AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Applied Psychology

ISSN: 0021-9010

2024, Vol. 109, No. 10, 1611–1634
https://doi.org/10.1037/apl0001203

Insights From an Updated Personnel Selection Meta-Analytic Matrix: Revisiting General Mental Ability Tests' Role in the Validity–Diversity Trade-Off

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

[1] Department of Management and Entrepreneurship, Kelley School of Business, Indiana University
[2] Lee Kong Chian School of Business, Singapore Management University
[3] Amazon, Arlington, Virginia, United States
[4] Department of Psychology, University of Minnesota Twin Cities

General mental ability (GMA) tests have long been at the heart of the validity–diversity trade-off, with conventional wisdom being that reducing their weight in personnel selection can improve adverse impact, but that this results in steep costs to criterion-related validity. However, Sackett et al. (2022) revealed that the criterion-related validity of GMA tests has been considerably overestimated due to inappropriate range restriction corrections. Thus, we revisit the role of GMA tests in the validity–diversity trade-off using an updated meta-analytic correlation matrix of the relationships six selection methods (biodata, GMA tests, conscientiousness tests, structured interviews, integrity tests, and situational judgment tests) have with job performance, along with their Black–White mean differences. Our results lead to the conclusion that excluding GMA tests generally has little to no effect on validity, but substantially decreases adverse impact. Contrary to popular belief, GMA tests are not a driving factor in the validity–diversity trade-off. This does not fully resolve the validity–diversity trade-off, though: Our results show there is still some validity reduction required to get to an adverse impact ratio of .80, although the validity reduction is less than previously thought. Instead, it shows that the validity–diversity trade-off conversation should shift from the role of GMA tests to that of other selection methods. The present study also addresses which selection methods now emerge as most valid and whether composites of selection methods can result in validities similar to those expected prior to Sackett et al. (2022).

Keywords: validity–diversity trade-off, general mental ability tests, meta-analytic correlation matrix, personnel selection

Supplemental materials: https://doi.org/10.1037/apl0001203.supp

Meta-analyses of the criterion-related validities and intercorrelations between various personnel selection methods have allowed researchers to create meta-analytic correlation matrices of the relationships between these selection methods and job performance (e.g., Bobko et al., 1999; Roth et al., 2011). These meta-analytic correlation matrices have been paired with information about racial/ethnic subgroup mean differences on the selection methods, which have then been used to provide answers to important questions related to the interplay of validity and adverse impact (AI) of personnel selection methods. In particular, these matrices have been used to quantify trade-offs in validity and diversity and test strategies to address these trade-offs. The validity–diversity trade-off reflects the finding that some of the selection methods that have the strongest validity for predicting job performance also have the largest racial/ethnic subgroup mean differences (Ployhart & Holtz, 2008). General mental ability (GMA) tests are the prototypic high-validity/large-mean differences selection method, and they have long been at the heart of the validity–diversity trade-off (Sackett et al., 2001). Given this dynamic, maximizing predictive validity by using the most valid selection methods (e.g., GMA tests) results in greater adverse impact, reducing diversity in hires; increasing diversity by using selection methods with less adverse impact results in lower predictive validity. To be clear, this does not mean that one cannot simultaneously have a diverse workforce and valid selection methods; rather, given what is known about validity and group mean differences for selection methods, increases toward one goal (e.g., maximizing diversity) result in some trade-off for the other goal. This is the essence of the validity–diversity trade-off and a lot of research has focused on

This article was published Online First May 2, 2024.

Christopher M. Berry https://orcid.org/0000-0001-5128-147X

An earlier version of this article was presented at the 2023 annual conference of the Society for Industrial and Organizational Psychology.

Christopher M. Berry played a lead role in data curation, investigation, methodology, project administration, supervision, writing–original draft, and writing–review and editing and a supporting role in formal analysis. Filip Lievens played a supporting role in conceptualization, data curation, methodology, and writing–review and editing. Charlene Zhang played a lead role in formal analysis and software and a supporting role in writing–review and editing. Paul R. Sackett played a supporting role in conceptualization, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Christopher M. Berry, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, 1275 East 10th Street, Bloomington, IN 47405, United States. Email: cmberry2@indiana.edu

various strategies to address it (e.g., Ployhart & Holtz, 2008; Sackett et al., 2001).

Much of this research has made use of the aforementioned meta-analytic correlation matrices. Two such meta-analytic correlation matrices have been most widely used in the personnel selection literature. The first is a matrix created by Bobko et al. (1999) of the relationships between GMA tests, structured interviews, conscientiousness tests, biodata, and job performance; this matrix also included Black–White mean differences on each selection method. The second is Roth et al.'s (2011) update to the Bobko et al. matrix. Among other changes, Roth et al. added integrity tests to the matrix and corrected the Bobko et al. correlations for direct range restriction. The updated Roth et al. matrix, or slightly edited versions of it (e.g., Song et al., 2017), have been used in a wide range of applications. For example, it has been used to study the effects of multipredictor selection systems on predictive bias (Dahlke & Sackett, 2022), the utility of compensatory versus multiple-hurdle selection models (Ock & Oswald, 2018), the use of heuristics in selection decisions (Luan et al., 2019), and the accuracy of dominance analysis (Braun et al., 2019) and range restriction correction methods (e.g., Fife et al., 2013). Most relevant to the present study, it has also been used in numerous studies utilizing Pareto-optimization to document the trade-offs in validity and diversity resulting from using various combinations of selection methods (e.g., De Corte et al., 2007, 2008; Song et al., 2017; Wee et al., 2014). A general conclusion from these Pareto-optimization studies has been that by using Pareto-optimal weighting of alternative predictors instead of regression weighting, it is possible to mitigate the adverse impact caused by GMA tests in personnel selection, but that this entails significant costs to criterion-related validity (i.e., the validity–diversity trade-off).

However, despite the value that the Roth et al.'s (2011) meta-analytic matrix has provided, and despite it using what at the time were state-of-the-art input values, we suggest there are at least three reasons why the matrix has become outdated. We believe the updates to this matrix will substantially change some of the conclusions of previous Pareto-optimization research using that matrix, particularly conclusions about the validity–diversity trade-off. First and most importantly, Sackett et al. (2022) demonstrated that, in large part due to inappropriate application of range restriction corrections, various meta-analyses overestimated the criterion-related validity of many personnel selection methods. Sackett et al. (2022) provided more appropriate, and often lower, estimates of criterion-related validity. As Roth et al. was carried out long before Sackett et al. (2022) was published, the inflated estimates of validity were included in Roth et al.'s meta-analytic matrix. Substituting in Sackett et al.'s (2022) new validity estimates is likely to significantly change conclusions. For example, Roth et al.'s validities for biodata, GMA tests, conscientiousness tests, structured interviews, and integrity tests were .32, .52, .22, .48, and .42, respectively. In comparison, Sackett et al.'s (2022) validities for those same selection methods were .38, .31, .19, .42, and .31. In particular, note that the validity for GMA tests is reduced from .52 to .31 and that GMA tests are no longer the strongest predictor of job performance, but instead lag behind structured interviews and biodata. A major driver of the validity–diversity trade-off has been that GMA tests had both exceptionally high validity (especially in comparison to most other selection methods) and large Black–White mean differences. Thus, to maximize validity of personnel selection models, GMA tests were given great weight, which resulted in substantial adverse impact; to mitigate this adverse impact, GMA tests were given less weight, but this substantially reduced validity. In light of Sackett et al.'s (2022) findings, this dynamic no longer holds. Yet, it remains unclear exactly what weight, if any, GMA tests should still have in personnel selection models, and what implications this has for the validity–diversity trade-off.

Second, in addition to updating the criterion-related validities in Roth et al.'s (2011) meta-analytic correlation matrix, there are also reasons to update the predictor intercorrelations and Black–White mean differences, as well. In some cases, there are new, better data available since Roth et al. For example, Dahlke and Sackett (2017) have since provided updated estimates of Black–White mean differences on some of the selection methods, so these improved estimates can be substituted in. In other cases, insights from Sackett et al. (2022) about when and how one should correct for range restriction lead us to revise some predictor intercorrelations in Roth et al. For example, Sackett et al. (2022) demonstrated that range restriction artifact distributions typically come from predictive validity studies, where range restriction can be sizable, but that applying range restriction corrections based on these predictive validity studies to concurrent validity studies, which make up the vast majority of samples in meta-analyses and are only minimally affected by range restriction, can result in substantial overcorrection. This issue, and others related to range restriction corrections, led to overestimates of some of the predictor intercorrelations in the Roth et al. matrix.

Third, although the selection methods included in the Roth et al. (2011) matrix represent some of the most widely used, traditional applied psychology selection methods, we believe there is value in adding another commonly used selection method to the matrix, namely situational judgment tests (SJTs). Like the other selection methods included in Roth et al.'s matrix, SJTs are one of the most common selection methods. For example, in a survey of 1,406 human resources professionals from companies around the world, 43% reported that their organizations used SJTs in personnel selection (Kantrowitz, 2014). Also like the other selection methods included in Roth et al.'s matrix, SJTs (and especially more generic SJTs, Lievens & Motowidlo, 2016) can be used with applicants regardless of whether they possess relevant job knowledge or experience, making them more widely applicable than other highly valid selection methods such as job knowledge tests, work samples, or assessment centers.[1] Moreover, SJTs typically have low to

[1] One might ask why only SJTs were added to the matrix, and not other selection methods that can be used regardless of applicant job knowledge/experience. Part of the answer is a practical consideration, namely that we wished to limit the matrix to a manageable size while still including a wide range of common selection methods, which we feel the current matrix with SJTs added achieves. Crucially, for any method added to the matrix the correlations with all other variables in the matrix are needed. We considered personality-based emotional intelligence measures as a possible addition to the matrix but were unable to obtain the full set of correlations with other predictors. One other addition we considered, but ultimately decided against, was the other Big Five personality traits in addition to just conscientiousness. However, because each of the other Big Five traits have such low validities compared to the other selection methods for predicting overall job performance, they would not play a significant role in the Pareto-optimal validity-maximization solutions; and because there were already selection methods with small Black–White mean differences (e.g., conscientiousness), additional selection methods with small mean differences were not needed for the Pareto-optimal diversity-maximization solutions. Thus, adding the rest of the Big Five was unlikely to impact any of the study results, and therefore the unnecessary complexity of adding them to the matrix outweighed any value they would provide.

moderate subgroup differences (Whetzel et al., 2008). Thus, adding SJTs to the matrix provides practitioners with useful information about what effects SJTs have on validity and diversity, when used in conjunction with the other popular selection methods included in Roth et al.'s original matrix. As such, we carried out new literature searches to identify studies relating SJTs to the other selection methods in the matrix.

This study makes two main contributions. First and most importantly, we will use the updated meta-analytic correlation matrix to carry out a series of analyses that will provide new insights on how the use of these selection methods impacts validity and diversity. For example, Sackett et al.'s (2022) validities for these selection methods are mostly lower than in Roth et al.'s (2011) matrix, and the rank order of the selection methods has changed. Does applied psychology simply have to resign itself to the idea that the criterion-related validities of our personnel selection methods are much lower than we previously thought, or can multiple-predictor composites yield validities that are comparable to the levels we expected before Sackett et al. (2022)? What effect will the changes in validity levels and rank order of selection methods have on the regression weights for each method, and what effect will this then have on validity and diversity implications of using such regression weights? When carrying out Pareto-optimization analyses that examine optimal trade-offs between validity and diversity, which predictors now emerge as the most important? Will GMA tests, long considered the most pivotal predictor when attempting to balance validity and diversity concerns, still play a substantial role? The present study will address each of these questions, and more.

A second contribution of the present study is simply the presentation of the new, updated meta-analytic correlation matrix of the relationships between these selection methods, overall job performance, and the Black–White mean differences on the selection methods. This updated matrix can be found in Table 1, alongside Roth et al.'s (2011) matrix. As mentioned above, Bobko et al.'s (1999) matrix and then Roth et al.'s matrix have been used in a wide range of studies on various topics related to personnel selection. The new matrix provided by the present study is based on recent insights about the validity of selection methods. So, we feel this updated matrix will be useful to future research on a broad range of personnel selection topics just like the previous meta-analytic matrices. In the following sections, we outline our updates to the meta-analytic correlation matrix and then describe the ways in which we will use the updated matrix to provide important, new knowledge and insights.

Updates to Selection Method Criterion-Related Validities

We refer readers to Appendix A for definitions of the six selection methods in the updated matrix. We relied on secondary data, drawing the estimates for the meta-analytic correlation matrix from existing studies. We drew the updated criterion-related validities for each of the selection methods from Sackett et al. (2022). These are operational validities, so they are corrected for range restriction, where applicable, and criterion measurement error using interrater reliability of .60 for supervisor ratings of overall job performance, but not for predictor measurement error. Sackett et al. (2022) provided details on how they arrived at each criterion-related validity estimate, so we direct readers to that article for those details. Here, we will just briefly note differences between the Roth et al. (2011) and Sackett et al. (2022) estimates, along with explanations for those differences.

Roth et al.'s (2011) biodata validity estimate was .32. In the Table 1 matrix we use Sackett et al.'s (2022) empirically keyed biodata estimate of .38. The modest difference is due to new data. That is, Roth et al.'s estimate was based on a study by Rothstein et al. (1990), Sackett et al.'s (2022) estimate was based on Rothstein et al. plus a new, independent meta-analysis by Speer et al. (2021).

Roth et al.'s (2011) GMA test validity estimate was .52, which was the average operational validity for medium complexity jobs from Hunter (1986) and Salgado et al.'s (2003) meta-analyses. Sackett et al.'s (2022) validity estimate of .31 was based on a wider range of meta-analyses, but the primary reason for the validity

Table 1

Predictor Intercorrelations, Criterion-Related Validities, and Black–White Subgroup d-Values for the Updated Matrix and Roth et al. (2011) Matrix

Variable	1	2	3	4	5	6
1. Biodata	—					
2. GMA tests	.13[a] (.08)/.37[b]	—				
3. Conscientiousness	.54[a] (.18)/.51[b]	.03[b] (.05)/.03[b]	—			
4. SI	*.21*[b] (.03)/.16[b]	.18[c] (.04)/.31[b]	*.08*[b] (.04)/.13[b]	—		
5. Integrity test	.25[b] (.11)/.25[b]	*.01*[b] (.11)/.02[b]	*.28*[b] (.11)/.34[b]	−.02[b] (.11)/−.02[b]	—	
6. SJT	.42[d] (.05)/NA	.29[e] (.17)/NA	.23[e] (.12)/NA[b]	.45[d] (.10)/NA	.16[d] (.10)/NA	—
Validities	.38[f] (.09)/.32[b]	.31[f] (.14)/.52[b]	.19[f] (.15)/.22[b]	.42[f] (.19)/.48[b]	.31[f] (.20)/.42[b]	.26[f] (.11)/NA
Black–White *d*-values	.32[g]/.57[b]	.79[f]/.72, .86[b]	−.07[h]/.06[b]	*.24*[b]/.32[b]	.10[i]/.04[b]	.37[i]/NA

Note. Numbers before the slashes are for the updated matrix; numbers after the slashes are from Roth et al. (2011) matrix. GMA tests have three Black–White *d*-values instead of two because Roth et al. used two *d*-values in their matrix (.72 for medium complexity jobs and .86 for low complexity jobs), while the present study used only one (.79, which is the average of Roth et al.'s two *d*-values). Numbers in plain text mean that value was taken as is from the source indicated by its superscript. Numbers in italics mean that value was taken from the source indicated by its subscript, but we changed that value based on new information. Numbers in parentheses are the standard deviations of each correlation with the estimated effects of relevant statistical artifacts removed. NA = not applicable; Roth et al.'s matrix did not include SJTs, so it does not have entries for these cells. GMA = general mental ability; SI = structured interview; SJT = situational judgment tests.

