sizes: It's time for more sensible benchmarks. *Industrial and Organizational Psychology*, **16**(3), 332–335.

Hough, L. M, & Oswald, F. L. (2023). Revisiting predictor-criterion construct congruence: Implications for designing personnel selection systems. *Industrial and Organizational Psychology*, **16**(3), 307–312.

Huffcutt, A., & Murphy, S. (2023). Structured interviews: Moving beyond mean validity. . . . . *Industrial and Organizational Psychology*, **16**(3), 344–348.

Hunter, J. E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization to the General Aptitude Test Battery (USES Test Research Report No. 45)* US Department of Labor, Employment and Training Administration.

Jones, J. W., & Cunningham, M. R. (2023). Going beyond a validity focus to accommodate megatrends in selection system design. *Industrial and Organizational Psychology*, **16**(3), 336–340.

Kulikowski, K. (2023). It takes more than meta-analysis to kill cognitive ability. *Industrial and Organizational Psychology*, **16**.

Morris, S. B. (2023). Meta-analysis in organizational research: A guide to methodological options. *Annual Review of Organizational Psychology and Organizational Behavior*, **10**, 225–259.

Murphy, K. R. (2023). Interpreting validity evidence: It is time to end the horse race. *Industrial and Organizational Psychology*, **16**(3), 341–343.

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

Oh, I. S., Mendoza, J. L., & Le, H. (2023). To correct or not to correct for range restriction, that is the question: Looking back and ahead to move forward. *Industrial and Organizational Psychology*, **16**(3), 322–327.

Olenick, J., & Somaraju, A. V. (2023). On the undervaluing of diversity in the validity-diversity tradeoff consideration. *Industrial and Organizational Psychology*, **16**(3), 353–357.

Ones, D. S., & Viswesvaran, C. (2023). A response to speculations about concurrent validities in selection: Implications for cognitive ability. *Industrial and Organizational Psychology*, **16**(3), 358–365.

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**, 373–406.

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**, 680–684.

Sackett, P. R., Berry, C. M., Lievens, F., & Zhang, C. (in press). Correcting for range restriction in meta-analysis: A reply to Oh et al. *Journal of Applied Psychology*.

Sackett, P. R., Berry, C.M., Zhang, C., & 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**(3), 283–300

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.

Salgado, J. F., & Moscoso, S. (2019). Meta-analysis of the validity of general mental ability for five performance criteria: Hunter and Hunter (1984) revisited*Frontiers in Psychology*, **10**. https://doi.org/10.3389/fpsyg.2019.02227.

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

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, **37**, 407–422.

Schoen, J. (2023). Hocus-pocus and hydraulics functions: Anything not worth doing is not worth doing well. *Industrial and Organizational Psychology*, **16**(3), 328–331.

Cite this article: Sackett, P. R., Berry, C. M., Lievens, F., & Zhang, C. (2023). A reply to commentaries on "Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors". *Industrial and Organizational Psychology* **16**, 371–377. https://doi.org/10.1017/iop.2023.47