Diversity Shrinkage: Cross-Validating Pareto-Optimal Weights to Enhance Diversity via Hiring Practices

Q. Chelsea Song  
University of Illinois at Urbana-Champaign

Serena Wee  
Singapore Management University

Daniel A. Newman  
University of Illinois at Urbana-Champaign

To reduce adverse impact potential and improve diversity outcomes from personnel selection, one promising technique is De Corte, Lievens, and Sackett’s (2007) Pareto-optimal weighting strategy. De Corte et al.’s strategy has been demonstrated on (a) a composite of cognitive and noncognitive (e.g., personality) tests (De Corte, Lievens, & Sackett, 2008) and (b) a composite of specific cognitive ability subtests (Wee, Newman, & Joseph, 2014). Both studies illustrated how Pareto-weighting (in contrast to unit weighting) could lead to substantial improvement in diversity outcomes (i.e., diversity improvement), sometimes more than doubling the number of job offers for minority applicants. The current work addresses a key limitation of the technique—the possibility of shrinkage, especially diversity shrinkage, in the Pareto-optimal solutions. Using Monte Carlo simulations, sample size and predictor combinations were varied and cross-validated Pareto-optimal solutions were obtained. Although diversity shrinkage was sizable for a composite of cognitive and noncognitive predictors when sample size was at or below 500, diversity shrinkage was typically negligible for a composite of specific cognitive subtest predictors when sample size was at least 100. Diversity shrinkage was larger when the Pareto-optimal solution suggested substantial diversity improvement. When sample size was at least 100, cross-validated Pareto-optimal weights typically outperformed unit weights—suggesting that diversity improvement is often possible, despite diversity shrinkage. Implications for Pareto-optimal weighting, adverse impact, sample size of validation studies, and optimizing the diversity-job performance tradeoff are discussed.

Keywords: adverse impact, cognitive ability/intelligence, cross-validation, diversity, Pareto-optimal weighting

When implementing personnel selection procedures to hire new employees, some managers and executives place value on the diversity of the new hires (Avery, McKay, Wilson, & Tonidandel, 2007). Diversity can be valued for a variety of reasons, including moral and ethical imperatives related to fairness and equality of the treatment, rewards, and opportunities afforded to different demographic subgroups (Bell, Connerley, & Cocchiara, 2009; Gilbert & Ivancevic, 2000; Gilliland, 1993; Newman, Hanges, & Outtz, 2007). Some scholars and professionals have also articulated a business case for enhancing organizational diversity, based on the ideas that diversity initiatives improve access to talent, help the organization understand its diverse customers better, and improve team performance (see reviews by Jayne & Dipboye, 2004, Kochan et al., 2003, and Konrad, 2003, which suggest only mixed and contingent evidence for diversity’s effects on organizational performance; as well as more recent and nuanced reports on race and gender diversity’s performance benefits from Avery et al., 2007; Joshi & Roh, 2009; Roberson & Park, 2007). In short, whereas the jury is still out on whether and when diversity enhances group performance, diversity remains an objective in itself that is valued by many organizations.

The attainment of organizational diversity is, in large part, a function of hiring practices. There are thus potentially far-reaching negative consequences for attaining diversity when an organization uses a selection measure demonstrating adverse impact potential. To elaborate, “adverse impact has been defined as subgroup differences in selection rates (e.g., hiring, licensure and certification, college admissions) that disadvantage subgroups protected under Title VII of the 1964 Civil Rights Act. Protected subgroups are defined on the basis of a number of demographics, including race, sex, age, religion, and national origin (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor & Department of Justice, 1978)” (Outtz & Newman, 2010, p. 53). Adverse impact is evidenced by the ad-
verse impact ratio (AI ratio)—which is calculated as the selection ratio for one subgroup (e.g., number of selected minority applicants divided by total number of minority applicants) divided by the selection ratio for the subgroup with the highest selection ratio (e.g., number of selected majority applicants divided by total number of majority applicants). As a rule of thumb, when the AI ratio falls below 4/5ths (i.e., when the minority selection ratio is 80% or less of the majority selection ratio), this is generally considered prima facie evidence of adverse impact discrimination (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor & Department of Justice, 1978). To be more specific, the “4/5ths rule” is not a legal definition of adverse impact, but rather provides a guideline for practitioners and for Federal enforcement agencies when deciding whether to further investigate discrepant selection rates. When the adverse impact ratio is low, organizations can face reputational risk and legal charges of discriminatory practice (Outtz & Newman, 2010; Sackett, Schmitt, Ellingson, & Kabin, 2001).

The issue of adverse impact often arises from using one particular selection device (i.e., the cognitive ability test) that exhibits among the strongest criterion validities (Schmidt & Hunter, 1998) and also large mean differences between racial groups (Goldstein, Scherbaum, & Yusko, 2010; Hough, Oswald, & Ployhart, 2001). The classic problem is that subgroup mean differences on valid tests result in differences in selection rates between the majority and the minority group, which often disadvantage the minority group in gaining access to jobs (Bobko, Roth, & Potosky, 1999; Outtz & Newman, 2010).

For example, cognitive tests have a meta-analytic estimated operational relationship with job performance of r = .52 (medium-complexity jobs; corrected for range restriction and unreliability in the criterion only; Roth, Switzer, Van Iddekinge, & Oh, 2011, p. 905), but also an average Black–White mean difference of d = .72 (within-jobs; for medium-complexity jobs; Roth, Bevier, Bobko, Switzer, & Tyler, 2001, p. 314; though we note the average Black–White difference on actual job performance is less than half as large as the Black–White difference on cognitive tests [see McKay & McDaniell, 2006]). Hence, organizations often face the seemingly incompatible goals of selecting either for high levels of expected job performance (by maximizing criterion validity) or selecting for greater racial diversity (by maximizing the AI ratio). That is, organizations often face a “performance versus diversity dilemma” (Sackett et al., 2001, p. 306; Ployhart & Holtz, 2008).

The trade-off required between fulfilling job performance and diversity objectives has driven research efforts to develop selection systems with substantial criterion validity but less adverse impact potential. In this enterprise, one important constraint is that the Civil Rights Act of 1991 prohibits treating predictor scores differently for different racial subgroups (e.g., using different cut scores for different racial groups, or including applicant minority status/race in the predictor composite). As such, we highlight that contemporary hiring strategies designed to enhance diversity—including the Pareto-weighting strategy described in the current paper—in no way include minority status or race of individual applicants in the predictor composite. Under this legal constraint, several alternative strategies have been traditionally recommended to accomplish both job performance and diversity objectives simultaneously: (a) using “low impact” predictors that exhibit smaller subgroup differences than cognitive ability tests do (e.g., Hough et al., 2001), (b) reducing irrelevant predictor variance (e.g., Arthur, Edwards, & Barrett, 2002), and (c) recruitment strategies (e.g., Newman & Lyon, 2009; for a review, see Ployhart & Holtz, 2008). Nonetheless, most of the research on adverse impact reduction has investigated different ways of combining predictors (e.g., Bobko et al., 1999; De Corte, Lievens, & Sackett, 2007; Pulakos & Schmitt, 1996; Sackett & Ellingson, 1997; Schmitt, Rogers, Chan, Sheppard, & Jennings, 1997; Wee, Newman, & Joseph, 2014; see Ryan & Ployhart, 2014, for a review). Although these studies indicate that predictor combination is one useful way to deal with the performance-diversity dilemma, many of these studies (Pulakos & Schmitt, 1996; Sackett & Ellingson, 1997; Schmitt et al., 1997) used approaches based on “trial-and-error” (as characterized by De Corte et al., 2007, p. 1381) to examine the effect of weighting strategies in generating acceptable trade-offs between job performance and diversity objectives.

More recently however, De Corte and colleagues (De Corte et al., 2007; De Corte, Sackett, & Lievens, 2011) proposed a combinatorial approach using Pareto-optimally derived predictor weights that allowed the examination of all possible trade-offs between job performance and diversity outcomes. Instead of optimizing on a single criterion (as done when using regression weights), or ignoring optimization (as done when using unit weights), Pareto-optimal weights optimize on multiple criteria simultaneously, providing an optimal solution on one objective, at a given value of another objective. Specifically, a Pareto-optimal solution provides the best possible diversity outcome (AI ratio) at a given level of job performance (criterion validity); it also provides the best possible job performance outcome at a given level of diversity. Using Pareto-optimal weights, De Corte and colleagues (De Corte et al., 2007; De Corte, Sackett, & Lievens, 2008) showed how improvements in both criterion validity and the AI ratio could be achieved via a Pareto-weighted predictor composite formed with a cognitive test and noncognitive predictors (e.g., structured interview), as compared with a cognitive test alone. Also building on De Corte et al.’s (2007) work, Wee et al. (2014) further illustrated the utility of Pareto-optimal weighting by showing how Pareto-optimal weighting of cognitive ability subtests could result in substantial diversity improvement (i.e., adverse impact reduction resulting from using Pareto-optimal weights rather than unit weights, with no loss in expected job performance).

Although the aforementioned studies provided evidence that Pareto-optimal weighting could be an effective strategy to deal with the trade-off between job performance and diversity, a key caveat remains: we do not know the extent to which the Pareto-optimally derived diversity improvement obtained in a particular calibration sample will cross-validate. Before we can determine the practical value of using Pareto-optimal weighting to enhance diversity, we should determine whether diversity improvements generalize beyond the sample on which the weights were obtained. That is the purpose of the current article—to examine the cross-validity of Pareto-optimal solutions. Specifically, by conducting Monte Carlo simulations, we examined the extent to which Pareto-optimal solutions shrink (i.e., the extent to which expected job performance and expected diversity decrease in size when the

---

1 Both De Corte et al. (2007) and Wee et al. (2014) examined sensitivity (i.e., sample-to-sample variation) in the Pareto-optimal results that were presented, but neither investigated the cross-validation/shrinkage of Pareto-optimal solutions.
predictor weights obtained in the original sample are applied to new samples), under different predictor combinations and sample sizes.