[a] Speer et al. (2022). [b] Roth et al. (2011). [c] Berry, Sackett, and Landers (2007). [d] Present study. [e] McDaniel et al. (2007). [f] Sackett et al. (2022). [g] Tenbrink et al. (2021). [h] Foldes et al. (2008). [i] Dahlke and Sackett (2017).

estimate being so much lower was Sackett et al.'s (2022) demonstration that there was minimal range restriction in previous meta-analyses such as Hunter and Salgado et al., and thus their large range restriction corrections resulted in substantial overestimates of validity.

Roth et al.'s (2011) conscientiousness test validity estimate was .22, while it was .19 in Sackett et al. (2022). The small difference in validities is mostly due to Sackett et al.'s (2022) inclusion of a larger number of meta-analyses than Roth et al.

Roth et al.'s (2011) structured interview validity estimate was .48, while it was .42 in Sackett et al. (2022). There were two reasons the validities differed. First, Sackett et al. (2022) included an updated meta-analysis (Huffcutt et al., 2014) that had been carried out since Roth et al. Second, Sackett et al. (2022) noted that the range restriction corrections that had been applied in the structured interview meta-analyses were inappropriate (see Sackett et al., 2022 for details) and provided more appropriate validity estimates.

Roth et al.'s (2011) integrity test validity estimate was .42 based on Ones et al. (1993), while it was .31 in Sackett et al. (2022). The major reason for the lower validity in Sackett et al. (2022) is that, in addition to Ones et al. (1993), they included a new meta-analysis of integrity test validity (Van Iddekinge et al., 2012) that had been carried out since Roth et al. and found considerably lower validity for integrity tests.

Roth et al. (2011) did not include SJTs in their meta-analytic matrix. Sackett et al.'s (2022) SJT validity estimate of .26 was based on McDaniel et al.'s (2007) meta-analysis.

Updates to Selection Method Intercorrelations and Black–White Mean Differences

Due to the existence of new data since the publication of Roth et al. (2011) and/or new insights (e.g., Sackett et al., 2022 insights about range restriction corrections), we also updated the selection method intercorrelations and Black–White mean differences in the meta-analytic correlation matrix. Given the focus on operational validity, these selection method intercorrelations and Black–White mean differences are corrected for range restriction, where applicable, but there were no corrections for predictor measurement error. With some noteworthy exceptions (e.g., the correlation between GMA tests and biodata, the Black–White mean difference on biodata, our addition of SJTs to the matrix), these updates resulted in relatively modest changes to the values in the matrix. Thus, we simply report the updated values in Table 1, and refer the interested reader to Appendix B for details on whether, how, and why we updated every entry in the matrix.

Answering Substantive Questions on the Basis of the Updated Meta-Analytic Correlation Matrix

As can be seen in the previous sections, there are mostly modest differences between the updated matrix and Roth et al.'s (2011) matrix in terms of selection method intercorrelations and Black–White mean differences, but more substantial differences in criterion-related validities, both in terms of levels of validity and rank order of the selection methods. This begs five questions that we will address in the following sections with various regression and Pareto-optimization analyses using this updated matrix.

Do Some Selection Methods Become More or Less Important Predictors When Their Shared Variance With the Other Selection Methods Is Accounted for?

Sackett et al. (2022) focused only on the bivariate correlations between each selection method and job performance. This highlights how well they can be expected to predict job performance, on average, when used in isolation. However, it does not account for the intercorrelations between the selection methods. Additionally, organizations do not typically use only a single selection method, so focusing only on the bivariate relationships with job performance does not address how well the selection methods can be expected to predict job performance when used together.

Method and Results

We used the updated meta-analytic matrix to regress job performance on the six selection methods: biodata, GMA tests, conscientiousness tests, structured interviews, integrity tests, and SJTs. We then carried out a dominance analysis in R to determine the relative weights for each selection method.

Table 2 provides a side-by-side comparison of the selection methods' bivariate criterion-related validities (taken from Table 1) to their standardized regression coefficients and relative weights from a dominance analysis when job performance is simultaneously regressed on all six selection methods.[2] The relative standing of biodata, GMA tests, structured interviews, and integrity tests is similar for the bivariate and multiple regression results, with the exception that structured interviews carry somewhat more weight compared to the other predictors in the multiple regression analysis than in the bivariate analysis. In contrast, the standing of conscientiousness tests and SJTs is substantially reduced in the multiple regression analysis. Both have very small relative weights; the regression weight for conscientiousness tests (−.04) is practically zero and the regression weight for SJTs is fairly negative (−.13).

Transparency and Openness

We adhered to the *Journal of Applied Psychology* methodological checklist. Details on whether, how, and why we updated every entry in the meta-analytic correlation matrix are provided in Appendix B. We conducted regression analyses using the R package psychmeta and dominance analyses using the R package dominanceanalysis. We ran the Pareto-optimization models (described in later sections of the article) using De Corte et al.'s (2023) multiple objective Pareto-optimization program. The Substantive Question 5 simulation was carried out using the R packages stats (R Core Team, 2022) and ParetoR (Song, 2022). A subset of the results in Tables 2 and 3 of the present study was presented in Sackett, Zhang, et al. (2023).

[2] A subset of the results in Table 2 was presented in Sackett, Zhang, et al. (2023).

Table 2

Comparison of the Selection Methods' Bivariate Criterion-Related Validities to Their Standardized Regression Coefficients and Dominance Analysis Relative Weights When Regressing Job Performance on the Selection Methods

Selection method	r	β	RW_{Raw}	$RW_{\%}$
Biodata	.38	.26	.0762	20.03
GMA tests	.31	.23	.0627	16.48
Conscientiousness	.19	−.04	.0117	3.06
Structured interviews	.42	.39	.1339	35.18
Integrity tests	.31	.28	.0751	19.73
Situational judgment tests	.26	−.13	.0210	5.52

Note. r = bivariate criterion-related validity for predicting job performance; β = standardized regression coefficient when job performance is regressed on all six selection methods; RW_{Raw} = raw relative weight; $RW_{\%}$ = relative weights rescaled as a percentage of predictable variance; GMA = general mental ability.

Can Combinations of Selection Methods Provide Levels of Criterion-Related Validity That Are More Comparable to What Was Expected Before Sackett et al. (2022)?

A key finding from Sackett et al. (2022) was that the criterion-related validity of most selection methods was considerably lower than the field previously thought. However, most selection methods are not used in isolation. Therefore, it is possible that some composites of multiple predictors will still have criterion-related validities comparable to those that were expected before Sackett et al. (2022).

Method and Results

We regressed job performance on all possible predictor combinations of the six selection methods using the old meta-analytic matrix (Roth et al., 2011) versus the updated matrix. The old matrix did not include SJTs, so we added the SJT validities and predictor intercorrelations to the old matrix to allow a fair comparison (see Table 1 for a side-by-side comparison of the two matrices). We computed mean multiple correlations (i.e., R) for one-, two-, three-, four-, five-, and six-predictor combinations. A handful of predictor combinations resulted in at least one of the predictors having a negative regression weight. As predictors with negative regression weights are unlikely to be used in applied settings, we set the weight for such predictors to zero.

Table 3 contains the multiple correlations of all possible predictor combinations of biodata, GMA tests, conscientiousness tests, structured interviews, integrity tests, and SJTs, from one selection method to all six.[3] As can be seen at the bottom of the table, using all six selection methods, the old meta-analytic correlation matrix (Roth et al., 2011) yielded a multiple correlation of .66, whereas the updated matrix yielded .61. Similarly, the mean multiple correlations yielded by the updated meta-analytic matrix were lower than that yielded by the old matrix for all possible one-, two-, three-, four-, and five-predictor combinations (see the Mean R rows in Table 3). In terms of magnitude, however, the multiple correlations using the updated matrix are only a few validity points lower than those with the old matrix. Thus, importantly, combinations of selection methods provide levels of criterion-related validity that are more comparable to what was expected before Sackett et al. (2022).

What Effects Will Excluding GMA Tests From the Selection Battery Have on Validity?

Prior to Sackett et al. (2022), the conventional wisdom was that GMA tests were one of the strongest predictors of job performance and that leaving GMA tests out of a selection battery would come at a steep cost to the criterion-related validity of that selection battery. However, Sackett et al. (2022) demonstrated that the criterion-related validity of GMA tests is considerably more modest than was previously thought, and that it even lags behind some of the other predictors in the meta-analytic matrix, such as structured interviews and biodata. This begs the question of whether criterion-related validity will still be substantially decreased when GMA tests are excluded from the selection battery. Thus, we will compare the validity of composites of selection methods including versus excluding GMA tests.

Method and Results

As described in the section above addressing Substantive Question 2, the analyses contributing to Table 3 produced multiple correlations for all possible combinations of the six selection methods. To highlight the impact of the much lower validity of GMA tests, we computed the mean multiple correlation of selection method combinations in Table 3 that contained GMA tests with that of selection method combinations without. These are listed in Table 3 for both the old meta-analytic correlation matrix and the updated matrix (see the "Mean R with GMA" and "Mean R without GMA" rows in the table). In order to hold constant the number of selection methods in each composite with and without GMA tests, we calculated the mean multiple correlations for all possible combinations of one selection method with and without GMA tests, all possible combinations of two selection methods with and without GMA tests, and so forth on up to all possible combinations of five selection methods with and without GMA tests (a six-predictor composite would, by definition, have to include GMA tests because there are only six selection methods). Using the old meta-analytic matrix, for composites of any number of selection methods, excluding GMA tests resulted in substantially lower validity (e.g., a mean multiple correlation of .59 for all three-predictor composites with GMA tests included, which reduced to .46 for all three-predictor composites with GMA tests excluded). Results are markedly different for the updated matrix, wherein for composites of any number of selection methods, the multiple correlations are virtually identical whether GMA tests are excluded or not. Thus, including GMA tests in the composites does not markedly improve criterion-related validity.

Using Pareto-Optimization to Provide Information About the Trade-Offs in Validity and Adverse Impact of Various Combinations of Selection Methods

De Corte and colleagues (e.g., De Corte et al., 2007, 2008) introduced Pareto-optimal weighting to the personnel selection

[3] A subset of the results in Table 3 was referenced in Sackett, Zhang, et al. (2023).

Table 3

Multiple Correlations With Job Performance Using All Possible Selection Method Combinations, Along With a Comparison of Combinations With and Without GMA Tests

Number of predictor	Selection method	R based on old matrix	R based on new matrix
1	BD	0.32	0.38
1	GMA	0.52	0.31
1	C	0.22	0.19
1	SI	0.48	0.42
1	I	0.20	0.31
1	SJT	0.26	0.26
	Mean R	0.33	0.31
	Mean R with GMA	0.52	0.31
	Mean R without GMA	0.30	0.31
2	BD + GMA	0.54	0.45
2	BD + C	0.33	0.38
2	BD + SI	0.54	0.52
2	BD + I	0.34	0.44
2	BD + SJT	0.35	0.40
2	GMA + C	0.56	0.36
2	GMA + SI	0.62	0.48
2	GMA + I	0.55	0.44
2	GMA + SJT	0.53	0.36
2	C + SI	0.51	0.45
2	C + I	0.26	0.33
2	C + SJT	0.31	0.29
2	SI + I	0.52	0.53
2	SI + SJT	0.48	0.43
2	I + SJT	0.31	0.38
	Mean R	0.45	0.41
	Mean R with GMA	0.56	0.42
	Mean R without GMA	0.39	0.41
3	BD + GMA + C	0.56	0.45
3	BD + GMA + SI	0.63	0.55
3	BD + GMA + I	0.56	0.51
3	BD + GMA + SJT	0.54	0.46
3	BD + C + SI	0.54	0.52
3	BD + C + I	0.35	0.44
3	BD + C + SJT	0.35	0.40
3	BD + SI + I	0.56	0.57
3	BD + SI + SJT	0.54	0.52
3	BD + I + SJT	0.37	0.45
3	GMA + C + SI	0.64	0.51
3	GMA + C + I	0.57	0.45
3	GMA + C + SJT	0.56	0.39
3	GMA + SI + I	0.65	0.58
3	GMA + SI + SJT	0.62	0.48
3	GMA + I + SJT	0.56	0.45
3	C + SI + I	0.53	0.53
3	C + SI + SJT	0.51	0.45
3	C + I + SJT	0.33	0.38
3	SI + I + SJT	0.52	0.53
	Mean R	0.53	0.48
	Mean R with GMA	0.59	0.48
	Mean R without GMA	0.46	0.48
4	BD + GMA + C + SI	0.64	0.55
4	BD + GMA + C + I	0.57	0.51
4	BD + GMA + C + SJT	0.56	0.46
4	BD + GMA + SI + I	0.65	0.61
4	BD + GMA + SI + SJT	0.63	0.55
4	BD + GMA + I + SJT	0.56	0.51
4	BD + C + SI + I	0.56	0.57
4	BD + C + SI + SJT	0.54	0.52
4	BD + C + I + SJT	0.37	0.45
4	BD + SI + I + SJT	0.56	0.57
4	GMA + C + SI + I	0.66	0.58
4	GMA + C + SI + SJT	0.64	0.51
4	GMA + C + I + SJT	0.58	0.46
4	GMA + SI + I + SJT	0.65	0.58

(*table continues*)

Table 3 (*continued*)

Number of predictor	Selection method	R based on old matrix	R based on new matrix
4	C + SI + I + SJT	0.53	0.53
	Mean R	0.58	0.53
	Mean R with GMA	0.62	0.53
	Mean R without GMA	0.51	0.53
5	BD + GMA + C + SI + I	0.66	0.61
5	BD + GMA + C + SI + SJT	0.64	0.55
5	BD + GMA + C + I + SJT	0.58	0.51
5	BD + GMA + SI + I + SJT	0.65	0.61
5	BD + C + SI + I + SJT	0.56	0.58
5	GMA + C + SI + I + SJT	0.66	0.57
	Mean R	0.62	0.57
	Mean R with GMA	0.64	0.57
	Mean R without GMA	0.56	0.57
6	BD + GMA + C + SI + I + SJT	0.66	0.61

Note. GMA = general mental ability; BD = biodata; C = conscientiousness tests; SI = structured interviews; I = integrity tests; SJT = situational judgment tests.

literature as a way to systematically identify how various predictors should be weighted when optimizing both criterion-related validity and adverse impact. That is, Pareto-optimal weighting of selection methods can be used to quantify the minimum level of adverse impact at a given level of validity or the maximum level of validity at a given level of adverse impact. In contrast with unit weights or regression weights typically used, Pareto-optimal weighting schemes are determined by simultaneously optimizing multiple criteria (e.g., validity and adverse impact).

We carry out Pareto-optimization analyses to answer questions such as "how much weight does each selection method carry in validity-maximizing solutions versus solutions resulting in an adverse impact ratio that would satisfy an adverse impact ratio of .80+?"[4] and "how much is validity compromised for solutions that greatly reduce adverse impact and how much is adverse impact increased for solutions that greatly improve validity?" We also carry out Pareto-optimization analyses that address whether and how much the answers to these questions change in various common scenarios. For example, what if one only focuses on "off-the-shelf" selection methods such as GMA tests, conscientiousness tests, and integrity tests; excluding more "customized" selection methods such as structured interviews or biodata that require substantially more time and resources? What if one only focuses on selection methods that are suitable for mass screening of applicants? That is, structured interviews are an important predictor of job performance, but are not generally feasible for mass screening. Most companies might mass screen with some of the other predictors in the meta-analytic matrix, with only those applicants who pass the screen moving on to the structured interview.

Method

We used De Corte et al.'s (2023) program for the analyses. This program expands on approaches used in prior work in the personnel selection field (e.g., De Corte et al., 2007; Song et al., 2017) which make use of Das and Dennis's (1998) normal boundary intersection (NBI) approach in identifying a Pareto front. De Corte et al. noted that in expanding Pareto-optimization approaches to examining more than two subgroups and/or more than two criteria, the NBI approach has limitations, and thus they introduced a new hybrid approach that blends the NBI method of Das and Dennis (1998), the enhanced normalized normal constraint method of Messac and Mattson (2004) and the successive boundary generation method of Mueller-Gritschneder et al. (2009). Appendix A of De Corte et al. details this novel hybrid procedure.

We conducted a series of Pareto-optimization analyses. First, we included all six predictors in a single-stage Pareto-optimization model. A single-stage model considers all six predictors for possible simultaneous use in a selection composite. The Pareto-optimization process generates a set of possible solutions, on a continuum from a validity-maximizing solution to a diversity-maximizing solution. For each solution, there is a resulting quality (composite validity or expected performance of the selected) and diversity value (here expressed as an adverse impact ratio), as well as a set of the weights applied to each predictor to produce the solution. Next, in-person structured interviews are not generally amenable to mass screening in high applicant volume settings. So, to represent the result of an initial screening including only predictors amenable to use in mass screening, we reran the single-stage Pareto-optimization model with all predictors other than structured interviews, as the interview is generally used with a smaller set of candidates who passed initial screens. Then, we ran a two-stage Pareto-optimization model with all predictors other than structured interviews at the first stage and structured interviews at the second stage. Last, we ran a single-stage Pareto-optimization model with only the off-the-shelf selection methods: GMA tests, conscientiousness tests, and integrity tests.

For all models, we used a Black-White standardized mean job performance difference of .38 (Roth et al., 2011), a minority proportion of .15 as the current U.S. workforce is 13% Black and 18% Hispanic (U.S. Bureau of Labor Statistics, 2021), and an overall selection ratio of .20 for all models.[5] For the two-stage model, the selection ratio was .50 for Stage 1 and .40 for Stage 2. Results for

[4] We focus on an adverse impact ratio of .80 because this is the benchmark the Uniform Guidelines on Employee Selection Procedures suggests for determining whether there is adverse impact. Of course, an adverse impact ratio is a continuum and the choice of the best solution will depend on the specifics of the situation and professional judgment.

[5] As described in the Results, we also carried out sensitivity analyses using a different selection ratio to ensure conclusions were not affected by the choice of selection ratio.

single-stage models are presented with the composite validity as the quality criterion. For multiple-stage models, we used expected performance of the selected applicants as the quality criterion instead, as a single composite validity value cannot be estimated.

Results

Single-Stage Pareto-Optimization Models. Table 4 contains the single-stage Pareto-optimization model using all six selection methods based on the old meta-analytic matrix. Using the old matrix, the validity-maximizing solution yields a validity composite of .66 and an adverse impact ratio of .35. The validity-maximizing solution gives the greatest weight to GMA tests and structured interviews, in that order; with smaller weights for integrity tests, conscientiousness tests, and biodata; and a zero weight for SJTs. When the weight of GMA tests gets dropped to 0, validity decreases to .46 and the adverse impact ratio increases to .78. To get the adverse impact ratio above the threshold of .80, validity drops further to .44 and conscientiousness tests, structured interviews, and to a lesser extent integrity tests have the greatest weights; with zero weights given to other predictors.

In contrast, Table 5 contains the single-stage Pareto-optimization model using the updated meta-analytic matrix and group difference values. The validity-maximizing solution has a validity composite of .61 and an adverse impact ratio of .42. It gives the largest weight to structured interviews, with smaller weights given to integrity tests, GMA tests, and biodata, in that order, and zero weights to conscientiousness tests and SJTs. When the weight of GMA tests gets dropped to 0, validity only decreases to .56 and the adverse impact ratio increases to .67. To get the adverse impact ratio above the threshold of .80, validity drops further to .48 and this solution gives similar weights to conscientiousness tests, structured interviews, and integrity tests (in that order), and zero weights to all other predictors.

Table 6 contains the single-stage Pareto-optimization model with all predictors other than structured interviews, representing the result of an initial screening using only predictors amenable to use in mass screening (i.e., a combination of biodata, GMA tests, conscientiousness tests, integrity tests, and SJTs). The validity-maximizing solution produces a validity of .51 and an adverse impact ratio of .37 and gives greatest and similar weights to biodata, GMA tests, and integrity tests; a small weight to SJTs, and a zero weight to conscientiousness tests. The validity and adverse impact ratios become .42 and .74, respectively, when GMA tests' weight drops to zero. The solution that has an adverse impact ratio just over .80 yields a validity of .39, and gives the greatest weight to integrity tests, moderate weights to conscientiousness tests and biodata, and zero weights to the other predictors.

Table 7 contains the single-stage Pareto-optimization model with only the three "off-the-shelf" selection methods: GMA tests, conscientiousness tests, and integrity tests. The three off-the-shelf methods together achieve a maximum validity of .45, which corresponds to an adverse impact ratio of .40; the greatest weight is given to GMA tests and then integrity tests, with a relatively small weight to conscientiousness tests. When GMA tests' weight drops to zero, validity reduces to .32 and the adverse impact ratio increases to .96. To reach the .80 adverse impact threshold, validity drops to .36, with integrity tests getting the largest weight, followed by conscientiousness tests, and a relatively small weight for GMA tests.

Figure 1 synthesizes all the single-stage Pareto-optimization model results. Comparing the six-predictor model based on the updated matrix with that based on the old matrix (green and red lines, respectively), there is a smaller rate of reductions in composite validity as the adverse impact ratio increases throughout most of the curve for the updated matrix (i.e., the curve is flatter overall), and despite the validities for the individual selection

Table 4
Single-Stage Pareto-Optimization Model Using All Six Selection Methods Based on the Old Meta-Analytic Matrix

		Predictor weight					
Composite validity	Adverse impact ratio	Biodata	GMA test	Conscientiousness test	Structured interview	Integrity test	Situational judgment test
.66	0.35	.03	.39	0.09	.34	.16	.00
.65	0.41	.00	.31	0.17	.35	.17	.00
.63	0.46	.00	.25	0.22	.36	.18	.00
.61	0.50	.00	.20	0.25	.37	.18	.00
.59	0.54	.00	.17	0.28	.37	.18	.00
.57	0.58	.00	.13	0.30	.38	.19	.00
.55	0.62	.00	.10	0.32	.38	.19	.00
.53	0.66	.00	.07	0.35	.39	.19	.00
.51	0.70	.00	.04	0.37	.39	.20	.00
.48	0.74	.00	.01	0.39	.40	.20	.00
.46	0.78	.00	.00	0.43	.37	.20	.00
.44	0.82	.00	.00	0.48	.33	.19	.00
.42	0.86	.00	.00	0.52	.29	.19	.00
.39	0.89	.00	.00	0.56	.26	.18	.00
.37	0.93	.00	.00	0.60	.22	.18	.00
.35	0.96	.00	.00	0.65	.18	.17	.00
.32	1.00	.00	.00	0.69	.14	.16	.00
.30	1.03	.00	.00	0.74	.10	.16	.00
.27	1.07	.00	.00	0.79	.06	.15	.00
.25	1.10	.00	.00	0.84	.02	.15	.00
.22	1.13	.00	.00	1.00	.00	.00	.00

Note. GMA = general mental ability.

Table 5
Single-Stage Pareto-Optimization Model Using All Six Selection Methods Based on the Updated Meta-Analytic Matrix

		Predictor weight					
Composite validity	Adverse impact ratio	Biodata	GMA test	Conscientiousness test	Structured interview	Integrity test	Situational judgment test
.61	0.42	.20	.20	0.00	.34	.26	.00
.60	0.48	.21	.14	0.00	.37	.29	.00
.59	0.53	.20	.09	0.01	.40	.30	.00
.58	0.58	.19	.05	0.03	.41	.31	.00
.57	0.63	.19	.01	0.05	.43	.32	.00
.56	0.67	.13	.00	0.11	.43	.33	.00
.54	0.71	.07	.00	0.17	.43	.33	.00
.52	0.75	.01	.00	0.22	.43	.34	.00
.50	0.79	.00	.00	0.29	.39	.32	.00
.48	0.82	.00	.00	0.35	.34	.31	.00
.45	0.85	.00	.00	0.40	.30	.29	.00
.43	0.88	.00	.00	0.45	.27	.28	.00
.40	0.91	.00	.00	0.50	.23	.27	.00
.38	0.93	.00	.00	0.55	.20	.25	.00
.35	0.96	.00	.00	0.59	.17	.24	.00
.33	0.98	.00	.00	0.64	.13	.23	.00
.30	1.01	.00	.00	0.69	.09	.21	.00
.27	1.03	.00	.00	0.75	.05	.20	.00
.25	1.06	.00	.00	0.80	.01	.18	.00
.22	1.08	.00	.00	0.89	.00	.11	.00
.19	1.10	.00	.00	1.00	.00	.00	.00

Note. GMA = general mental ability.

methods being overall smaller than in the old matrix, the new matrix produces composite validity higher than the old matrix at most adverse impact ratios (i.e., the portions where the green line is above the red line). This primarily reflects the impact of the inflated validity estimate for GMA tests in the old matrix and indicates that the validity–diversity trade-off is less severe than previously thought. Comparing the three models using the updated matrix, reducing the number of selection methods decreases composite validity. Excluding structured interviews substantially reduces overall validity for most of the curve (blue line). Further excluding biodata and SJTs reduces validity even more, although to a smaller extent (purple line).

Table 6
Single-Stage Pareto-Optimization Model Using All Selection Methods Other Than Structured Interviews for Initial Screening Based on the Updated Meta-Analytic Matrix

		Predictor weight				
Composite validity	Adverse impact ratio	Biodata	GMA test	Conscientiousness test	Integrity test	Situational judgment test
.51	0.37	.33	.32	0.00	.30	.05
.50	0.43	.37	.24	0.00	.35	.05
.49	0.48	.38	.19	0.01	.38	.04
.48	0.53	.39	.14	0.03	.40	.04
.47	0.57	.39	.11	0.05	.41	.03
.46	0.62	.39	.07	0.07	.43	.03
.44	0.66	.40	.04	0.09	.45	.03
.43	0.70	.40	.01	0.11	.46	.02
.42	0.74	.37	.00	0.16	.47	.00
.40	0.78	.30	.00	0.22	.48	.00
.39	0.82	.24	.00	0.28	.48	.00
.37	0.86	.18	.00	0.34	.48	.00
.35	0.89	.13	.00	0.39	.49	.00
.34	0.93	.07	.00	0.44	.49	.00
.32	0.96	.02	.00	0.48	.49	.00
.30	0.99	.00	.00	0.56	.44	.00
.28	1.02	.00	.00	0.65	.35	.00
.26	1.04	.00	.00	0.74	.26	.00
.24	1.06	.00	.00	0.82	.18	.00
.21	1.08	.00	.00	0.91	.09	.00
.19	1.10	.00	.00	1.00	.00	.00

Note. GMA = general mental ability.

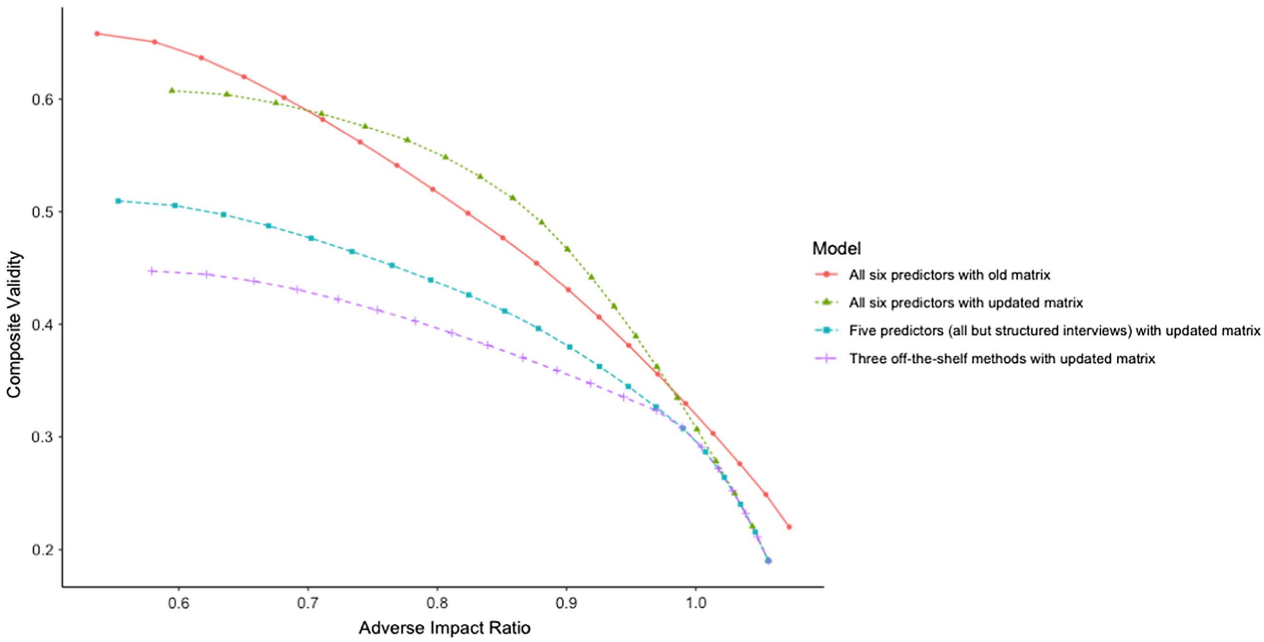
Table 7
Single-Stage Pareto-Optimization Model With Three Off-the-Shelf Selection Methods Based on the Updated Meta-Analytic Matrix

		Predictor weight		
Composite validity	Adverse impact ratio	GMA test	Conscientiousness test	Integrity test
.45	0.40	.44	0.15	.41
.44	0.46	.37	0.19	.44
.44	0.51	.33	0.22	.46
.43	0.55	.29	0.24	.47
.42	0.60	.25	0.26	.49
.41	0.64	.22	0.28	.50
.40	0.68	.19	0.29	.52
.39	0.72	.16	0.31	.53
.37	0.76	.13	0.33	.54
.36	0.80	.10	0.34	.56
.35	0.84	.07	0.36	.57
.34	0.88	.05	0.37	.58
.33	0.92	.02	0.39	.59
.32	0.96	.00	0.45	.55
.30	0.99	.00	0.55	.45
.29	1.01	.00	0.63	.37
.27	1.03	.00	0.71	.29
.25	1.05	.00	0.78	.22
.23	1.07	.00	0.85	.15
.21	1.09	.00	0.92	.08
.19	1.10	.00	1.00	.00

Note. GMA = general mental ability.