We attempt to make three contributions to the personnel selection literature. First, we extend the concept of cross-validation from a context where only a single objective is optimized (as in multiple regression), to a context where multiple objectives are optimized. We distinguish between the notions of validity shrinkage (i.e., loss of predicted job performance) and diversity shrinkage (i.e., loss of predicted diversity). This distinction is important because we hypothesize that shrinkage covaries with optimization. That is, when two objectives are simultaneously considered, we expect that the extent of shrinkage on one objective (e.g., job performance) depends on the extent to which it is maximized in relation to the second objective (e.g., diversity). Such phenomena are central to understanding the generalizability and practical value of the Pareto-optimal weighting strategy. Second, we investigate the generalizability of Pareto-optimal weights by examining the extent to which different predictor combinations and sample sizes impact validity shrinkage and diversity shrinkage. If the cross-validity of Pareto-optimal weights shrinks drastically across various predictor combinations and sample sizes, the benefit of this strategy for addressing diversity concerns is constrained; the weights generated in a given sample would be less useful in making predictions in other samples, except in rare cases. In contrast, if Pareto-optimal weights demonstrate little shrinkage under certain conditions, then the strategy would be promising under those conditions. Thus, the current article seeks to ascertain the boundary conditions for when Pareto-optimal weighting may be used in organizational settings. Third, we compare the shrunk Pareto-optimal solutions with unit-weighted solutions. Unit-weighted solutions are often recommended (Bobko, Roth, & Buster, 2007; see Einhorn & Hogarth, 1975) and widely used in selection practice, and thus serve as a natural point of comparison. In addition, because unit-weighted solutions are not based on any kind of optimization technique, they have the desirable property that they do not demonstrate shrinkage when subject to cross-validation (Bobko et al., 2007).

In what follows, we first describe the concept of cross-validation as a way of assessing the generalizability of a predictive model, and very briefly summarize extant approaches to cross-validation in the personnel selection context (e.g., Cascio & Aguinis, 2005; Cattin, 1980; Raju, Bilgic, Edwards, & Fleer, 1999). Because cross-validation research in personnel selection has thus far focused only on single-objective optimization (e.g., maximizing expected job performance), we leave our discussion of shrinkage when multiple objectives are optimized until after we describe optimization and the Pareto technique. Then, we outline how we expect validity shrinkage and diversity shrinkage to occur, and change, across the entire Pareto-optimal surface.

We conducted a simulation to examine the conditions under which Pareto-optimal weights demonstrate substantial cross-validity. To compare our Pareto shrinkage results against more widely known regression-based results, we first examined the degree of validity shrinkage and diversity shrinkage when only one objective was maximized. We then extended the results to multicriterion optimization, and we also compared cross-validated Pareto-optimal solutions to unit-weighted solutions.

### Shrinkage and Cross-Validation Using Multiple Regression Weights

Multiple regression, where predictors are weighted to maximize the variance accounted for (\( R^2 \)) in a single criterion (e.g., job performance), is typically used because it provides the best linear unbiased estimates of the predictor weights for predicting the criterion. However, we must consider shrinkage when these weights from one sample are used to predict the criterion in another sample, or to generalize to the population (Larson, 1931; Wherry, 1931). Because the weights provided by any model-fitting procedure capitalize on the unique characteristics of the sample, these weights do not generate an optimal solution when applied to a different sample. That is, the predictive power of the model (i.e., the sample’s squared multiple correlation coefficient, \( R^2 \)) is optimized in the sample from which the weights were derived; but the model is overfitted (Yin & Fan, 2001) and shrinkage will likely occur when the weights are applied to a new sample. Shrinkage describes the extent to which the variance explained in the population or in a new sample is smaller than the variance explained in the original sample. Shrinkage parameters such as the squared population multiple correlation coefficient (\( \rho^2 \)) and the squared cross-validated multiple correlation coefficient (\( \rho_{cv}^2 \)) index the relationship between the set of predictors and the criterion in the population, and in a sample from the population, respectively (see Appendix A for details).

The procedure for assessing shrinkage is referred to as cross-validation. There are two approaches to estimate cross-validity: an empirical approach and a statistical approach (Cascio & Aguinis, 2005). The empirical approach consists of fitting a regression model to data from a calibration sample, and then applying the obtained predictor weights to an independent, validation sample. The obtained result, \( \rho_{cv}^2 \), does not suffer from overfitting and thus provides a generalizable estimate of the variance accounted for in the criterion by the predictors. The statistical approach uses formulas to adjust the squared sample multiple correlation coefficient, \( R^2 \), by the sample size (\( N \)) and the number of predictors (\( k \)) used to obtain the sample estimate (\( \hat{R}^2 \)). Larger adjustments are made when estimates are based on smaller samples, more predictors, or smaller coefficient values (Cattin, 1980). The statistical approach is generally preferred over the empirical approach as it provides results as accurate as the empirical approach, but more cost-effectively, given that a second sample is not required (Cascio & Aguinis, 2005).

### The Pareto-Optimal Weighting Strategy

De Corte et al. (2007) proposed a Pareto-optimal weighting strategy as a way to systematically estimate (locally) optimal solutions. Under Pareto-optimal weighting, predictors are weighted to optimize two criteria: job performance and diversity. Instead of a single, globally optimal outcome (as with multiple regression), Pareto-optimal weights produce a frontier (e.g., see Figure 1) of possible selection outcomes simultaneously involving two criteria. This is represented as a downward-sloping trade-off curve with job performance (e.g., criterion validity) plotted against diversity (e.g., AI ratio). The negative slope of the Pareto curve in Figure 1 indicates that additional diversity can only be obtained at some cost in terms of job performance (cf. De Corte, 1999). A
steeper slope of the trade-off curve would indicate that the job performance cost is high for obtaining additional diversity, whereas a flatter slope would indicate that the job performance cost is low for obtaining additional diversity.

Figure 1 shows the Pareto-optimal trade-off curve (left panel) and the corresponding weights for each of three predictors: A, B, and C (right panel). As shown in Figure 1 (left panel), the trade-off curve represents a range of possible selection outcomes, in terms of both job performance and diversity. Each point on the curve represents a unique solution given by a set of Pareto-optimal predictor weights. As an example, the set of weights corresponding to Point 2 in Figure 1 (left panel) shows a Pareto-optimal solution where job performance is maximized (that is, where only job performance was maximized, and diversity was not considered). Similarly, Point 11 in Figure 1 (left panel) shows a Pareto-optimal solution where diversity is maximized; job performance was not considered. In contrast, Point 6 in Figure 1 (left panel) shows a Pareto-optimal solution where job performance and diversity objectives were both considered—corresponding to a selection scenario where job performance and diversity are both explicitly valued. It should be noted that every point along the Pareto-optimal curve (i.e., Points 2 through 11) is locally optimal, whereas points below the curve (e.g., Point 1) are suboptimal. That is, for Point 1 (which corresponds to unit-weighting), there will be a Pareto-optimal solution with the same level of job performance, but with higher levels of diversity. Thus, for Point 1, diversity improvements could be made by using Pareto-optimal rather than unit weights.

Shrinkage and Cross-Validation Using Pareto-Optimal Weights

As highlighted earlier, the predictor weights provided by any model-fitting procedure capitalize on the unique characteristics of the sample. Thus, as with multiple regression, variance accounted for under Pareto-optimal weighting should also be adjusted for shrinkage. Although a statistical approach would be preferred over an empirical approach, the currently available shrinkage formulas are based on least-squares assumptions (Dorans & Drasgow, 1980) that do not form the basis for Pareto-optimization. Therefore, in the current study we examined the issue of shrinkage in Pareto weights by using an empirical approach (i.e., Monte Carlo simulation).

When using Pareto-optimal weighting to reduce adverse impact, the weights simultaneously optimize both job performance and diversity. Because both objectives are optimized by the given set of Pareto weights (although the two objectives are achieved to differing extents at different points on the Pareto frontier), the issue of shrinkage is more complicated than in the single objective, multiple regression case. That is, optimizing two objectives at the same time complicates shrinkage. For each set of Pareto weights, we must consider the degree of shrinkage on both of the objectives being optimized. Specifically, we refer to shrinkage on the job performance objective as validity shrinkage, and to shrinkage on the diversity objective as diversity shrinkage; and we assess each in turn. As illustrated in Figure 1, Points 2 and 11 indicate Pareto-optimal solutions where only one objective is maximized.
(i.e., job performance is maximized at Point 2, and diversity is maximized at Point 11). Shrinkage in these situations should be akin to shrinkage in the multiple regression case. When maximizing job performance (e.g., Point 2 in Figure 1), we expect considerable validity shrinkage and negligible diversity shrinkage. Likewise, when maximizing diversity (Point 11 in Figure 1), we expect considerable diversity shrinkage and negligible validity shrinkage. This is because—at the endpoints of the Pareto curve—the predictor weights should be overfitted for the maximized objective, but not directly overfitted for the ignored objective.

By contrast, Points 3 through 10 in Figure 1 indicate solutions where both objectives are being valued. Whereas Point 3 indicates a solution where job performance is maximized to a greater extent than diversity is, and Point 10 indicates a solution where diversity is maximized to a greater extent than job performance is, Points 6 and 7 indicate solutions where job performance and diversity are considered similarly important and maximized to a similar extent. Extending the concept of shrinkage to these situations, we expect greater validity shrinkage at Point 3 than at Point 10, and greater diversity shrinkage at Point 10 than at Point 3. More generally, we expect shrinkage to be proportionally larger on the more valued objective (i.e., the objective to which more importance is attached when selecting a point on the Pareto curve). As such, to the extent diversity is being maximized, we expect greater diversity shrinkage.