Two-Stage Pareto-Optimization Model. Table 8 contains the two-stage Pareto-optimization model with all predictors other than structured interviews used for Stage 1 and then structured interviews used for Stage 2. Because there is no clear way to express a single validity value for a multistage selection system, the expected performance of the selected is used as the quality criterion. The validity-maximizing solution yields an expected performance of .72 with an adverse impact ratio of .49 and gives greatest weight to integrity tests, moderate and similar weights to GMA tests and biodata, and zero weights to the other predictors. When the weight of

Figure 1
Composite Validity–Adverse Impact Ratio Curves for All Single-Stage Pareto-Optimization Models



Note. See the online article for the color version of this figure.

Table 8

Two-Stage Pareto-Optimization Model With All Predictors Other Than Structured Interviews Used for Stage 1 and Structured Interviews Used for Stage 2 Based on the Updated Meta-Analytic Matrix

		Predictor weight				
Expected performance of the selected	Adverse impact ratio	Biodata	GMA test	Conscientiousness test	Integrity test	Situational judgment test
.72	0.49	.29	.31	0.00	.40	.00
.72	0.52	.32	.24	0.00	.44	.00
.71	0.55	.34	.19	0.00	.47	.00
.71	0.57	.34	.15	0.02	.49	.00
.70	0.59	.34	.11	0.04	.50	.00
.69	0.62	.34	.08	0.06	.52	.00
.68	0.64	.34	.04	0.08	.53	.00
.68	0.66	.34	.01	0.09	.55	.00
.67	0.68	.30	.00	0.15	.56	.00
.66	0.70	.23	.00	0.21	.56	.00
.65	0.71	.17	.00	0.27	.56	.00
.63	0.73	.11	.00	0.33	.57	.00
.62	0.74	.06	.00	0.37	.57	.00
.61	0.76	.00	.00	0.42	.57	.00
.60	0.77	.00	.00	0.52	.48	.00
.58	0.78	.00	.00	0.61	.39	.00
.57	0.79	.00	.00	0.68	.32	.00
.55	0.80	.00	.00	0.76	.24	.00
.53	0.81	.00	.00	0.84	.16	.00
.52	0.82	.00	.00	0.91	.09	.00
.50	0.82	.00	.00	1.00	.00	.00

Note. GMA = general mental ability.

GMA tests is dropped to zero, expected performance changes to .67 with an adverse impact ratio of .68. Expected performance drops further to .55 to meet the adverse impact threshold of .80, with the greatest weight given to conscientiousness tests, a moderate weight given to integrity tests, and zero weights for all other predictors. Thus, the general conclusions from the single-stage model including all six selection methods hold in the perhaps more realistic scenario wherein only a subset of applicants move on to the in-person structured interviews after passing initial screening with the other less time- and resource-intensive selection methods.

Sensitivity Analysis. To test whether our conclusions were affected by the choice to use a selection ratio of .20 in the above Pareto-optimization analyses, we also repeated all Pareto-optimization analyses using an overall selection ratio of .50 for the single-stage models, and .70 for the first stage and .71 for the second stage of the two-stage model. While a higher selection ratio results in larger adverse impact ratios, individual solutions (i.e., predictor weights at each level of validity) remain essentially the same. Therefore, conclusions about the validity–diversity trade-off are the same regardless of the selection ratio used. Results are available upon request.

How Much Do Results Change When Variability Around the Meta-Analytic Means Is Modeled?

The analyses addressing Substantive Question 4 relied on a single meta-analytic correlation matrix that includes mean meta-analytic values for the selection methods' validities, intercorrelations, and Black–White mean differences. This is valuable because it gets at what one can expect, on average. However, there is variability around each of those meta-analytic averages. That is, even after subtracting out the effects of statistical artifacts, some of the variance across primary studies for each of these meta-analytic estimates remains unexplained. For example, although the predictor with the strongest average validity is structured interviews (validity of .42), the 80% credibility interval around that validity ranges from .24 to .66 (Sackett et al., 2022). This likely reflects at least in part that structured interviews (along with some of the other selection methods used in the present study such as biodata and SJTs) are methods that can be designed well or poorly or to measure constructs that are highly or less predictive of overall job performance. Similarly, the credibility interval around GMA tests' validity of .31 ranges from .13 to .49. Where validity falls in that range may be a function of things such as the specificity of the cognitive ability (or abilities) being tested, the type(s) of cognitive ability required by the job, and the choice of criterion. So, there will be some settings in which GMA tests could have higher validity than structured interviews (i.e., below average validity for structured interviews paired with above average validity for GMA tests). This is true of all the selection methods. How will results change if this variability is modeled? For example, although GMA tests received moderate weight at best in the Pareto-optimization analyses when using the meta-analytic means matrix, how often might GMA tests receive greater weight if variability around those meta-analytic means is modeled? We address each of these issues in our investigation of Substantive Question 5.

Method

We first attempted to locate variability estimates for each of the entries in the meta-analytic correlation matrix (Table 1). Variability estimates were not reported for the majority of meta-analyses of the Black–White mean differences, so we did not model variability in Black–White mean differences. We were able to locate variability estimates for most of the selection method validities (SD_ρ, the residual standard deviation of the operational validity after

subtracting out the expected effects of sampling error, criterion reliability, and range restriction where appropriate) and intercorrelations (SD_{Res}, the residual standard deviation around the intercorrelation after subtracting out only the expected effects of sampling error because there are no corrections for predictor measurement error in operational validity), and thus we modeled variability around these values in the matrix. See Table 1 for each of the estimates of SD_{ρ} and SD_{Res} (i.e., numbers in parentheses). Each of the estimates of SD_{ρ} for the selection methods were taken from Table 3 of Sackett et al. (2022). See Appendix B for descriptions of how we arrived at each of the SD_{Res} estimates for the selection method intercorrelations.

We conducted a 10,000-iteration simulation on the single-stage Pareto-optimization model with all six predictors (i.e., the model in Table 5). For each iteration, each value in the correlation matrix was randomly drawn based on a normal distribution with the respective meta-analytic mean and standard deviation using the R package stats (R Core Team, 2022), which resulted in 10,000 different matrices. We then used these correlation matrices as input to conduct 10,000 Pareto-optimization models using the R package ParetoR (Song, 2022). The Black–White mean differences, proportion of minority, and selection ratio remained constant for all iterations. Each of the 10,000 Pareto-optimization models produced 21 solutions, ranging from validity-maximizing to diversity-maximizing. We saved all output from each solution including composite validity, adverse impact ratio, as well as weights for the six predictors. The code for this simulation is included in Supplemental Material B.

Results

Table 9 lists the number of times out of the 10,000 iterations that each selection method was at each of the six ranks for the validity-maximizing (top half of the table) and adverse impact ratio ≥.80 (bottom half of the table) solutions. For example, for the validity-maximizing solutions, structured interviews had the greatest weight and therefore the top rank (i.e., rank of 1) 4,245 times (42.45%).

Table 9

Number of Times Out of 10,000 That Each Selection Method Was at Ranks 1–6 in the Pareto-Optimal Validity-Maximizing Solutions Versus the Adverse Impact Ratio ≥.80 Solutions

	Rank					
Selection method	1	2	3	4	5	6
Validity-maximizing solution						
Biodata	1,196	1,854	2,150	2,082	2,384	334
GMA tests	1,527	2,249	2,938	2,000	1,209	77
Conscientiousness	324	697	1,057	1,385	5,700	837
Structured interviews	4,245	2,275	1,521	974	868	117
Integrity tests	2,555	2,611	1,608	1,448	1,598	180
SJTs	178	325	611	933	6,364	1,589
Adverse impact ratio ≥.80 solution						
Biodata	92	837	1,117	2,271	5,631	334
GMA tests	0	79	391	2,252	7,211	77
Conscientiousness	2,752	4,688	1,977	435	147	837
Structured interviews	3,134	3,563	1,694	725	862	117
Integrity tests	4,046	626	2,713	1,561	1,042	180
SJTs	0	221	360	1,933	7,352	1,589

Note. GMA = general mental ability; SJT = situational judgment tests.

GMA tests had the highest rank 1,527 times (15.27%), demonstrating that because of variability around the meta-analytic estimates, it is possible for GMA tests (and each of the other predictors) to get the greatest weight in the validity-maximizing solutions at least some of the time. For the adverse impact ratio ≥.80 solutions, integrity tests were the most common top-ranked method, with structured interviews and conscientiousness also often being top-ranked. GMA tests were never the top-ranked method in any of the 10,000 adverse impact ≥.80 solutions.

Discussion

Because of the availability of new meta-analytic estimates, and particularly because of the insights provided by Sackett et al. (2022) about overcorrection for range restriction in selection method meta-analyses, we provided an updated meta-analytic correlation matrix of the criterion-related validity, intercorrelations, and Black–White mean differences for six commonly used selection methods. We then used that matrix to provide new and important practical and theoretical insights, particularly related to the role of GMA tests in the validity–diversity trade-off.

Theoretical and Practical Implications

Personnel selection researchers and practitioners have long grappled with the validity–diversity trade-off. At the heart of this trade-off was the supposed reality that one of the most valid selection methods, GMA tests, also resulted in large race/ethnicity subgroup mean differences (Ployhart & Holtz, 2008; Sackett et al., 2001). So, the trade-off was that if one wanted to reduce adverse impact by limiting the weight given to GMA tests in personnel selection, this would come at a steep validity cost. The results of the present study show this is no longer the case and was primarily an artifact of the inflated criterion-related validity of GMA tests due to inappropriate range restriction corrections in past meta-analyses. When the updated meta-analytic correlation matrix is used, with Sackett et al.'s (2022) lower criterion-related validity estimate for GMA tests, it shows that there is nothing special about GMA tests, at least in the context of predicting overall job performance. GMA tests have middling criterion-related validity compared to the other selection methods. Excluding GMA tests from the selection battery has little to no effect on criterion-related validity, but substantially reduces adverse impact. There are small exceptions to this conclusion. For example, in the single-stage Pareto-optimization analyses that did not include structured interviews or that only focused on "off-the-shelf" selection methods, there were appreciable decreases in validity if GMA tests were completely excluded from the selection batteries (from .51 to .42 and from .45 to .32, respectively). However, for most purposes, and especially if structured interviews are a part of the selection battery, simply excluding GMA tests has almost no validity downside and results in improved adverse impact.

Another noteworthy exception to this conclusion is when variability was modeled around the meta-analytic averages in the correlation matrix (Substantive Question 5). For example, the results in Table 9 showed that GMA tests were the selection method with the largest weight in the validity-maximizing solutions about 15% of the time and were one of the top two ranked selection methods about 38% of the time (i.e., 3,776 out of 10,000 iterations). Thus, simply due to the substantial variability around the meta-analytic estimates

in the matrix, there are certainly scenarios in which GMA tests will carry substantial weight, at least in validity-maximizing solutions (in adverse impact ratio greater than .80 solutions, GMA tests were never the selection method with the largest weight and were usually the fourth- or fifth-ranked selection method). This highlights the importance of attending to variability, and not just the mean, in meta-analytic estimates.

However, there are two important caveats to these Substantive Question 5 results. First, it has long been lamented (e.g., Berry, 2015) that much of what we know about the validity of GMA tests is based on very old data that is mostly from the 20th century, with much of it even from the 1980s or before. A recent, updated meta-analysis of the validity of GMA tests using only studies carried out in the 21st century (Sackett, Demeke, et al., 2023) found a somewhat lower operational validity and SD_ρ (operational validity of .22, with SD_ρ of 0.11) than in the Sackett et al. (2022) values we used in the present study (operational validity of .31 and SD_ρ of 0.14), which were based on much older data, mostly from the 20th century. Sackett, Demeke, et al. (2023) speculated that the lower validity in the 21st century validity studies may be due to the increased emphasis on teamwork and the resulting interpersonal aspects of work compared to the validity studies of the 20th century which were dominated by manufacturing jobs in which the focus was mostly on the quantity and quality of task performance. We carried out the same Substantive Question 5 analysis, but using the 21st century validity and SD_ρ for GMA tests instead of the older Sackett et al. (2022) values and results changed markedly; full results are in Supplemental Table S1. In this case, GMA tests were only the top-ranked selection method in the validity-maximizing solutions 4% of the time (i.e., 400 out of 10,000 iterations) and were one of the top two ranked selection methods about 17% of the time (1,684 out of 10,000 iterations); they were never the top-ranked selection method in the adverse impact ratio greater than or equal to .80 solutions and were most commonly the fifth-ranked selection method. Thus, when one focuses on more contemporary validity data for GMA tests, they carry even less weight.

Second, we note that it is not clear that this variability in the Pareto-optimal solutions demonstrated in the Substantive Question 5 analyses is actionable in practice. Absent knowing when and why one should expect larger versus smaller validities around the meta-analytic averages, one would typically have to rely on the mean validity estimates in the design of selection systems. If meta-analyses reveal useful moderators of validity (e.g., job complexity in the relationship between GMA tests and job performance), then users can and should insert validities applicable to their setting when carrying out Pareto-optimization analyses. This highlights the importance of continuing to identify the reasons (e.g., moderators) for why there is so much variability around some meta-analytic estimates. But, again, absent knowing what those moderators are, one must rely on the mean validity estimates, and our results show that GMA tests generally carry little weight in the results based on meta-analytic means. In all, contrary to prior belief, GMA tests are generally no longer a driving factor in the validity–diversity trade-off.

However, this does not necessarily mean there is no longer a validity–diversity trade-off. For example, although giving GMA tests a zero weight in the Pareto-optimization analysis including all six selection methods resulted in an improvement in the adverse impact ratio, it did not result in the adverse impact ratio reaching the rule of thumb adverse impact ratio of .80. This required a decrease in validity from .61 in the validity-maximizing solution to .48 in the solution that resulted in the adverse impact ratio being above .80. So, getting the adverse impact ratio to .80 still results in some cost to validity, even if this cost is not as steep as when using the old meta-analytic correlation matrix. This is because not all adverse impact was a result of GMA tests. Table 1 shows that there are also mean differences, albeit smaller ones, in favor of the White subgroup on some of the other selection methods, most notably biodata. So, while the validity-maximizing solution gave substantial weight to biodata, GMA tests, structured interviews, and integrity tests; the diversity-enhancing solutions gave less (and eventually zero) weight to biodata and GMA tests, and increasing weight to structured interviews, integrity tests, and especially conscientiousness tests, which have the lowest validity of the six selection methods. Patterns were similar in most of the other Pareto-optimization analyses, particularly in the two-stage analysis that models what we view as probably the most prototypic selection battery: one that uses selection methods other than interviews as an initial screen, with applicants passing that screen moving on to an interview. Thus, although the present study's results do not completely resolve the validity–diversity trade-off, in that there is still some validity cost for reduced adverse impact, it does show that those validity costs are not as steep as was previously thought.