As reviewed earlier, cross-validation research using multiple regression indicates greater shrinkage should be expected when estimates are based on small samples (N), a large number of predictors (k), and predictor composites with low predictive power (low R²; Cattin, 1980). We expect these same factors would also influence the degree of shrinkage in Pareto-optimal solutions. What remains unknown, however, is the extent to which each of these factors independently influences validity shrinkage and diversity shrinkage, along the entire frontier of Pareto-optimal selection solutions.

To ensure the simulation conditions we considered were applicable to actual selection contexts, we examined Pareto-optimal shrinkage using two distinct sets of predictor combinations: (a) a set of cognitive and noncognitive predictors comprising a cognitive ability test, a biodata form, a conscientiousness measure, a structured interview, and an integrity test (see Table 1; similar to that used by De Corte et al., 2008); and (b) a set of cognitive ability test predictors comprising tests of verbal ability, mathematical ability, technical knowledge, and clerical speed (see Table 2; as used by Wee et al., 2014). We chose these particular predictor combinations because, to our knowledge, they are the primary sets of predictor combinations for which Pareto-optimization has previously been demonstrated. Our study thus provides an important test of the extent to which these published results would generalize under cross-validation.

Given that shrinkage is affected by sample size, and that sample sizes vary widely across selection contexts, we also examined the extent of validity shrinkage and diversity shrinkage across a range of sample sizes. Finally, we investigated the conditions under which cross-validated Pareto-optimal solutions reliably outperformed unit-weighted solutions. In summary, we addressed the following research questions:

Research Question 1: What is the extent of validity shrinkage and diversity shrinkage along the entire frontier of the Pareto-optimal trade-off curve, as a function of (a) sample size (i.e., number of applicants in the calibration sample) and (b) type of predictor combination (i.e., cognitive-noncognitive combination vs. a combination of cognitive subtests)?

Research Question 2: Is validity shrinkage larger when job performance is maximized, and is diversity shrinkage larger when diversity is maximized?

Research Question 3: When are unit weights outperformed by cross-validated Pareto-optimal weights? That is, under what conditions (sample size and predictor combination) would cross-validated Pareto-optimal weights result in greater diversity improvement than unit weights?

Method

In the current study, validity shrinkage and diversity shrinkage of Pareto-optimal weights were investigated via a Monte Carlo simulation with two factors: calibration sample size and predictor combination. The design incorporates five calibration sample sizes (N = 40, 100, 200, 500, and 1,000), and two predictor combinations: (a) a set of cognitive and noncognitive predictors (we refer to these as “Bobko-Roth predictors”; see Table 1 for details), and (b) a set of cognitive subtest predictors (from Wee et al., 2014; see Table 2 for details).

When demonstrating diversity shrinkage on the cognitive subtest predictors, a helpful reviewer encouraged us to limit our illustration to only two prototypical job families from the original seven job families investigated by Wee et al. (2014). We chose two job families that theoretically should differ in how much diversity improvement is possible. First, we chose one job family (Machin-
DIVERSITY SHRINKAGE

Table 2

| Cognitive Subtests Predictor Intercorrelations, Black–White Subgroup ds, and Criterion Validities |
|----------------------------------|---|---|---|---|
| Variable                          | 1   | 2   | 3   | 4   |
| ASVAB cognitive subtests          |     |     |     |     |
| 1. Verbal ability                 | 1.00|     |     |     |
| 2. Mathematical ability           | .79 | 1.00|     |     |
| 3. Technical knowledge            | .78 | .68 | 1.00|     |
| 4. Clerical speed                 | .46 | .64 | .23 | 1.00|
| Criterion validities (by job families) |     |     |     |     |
| Machinery repair                  | .52 | .58 | .54 | .36 |
| Control & communication           | .40 | .55 | .37 | .43 |
| Black–White subgroup d            | .91 | .69 | 1.26| .06 |

Note. Values obtained from Wee et al. (2014). Performance Black–White subgroup d was .30. The predictor intercorrelation matrix, criterion validities, and subgroup ds for the cognitive subtests condition (i.e., verbal ability, mathematical ability, technical knowledge, clerical speed) comes from the ASVAB matrix in Wee et al. (2014).

The procedure includes eight steps. The simulation was conducted using the TROFSS program (De Corte, 2006) and the R programming language (R version 3.2.0; R Core Team, 2015). For future users, we also developed an easy-to-use R package (https://github.com/Diversity-ParetoOptimal/ParetoR), as well as a simple web application (i.e., a ShinyApp), that both implement the Pareto-optimal weighting technique described in the current paper (see Appendix C.).

1. Generate calibration samples. One thousand calibration samples were generated for each condition, based on each corresponding population correlation matrix (Tables 1 and 2). For example, samples of a given size (e.g., N = 40) were drawn from a multivariate normal population distribution, and this process was repeated 1,000 times to form 1,000 calibration sample data sets. Each calibration sample data set consisted of predictor scores, job performance scores, and a binary variable indicating race. The sample correlation matrix and subgroup ds were then calculated for each calibration sample.

We note that race is represented as a binary variable in the current simulation. Thus, to maintain multivariate normality during the simulation step where we generate samples from the population correlation matrix, we needed to first convert the population subgroup ds into population correlations that included race.

Then we generated the sample data based on the population correlation matrix, and lastly converted the sample race correlations back into sample subgroup ds. Specifically, the following steps were used to generate the samples.

First, subgroup ds (i.e., for race) were converted into point-biserial correlations using the following formula (see Lipsy & Wilson, 2001):

\[ r_{pb} = d/\sqrt{d^2 + 1/(p(1-p))} \]

where \( p \) is the proportion of minority applicants (e.g., Black applicants), \((1-p)\) is the proportion of majority applicants (e.g., White applicants), and \( d \) is the standardized mean difference between the majority and minority applicant groups.

Second, point-biserial correlations were converted into biserial correlations using the following formula (see Bobko, 2001, p. 38):

\[ r_{bis} = r_{pb} \sqrt{p(1-p)/\lambda} \]

again where \( p \) is the proportion of minority applicants, \((1-p)\) is the proportion of majority applicants, and \( \lambda \) is the height (i.e., density) of the standard normal distribution at the point where the sample is divided into majority and minority applicant groups.

Third, the race biserial correlations were included in the population correlation matrix and random samples were drawn from a multivariate normal distribution. Each sample consists of predictor scores, job performance scores, and a continuous variable for race.

Fourth, the continuous race variable was dichotomized. To achieve this, the sample was first sorted based on values of the continuous race variable. Then, the first \((1-p) \times N\) simulated applicants were labeled “2” (indicating White applicant), with the

---

2 We note that one of the job families that was removed from the manuscript, the Weapons Operation job family, exhibited a set of shrunken Pareto weights that were similar to—but no better than—unit weights. In other words, for one of the seven ASVAB job families from Wee et al. (2014), unit weights were not suboptimal. This likely occurred because the bivariate criterion validities (as well as the optimal regression weights) were very similar across predictors (i.e., approximately equally weighted) for the Weapons Operation job family.
rest of the simulated applicants labeled “1” (indicating Black applicant). This produced a binary variable representing race.

Finally, we calculated the sample subgroup $d$ for each predictor from the raw data in each sample. As a check on the accuracy of the simulation procedures, the current simulation produced sample correlation matrices and sample subgroup $d$ values whose average was similar—within .01 on average—to the population correlation matrices and population subgroup $d$s shown in Tables 1 and 2. The current simulation procedure contrasts with previous procedures that have been used to generate separate multivariate normal distributions for each race (e.g., De Corte et al., 2007; Newman & Ostroff, 1993). The TROFSS program also requires as input the overall selection ratio (SR), the proportion of minority applicants ($p$), and the number of points along the Pareto-optimal curve to be generated (these points are spread evenly along the curve). To be consistent as possible with previous studies (De Corte et al., 2007; Wee et al., 2014), we set $SR = 0.15$ and the number of Pareto points at 21. We used $p = 0.157$ for the current simulations (i.e., 1/6 of applicants are Black), which is between the $p = 1/4$ example from De Corte et al. (2007, p. 1384) and the $p = 1/8$ used by De Corte et al. (2008, p. 187). There were 21 sets of Pareto-optimal weights estimated for each calibration sample. For a given set of weights, the calibration sample criterion validity (denoted $R_{perf, cal}$) is given by the correlation between the weighted predictor composite scores and the job performance scores. Similarly, the calibration sample diversity (denoted $AI_{ratio, cal}$) is monotonically related to the cross-validated diversity ($AI_{ratio, val}$). The weighted prediction composite score is obtained by using the Pareto-optimal weights on the individual predictors (e.g., biodata, cognitive ability, etc.). The weighted predictor composite score is obtained by using the Pareto-optimal weights on the individual predictors (e.g., biodata, cognitive ability, etc.).

2. Generate validation samples. Validation samples were generated in the same manner as calibration samples; 1,000 validation samples were generated. However, validation sample size was always set at $N = 10,000$. We chose an extremely large sample size to minimize the effect of random sample-to-sample variation due to validation sample size. A helpful reviewer also pointed out that our validation sample size of 10,000 is sufficient to make results reported to the second decimal place meaningful (Bedeian, Sturman, & Streiner, 2009). Thus, in our simulation, we were able to closely estimate the shrinkage from a calibration sample to the true local population.

3. For each replication, calculate calibration sample Pareto-optimal predictor weights, criterion validity, and diversity. The calibration sample correlation matrices and subgroup $d$s obtained in Step 1 were used as input for the TROFSS program (De Corte, 2006). The TROFSS program also requires as input the overall selection ratio ($SR$), the proportion of minority applicants ($p$), and the number of points along the Pareto-optimal curve to be generated (these points are spread evenly along the curve). To be consistent as possible with previous studies (De Corte et al., 2007; Wee et al., 2014), we set $SR = 0.15$ and the number of Pareto points at 21. We used $p = 0.157$ for the current simulations (i.e., 1/6 of applicants are Black), which is between the $p = 1/4$ example from De Corte et al. (2007, p. 1384) and the $p = 1/8$ used by De Corte et al. (2008, p. 187). There were 21 sets of Pareto-optimal weights estimated for each calibration sample. For a given set of weights, the calibration sample criterion validity (denoted $R_{perf, cal}$) is given by the correlation between the weighted predictor composite scores and the job performance scores. Similarly, the calibration sample diversity (denoted $AI_{ratio, cal}$) is monotonically related to the cross-validated diversity ($AI_{ratio, val}$). The weighted prediction composite score is obtained by using the Pareto-optimal weights on the individual predictors (e.g., biodata, cognitive ability, etc.).

The weighted prediction composite score is obtained by using the Pareto-optimal weights on the individual predictors (e.g., biodata, cognitive ability, etc.). A helpful reviewer also pointed out that our validation sample size of 10,000 is sufficient to make results reported to the second decimal place meaningful (Bedeian, Sturman, & Streiner, 2009). Thus, in our simulation, we were able to closely estimate the shrinkage from a calibration sample to the true local population.

That is, for each calibration sample, the inputs to the simulation include: (a) the correlation matrix among predictors and job performance, (b) the standardized Black–White subgroup difference on each of the predictors, (c) the overall selection ratio ($SR$), and (d) the proportion of minority applicants ($p$). The weighted predictor composite score is obtained by using the Pareto-optimal weights on the individual predictors (e.g., biodata, cognitive ability, etc.). We should be clear that the predictor composite only includes preemployment tests; it does not include minority status of the candidates. The Pareto optimal predictor composites were then correlated with job performance to obtain the calibration sample criterion validities (i.e., $R_{perf, cal}$), and correlated with applicant minority status (race) to obtain the correlation sample point-biserial multiple correlations (i.e., $R_{point-bis, cal}$), which were subsequently converted to $AI_{ratio, cal}$ using the formula above.

4. For each replication, apply calibration sample Pareto-optimal predictor weights to the validation sample, to obtain cross-validated (shrunken) criterion validity and diversity estimates. The calibration sample Pareto-optimal predictor weights obtained in Step 3 were applied to the validation sample raw data obtained in Step 2. For each of the 21 sets of predictor weights for each sample, the cross-validated criterion validity ($R_{perf, val}$) is estimated as the correlation, in the validation sample, between the weighted predictor composite scores and the job performance scores. Similarly, the cross-validated diversity ($AI_{ratio, val}$) which is monotonically related to the cross-validated diversity ($AI_{ratio, cal}$) to obtain the calibration sample
race point-biserial multiple correlation, $R_{race.val}$) is estimated as the correlation in the validation sample, between the weighted predictor composite scores and the binary race variable, at the given SR and $p$ (as described in Step 3 above).

5. For each replication, calculate unit-weighted criterion validity and diversity in the validation sample. Step 4 was repeated, using unit weights for the predictors instead of using the calibration sample Pareto-optimal weights. By giving each predictor a weight of 1.0, the unit-weighted criterion validity ($R_{unit}$) and diversity ($AI_{ratio}$) were obtained.

6. Aggregate data across replications. We aggregated the data to obtain the calibration and validation samples’ mean Pareto-optimal curves. To do so, we averaged across the 1,000 replications for each of the 21 Pareto points. For example, the mean calibration sample criterion validity ($\bar{R}_{pref.cal}$) at the Pareto point where criterion validity is maximized—that is, the first or leftmost point on the Pareto-optimal curve—is calculated by averaging the criterion validities obtained when criterion validity is maximized, across the 1,000 calibration samples. The mean calibration sample diversity ($AI_{ratio.cal}$) at the Pareto point where criterion validity is maximized—also the first point on the Pareto curve—is calculated by first averaging the race point-biserial multiple correlations ($R_{race.cal}$) obtained when criterion validity is maximized, across the 1,000 calibration samples, then converting $R_{race.cal}$ to $AI_{ratio.cal}$. This procedure was repeated for each of the 21 Pareto points in the calibration sample, and then again for each of the 21 Pareto points in the validation sample. We also obtained the mean unit-weighted criterion validity ($\bar{R}_{pref,unit}$) and diversity ($AI_{ratio,unit}$), across replications.

7. Plot the Pareto-optimal trade-off curves and unit-weighted solutions. Each Pareto-optimal curve (defined by 21 Pareto solutions) was graphed, by plotting the mean criterion validities (i.e., the job performance objective) against the mean AI ratios (i.e., the diversity objective). For each condition, separate curves were plotted for the calibration and validation samples. The unit-weighted solution was obtained by plotting the mean unit-weighted criterion validity ($\bar{R}_{pref,unit}$) against the mean unit-weighted diversity ($AI_{ratio,unit}$).

8. Calculate validity shrinkage and diversity shrinkage. For each point along the Pareto-optimal curve, the amount of validity shrinkage ($\Delta R$) was calculated by subtracting the mean criterion validity of the validation sample from the mean criterion validity of the calibration sample. Likewise, the amount of diversity shrinkage ($\Delta AI$ ratio) was calculated by subtracting the mean AI ratio of the validation sample from the mean AI ratio of the calibration sample.

Results

The Pareto-optimal trade-off curves for different sample size conditions are presented in the “rainbow graphs” in Figures 2, 3 and 4. The Pareto curves for the Bobko-Roth predictors are presented in Figure 2, and the curves for the cognitive subtest predictors, for each of the two job families, are presented in Figures 3 and 4. To understand how to interpret these figures, look at the example graph in Figure 5. In Figure 5, the Pareto curve for the calibration sample (i.e., solid line) and the Pareto curve for the validation sample (i.e., dotted line) are shown. The Pareto curve for the validation sample (dotted line in Figure 5) is the shrunk

Pareto curve. That is, each shrunk Pareto curve is the set of cross-validated Pareto-optimal solutions (i.e., the criterion validities and diversity/AI ratio values obtained by applying the calibration sample Pareto-optimal predictor weights to the validation sample).

In Figure 5, the fact that the validation Pareto curve consistently lies below and to the left of the calibration Pareto curve means that there is shrinkage in both criterion validity (expected job performance) and diversity (expected AI ratio) across the entire range of the Pareto curve. We use the term Pareto shrinkage to describe this shift of the Pareto curve (i.e., the shift of each Pareto solution, or each Pareto point) to a lower expected value of both criterion validity and diversity. In Figure 5, the diagonal line from calibration solution Point 4 to validation solution Point 4 is the Pareto shrinkage for solution Point 4. It is important to realize that Pareto shrinkage can be decomposed into two component vectors: (a) validity shrinkage—which is the cross-validation shrinleng along the vertical (job performance or criterion validity) axis, and (b) diversity shrinkage—which is the cross-validation shrinkage along the horizontal (diversity or AI ratio) axis. To restate, for a given sample size (e.g., $N = 40$), validity shrinkage is indicated by the vertical distance (e.g., dashed vertical line in Figure 5) between the criterion validities of the calibration (e.g., solid curved line) and validation (e.g., dotted curved line) samples for the same Pareto-point (e.g., solution Point 4). Similarly, diversity shrinkage is indicated by the horizontal distance (e.g., dashed horizontal line in Figure 5) between the diversity values (i.e., AI ratios) of the calibration and validation samples (e.g., horizontal distance between solid and dotted curved lines) for the same Pareto-point (e.g., solution Point 4). In Figure 5’s illustration, validity shrinkage and diversity shrinkage for Point 4 are two component vectors of Pareto shrinkage. In Figures 2 through 4, in addition to plotting the calibration and validation Pareto curves, we also plotted the unit-weighted solution (i.e., a black square).

Sample Size Effects on Validity Shrinkage and Diversity Shrinkage

Research Question 1 involves sample size effects on diversity shrinkage. Figures 2 through 4 showed consistent and intuitive findings regarding sample size: validity shrinkage and diversity shrinkage both decreased as sample size increased. For the Bobko-Roth (see Figure 2) and cognitive subtest (Figures 3 and 4) predictors, validity shrinkage when $N = 40$ (i.e., vertical distances between the solid and dotted lines) was substantially larger than validity shrinkage when $N = 1,000$ (i.e., vertical distances between the solid and dotted lines), at all points along the Pareto-optimal curve. The same pattern of results held for diversity shrinkage. For

---

6 We note that our procedure of averaging the multiple correlation ($R_{unit}$) across replications, rather than averaging the AI ratio across replications, has a distinct advantage. Our procedure minimizes the influence of extreme AI ratios on our results, which would have been obtained for individual replications where the denominator of the AI ratio was close to zero. The race multiple correlation ($R_{race}$) is bounded on the interval $[-1, +1]$, whereas the AI ratio has no upper bound. Extremely large and implausible AI ratios sometimes occur, especially when simulations are based on small sample sizes.

7 Results for the other five ASVAB job families from Wee et al. (2014) may be obtained from the first author.
the Bobko-Roth (see Figure 2) and cognitive subtest (Figures 3 and 4) predictors, diversity shrinkage when $N = 40$ (i.e., horizontal distances between the solid and dotted lines) was larger than when $N = 1,000$, for all points along the Pareto-optimal curve.

In general, the incremental benefit of sample size was also greater when sample sizes were small (i.e., when attempting to control shrinkage, there are diminishing returns to sample size). That is, increasing sample size from 40 to 100 ameliorated shrinkage more than increasing sample size from 100 to 200, because of a nonlinear relationship between shrinkage and sample size. One part of the sample size results that might not be entirely intuitive is the thresholds for calibration sample size that render shrinkage negligible, which we discuss in the sections below.