This shifts the emphasis of the validity–diversity trade-off away from the role of GMA tests and toward the role that other selection methods play in this relatively smaller validity–diversity trade-off. For example, Ployhart and Holtz (2008) evaluated 16 different strategies for addressing the validity–diversity trade-off. Our results have the most direct implications for two strategies that Ployhart and Holtz concluded were the most effective: (a) using alternative measurement methods and (b) assessing the entire range of knowledge, skills, abilities, and other constructs (KSAOs). Regarding (a), our results show this strategy is now more effective than was previously thought. Although Ployhart and Holtz suggested criterion-related validity may be lower than for GMA tests when using alternative predictors, we show this is not the case. Thus, this suggests that an important new focus of research on resolving the validity–diversity trade-off should be understanding why these "alternative methods" cause adverse impact. There are numerous reasons the alternative methods might cause adverse impact (e.g., bias in interview ratings, biodata inventories including items for which there is not equal access, the cognitive load of the alternative predictors). Although the data in the present study cannot shed light on the first two example reasons, there are signs in the present study's data that cognitive load plays a role. As can be seen in Table 1, many of the other selection methods have nonzero relationships with GMA tests and also have the largest Black–White *d*-values. The two selection methods with the smallest correlations with GMA tests (conscientiousness and integrity tests) also have the smallest Black–White *d*-values. In fact, there is a correlation of .89 between the five alternative methods' Black–White *d*-values and their correlations with GMA tests. This is in line with Dahlke and Sackett's (2017) similar correlation of .84 for a wider range of selection methods. This also suggests that another important new focus of research on resolving the validity–diversity trade-off should be understanding ways to mitigate the effects of the "alternative methods" on diversity outcomes. Relatedly, these selection methods have long been referred to as "alternative methods" because they were thought of as alternatives to the default with the highest validity: GMA tests. Our results demonstrate that these

"alternative methods" actually should be thought of as the defaults, with GMA tests being an "alternative method" that should only be considered for use in very specific circumstances (e.g., when only off-the-shelf selection methods are an option). Regarding (b), assessing the full range of KSAOs remains an important strategy for mitigating the validity–diversity trade-off. However, our results show that for most purposes, assessing the full range of KSAOs that meaningfully predict overall job performance can be done without GMA tests.

Our results also provide guidance to those designing selection systems. Selection systems emphasizing biodata, structured interviews, integrity tests, and perhaps GMA tests will maximize validity. If one wants more of a balance of validity and diversity, reducing the role of biodata and GMA tests is necessary, while putting greater emphasis on conscientiousness tests, structured interviews, and integrity tests. Structured interviews, in particular, play a key role. Of the six selection methods, they have the highest criterion-related validity, were the method given the highest weights in the Pareto-optimization validity-maximization solutions in which they were included, and even carried substantial weight in solutions that resulted in adverse impact ratios above .80. Of course, using structured interviews in an initial screening battery may prove difficult (although future research could investigate whether more modern versions such as asynchronous video interviews might be more feasible, while still exhibiting the strong validity of in-person structured interviews), but their value at some stage of the selection process both for validity and diversity outcomes is substantial (e.g., see the two-stage Pareto-optimization model).

Interestingly, SJTs carried zero weight in most of the Pareto-optimization solutions. This does not necessarily mean that SJTs have no value in personnel selection. As can be seen in Table 1, on their own SJTs have appreciable criterion-related validity. However, among the six selection methods, SJTs have one of the lowest validities and one of the highest Black–White mean differences. So, in selection batteries with the other selection methods, SJTs did little to improve validity or adverse impact. In fact, in the Table 2 multiple regression analysis, despite SJTs' positive bivariate correlation with job performance ($r = .26$), SJTs' standardized regression coefficient was weak and negative ($\beta = -.13$). This is likely because SJTs were one of the weakest bivariate predictors, with only conscientiousness tests being weaker, and SJTs had their strongest bivariate relationships with the two strongest bivariate predictors: structured interviews and biodata (correlations of .45 and .42, respectively, with SJTs). So, the two strongest bivariate predictors share quite a bit of variance with SJTs. This pattern of results suggests that SJTs' positive bivariate relationship with job performance was due to it sharing variance with structured interviews and biodata, and once that variance had already been accounted for by structured interviews and biodata, SJTs became a weak (negative) predictor. This may be because the things that would help one successfully respond to an SJT have much in common with the things tapped by structured interviews and biodata. For example, SJTs present applicants with scenarios and ask them to determine what they would or should do. This is also common practice in structured interviews. In particular, a common form of structured interview is the situational interview, which is in many ways an interview form of an SJT. Additionally, having past relevant experience would help one when responding to an SJT and the point of biodata is to assess applicants' past relevant experience. Of course, an important caveat to all these points about SJTs is that SJTs represent a selection method that can be used to measure various constructs. Thus, the above conclusions about SJTs can only be known to hold for SJTs measuring constructs similar to those reflecting the meta-analytic averages in the published literature to date. It is possible that results could differ for SJTs measuring different constructs.

Because of the well-known distinction between constructs and methods in the personnel selection literature (Arthur & Villado, 2008), and because some of the predictors included in the present study are more to the construct side of the continuum (GMA tests, conscientiousness, integrity tests) while the other predictors are more to the method side of the continuum (structured interviews, biodata, SJTs), we believe this issue of constructs and methods deserves more discussion. For one, the fact that the predictor methods can be designed to measure a wide range of constructs may be part of the reason for variability around the meta-analytic averages. Structured interviews provide a good example. They are one of the predictors in the present study with the largest standard deviation (.19) around their meta-analytic validity. This is likely at least in part due to different interviews measuring different combinations of constructs (Huffcutt & Murphy, 2023). Our Substantive Question 5 analyses were designed to assess the implications of the substantial variability around the meta-analytic averages, but they cannot explain what is causing that variability. We hope that there will be a point in the future when, rather than focusing on the validity of structured interviews (or biodata or SJTs), the field focuses on the validity of structured interviews that measure "these constructs" versus "those constructs."

This construct-method issue is also important because the constructs measured by the predictor methods and the degree to which they overlap with the constructs in the other predictors affect results. For example, one of the reasons structured interviews fared so well in our results is that they are relatively weakly related to the three predictor constructs (i.e., correlations of .18, .08, and −.02 with GMA tests, conscientiousness, and integrity tests, respectively), giving them more opportunity for incremental validity. One of the reasons SJTs fare relatively poorly is that they have stronger relationships with the three predictor constructs (i.e., correlations of .29, .23, and .16 with GMA tests, conscientiousness, and integrity tests, respectively), giving them less opportunity for incremental validity. However, structured interviews and SJTs (and biodata) are predictor methods that could measure a wide range of constructs. For instance, it is at least hypothetically possible to design an SJT with weaker relationships GMA tests, conscientiousness, and integrity tests. This would increase opportunity for incremental validity if the criterion-related validity of such an SJT was still appreciable. Similarly, one could design a structured interview that has much stronger correlations with GMA tests, conscientiousness, and integrity tests, and this would affect its opportunity for incremental validity. We think there is still value in knowing how the average structured interview, biodata inventory, or SJT in the published literature to date fares in comparison to other selection predictors, which is what the analyses based on our meta-analytic correlation matrix reflect. Still, it is also important to recognize and continue to explore the variability around those averages.

Additionally, this construct-method issue has implications for the role of GMA tests versus the construct of GMA in personnel selection. Our results show that, in the presence of the other predictors, GMA tests generally carry little weight and excluding them from the selection battery has little effect on validity. However,

this does not necessarily mean that the construct of GMA has little or no role. We note that GMA tests, and presumably the construct being measured by these tests, correlate positively with a number of the other predictors in our analyses (e.g., correlations of .13, .18, and .29 with biodata, structured interviews, and SJTs, respectively). Thus, the construct of GMA will still play at least some role in personnel selection due to its correlations with these other predictors.

Another important result in the present study is the demonstration that, although Sackett et al. (2022) showed that the criterion-related validity of many individual selection methods had been overestimated due to inappropriate range restriction corrections, it is still possible for composites of multiple selection methods to have criterion-related validity similar to what was expected before Sackett et al. (2022). For example, using Roth et al.'s (2011) meta-analytic correlation matrix resulted in a validity of .66 for a composite of all six selection methods, and this only reduced to .61 when using the updated matrix with its more modest validities for most selection methods. This same pattern was repeated for composites using smaller numbers of selection methods. Thus, despite Sackett et al.'s (2022) results showing that validity estimates of many individual selection methods have been substantially overestimated, the selection methods remain useful and valid, especially when used in conjunction with each other. This is particularly important because selection methods are almost always used together in multipredictor batteries in actual selection practice.

Limitations and Directions for Future Research

One limitation is that the criterion-related validities, selection method intercorrelations, and Black–White mean differences in the updated meta-analytic matrix are simply the best estimates available at this point in time. For example, the present study focused on operational validities which correct only for criterion measurement error and range restriction, where applicable, because this provides an estimate of validity for selection methods in operational use. However, Hunter and Schmidt (2004) outlined a number of other statistical artifacts that can affect validity estimates (e.g., imperfect construct validity). To the degree that such artifacts exist, if future research could correct for them, validity estimates may increase. Further, as new data become available, it will be necessary to continue updating this matrix and it is possible that some of the conclusions of the present study could be affected. For example, much of the data from the meta-analyses that contributed to the present study's validity estimate for GMA tests is quite old. As noted earlier, Sackett, Demeke, et al. (2023) recently meta-analyzed the relationship between GMA tests and job performance including only studies carried out in the 21st century, finding an average criterion-related validity somewhat lower than that which was used in the present study. We demonstrated that when Sackett, Demeke, et al.'s (2023) updated and lower 21st century validity estimate is substituted for the validity estimate based on older data from Sackett et al. (2022), results differed in that GMA tests received substantially lower weight in the Pareto-optimal solutions wherein we modeled variability around the meta-analytic averages in the correlation matrix (i.e., Substantive Question 5). Our other results would also differ if we used the updated 21st century operational validity. For example, we reran our single-stage Pareto-optimal model using all six selection methods (i.e., the model in Table 5), but substituting the 21st century validity for GMA tests; full results are in Supplemental Table S2. The general pattern of results is similar, but GMA tests received less weight in the validity-maximizing solution (a weight of .12 compared to the largest weight of .38 for structured interviews in Supplemental Table S2 vs. a weight of .20 compared to the largest weight of .34 for structured interviews in Table 5), received a weight of zero earlier (i.e., in the fourth row of Supplemental Table S2 vs. the sixth row of Table 5), and composite validity was reduced less when GMA tests received a weight of zero (composite validity only reduces from .58 to .57 in Supplemental Table S2 vs. from .61 to .56 in Table 5). Using Sackett, Demeke, et al.'s (2023) 21st century criterion-related validity estimate just strengthens our conclusion about the reduced role of GMA tests. Still, this highlights the importance of future research to continue to examine and update the relationships and mean differences included in the new meta-analytic matrix.

It is also important to note that this study focuses on how well these selection methods predict overall job performance. This reflects that a major motivation for this study was Sackett et al.'s (2022) revised and mostly lower estimates of these selection methods' criterion-related validity for predicting overall job performance. However, job performance is known to be multidimensional, including dimensions such as task performance, organizational citizenship behavior, and counterproductive work behavior, among others (Rotundo & Sackett, 2002). Thus, it is likely that conclusions about the relative validity of these selection methods could change if the focus shifted from overall job performance to some of its dimensions. It is also the case that selection methods outside of the six included in the present study might be important additions for predicting some job performance dimensions. For example, in addition to conscientiousness, agreeableness and emotional stability are important Big Five personality predictors of counterproductive work behavior (Berry, Ones, et al., 2007). Furthermore, job performance is not the only criterion that is of interest to organizations. For example, in jobs requiring significant training of employees, training performance is an important criterion and GMA tests have long been found to be a strong predictor of training performance. Thus, it may be the case that GMA tests remain an important (perhaps even the most important) predictor when training performance is the criterion of interest. However, it is important to note that the validity of GMA tests for predicting training performance has not been revisited in light of the range restriction correction issues that Sackett et al. (2022) highlighted. In any case, there would be value in future research creating meta-analytic correlation matrices of various selection methods' criterion-related validity for predicting training performance or dimensions of job performance.

Another caveat for the present study's conclusions about GMA tests is that they are derived from estimates of criterion-related validities and Black–White mean differences for tests of GMA. There are also narrower facets of GMA, such as verbal, quantitative, and spatial abilities. There is evidence that criterion-related validity (e.g., Salgado et al., 2003) and Black–White mean differences (e.g., Dahlke & Sackett, 2017) differ somewhat across tests of facets of GMA. Thus, results could differ if one focuses on facet tests instead of GMA tests. However, the driving factor behind this study's conclusions about GMA tests is that they have lower criterion-related validity than was thought before Sackett et al. (2022), but still have large Black–White mean differences. The general pattern for tests of most of the facets of GMA is that they also have large Black–White differences (Dahlke & Sackett, 2017), and have

similar or lower criterion-related validity than GMA, at least for predicting overall job performance (Salgado et al., 2003). Therefore, we would expect similar overall conclusions for GMA tests and most facet tests.

A final limitation is that the present study only focused on these selection methods' Black–White mean differences and their implications for adverse impact and the validity–diversity trade-off. This reflects that the validity–diversity trade-off has often focused on those two subgroups, and that good estimates of mean differences for other subgroups do not exist for many of the selection methods. So, we see value in future research focused on the validity and diversity implications of selection methods for other relevant subgroups.

Conclusion

We provided an updated meta-analytic correlation matrix of six selection methods' criterion-related validity for predicting overall job performance. Most importantly, this matrix included Sackett et al.'s (2022) criterion-related validities for these selection methods. Although Sackett et al.'s (2022) criterion-related validity estimates were in many cases substantially lower than those included in previous meta-analytic matrices (e.g., Roth et al., 2011), the present study demonstrated that composites of multiple selection methods can have validity nearly as high as was expected before Sackett et al. (2022). The present study also demonstrated that these composites' validity does not hinge on inclusion of GMA tests. Due to Sackett et al.'s (2022) much lower estimate of the criterion-related validity of GMA tests, GMA tests can be left out of most selection batteries and this will have little effect on validity, but will greatly improve adverse impact. This essentially resolves a major dilemma that personnel selection researchers and practitioners had long faced that improving adverse impact by excluding GMA tests came at a steep validity cost. The present study demonstrates this is no longer the case.

References

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435–442. https://doi.org/10.1037/0021-9010.93.2.435

Barends, A. J., de Vries, R. E., & van Vugt, M. (2022). Construct and predictive validity of an assessment game to measure honesty–humility. *Assessment*, *29*(4), 630–650. https://doi.org/10.1177/1073191120985612

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 435–463. https://doi.org/10.1146/annurev-orgpsych-032414-111256

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, *92*(2), 410–424. https://doi.org/10.1037/0021-9010.92.2.410

Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology*, *60*(4), 837–874. https://doi.org/10.1111/j.1744-6570.2007.00093.x

Berry, C. M., Sackett, P. R., & Wiemann, S. A. (2007). A review of recent developments in integrity test research. *Personnel Psychology*, *60*(2), 271–301. https://doi.org/10.1111/j.1744-6570.2007.00074.x

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, *52*(3), 561–589. https://doi.org/10.1111/j.1744-6570.1999.tb00172.x

Braun, M. T., Converse, P. D., & Oswald, F. L. (2019). The accuracy of dominance analysis as a metric to assess relative importance: The joint impact of sampling error variance and measurement unreliability. *Journal of Applied Psychology*, *104*(4), 593–602. https://doi.org/10.1037/apl0000361

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology*, *63*(1), 83–117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology*, *80*(5), 565–579. https://doi.org/10.1037/0021-9010.80.5.565

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI)*. Psychological Assessment Resources.