### Validity Shrinkage and Diversity Shrinkage When One Objective Is Maximized

As shown across both the Bobko-Roth (see Figure 2) and the cognitive subtest (Figures 3 and 4) predictors, the degree of validity shrinkage and diversity shrinkage changes from one end of the Pareto curve to the other. For Research Question 2, we are interested in whether validity shrinkage is larger when job performance is being maximized, and whether diversity shrinkage is larger when diversity is being maximized. To answer Research Question 2, we refer to Tables 3 and 4. To highlight the differences in validity shrinkage and diversity shrinkage when each objective was maximized, we compared the criterion validities and diversity values obtained at the endpoints of the Pareto-optimal curve. These results are presented for the Bobko-Roth predictors in Table 3, and for the cognitive subtest predictors (Machinery Repair and Control & Communication) in Table 4. The top halves of Tables 3 and 4 show the criterion validity ($R$), validity shrinkage ($\Delta R$), diversity value (i.e., AI ratio), and diversity shrinkage ($\Delta AI$ ratio) obtained when job performance (i.e., criterion validity) was maximized. The bottom halves of Tables 3 and 4 show the same parameters, but obtained under conditions when diversity was maximized. In each section (top half and bottom half) of Tables 3 and 4, the criterion validities and diversity values obtained in the calibration and validation samples are provided, for each calibration sample size (i.e., from $N = 40$ to 1,000). The amount of validity shrinkage ($\Delta R$) was calculated by subtracting the mean criterion validity obtained in the validation sample (i.e., cross-validated validity) from the mean criterion validity obtained in the calibration sample. The amount of diversity shrinkage ($\Delta AI$ ratio) was calculated by subtracting the mean AI ratio obtained in the validation sample (i.e., cross-validated AI ratio) from the mean AI ratio obtained in the calibration sample.

For the Bobko-Roth (see Table 3) and cognitive subtest (see Table 4) predictors, the results indicated greater shrinkage on the maximized objective. Specifically, diversity shrinkage was larger when diversity was maximized (i.e., $\Delta AI$ ratio values in the bottom halves of Tables 3 and 4), as compared with when criterion validity was maximized (i.e., $\Delta AI$ ratio values in the top halves of Tables 3 and 4). For example, for the Bobko-Roth predictors (Table 3, when $N = 40$), diversity shrinkage was $\Delta AI = 1.24$ when

---

**Figure 2.** Validity and diversity shrinkage (original/calibration and shrunken Pareto curves) for Bobko-Roth predictors using meta-analytically derived values presented in Table 1. See the online article for the color version of this figure.

**Figure 3.** Validity and diversity shrinkage (original/calibration and shrunken Pareto curves) for Machinery Repair job family using cognitive subtest predictors (Wee et al., 2014). See the online article for the color version of this figure.

---

8 As noted by Bobko (2001, p. 244), in Wherry’s (1931) classic shrinkage formula, the only factor by which validity ($R^2$) is reduced during shrinkage is $(n - 1)/(n - k - 1)$—that is, the ratio of $df$ total to $df$ error, which is roughly equivalent to $1/(1 - (k/n))$. In other words, increasing $n$ has a much bigger effect on shrinkage (i.e., a big effect on $1/(1 - (k/n))$ when $n$ is small.
diversity was maximized, but only $\Delta AI = 0.00$ when criterion validity was maximized. For the cognitive subtest predictors (Table 4, when $N = 40$), diversity shrinkage was $\Delta AI = 0.05$ (Machinery Repair) and $\Delta AI = 0.03$ (Control & Communication) when diversity was maximized, but was smaller at $\Delta AI = 0.01$ (Machinery Repair) and $\Delta AI = 0.00$ (Control & Communication) when criterion validity was maximized.

Likewise, validity shrinkage was larger when criterion validity was maximized (i.e., $\Delta R$ values in the top halves of Tables 3 and 4), as compared with when diversity was maximized (i.e., $\Delta R$ values in the bottom halves of Tables 3 and 4). For example, for the Bobko-Roth predictors (Table 3, when $N = 40$), validity shrinkage was $\Delta R = .06$ when criterion validity was maximized, and only $\Delta R = -.02$ when diversity was maximized. Similarly, for the cognitive subtest predictors (Table 4, when $N = 40$), validity shrinkage was $\Delta R = .04$ (for both Machinery Repair and Control & Communication job families) when criterion validity was maximized, but was smaller at $\Delta R = -0.01$ (for both Machinery Repair and Control & Communication job families) when diversity was maximized.

To summarize our answer to Research Question 2, when examining Table 3 (Bobko-Roth predictors) and Table 4 (cognitive subtest predictors), we see negligible amounts of shrinkage on the objective that was not maximized. That is, for both sets of predictors, there was little to no validity shrinkage when diversity was maximized, and little to no diversity shrinkage when criterion validity/job performance was maximized. To restate, diversity shrinkage is greatest when diversity is maximized, and validity shrinkage is greatest when validity is maximized.

Validity Shrinkage and Diversity Shrinkage Along the Entire Pareto-Optimal Curve

Aside from the endpoints of the Pareto curve, we note that results in the middle of the curve can also be easily interpreted. That is, diversity shrinkage was smaller in the middle of the curve than it was at the endpoint where diversity was maximized, but diversity shrinkage was also larger in the middle of the curve than at the endpoint where criterion validity was maximized. Similarly, validity shrinkage in the middle of the curve was smaller than at the endpoint where validity was maximized, but validity shrinkage in the middle of the curve was larger than at the endpoint where diversity was maximized. This pattern of results generalized across the different predictor combinations, as shown in Figures 2 through 4. To restate, within a given selection scenario, diversity shrinkage increases to the extent diversity is being maximized. Also, validity shrinkage increases to the extent validity is being maximized.

Diversity Shrinkage When Using Low-Impact Predictors

Research Question 1b dealt with the effects of different types of predictor combinations on diversity shrinkage. Two major differences were observed between the pattern of results obtained for the Bobko-Roth (cognitive and noncognitive) predictors versus the cognitive subtest predictors. First, the Bobko-Roth predictors were more strongly affected by diversity shrinkage than were the cognitive...
subtest predictors. This can be seen by observing the larger horizontal distances between the solid and dotted lines for the Bobko-Roth predictors (Figure 2 and Table 3) than for the cognitive subtest predictors (Figures 3 and 4, and Table 4). Second, as compared with the cognitive subtest predictors, the Bobko-Roth predictors required larger sample sizes before diversity shrinkage was negligible.

Let us arbitrarily define “negligible” diversity shrinkage as ΔAI ratio <.05 (i.e., where diversity shrinkage rounds to 0.0, rather than 0.1), for the sake of interpretation. In Figure 2 (i.e., Bobko-Roth predictors), the Pareto-optimal solutions based on calibration sample sizes of N = 500 or fewer all demonstrated substantial (non-negligible) diversity shrinkage. In contrast, in Figure 3 and Figure 4 (i.e., cognitive subtest predictors in the Machinery Repair and Control & Communication job families, respectively), there was negligible diversity shrinkage whenever the calibration sample size was at N = 100 or more (i.e., the horizontal distance between the solid and dotted lines is small).

In general, these results suggest that including low impact predictors tended to increase diversity shrinkage. As can be seen in Tables 1 and 2, the average subgroup d value of the Bobko-Roth predictors tended to increase diversity shrinkage.

### Table 3

**Criterion Validity and Diversity for Pareto Curve Endpoints Using Bobko-Roth Predictors**

<table>
<thead>
<tr>
<th>Calibration sample size</th>
<th>Criterion validity (R)</th>
<th>Diversity (AI ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unit-weighted validity</td>
<td>Cross-validated validity</td>
</tr>
<tr>
<td>N = 40</td>
<td>.58</td>
<td>.69</td>
</tr>
<tr>
<td>N = 100</td>
<td>.58</td>
<td>.67</td>
</tr>
<tr>
<td>N = 200</td>
<td>.58</td>
<td>.67</td>
</tr>
<tr>
<td>N = 500</td>
<td>.58</td>
<td>.66</td>
</tr>
<tr>
<td>N = 1,000</td>
<td>.58</td>
<td>.66</td>
</tr>
</tbody>
</table>

**Endpoint where criterion validity is maximized**

<table>
<thead>
<tr>
<th>Calibration sample size</th>
<th>Criterion validity (R)</th>
<th>Diversity (AI ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unit-weighted validity</td>
<td>Cross-validated validity</td>
</tr>
<tr>
<td>N = 40</td>
<td>.58</td>
<td>.25</td>
</tr>
<tr>
<td>N = 100</td>
<td>.58</td>
<td>.22</td>
</tr>
<tr>
<td>N = 200</td>
<td>.58</td>
<td>.22</td>
</tr>
<tr>
<td>N = 500</td>
<td>.58</td>
<td>.21</td>
</tr>
<tr>
<td>N = 1,000</td>
<td>.58</td>
<td>.21</td>
</tr>
</tbody>
</table>

**Endpoint where diversity is maximized**

<table>
<thead>
<tr>
<th>Calibration sample size</th>
<th>Criterion validity (R)</th>
<th>Diversity (AI ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unit-weighted validity</td>
<td>Cross-validated validity</td>
</tr>
<tr>
<td>N = 40</td>
<td>.60/.52</td>
<td>.64/.57</td>
</tr>
<tr>
<td>N = 100</td>
<td>.60/.52</td>
<td>.62/.57</td>
</tr>
<tr>
<td>N = 200</td>
<td>.60/.52</td>
<td>.62/.56</td>
</tr>
<tr>
<td>N = 500</td>
<td>.60/.52</td>
<td>.62/.56</td>
</tr>
<tr>
<td>N = 1,000</td>
<td>.60/.52</td>
<td>.62/.56</td>
</tr>
</tbody>
</table>

### Table 4

**Criterion Validity and Diversity for Pareto Curve Endpoints Using Cognitive Subtest Predictors (Machinery Repair Job Family/Control and Communication Job Family)**

<table>
<thead>
<tr>
<th>Calibration sample size</th>
<th>Criterion validity (R)</th>
<th>Diversity (AI ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unit-weighted validity</td>
<td>Cross-validated validity</td>
</tr>
<tr>
<td>N = 40</td>
<td>.60/.52</td>
<td>.37/.42</td>
</tr>
<tr>
<td>N = 100</td>
<td>.60/.52</td>
<td>.36/.43</td>
</tr>
<tr>
<td>N = 200</td>
<td>.60/.52</td>
<td>.35/.43</td>
</tr>
<tr>
<td>N = 500</td>
<td>.60/.52</td>
<td>.36/.43</td>
</tr>
<tr>
<td>N = 1,000</td>
<td>.60/.52</td>
<td>.36/.43</td>
</tr>
</tbody>
</table>

**Note.** Validity shrinkage (ΔR) refers to the mean criterion validity difference between calibration sample validity and validation sample (i.e., cross-validated) validity, and indicates job performance shrinkage. Diversity shrinkage (ΔAI ratio) refers to the mean diversity/AI ratio difference between calibration sample and validation sample values. AI = adverse impact.
DIVERSITY SHRINKAGE

1647

...predictors is lower than the average subgroup d value of the cognitive subtest predictors. Although the low-impact predictors (e.g., personality, structured interviews) might suggest that large diversity improvements are possible in the original calibration sample from which the Pareto-optimal weights were obtained, our current results suggest that caution is warranted in applying these weights to new samples (because of diversity shrinkage), especially if the Pareto-optimal weighting was conducted using a calibration sample size of 500 or smaller. In specific, when applying Pareto weights to Bobko-Roth predictors, using calibration samples of \( N = 500 \) or smaller, there is risk of non-negligible diversity shrinkage; the diversity improvement estimated using Pareto weighting in the calibration sample will be an overestimate. Nonetheless, even when diversity shrinkage is large and thus the expected diversity benefits are overstated by the calibration sample (when calibration \( N \leq 500 \)), it is still plausible that Pareto weights will yield better selection diversity outcomes in contrast to unit weighting. So, even when diversity improvement (i.e., Pareto-weighted diversity minus unit-weighted diversity; Figure 1) is overestimated, it can still be positive. That is, Pareto weighting can still potentially outperform unit weighting, despite diversity shrinkage.

Diversity Shrinkage and the 4/5ths Threshold

We should note that, when shrunken diversity improvement is positive (which it typically is), this means that some diversity enhancement is possible—not that adverse impact (in terms of the 4/5ths rule) can be completely removed. When using the current set of predictors (i.e., a set of predictors that includes cognitive tests), AI ratios above .8 (meeting the 4/5ths rule) could not be achieved without considerable loss of some criterion validity and expected job performance. Further, because of diversity shrinkage, the expected number of violations of the 4/5ths rule sometimes increased (i.e., the AI ratio shifted from “greater than .8” to “less than .8”) when calibration sample sizes were small. For example (as seen in Figure 3), when \( N = 40 \), a full 10 of the 21 Pareto solution points exhibit AI ratios safely above .8 in the calibration sample, but with corresponding shrunken AI ratios that fall below .8. So almost half of the Pareto solution points are too optimistic in regard to the 4/5ths rule, when \( N \) is very small (\( N = 40 \)). In contrast, when \( N = 100 \), only 5 of the 21 Pareto solutions show a change in 4/5ths rule violation due to diversity shrinkage. When \( N = 200 \), the number of changes in 4/5ths rule violations drops to only 3 out of 21; and when \( N \geq 500 \), there is only 1 of 21 Pareto solutions that exhibits changes in outcomes of the 4/5ths rule due to diversity shrinkage. Our interpretation is that the use of Pareto weighting to avoid violations of the 4/5ths rule can sometimes produce inflated expectations (under small sample size conditions), unless diversity shrinkage is taken into account.

Diversity Improvements Using Cross-Validated Pareto Weights Instead of Unit Weights

Can we reliably get diversity improvement, after shrinkage? To examine the practical utility of Pareto-optimal weighting, we compared the cross-validated Pareto-optimal solutions (i.e., dotted lines in Figures 2 through 4) with the unit-weighted solution (i.e., the black square in each figure). In terms of the cross-validated diversity improvement (i.e., the horizontal distance between the cross-validated diversity value and the unit-weighted diversity value; see Figure 1, left panel), the simulation results (Figures 2 through 4) showed that the cross-validated Pareto-optimal solutions outperformed the unit-weighted solution in all cases simulated, whenever the calibration sample size was at least 100. Thus, unit-weighting was suboptimal. That is, even after accounting for diversity shrinkage, Pareto weights yielded greater diversity outcomes (as compared with unit weights), while continuing to produce shrunken criterion validities/job performance that were as good as (and sometimes better than) unit weights. In other words, reliable diversity improvement was typically possible via Pareto weighting.

To explain this in a different way, we note the sample size required before cross-validated Pareto weights reliably outperformed unit weights. This critical sample size can be contrasted depending on the types of predictors (e.g., Figure 2 vs. Figure 3) and the particular job family (e.g., Figure 3 vs. Figure 4) considered. Based on our simulation results, the Bobko-Roth predictors (see Figure 2) required sample sizes of at least 100 before Pareto-optimal weights reliably outperformed unit weights in providing diversity improvement. This was similar to the sample size required for using cognitive subtests in the Machinery Repair job family (see Figure 3). When using the cognitive subtests in the Control and Communication job family, even the minimum sample size of \( N = 40 \) produced cross-validated Pareto weights that reliably outperformed unit weights.

Across all conditions, the unit-weighted solutions had fairly high criterion validities (relatively good job performance outcomes) but fairly low AI ratio values (relatively poor diversity outcomes). Importantly, compared to the Pareto-optimal solutions, the unit-weighted solutions tended to be suboptimal, suggesting diversity improvements over the unit-weighted solution would often be possible in many of these conditions.

Sample size and diversity improvement. To highlight the effect of sample size on diversity improvement, we plotted the diversity improvement (i.e., \( \Delta AI \) ratio from unit weights to Pareto weights, holding job performance constant) that could be obtained when using Pareto-optimal weights, at each distinct sample size condition. We plotted both shrunken and unshrunken diversity improvement for the Bobko-Roth predictors (Figure 6, left panel) and the cognitive subtest predictors (i.e., Figure 6, right panel). For the Bobko-Roth predictors (Figure 6, left panel), the solid line represents the unshrunken diversity improvement (i.e., diversity improvement obtained in the calibration sample, by using Pareto weights rather than unit weights). The dotted line represents the cross-validated (shrunken) diversity improvement (i.e., diversity improvement obtained in the validation sample, by using cross-validated Pareto weights rather than unit weights). For example, at small sample sizes, diversity improvement is not only negligible, but it can also be greatly overstated by the unshrunken estimates. As can be seen from the bottom two lines in Figure 6 (left panel), at smaller sample sizes (e.g., \( N = 40 \)), the unshrunken diversity improvement value (\( \Delta AI \) ratio = 0.28) overestimates the actually

\footnote{9 Results in Figure 6 right panel pertain to the Machinery Repair job family; results for other job families may be obtained from the first author.}
attainable shrunken/cross-validated diversity improvement value \( \Delta \text{AI ratio} = -0.01 \). In contrast, at larger sample sizes (e.g., \( N = 500 \)), the unshrunken diversity improvement \( \Delta \text{AI ratio} = 0.08 \) and the cross-validated diversity improvement \( \Delta \text{AI ratio} = 0.06 \) do not differ substantially. A similar pattern of results is observed for the cognitive subtest predictors (bottom two lines in Figure 6, right panel).

In Figure 6, we also plotted the diversity improvement that could be obtained if an organization were willing to accept a small decrement in performance (i.e., criterion validity) in order to obtain diversity improvement. The top two lines in Figure 6 (both left and right panels) indicate diversity improvement calculated as the difference between the AI ratio of Pareto weights at Unit-Weighted \( R \) versus AI ratio of Unit Weights [in both the calibration sample (solid red lines) and in the validation sample (dotted red lines)]. Blue lines indicate the diversity improvement calculated as the difference between AI ratio of Pareto Weights at \( R \) versus AI ratio of Unit Weights [in the calibration sample (solid blue lines) and in the validation sample (dotted blue lines)]. As such, red lines indicate diversity improvement at no cost in terms of job performance, and blue lines indicate diversity improvement at a small cost in terms of job performance (i.e., dropping from unit-weighted \( R \) to \( R = 0.50 \)). See the online article for the color version of this figure.

Figure 6. Diversity improvement (\( \Delta \text{AI ratio} \)) by calibration sample size, for the Bobko-Roth predictors (left panel) and the cognitive subtest predictors in the Machinery Repair job family (right panel). Red lines indicate diversity improvement calculated as the difference between AI ratio of Pareto Weights at Unit-Weighted \( R \) versus AI ratio of Unit Weights [in both the calibration sample (solid red lines) and in the validation sample (dotted red lines)]. Blue lines indicate the diversity improvement calculated as the difference between AI ratio of Pareto Weights at \( R = 0.50 \) versus AI ratio of Unit Weights [in the calibration sample (solid blue lines) and in the validation sample (dotted blue lines)]. As such, red lines indicate diversity improvement at no cost in terms of job performance, and blue lines indicate diversity improvement at a small cost in terms of job performance (i.e., dropping from unit-weighted \( R \) to \( R = 0.50 \)). See the online article for the color version of this figure.