Dahlke, J. A., & Sackett, P. R. (2017). The relationship between cognitive-ability saturation and subgroup mean differences across predictors of job performance. *Journal of Applied Psychology*, *102*(10), 1403–1420. https://doi.org/10.1037/apl0000234

Dahlke, J. A., & Sackett, P. R. (2022). On the assessment of predictive bias in selection systems with multiple predictors. *Journal of Applied Psychology*, *107*(11), 1995–2012. https://doi.org/10.1037/apl0000996

Dalessio, A., & Silverhart, T. (1994). Combining biodata test and interview information: Predicting decisions and performance criteria. *Personnel Psychology*, *47*(2), 303–315. https://doi.org/10.1111/j.1744-6570.1994.tb01726.x

Das, I., & Dennis, J. E. (1998). Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM Journal on Optimization*, *8*(3), 631–657. https://doi.org/10.1137/S1052623496307510

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, *92*(5), 1380–1393. https://doi.org/10.1037/0021-9010.92.5.1380

De Corte, W., Lievens, F., & Sackett, P. R. (2008). Validity and adverse impact potential of predictor composite formation. *International Journal of Selection and Assessment*, *16*(3), 183–194. https://doi.org/10.1111/j.1468-2389.2008.00423.x

De Corte, W., Sackett, P. R., & Lievens, F. (2023). Developing Pareto-optimal selection systems for multiple protected groups. *Journal of Applied Psychology*, *109*(4), 513–533. https://doi.org/10.1037/apl0001145

de Leng, W. E., Stegers-Jager, K. M., Born, M. P., & Themmen, A. P. N. (2018). Integrity situational judgement test for medical school selection: Judging 'what to do' versus 'what not to do'. *Medical Education*, *52*(4), 427–437. https://doi.org/10.1111/medu.13498

de Meijer, L. A. L., Born, M. P., van Zielst, J., & van der Molen, H. T. (2010). Construct-driven development of a video-based situational judgment test for integrity: A study in a multi-ethnic police setting. *European Psychologist*, *15*(3), 229–236. https://doi.org/10.1027/1016-9040/a000027

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

Dean, M. A. (2013). Examination of ethnic group differential responding on a biodata instrument. *Journal of Applied Social Psychology*, *43*(9), 1905–1917. https://doi.org/10.1111/jasp.12212

Fife, D. A., Mendoza, J. L., & Terry, R. (2013). Revisiting case IV: A reassessment of bias and standard errors of case IV under range restriction.

British Journal of Mathematical and Statistical Psychology, *66*(3), 521–542. https://doi.org/10.1111/j.2044-8317.2012.02060.x

Foldes, H. J., Duehr, E. E., & Ones, D. S. (2008). Group differences in personality: Meta-analyses comparing five US racial groups. *Personnel Psychology*, *61*(3), 579–616. https://doi.org/10.1111/j.1744-6570.2008.00123.x

Gandy, J., Dye, D., & MacLane, C. (1994). Federal government selection: The individual achievement record. In G. Stokes, M. Mumford, & W. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 275–310). Consulting Psychologists Press.

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

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. *Intelligence*, *24*(1), 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8

Heggarty, P., Teague, P. A., Alele, F., Adu, M., & Malau-Aduli, B. S. (2020). Role of formative assessment in predicting academic success among GP registrars: A retrospective longitudinal study. *BMJ Open*, *10*(11), Article e040290. https://doi.org/10.1136/bmjopen-2020-040290

Heimann, A. L., Ingold, P. V., Debus, M. E., & Kleinmann, M. (2021). Who will go the extra mile? Selecting organizational citizens with a personality-based structured job interview. *Journal of Business and Psychology*, *36*(6), 985–1007. https://doi.org/10.1007/s10869-020-09716-1

Heimann, A. L., Ingold, P. V., & Kleinmann, M. (2020). Tell us about your leadership style: A structured interview approach for assessing leadership behavior constructs. *The Leadership Quarterly*, *31*(4), Article 101364. https://doi.org/10.1016/j.leaqua.2019.101364

Huffcutt, A. I., & Arthur, W., Jr. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, *79*(2), 184–190. https://doi.org/10.1037/0021-9010.79.2.184

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology: Employment interview validity. *International Journal of Selection and Assessment*, *22*(3), 297–309. https://doi.org/10.1111/ijsa.12078

Huffcutt, A. I., & Murphy, S. A. (2023). Structured interviews: Moving beyond mean validity. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *16*(3), 344–348. https://doi.org/10.1017/iop.2023.42

Huffcutt, A. I., & Roth, P. L. (1998). Racial group differences in interview evaluations. *Journal of Applied Psychology*, *83*(2), 179–189. https://doi.org/10.1037/0021-9010.83.2.179

Huffcutt, A. I., Roth, P. L., & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, *81*(5), 459–473. https://doi.org/10.1037/0021-9010.81.5.459

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, *29*(3), 340–362. https://doi.org/10.1016/0001-8791(86)90013-8

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Sage Publications. https://doi.org/10.4135/9781412985031

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594–612. https://doi.org/10.1037/0021-9010.91.3.594

Husbands, A., Dowell, J., Homer, M., McAndrew, R., & Greatrix, R. (2018). *Exploring the relationship between the UKCAT situational judgement test and the multiple mini interview*. University Clinical Aptitude Test Board. https://www.ucat.ac.uk/media/1277/ukcat-sjt-mmi-report-march-2018.pdf

Husbands, A., Rodgerson, M. J., Dowell, J., & Patterson, F. (2015). Evaluating the validity of an integrity-based situational judgement test for medical school admissions. *BMC Medical Education*, *15*(1), Article 144. https://doi.org/10.1186/s12909-015-0424-0

Kantrowitz, T. M. (2014). *2014 global assessment trends report* [Unpublished technical report]. CEB.

Knorr, M., Schwibbe, A., Ehrhardt, M., Lackamp, J., Zimmermann, S., & Hampe, W. (2018). Validity evidence for the Hamburg multiple mini-interview. *BMC Medical Education*, *18*(1), Article 106. https://doi.org/10.1186/s12909-018-1208-0

Kriska, S. D. (2001, April). *The validity-adverse impact trade-off: Real data and mathematical model estimates* [Paper presentation]. Society for Industrial and Organizational Psychology, San Diego, CA, United States.

Lee, Y., Berry, C. M., & Gonzalez-Mulé, E. (2019). The importance of being humble: A meta-analysis and incremental validity analysis of the relationship between honesty–humility and job performance. *Journal of Applied Psychology*, *104*(12), 1535–1546. https://doi.org/10.1037/apl0000421

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *9*(1), 3–22. https://doi.org/10.1017/iop.2015.71

Luan, S., Reb, J., & Gigerenzer, G. (2019). Ecological rationality: Fast-and-frugal heuristics for managerial decision making under uncertainty. *Academy of Management Journal*, *62*(6), 1735–1759. https://doi.org/10.5465/amj.2018.0172

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, *60*(1), 63–91. https://doi.org/10.1111/j.1744-6570.2007.00065.x

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*(5), 812–821. https://doi.org/10.1037/0021-9010.85.5.812

Messac, A., & Mattson, C. A. (2004). Normal constraint method with guarantee of even representation of complete Pareto frontier. *AIAA Journal*, *42*(10), 2101–2111. https://doi.org/10.2514/1.8977

Morgeson, F. P., Reider, M. H., & Campion, M. A. (2005). Selecting individuals in team settings: The importance of social skills, personality characteristics, and teamwork knowledge. *Personnel Psychology*, *58*(3), 583–611. https://doi.org/10.1111/j.1744-6570.2005.655.x

Mueller-Gritschneder, D., Graeb, H., & Schlichtmann, U. (2009). A successive approach to compute the bounded Pareto front of practical multiobjective optimization problems. *SIAM Journal on Optimization*, *20*(2), 915–934. https://doi.org/10.1137/080729013

Ock, J., & Oswald, F. L. (2018). The utility of personnel selection decisions: Comparing compensatory and multiple-hurdle selection models. *Journal of Personnel Psychology*, *17*(4), 172–182. https://doi.org/10.1027/1866-5888/a000205

Ones, D. S. (1993). *The construct validity of integrity tests* [Unpublished doctoral dissertation]. University of Iowa.

Ones, D. S., & Viswesvaran, C. (1998). Gender, age, and race differences on overt integrity tests: Results across four large scale job applicant data sets. *Journal of Applied Psychology*, *83*(1), 35–42. https://doi.org/10.1037/0021-9010.83.1.35

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*(4), 679–703. https://doi.org/10.1037/0021-9010.78.4.679

Oostrom, J. K., de Vries, R. E., & de Wit, M. (2019). Development and validation of a HEXACO situational judgment test. *Human Performance*, *32*(1), 1–29. https://doi.org/10.1080/08959285.2018.1539856

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology*, *89*(2), 187–207. https://doi.org/10.1037/0021-9010.89.2.187

Patterson, F., Carr, V., Zibarras, L., Burr, B., Berkin, L., Plint, S., Irish, B., & Gregory, S. (2009). New machine-marked tests for selection into core

medical training: Evidence from two validation studies. *Clinical Medicine*, *9*(5), 417–420. https://doi.org/10.7861/clinmedicine.9-5-417

Patterson, F., Rowett, E., Hale, R., Grant, M., Roberts, C., Cousans, F., & Martin, S. (2016). The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia. *BMC Medical Education*, *16*(1), Article 87. https://doi.org/10.1186/s12909-016-0606-4

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153–172. https://doi.org/10.1111/j.1744-6570.2008.00109.x

Potosky, D. P., Bobko, P., & Roth, P. L. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Selection and Assessment*, *13*(4), 304–315. https://doi.org/10.1111/j.1468-2389.2005.00327.x

Pulakos, E., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance*, *9*(3), 241–258. https://doi.org/10.1207/s15327043hup0903_4

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roberts, C., Clark, T., Burgess, A., Frommer, M., Grant, M., & Mossman, K. (2014). The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training. *BMC Medical Education*, *14*(1), Article 169. https://doi.org/10.1186/1472-6920-14-169

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*(2), 297–330. https://doi.org/10.1111/j.1744-6570.2001.tb00094.x

Roth, P. L., Switzer, F. S., III, Van Iddekinge, C. H., & Oh, I. S. (2011). Toward better meta-analytic matrices: How input values can affect research conclusions in human resource management simulations. *Personnel Psychology*, *64*(4), 899–935. https://doi.org/10.1111/j.1744-6570.2011.01231.x

Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., Jr., & Bobko, P. (2002). Corrections for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology*, *87*(2), 369–376. https://doi.org/10.1037/0021-9010.87.2.369

Rothstein, H. R., Schimdt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable? *Journal of Applied Psychology*, *75*(2), 175–184. https://doi.org/10.1037/0021-9010.75.2.175

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology*, *87*(1), 66–80. https://doi.org/10.1037/0021-9010.87.1.66

Sackett, P. R., Demeke, S., Bazian, I. M., Griebie, A. M., Priest, R., & Kuncel, N. R. (2023). A contemporary look at the relationship between general cognitive ability and job performance. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0001159

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, *56*(4), 302–318. https://doi.org/10.1037/0003-066X.56.4.302

Sackett, P. R., 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*(11), 2040–2068. https://doi.org/10.1037/apl0000994

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: Perspectives on Science and Practice*, *16*(3), 283–300. https://doi.org/10.1017/iop.2023.24

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology*, *88*(6), 1068–1081. https://doi.org/10.1037/0021-9010.88.6.1068

Salgado, J. F., & Moscoso, S. (2002). Comprehensive meta-analysis of the construct validity of the employment interview. *European Journal of Work and Organizational Psychology*, *11*(3), 299–324. https://doi.org/10.1080/13594320244000184

Schmitt, N., Keeney, J., Oswald, F. L., Pleskac, T. J., Billington, A. Q., Sinha, R., & Zorzie, M. (2009). Prediction of 4-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, *94*(6), 1479–1497. https://doi.org/10.1037/a0016810

Song, Q. C. (2022). *ParetoR: Estimates Pareto-Optimal solution for diversity hiring* (R package Version 0.1.0) [Computer software]. https://rdrr.io/github/Diversity-ParetoOptimal/ParetoR/

Song, Q. C., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology*, *102*(12), 1636–1657. https://doi.org/10.1037/apl0000240

Speer, A. B., Sendra, C., & Shihadeh, M. (2021). *Meta-analysis of biodata in employment settings: Providing clarity to criterion and construct-related validity estimates* [Conference session]. Society for Industrial and Organizational Psychology, New Orleans, LA, United States.

Speer, A. B., Tenbrink, A. P., Wegmeyer, L. J., Sendra, C. C., Shihadeh, M., & Kaur, S. (2022). Meta-analysis of biodata in employment settings: Providing clarity to criterion and construct-related validity estimates. *Journal of Applied Psychology*, *107*(10), 1678–1705. https://doi.org/10.1037/apl0000964

Tenbrink, A. P., Wegmeyer, L., Rowley, S. J., Sendra, C., & Speer, A. (2021, April). *Group differences in biographical data: A meta-analysis* [Conference session]. Society for Industrial and Organizational Psychology, Virtual Conference.

U.S. Bureau of Labor Statistics. (2021). *Labor force characteristics by race and ethnicity, 2020*. Retrieved on September 20, 2022, from https://www.bls.gov/cps/demographics.htm#race

Van Iddekinge, C. H., Raymark, P. H., Eidson, C. E., Jr., & Attenweiler, B. (2004). What do structured interviews really measure? The construct validity of behavior description interviews. *Human Performance*, *17*(1), 71–93. https://doi.org/10.1207/S15327043HUP1701_4

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology*, *97*(3), 499–530. https://doi.org/10.1037/a0021196

Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than *g*: Selection quality and adverse impact implications of considering second-stratum cognitive abilities. *Journal of Applied Psychology*, *99*(4), 547–563. https://doi.org/10.1037/a0035183

Whetzel, D., McDaniel, M., & Nguyen, N. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*(3), 291–309. https://doi.org/10.1080/08959280802137820

Yingling, S., Park, Y. S., Curry, R. H., Monson, V., & Girotti, J. (2018). Beyond cognitive measures: Empirical evidence supporting holistic medical school admissions practices and professional identity formation. *MedEdPublish*, *7*, Article 274. https://doi.org/10.15694/mep.2018.0000274.1

(*Appendices follow*)

Appendix A

Definitions of Selection Methods

Selection method	Definition
Biodata	"Standardized measures" that deal with "describing behaviors and events occurring earlier in one's life, including personal background and life history events" (Speer et al., 2022, p. 4)
General mental ability tests	Tests designed to measure general mental ability, which is "a very general mental capacity that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience" (Gottfredson, 1997, p. 13).
Conscientiousness tests	Self-report inventories or scales designed to measure conscientiousness, which is the degree to which one is competent, orderly, dutiful, achievement striving, self-disciplined, and deliberate (Costa & McCrae, 1992).
Structured interviews	In-person interviews incorporating elements of structure, such as standardization of interview questions or standardization of response evaluation (Conway et al., 1995).
Integrity tests	Tests designed to measure honesty or integrity. They can be overt or personality-based. "Overt integrity tests commonly consist of two sections. The first is a measure of theft attitudes and includes questions pertaining to beliefs about the frequency and extent of theft, punitiveness toward theft, ruminations about theft, perceived ease of theft, endorsement of common rationalizations for theft, and assessments of one's own honesty. The second involves requests for admissions of theft and other wrongdoing … Personality-oriented measures are closely linked to normal-range personality devices" and "include items dealing with dependability, Conscientiousness, social conformity, thrill seeking, trouble with authority, and hostility" (Berry, Sackett, & Wiemann, 2007, pp. 271–272).
Situational judgment tests	"Situational judgment tests (SJTs) are personnel selection instruments that present job applicants with work-related situations and possible responses to the situations. There are typically 2 types of instructions: behavioral tendency and knowledge. Behavioral tendency instructions ask respondents to identify how they would likely behave in a given situation. Knowledge instructions ask respondents to evaluate the effectiveness of possible responses to a given situation" (McDaniel et al., 2007, p. 63).