Altogether, these results demonstrate that many—but not all—of the claims made by De Corte et al. (2008) and Wee et al. (2014) about diversity improvements that can be gleaned from Pareto weighting are fairly robust to diversity shrinkage (i.e., cross-validation of the Pareto curve), under the crucial boundary condition that the calibration sample size is adequate, given the mix of high-impact and low-impact predictors at hand. And, even greater diversity improvements are possible if modest decrements in criterion validity/job performance are accepted.

**Discussion**

The current study examining the cross-validity of Pareto-optimal weights supports the following conclusions. First, along the Pareto-optimal curve, diversity shrinkage increases to the extent that diversity is being maximized. That is, validity shrinkage was greater for Pareto solutions where maximizing job performance was more important than maximizing diversity (i.e., left side of the Pareto-optimal curve); whereas diversity shrinkage was greater at points where maximizing diversity was more important than maximizing job performance (i.e., right side of the Pareto-optimal curve). Both validity and diversity shrinkage decreased with increasing sample size, although larger diversity shrinkage was seen for the Bobko-Roth predictors than for the cognitive subtest predictors, across all sample sizes examined (Research Question 1). Second, consistent with this general finding, when only one objective was maximized (akin to multiple regression, and represented by the endpoints of the Pareto-optimal curve), shrinkage was larger on the maximized objective and negligible on the ignored objective. Specifically, validity shrinkage existed to
the extent that job performance was being maximized, and diversity shrinkage existed to the extent that diversity was being maximized. For both the Bobko-Roth and cognitive subtest predictors, validity shrinkage when job performance was maximized became fairly negligible (i.e., we found $\Delta R \leq .03$) when sample size was at least 100. With a sample size of at least 100, cognitive subtest predictors also typically demonstrated negligible diversity shrinkage when diversity was maximized (i.e., we found $\Delta AI$ ratio $\leq .01$). However, sample sizes greater than 500 were required for the Bobko-Roth predictors before the cross-validated Pareto weights demonstrated negligible diversity shrinkage (Research Question 2). Third, cross-validated Pareto-optimal weights typically evidenced diversity improvement over unit weights when the original sample had at least 100 applicants (Research Question 3).

In sum, when using the Pareto-optimal weighting strategy, diversity improvements over unit weights are still possible even after accounting for diversity shrinkage. However, it must also be reiterated that these cross-validated diversity improvements never produced AI ratios exceeding four-fifths, without some loss of job performance compared to unit weights. Thus, to further our efforts at enhancing organizational diversity—and ameliorating the performance versus diversity dilemma—the Pareto-optimal weighting strategy should also be complemented with other approaches that seek to develop predictor measures that are less race-loaded (e.g., Avery & McKay, 2006), and conceptualize and measure performance as a multifaceted rather than monolithic construct (Campbell & Wiernik, 2015; Hattrup, Rock, & Scalia, 1997).

**Theoretical Implications**

In general, results from the current study extended previous research on cross-validation/shrinkage to the multicriterion optimization context. To recap, previous research (i.e., using single-criterion optimization) indicated that smaller sample sizes and smaller observed multiple correlation coefficients result in greater shrinkage of the multiple correlation.10 We expected, and found, that these validity shrinkage factors, which are well-known when only the one objective—job performance/criterion validity—is being maximized (i.e., at the leftmost endpoint of the Pareto-optimal curve), would generalize to diversity shrinkage in the case where only the one objective of diversity was being maximized (i.e., at the rightmost endpoint of the Pareto-optimal curve).

One insight that emerged from applying our knowledge of single-criterion shrinkage (classic shrinkage formulas) to the diversity objective (i.e., diversity shrinkage) was this: using low-impact predictors (e.g., personality, structured interviews) yields larger diversity shrinkage. The logic for this finding can be gleaned from applying classic shrinkage formulations to diversity, as opposed to criterion validity. First, we know that, all else equal, validity shrinkage decreases as $R^2$ increases (see footnote 10). This means that validity shrinkage decreases as we include more high-validity predictors in the regression model (when predicting job performance). The insight comes when we switch from predicting job performance to predicting race (as we essentially are doing when we estimate diversity shrinkage). For a regression model that predicts race, the analog to high $R^2$ (i.e., high racial variance accounted for by the predictors) would be a regression scenario in which we are using high-impact predictors (e.g., predictors with large Black–White subgroup differences, such as cognitive subtests, for which the predictors have large correlations with race). According to classic shrinkage formulas (and as confirmed in the current simulation), such a scenario with high-impact predictors should yield small diversity shrinkage. Conversely, when the selection scenario involves low-impact predictors, we would expect larger diversity shrinkage. This is consistent with what we see in the current study when we compare Bobko-Roth predictors against cognitive subtest predictors—the Bobko-Roth predictors (which include more low-impact predictors; e.g., personality) suffer greater diversity shrinkage when cross-validating.

Another contribution made in the current paper, beyond extending single-criterion shrinkage notions to diversity shrinkage, involves shrinkage of both criterion validity and diversity across the entire Pareto-optimal curve (i.e., shrinkage when optimizing two criteria simultaneously). That is, classic work on validity shrinkage did not dictate how validity shrinkage and diversity shrinkage would both behave at all the points along the Pareto curve, for different predictor combinations (addressed in our response to Research Question 1). Our study thus provides evidence of the magnitude of shrinkage on two criteria—validity shrinkage and diversity shrinkage.

In addition, the current study also provides evidence that the extent of shrinkage is influenced by the extent of maximization: validity shrinkage was only observed to the extent that validity was being maximized, and diversity shrinkage was only observed to the extent that diversity was being maximized. Aside from the endpoints of the Pareto curve, it appears from the current simulation results that we can directly interpolate shrinkage on validity and diversity when both are being simultaneously optimized using the Pareto-optimization procedure. To restate, shrinkage follows the extent of maximization on each of the two criteria.

**Practical Implications**

Previous research has suggested that Pareto-optimal weighting could be an effective—and easily implemented—strategy to improve diversity hiring outcomes in organizations (De Corte et al., 2008; Wee et al., 2014). Although these studies showed that diversity improvements were theoretically possible, they did not explicitly estimate diversity shrinkage. Our study shows diversity improvements do generalize, as long as moderate sample sizes are used (i.e., based on at least 100 job applicants, in typical scenarios). More specifically, with a sample size of at least $N = 100$, cross-validated Pareto-optimal weights often provided notable diversity improvements over unit weights (e.g., notably increasing the number of job offers extended to minority applicants, at no cost in terms of job performance). A full set of practical implications is summarized in Table 5.

One practical caveat is that the size of the applicant pool/calibration sample can vary widely in practice, and is not always within the direct control of an organization. The current results thus suggest that organizations with large applicant pools (e.g., the Armed Services, the civil service, Google Inc., etc.) could attenuate the performance-diversity trade-off that they would typically

---

10 By algebraically manipulating Wherry’s (1931) formula, we get shrinkage $= R^2 - \rho^2 = (R^2 - 1)(1 - \frac{\rho^2}{R^2 - 1})$. Note that as $R^2$ gets larger, the absolute value [magnitude] of $(R^2 - 1)$ gets smaller. Hence shrinkage decreases as $R^2$ increases.
Table 5

Summary of Basic Findings

1. Sample size. Validity shrinkage and diversity shrinkage both decrease when sample size increases. The incremental benefit of sample size is greater when sample size is small.
   a. In the current simulation, diversity shrinkage was sizable when using the standard set of Bobko-Roth predictors (cognitive and noncognitive tests), whenever sample size was at or below 500. (Figure 2)
   b. In the current simulation, diversity shrinkage was typically negligible for cognitive subtest predictors, whenever sample size was at or above 100. (Figures 3 and 4)
2. Diversity shrinkage is greater when diversity is being maximized. Validity shrinkage is greater when job performance is being maximized. (Tables 3 and 4)
3. Including low impact predictors tends to increase diversity shrinkage.
   a. When maximizing a particular criterion, shrinkage decreases as $R^2$ increases (this can be clearly seen in classic shrinkage formulas).
   b. For this reason, validity shrinkage should decrease when you have high-validity predictors (because high-validity predictors increase $R^2$).
   c. Likewise, diversity shrinkage should decrease when you have high-impact predictors (because high-impact predictors increase $R^2$ for predicting race). Conversely, low impact predictors would increase diversity shrinkage.
4. Pareto-optimal weights typically outperform unit weights.
   a. Pareto-optimal weights generally (but not always) yield equal or greater job performance outcomes than unit-weights (even after accounting for shrinkage; when calibration sample size is $N > 100$), and
   b. Pareto weights can typically yield greater diversity outcomes compared to unit weights, at no cost in terms of job performance. (Figures 2, 3, and 4)
   c. Further, if practitioners are willing to sacrifice a moderate amount of job performance (e.g., using $R = .50$, instead of unit-weighted $R = .58$ [with Bobko-Roth predictors]), usually a notably greater diversity outcome can be achieved (e.g., AI ratio = .78, instead of unit-weighted AI ratio = .63 [with Bobko-Roth predictors]) by using Pareto weights. (Figures 2 and 6)

Limitations and Directions for Future Research

In the current study, we examined shrinkage in Pareto-optimal solutions using data based on two distinct predictor combinations. This allowed us to draw tentative conclusions regarding the use of Pareto weighting strategies in real-world contexts. By doing so, however, we were unable to systematically manipulate the effects of sample size ($N$), the number of predictors ($k$), and the size of the multiple correlation coefficients (i.e., $R_{\text{ perf}}$ and $R_{\text{ race}}$) in a crossed experimental design, which would have allowed us to more directly assess main and interaction effects on the shrinkage of the Pareto-optimal solutions. For example, how much larger must sample sizes be to compensate for low-impact predictors? We addressed this question with regard to the Bobko-Roth predictors, but not with regard to other possible predictor combinations involving low-impact predictors. Also, what increases in sample size become necessary as we add each additional predictor?