Appendix B

Descriptions of Updates to the Selection Method Intercorrelations and Black–White Subgroup Mean Differences

Updates to Selection Method Intercorrelations

Except for SJT's correlations with biodata, structured interviews, and integrity tests, which we added to the meta-analytic correlation matrix via new literature searches as detailed below, the selection method intercorrelations in Table 1 were drawn from previous studies. Some of the intercorrelations were simply taken from those previous studies as is (the intercorrelations in plain text in Table 1), while others changed somewhat due to new insights (italicized intercorrelations in Table 1). Because our focus is on operational validity, intercorrelations based on applicant pools are needed. Therefore, we carefully reviewed every study contributing selection method intercorrelations to ensure that they used applicant pool samples, or at least samples that would be minimally affected by range restriction compared to applicant pool samples. Thus, when possible, the intercorrelations are based on applicant pools before screening. In some instances, intercorrelations were available for applicant samples that had already been screened using one of the selection methods in question (e.g., a correlation between biodata and a structured interview for a sample that had been screened using the biodata measure), meaning the applicant sample was affected by range restriction; in these cases, the sample was only included if a credible correction for range restriction was possible. There were some instances in which job incumbent intercorrelations were included. Importantly, these incumbent intercorrelations were only included if the incumbents had not been selected into their job using the selection method in question (e.g., a biodata-conscientiousness correlation from a job incumbent sample that had not been selected into their jobs using the biodata or conscientiousness measures), as a key point from Sackett et al. (2022) was that range restriction is minimal if the selection method in question was not used to screen the incumbent sample. Further, because predictor measurement error is not corrected for in operational validity estimates, the selection method intercorrelations were not corrected for measurement error.

To get estimates of SD_{Res} (the standard deviation of the meta-analytic intercorrelation after subtracting out the expected effects of sampling error), where possible, we simply used the values of SD_{Res} reported in the meta-analyses. Many meta-analyses did not report SD_{Res}, but rather reported SD_r (the sample size-weighted standard deviation of the meta-analytic intercorrelation). In these cases, we calculated SD_{Res} ourselves from the sample size (N), number of samples (k), the meta-analytic correlation, and the reported SD_r. That is, to calculate SD_{Res}, one simply squares SD_r (to convert it to a variance) and then subtracts the sampling error variance from squared SD_r; the square root of that difference is SD_{Res}. The sampling error variance formula, Equation A1, is:

$$\frac{(1-\bar{r}^2)^2}{\bar{N}-1}, \tag{A1}$$

where $\bar{r}$ is the meta-analytic intercorrelation and $\bar{N}$ is the average sample size of the primary studies contributing to the meta-analysis (which can be calculated by N by k).

In the following sections, we describe how we arrived at each selection method intercorrelation and SD_{Res}, along with explanations for differences between our and Roth et al. (2011) intercorrelations (Roth et al. did not include values of SD_{Res}, and thus we do not compare these values to Roth et al.).

(*Appendices continue*)

Biodata and GMA Tests

Roth et al. (2011) drew their correlation of .37 between biodata and GMA tests from Potosky et al. (2005), who based their estimate on two studies: Dean (1999) and Kriska (2001). Since Roth et al. (2011), Speer et al. (2022) meta-analyzed the correlation between biodata and GMA tests including more samples. Speer et al. reported that they were only able to locate information about range restriction in a handful of their samples and that the amount of range restriction in those samples was very small (that is, the ratios of restricted to unrestricted standard deviations [u-ratios] were $u = .92$ and $u = .99$ for direct and indirect range restriction, respectively). Thus, Speer et al.'s uncorrected correlation of .13 ($k = 10$, $n = 12{,}691$) should approximate the applicant pool value and is therefore the correlation we use in Table 1. Speer et al. reported SD_r instead of SD_{Res}, so we calculated the SD_{Res} value of 0.053 using their $k = 10$, $N = 12{,}691$, $r = .13$, and $SD_r = 0.06$.

Biodata and Conscientiousness Tests

Roth et al. (2011) intercorrelation of .51 between biodata and conscientiousness tests was drawn from Bobko et al. (1999), who based their biodata-conscientiousness intercorrelation on two studies: Gandy et al. (1994) and Pulakos and Schmitt (1996). Speer et al. (2022) meta-analyzed the biodata-conscientiousness correlation including more studies. As we discussed earlier, any range restriction in Speer et al.'s samples was minimal, so we use their uncorrected correlation of $r = .54$ ($k = 9$, $n = 21{,}214$) in Table 1. Speer et al. reported SD_r instead of SD_{Res}, so we calculated the SD_{Res} value of 0.179 using their $k = 9$, $N = 21{,}214$, $r = .54$, and $SD_r = 0.18$.

Biodata and Structured Interviews

Roth et al.'s (2011) intercorrelation of .16 between biodata and structured interviews was drawn from Bobko et al. (1999), who based their biodata-interview correlation on two studies: Dalessio and Silverhart (1994) and Pulakos and Schmitt (1996). Although Pulakos and Schmitt's correlation is based on a job incumbent sample, the incumbents were not selected into their jobs using either the biodata measure or structured interview, so any range restriction should be minimal. Thus, we included Pulakos and Schmitt's biodata-interview correlation of .27. Bobko et al., and thus Roth et al., used a biodata-interview correlation of .08 from Dalessio and Silverhart. However, Dalessio and Silverhart reported that the correlation of .08 came from a sample that was first screened on the biodata measure before participating in the structured interview. Therefore, there was direct range restriction on the biodata measure. Dalessio and Silverhart corrected the correlation of .08 for direct range restriction in the biodata measure, which resulted in a corrected correlation of .17. We thus use the corrected correlation of .17 from Dalessio and Silverhart. We meta-analyzed Dalessio and Silverhart and Pulakos and Schmitt's correlations, resulting in $r = .21$ ($n = 1{,}042$), which is the correlation we use for the relationship between biodata and structured interviews in Table 1. This meta-analysis also provided the value of .027 for SD_{Res}.

Biodata and Integrity Tests

Roth et al. (2011) intercorrelation of .25 between biodata and integrity tests was drawn from McFarland and Ryan (2000) study in which a biodata measure and integrity test were administered to a sample of undergraduates. We are not aware of any other estimates of the relationship between biodata and integrity tests. Thus, we also use .25 ($n = 192$) as the intercorrelation between biodata and integrity tests in Table 1. Because the intercorrelation between biodata and integrity tests is based on a single study, SD_{Res} cannot be calculated. Therefore, we used the average of the other SD_{Res} values in the matrix (0.108).

Biodata and SJTs

Given that SJTs are a new addition to the matrix, we conducted a literature search to find studies that correlated biodata with SJTs. We were able to locate two primary studies that reported the correlation between biodata and SJTs (Oswald et al., 2004; Schmitt et al., 2009). These studies used both the biodata and SJT measures in a research context with college students. For both studies, students were not selected using biodata or SJT. Hence, range restriction was minimal and we therefore use these studies' uncorrected correlations. Both studies reported correlations between an SJT measure and multiple biodata dimension scores, so we used a composite formula (Ghiselli et al., 1981, p. 164) to estimate the correlations between the SJTs and composites of the biodata dimension scores. We then meta-analyzed (weighting by sample size) the resulting two composite correlations to arrive at a meta-analytic correlation of .42 ($n = 3{,}400$) between biodata and SJTs, which is the value we use in Table 1. SD_{Res} was 0.05. Because this a new meta-analysis calculated in the present study, we also report the full meta-analysis results in Table B1.

GMA Tests and Conscientiousness Tests

Roth et al. (2011) intercorrelation of .03 between GMA tests and conscientiousness tests was drawn from Potosky et al. (2005). Potosky et al.'s estimate was based on seven samples, one of which was a job applicant sample and the other six of which were general population samples that were unlikely to be affected by range restriction. One could make the case that self-selection into applicant pools might make the applicant pool sample more restricted in range than the general population samples. However, two things make us believe this is of little concern. First, the correlation in the applicant pool sample ($r = .12$) is actually slightly larger than in the general population samples. Second, the average correlation of .03 is so

Table B1
Meta-Analytic Relationships With Situational Judgment Tests

Selection method	k	n	r_m	SD_r	SD_{se}	SD_{Res}	%Var	95% CI *LL*	95% CI *UL*
Biodata	2	3,400	.42	0.05	0.02	0.05	15.4	.35	.49
Integrity test	5	742	.16	0.13	0.08	0.10	38.8	.04	.27
Structured interview	8	7,761	.45	0.12	0.03	0.12	4.4	.37	.54

Note. k = number of samples; n = sample size; r_m = meta-analytic correlation; SD_r = observed standard deviation; SD_{se} = standard deviation expected due to sampling error; SD_{Res} = residual standard deviation; % Var = percentage of variance accounted for by sampling error; CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

(*Appendices continue*)

small in the first place, that any plausible amount of range restriction would only minimally affect it anyway (e.g., even if there was enough range restriction such that correction for that restriction doubled the correlation, this would still only result in a very small correlation of .06). Thus, like Roth et al., we use .03 (n = 6,759) as the correlation between GMA tests and conscientiousness tests in Table 1.

Roth et al. (2011) did not report any estimates of variability, not even SD_r. Roth et al. reported their correlation of .03 from Potosky et al. (2005), but Potosky et al. also do not report any estimates of variability. However, Potosky et al. did report which seven samples contributed to their intercorrelation between GMA tests and conscientiousness. So, we located those seven samples and calculated SD_{Res} = 0.054 based on a meta-analysis of those samples' correlations.

GMA Tests and Structured Interviews

Roth et al. (2011) intercorrelation of .31 between GMA tests and structured interviews was drawn from Potosky et al. (2005). Potosky et al. arrived at r = .31 by correcting the uncorrected r = .23 from Huffcutt et al. (1996; Roth et al. reported this as r = .24, but it is reported as r = .23 in both Potosky et al. and Huffcutt et al.) for direct range restriction using u = .74 based on an interview range restriction artifact distribution of 15 studies from Huffcutt and Arthur (1994). We think this range restriction correction likely resulted in an overestimate of the correlation between GMA tests and structured interviews. Applying a direct range restriction correction to the uncorrected correlation of r = .23 is only appropriate if direct range restriction resulting from selection on the interview actually affected every study in Huffcutt et al.'s meta-analysis. This is certainly not the case. Berry, Sackett, and Landers (2007) reanalyzed the ability–interview relationship, including all of Huffcutt et al.'s data, along with additional data. Importantly, Berry et al. sorted studies into different categories based on the form of range restriction affecting each study. Out of 65 samples, Berry et al. only found direct range restriction in 12, and of these 12 a number were restricted due to selection on the GMA test, not the interview. In the majority of the other samples range restriction was indirect, but there were even 12 samples with no range restriction because all applicants completed the GMA test and interview. Thus, direct range restriction resulting from selection on the interview is quite uncommon in studies of the relationship between GMA tests and interviews. So, a blanket correction to all studies for direct range restriction resulting from selection on the interview would result in a misestimate of the ability–interview correlation. Berry et al. reported their coding in their Table 1. In that table, we located seven samples in which a GMA test and structured interview were administered to an entire applicant pool, meaning there is no range restriction. We meta-analyzed the correlations between GMA tests and structured interviews in those seven samples, resulting in a meta-analytic intercorrelation of .18 (n = 4,927), which is the intercorrelation we use in the Table 1 meta-analytic matrix. SD_{Res} was 0.038.

GMA Tests and Integrity Tests

Roth et al. (2011) intercorrelation of .02 between GMA tests and integrity tests was drawn from Ones (1993) meta-analysis. The correlation of .02 was corrected for range restriction and measurement error. Our focus is on operational validity, so there should be no correction for measurement error in the selection method. For reasons we discuss in detail below for the relationship between conscientiousness tests and integrity tests drawn from Ones, we do not think a credible range restriction correction can be made for the relationships that integrity tests have with conscientiousness tests or GMA tests. In any case, the uncorrected correlation between GMA tests and integrity tests (r = .01) is so small in the first place, that it remains r = .01 whether one corrects for range restriction or not. Thus, we use .01 (n = 23,306) as the intercorrelation between GMA tests and integrity tests in Table 1. Ones did not report any estimates of variability and we were unable to locate any variability estimates elsewhere. Therefore, we used the average of the other SD_{Res} values in the matrix (0.108).

GMA Tests and SJTs

McDaniel et al. (2007) meta-analyzed the relationship between GMA tests and SJTs. Although many of their 95 samples were likely incumbent samples, we view it as unlikely that most of those samples were selected into their jobs using the GMA tests or SJT measures being validated. Thus, range restriction is likely minimal. This, along with Sackett et al. (2022) principle of conservative estimation, led us to use McDaniel et al.'s uncorrected r = .29 (k = 95, n = 30,859) as the correlation between GMA tests and SJTs in Table 1. McDaniel et al. reported SD_r instead of SD_{Res}, so we calculated the SD_{Res} value of 0.173 using their k = 95, N = 30,859, r = .29, and SD_r = 0.18.

Conscientiousness Tests and Structured Interviews

Roth et al. (2011) intercorrelation of .13 between conscientiousness tests and structured interviews was drawn from Salgado and Moscoso (2002) meta-analysis. Salgado and Moscoso arrived at that intercorrelation by correcting the uncorrected r = .08 for direct range restriction on the interviews using a u-ratio of .61 from an artifact distribution they created based on a subset of their primary studies. The issue here is similar to the issue we highlighted with the relationship between GMA tests and structured interviews. That is, this blanket correction for direct range restriction is only appropriate if all studies in Salgado and Moscoso's meta-analysis were restricted in range due to selection only on the interview. However, Berry, Sackett, and Landers (2007) meta-analysis suggests this is unlikely to be the case and that such a blanket direct range restriction correction resulted in an overestimate. Lacking a way to know exactly how much range restriction affected all of the studies in Salgado and Moscoso, we argue for using their uncorrected r = .08 for two reasons. The first is that the uncorrected correlation is so small in the first place that even an overcorrection such as that used by Salgado and Moscoso only increased the relationship by a few correlation points. So, range restriction does not have a large effect on this relationship. That leads to the second reason: following Sackett et al. (2022) principle of conservative estimation, we would prefer a slight underestimate (the uncorrected correlation) to an overestimate. Thus, we used Salgado and Moscoso's uncorrected r = .08 (n = 1,497) for the intercorrelation between conscientiousness tests and structured interviews in Table 1. Salgado and Moscoso reported SD_r instead of SD_{Res}, so we calculated the SD_{Res} value of 0.037 using their k = 13, N = 1,497, r = .08, and SD_r = 0.10.

(*Appendices continue*)

Conscientiousness Tests and Integrity Tests

Roth et al. (2011) intercorrelation of .34 between conscientiousness tests and integrity tests was drawn from Ones (1993). The uncorrected $r = .28$ was corrected for direct range restriction on the integrity tests using Ones' u-ratio of .81, resulting in the corrected correlation of .34. However, there are two issues with correcting this correlation for range restriction. First, this correction for direct range restriction is only appropriate if the studies in Ones were all restricted in range due to selection only on the integrity test. This is highly unlikely because the majority of the studies in Ones' meta-analysis were likely concurrent validity studies[B1] and any range restriction in concurrent validity studies is typically indirect rather than direct (Hunter et al., 2006). Second, the u-ratio of .81 used to correct for range restriction was drawn solely from predictive validity studies, in which range restriction can be sizable (Sackett et al., 2022), but was then applied to studies that used some unknown mix of predictive and concurrent validity designs (likely a majority concurrent, per Footnote 2). Sackett et al. (2022) demonstrated that range restriction is typically minimal in concurrent validity studies, so applying a correction using a u-ratio based on predictive studies results in an overestimate of the correlation. Lacking information on exactly how many studies in Ones were predictive versus concurrent, Sackett et al. (2022) principle of conservative estimation leads us to use an estimate that may be a slight underestimate (i.e., the uncorrected correlation) rather than one that is an overestimate (the range restriction-corrected correlation). As such, we use Ones' uncorrected $r = .28$ ($n = 91{,}360$) as the intercorrelation between conscientiousness tests and integrity tests in Table 1. Ones did not report any estimates of variability and we were unable to locate any variability estimates elsewhere. Therefore, we used the average of the other SD_{Res} values in the matrix (0.108).

Conscientiousness Tests and SJTs

McDaniel et al. (2007) meta-analyzed the relationship between conscientiousness tests and SJTs. Similar to the correlation between GMA tests and SJTs in that meta-analysis, although many of their 53 samples were likely incumbent samples, we view it as unlikely that most of those sample were selected into their jobs using the conscientiousness or SJT measures. Thus, range restriction is likely minimal. This, along with Sackett et al. (2022) principle of conservative estimation, led us to use McDaniel et al.'s uncorrected $r = .23$ ($k = 53$, $n = 31{,}277$) as the correlation between conscientiousness tests and SJTs in Table 1. McDaniel et al. reported SD_r instead of SD_{Res}, so we calculated the SD_{Res} value of 0.124 using their $k = 53$, $N = 31{,}277$, $r = .23$, and $SD_r = 0.13$.

Structured Interviews and Integrity Tests

Roth et al. (2011) intercorrelation of −.02 between structured interviews and integrity tests was drawn from a primary study by Van Iddekinge et al. (2004). In this study job applicants were first screened using a personality test and integrity test before those passing that screen were interviewed. So, the correlation between the interview and integrity test is restricted in range, but that restriction and its effects on the interview–integrity relationship are minimal. That is, only 18% of applicants were screened out using the personality and integrity test and the uncorrected correlation of −.02 is so small in the first place that any range restriction correction would leave it almost unchanged. As such, like Roth et al., we use −.02 ($n = 427$) as the intercorrelation between structured interviews and integrity tests in Table 1. Because the intercorrelation between structured interviews and integrity tests is based on a single study, SD_{Res} cannot be calculated. Therefore, we used the average of the other SD_{Res} values in the matrix (0.108).

Structured Interviews and SJTs

Given that SJTs are a new addition to the matrix, we conducted a literature search to find studies that correlated structured interviews with SJTs. We located 11 studies including a correlation between an interview and SJT. We excluded four of these studies because we could not determine whether or how much range restriction affected the sample (Heggarty et al., 2020; Husbands et al., 2018), whether the interview was structured (Husbands et al., 2018; Yingling et al., 2018), or the constructs measured in the SJT were very different from the SJT constructs in the rest of the samples (i.e., emotional management SJT in Knorr et al., 2018). This left seven studies which included eight independent samples, which we describe in the following.

Morgeson et al. (2005) reported an uncorrected correlation of .23 between a structured interview and SJT in a sample of 90 job incumbents. The incumbents in this study were not hired using the study measures. So, range restriction was minimal. In Husbands et al. (2015), 200 medical school candidates completed a multiple mini interview (MMI; which was similar to a structured interview) and an SJT. They were not selected based on their SJT or interview scores. Hence, we assumed range restriction to be minimal and therefore included their uncorrected $r = .32$. In Patterson et al. (2016), an entire applicant pool of 1,594 general practitioner registrars took an MMI (i.e., past behavior structured interview) and SJT. Thus, there was no range restriction and we therefore included the uncorrected correlation of .53. Patterson et al. (2009) contained two samples with structured interview–SJT correlations. In both samples, there was minimal range restriction because applicants were not screened using the SJT or MMI. So, we included the uncorrected correlations of .52 ($n = 837$) and .53 ($n = 3{,}231$). In Roberts et al. (2014), 1,382 Australian applicants for specialty training completed an MMI and SJT. As this was an applicant sample, there was no range restriction. Thus, the uncorrected $r = .26$ correlation was included. Finally, we received unpublished data related to two published articles. Heimann et al. (2020) provided correlations between structured interview ratings of interviewees' Big Five personality dimensions and leadership SJT scores for 223 job incumbents. Neither the interview nor SJT were used in selection, so range restriction should be minimal. For the interview, there were both past behavior and situational interview ratings. The leadership SJT

[B1] Ones (1993) did not report how many primary studies were concurrent validity studies, but Sackett et al. (2022) demonstrated that most studies contributing to personnel selection method meta-analyses are concurrent. Additionally, other integrity test meta-analyses, including ones coauthored by Ones, have included a majority of concurrent studies. Specifically, in Ones et al. (1993), 63% and 68% of the studies relating integrity tests to job performance and counterproductive behaviors, respectively, were concurrent. Similarly, in Van Iddekinge et al. (2012), 54% and 79% of the studies relating integrity tests to job performance and counterproductive work behavior, respectively, were concurrent.

(*Appendices continue*)

measured dimensions describing positive (transformational) as well as negative (transactional) leadership behavior. When averaging correlations related to these SJT leadership dimensions, we took into account that for transactional leadership dimensions a negative correlation with interview scores is anticipated, whereas for transformational leadership dimensions a positive correlation is anticipated; and thus we changed the signs of correlations as necessary when forming these dimensions into an overall composite. We used a composite formula (Ghiselli et al., 1981, p. 174) to estimate the correlation between a composite of the past behavior and situational interview scores and a composite of the interview dimensions. The overall composite correlation was .18. In Heimann et al. (2021), 204 job incumbents (who were not selected using the study instruments, so range restriction was minimal) took the same leadership SJT and a structured interview (again containing both past behavior and situational interview questions) that included task-, relationship-, and change-oriented leadership dimension scores. The same composite formula approach was used to estimate the correlation between the overall interview and SJT scores. This led to a correlation of .26.

We then meta-analyzed the correlations from those eight samples. The resulting meta-analytic correlation was .45 ($n = 7{,}761$), which is the value we used in Table 1. SD_{Res} was 0.12. We also provide full meta-analytic results in Table B1.

Integrity Tests and SJTs

We carried out another literature search to find studies that correlated self-report integrity tests with SJTs. This led to five studies containing eight independent samples. All these studies correlated the SJT with the honesty–humility scale from the HEXACO model of personality. Past scholarship has made the case that measures of honesty–humility are similar to personality-based integrity tests (Berry, Sackett, & Wiemann, 2007). Further, we note that Ones et al. (1993) found that the mean uncorrected correlation between various personality-based integrity tests is .43; while Lee et al. (2019) found that the uncorrected correlation between honesty–humility and integrity tests was .44. So, honesty–humility correlates with integrity tests about as strongly as personality-based integrity tests correlate with each other. Thus, we believe it is reasonable to base our estimate of the correlation between integrity tests and SJTs on these honesty–humility studies.

Inspection of the five integrity–SJT studies indicated that some SJTs were specifically designed to target the construct (integrity) with which we want to correlate the SJT, whereas in other studies the SJT was not designed to explicitly capture integrity. There is consensus that (similar to structured interviews) SJTs are measurement methods that might capture a multitude of constructs (Christian et al., 2010; McDaniel et al., 2007). Hence, this is also how we conceptualized the SJT in our Table 1 matrix. In other words, similar to structured interviews, the "SJT" in the matrix is not an SJT specifically designed to target integrity or any other construct. Therefore, we excluded three of the five studies (i.e., de Leng et al., 2018; de Meijer et al., 2010; Husbands et al., 2015) because they designed an SJT to solely measure integrity and then correlated it with a self-report integrity test.

This left us with two studies containing five samples. Oostrom et al. (2019) contained three samples. Study 1 included a general population sample ($n = 72$), so range restriction should be minimal; Study 2 included a full applicant pool ($n = 157$) with no range restriction; and Study 3 included job incumbents ($n = 110$), but they were not selected using the integrity test or SJT, so range restriction should be minimal. In all three samples, honesty–humility was correlated with an SJT designed to measure the six HEXACO personality traits. Because separate scores were provided on the SJT for the six HEXACO dimensions, we used a composite formula (Ghiselli et al., 1981, p. 164) to estimate the correlations between honesty–humility and the overall SJT in each sample. These composite correlations were .26, .35, and .23 in Studies 1, 2, and 3, respectively. Barends et al. (2022) contains two samples. Study 1 included 116 Dutch graduates and Study 2 included 287 MTurkers; neither the SJT nor integrity test were used to select participants, so range restriction should be minimal. In their online supplemental material, Barend et al. reported correlations for each sample between a measure of honesty–humility and an SJT designed to measure the other five traits in the HEXACO model. Because they did not provide intercorrelations between the five SJT dimensions, we simply averaged the correlations between honesty–humility and the five SJT dimensions. The average correlations were .03 and .05 for Studies 1 and 2, respectively.

We then meta-analyzed the correlations from those five samples. The resulting meta-analytic correlation was .16 ($n = 742$), which is the value we used in Table 1. SD_{Res} was 0.10. We also provide full meta-analytic results in Table B1.

Updates to Black–White Standardized Mean Differences on the Selection Methods

Biodata

Roth et al. (2011) Black–White mean difference estimate for biodata was $d = .57$, drawn from Potosky et al. (2005). Potosky et al.'s estimate came from two studies: Dean (1999; $d = .73$) dissertation (subsequently published as Dean, 2013) and Kriska (2001; $d = .27$). We suggest the $d = .73$ from Dean is implausible. The uncorrected d-value in Dean was .33, which was then corrected for indirect range restriction due to selection on GMA tests to arrive at the corrected $d = .73$. However, the correlation between GMA tests and biodata reported in Dean was only .11, so it is not possible for indirect range restriction due to selection on GMA tests to reduce that d-value from .73 to .33. Given the very small correlation between biodata and GMA tests in that sample, the actual range restriction-corrected d-value would be very close to the uncorrected $d = .33$. In any case, since the publication of Roth et al. and Potosky et al., Tenbrink et al. (2021) reported a meta-analysis of the Black–White mean difference on biodata that is based on 10 studies, including Kriska and Dean (although Tenbrink et al. used Dean's uncorrected d-value, which we view as appropriate). We were able to locate and examine eight of the 10 studies included in Tenbrink et al. In none of the eight was the sample selected using the biodata measure, so any range restriction should be minimal. Thus, we use Tenbrink et al.'s meta-analytic Black–White uncorrected $d = .32$ ($k = 10$, $n = 24{,}359$) in Table 1.

GMA Tests

Roth et al. (2011) used two separate Black–White mean difference estimates for GMA tests: one for medium complexity

(*Appendices continue*)

jobs ($d = .72$, $n = 31{,}990$) and one for low complexity jobs ($d = .86$, $n = 125{,}654$), both of which were drawn from Roth et al. (2001). Sackett et al. (2022) averaged these two d-values to arrive at $d = .79$, which is the Black–White GMA test d-value we use in Table 1.

Conscientiousness Tests

Roth et al. (2011) Black–White mean difference estimate of $d = .06$ for conscientiousness tests was drawn from Potosky et al. (2005), which was based on three samples. We instead drew our $d = -.07$ estimate from Foldes et al. (2008) meta-analysis of 67 samples ($n = 180{,}478$). Although Foldes et al. included a mix of applicant, incumbent, and student samples, we think it is unlikely that range restriction significantly affects this d-value for a few reasons. First, we believe it is unlikely that most of Foldes et al.'s samples were selected using the conscientiousness measure. Second, past research has generally found minimal range restriction for personality traits (e.g., Sackett et al., 2022). Third, the uncorrected $d = -.07$ is so small in the first place that any range restriction correction would leave it almost unchanged.

Structured Interviews

Roth et al. (2011) Black–White mean difference estimate for structured interviews was $d = .32$. They arrived at $d = .32$ by adding a single $d = .44$ from a primary study by Roth et al. (2002) to the meta-analytic d-value of .31 reported by Potosky et al. (2005). To arrive at $d = .31$, Potosky et al. applied a direct range restriction correction to Huffcutt and Roth (1998) uncorrected $d = .23$ using a u-ratio of .74 from Huffcutt and Arthur (1994). We discussed above why such a range restriction correction was inappropriate for the correlation between GMA tests and structured interviews, and we think the same reasoning applies in this case. Dahlke and Sackett (2017) review of Black–White mean differences on selection methods simply used Huffcutt and Roth's uncorrected $d = .23$, which we view as more appropriate, based on the principle of conservative estimation. We added Roth et al. (2002) primary study $d = .44$ to Huffcutt and Roth's uncorrected meta-analytic $d = .23$, which resulted in $d = .24$ ($n = 9{,}175$), which is the Black–White mean difference for structured interviews that we include in Table 1.

Integrity Tests

Roth et al. (2011) Black–White mean difference estimate for integrity tests was .04 ($n = 481{,}523$), taken from a large-scale study of overt integrity tests carried out by Ones and Viswesvaran (1998). Dahlke and Sackett (2017) more recent review presented an estimate of $d = .10$ ($n = 882{,}781$), which was the average of their estimates of .04 for overt integrity tests (also taken from Ones and Viswesvaran) and .16 for personality-based integrity tests (taken from Ones, 1993). We thus use Dahlke and Sackett's $d = .10$ in Table 1.

Situational Judgment Tests

Roth et al. (2011) did not include a Black–White mean difference for SJTs in their meta-analytic matrix. We drew our Black–White mean difference on SJTs estimate from Dahlke and Sackett (2017), who obtained their estimates from Whetzel et al. (2008). We averaged Dahlke and Sackett's behavioral tendency SJT ($d = .34$, $k = 17$, $n = 5{,}380$) and knowledge SJT ($d = .39$, $k = 45$, $n = 36{,}348$) mean differences to arrive at $d = .37$, which is the d-value we use in Table 1.

Received November 9, 2022
Revision received March 6, 2024
Accepted March 18, 2024 ■