Practically speaking, a statistical approach to cross-validation would still be preferred over an empirical approach. We did not use a statistical approach because the classic statistical shrinkage formulas were based on least squares assumptions that do not hold for Pareto-optimization. However, future research might consider the extent to which current formulas can be adapted to provide reasonable approximations, even for Pareto-optimal solutions. Al-
ternatively, statistical formulas for Pareto-optimal shrinkage could be specifically developed.

Finally, we note that the need to examine shrinkage through cross-validation efforts extends even beyond Pareto-optimization techniques. All techniques for optimizing fit between sample data and an empirically derived prediction model should be subjected to cross-validation. A helpful reviewer asked whether this sort of shrinkage and need for cross-validation would also apply to prediction models estimated using Big Data methods and algorithms, such as LASSO regression. To elaborate, one family of Big Data methods is supervised learning methods (from machine learning or statistical learning) in which an algorithm tries to learn a set of rules in order to predict a criterion using observed data. Some such Big Data methods and algorithms do indeed maximize local fit (e.g., $R^2$) by capitalizing on chance, and we would expect considerable shrinkage for such methods (whether maximizing job performance or enhancing diversity). In contrast, LASSO regression (least absolute shrinkage and selection operator; Tibshirani, 1996), as briefly described by Oswald and Putka (2016), attempts to address shrinkage by using a tuning parameter that tries to "strike a balance" between, "local optimization of prediction and future generalization of prediction" (p. 52). In particular, James, Witten, Hastie, and Tibshirani (2013) explained that the tuning parameter itself should be selected using cross-validation, by choosing the tuning parameter that yields the smallest cross-validation error (p. 227). That is, LASSO-weights are designed to maximize cross-validated $R^2$. At present, however, LASSO-style techniques that can incorporate one or more tuning parameters into a Pareto-weighting algorithm (i.e., for use when simultaneously optimizing on two objectives—job performance and diversity) have not yet, to our knowledge, been invented.

Conclusion

The current study examined the extent of shrinkage in Pareto-optimal solutions. We found evidence for the magnitude of validity shrinkage and diversity shrinkage across the entire Pareto-optimal trade-off curve. Specifically, validity shrinkage was greater when the performance objective was more important (i.e., maximized to a greater extent), and diversity shrinkage was greater to the extent the diversity objective was being maximized. Diversity shrinkage was also larger when lower impact predictors were used (e.g., a combination of only cognitive subtest predictors), and was relatively small when higher impact predictors were used (e.g., a combination of only cognitive subtest predictors). As expected, both validity shrinkage and diversity shrinkage decreased as calibration sample size increased. Further, when sample sizes were at least 100, cross-validated Pareto-optimal weights typically afforded diversity improvements over unit weights, at no cost in terms of job performance. In sum, if the calibration sample is of sufficient size, practitioners should consider using Pareto-optimal weighting of predictors to achieve diversity gains.


References


Appendix A

Multiple Regression Shrinkage Formulas

The squared multiple correlation coefficient observed in a calibration sample ($R^2$) is typically an overestimate of both the population squared multiple correlation ($\hat{R}^2$) and the cross-validated squared multiple correlation ($\rho^2$). Commonly used shrinkage formulas for the population squared multiple correlation coefficient ($\hat{R}^2$), which estimates what $R^2$ would be in the general population, include Wherry’s (1931) shrinkage formula and Olkin and Pratt’s (1958, p. 211) shrinkage formula. Cattin (1980) showed that Wherry’s (1931) shrinkage formula provides a biased estimate of population shrinkage whereas Olkin and Pratt’s (1958) shrinkage formula provides an unbiased estimate of shrinkage from the calibration sample to the population (see Cattin, 1980, Table 1 for more details).

\textit{Olkin and Pratt’s shrinkage formula for $\rho^2$ (approximate formula, see Cattin, 1980, p. 409)}

\[
\hat{\rho}^2 = 1 - \frac{N-3}{N-k-1}(1-R^2) \times \left[ 1 + \frac{2(1-R^2)}{N-k+1} + \frac{8(1-R^2)^2}{(N-k-1)(N-k+3)} \right]
\]

where $\hat{\rho}^2$ is the corrected population squared multiple correlation coefficient, $R^2$ is the calibration sample squared multiple correlation coefficient, $N$ is the calibration sample size, and $k$ is the number of predictors.

A second set of shrinkage formulas has been used to describe shrinkage for the cross-validated squared multiple correlation coefficient ($\rho^2_c$), which is an estimate of what $R^2$ would be in a new sample. These formulas estimate $\rho^2_c$ based on the population squared multiple correlation coefficient ($\rho^2$). Browne’s (1975) shrinkage formula—which was based on Lord-Nicholson’s shrinkage formula (Lord, 1950; Nicholson, 1960)—is commonly used to estimate $\rho^2_c$. Yin and Fan (2001) conducted a simulation study to investigate the effectiveness of various shrinkage formulas and found that Olkin and Pratt’s (1958) shrinkage formula for $\rho^2$ and Browne’s (1975) shrinkage formula for $\rho^2_c$ outperformed other shrinkage formulas.

\textit{Browne’s shrinkage formula for $\rho^2_c$}

\[
\hat{\rho}^2_c = \frac{(N-k-3)\hat{\rho}^2 + \hat{\rho}^2}{(N-2k-2)\hat{\rho}^2 + k}
\]

where $\hat{\rho}^2_c$ is the corrected cross-validated squared multiple correlation coefficient; $\hat{\rho}^2$ is the corrected population squared multiple correlation coefficient obtained from a formula such as Olkin and Pratt’s (1958) shrinkage formula, $N$ is the calibration sample size, and $k$ is the number of predictors.

(Appendices continue)
Appendix B
Supplemental Analyses

Figure B1. Comparison of results based on different selection ratios (SR) and proportions of applicants from minority group (Prop) [cognitive subtests: Machinery Repair job family]. SR = selection ratio (overall), prop = proportion of applicants from minority group. [For example, prop = 1/6 means that among all the applicants, 1/6 of the applicants were minority applicants (e.g., Black applicants)] Machinery Repair. See the online article for the color version of this figure.

(Appendices continue)
Figure B2. Comparison of results based on different selection ratios (SR) and proportion of applicants from minority group (Prop) [Bobko-Roth predictors]. SR = selection ratio (overall), prop = Proportion of applicants from minority group. [For example, prop = 1/6 means that among all the applicants, 1/6 of the applicants were minority applicants (e.g., Black applicants)]. See the online article for the color version of this figure.

(Appendices continue)
Appendix C

ParetoR R Package and Web Application for Implementing Pareto Weighting in R

Pareto-Optimization via Normal Boundary Intersection Method in Diversity Hiring
Developer: Q. Chelsea Song
Contact: qianqisong@gmail.com
Last Update: 01/11/2017
The commands below can be found at: https://github.com/Diversity-ParetoOptimal/ParetoR

Objective
The current R program provides a set of Pareto-optimal solutions that simultaneously optimize both diversity and criterion validity in a personnel selection scenario [see Song, Wee, & Newman (in press); adapted from De Corte, Lievens, and Sackett (2007); also see Wee, Newman, and Joseph (2014) for more details]. Pareto-optimal solutions are estimated using the Normal-Boundary Intersection method (Das & Dennis, 1998).

Instructions
1. Open an R console or RStudio window. (R can be downloaded for free from https://cran.r-project.org; RStudio can be downloaded for free from https://www.rstudio.com/).
2. Install R package “ParetoR” through Github, by pasting and running the following commands in R console or RStudio:
   install.packages("devtools")
   library("devtools")
   install_github("Diversity-ParetoOptimal/ParetoR")
   library("ParetoR")
3. Specify four inputs (example from De Corte, Lievens, and Sackett (2007) is given below):
   # (1) Proportion of minority applicants (prop) = (# of minority applicants)/(total # of applicants)
   prop <- -1/4
   # (2) Selection ratio (sr) = (# of selected applicants)/(total # of applicants)
   sr <- -0.10
   # (3) Subgroup differences (d): standardized mean differences between minority and majority subgroups, on each predictor (in applicant pool)
   d <- c(1.00, 0.23, 0.09, 0.33)
   # (4) Correlation matrix (R) = criterion & predictor intercorrelation matrix (in applicant pool)
   ## Example:
   # Format: Predictor_1, . . . , Predictor_n, Criterion
   R <- matrix(c(1, .24, .00, .19, .30,
                 .24, 1, .12, .16, .30,
                 .00, .12, 1, .51, .18,
                 .19, .16, .51, 1, .28,
                 .30, .30, .18, .28, 1),
               (length(d)+1),(length(d)+1))
4. Paste and run the following command in R console or RStudio:
   out = ParetoR(prop, sr, d, R)

Output Description
1. Pareto Optimal solutions (i.e., 21 equally spaced solutions that characterize the Criterion validity—AI ratio trade-off curve, and Predictor Weights at each point along trade-off curve).
2. Plots (i.e., Criterion validity—AI ratio trade-off curve, and Predictor weights across trade-off points).

Note
The program is modeled after De Corte’s (2006) TROFSS Fortran program and Zhou’s (2006) NBI Matlab program (version 0.1.3). The current version only supports scenarios where AI ratio and one other criterion are being optimized.

References

(Appendices continue)
Acknowledgments
Great appreciation to Dr. Serena Wee, Dr. Dan Newman, and Dr. Wilfried De Corte for guidance and feedback on development of the program.

Web Application
We also developed a user-friendly web application to implement the Pareto-optimal technique described in the current paper (https://chelseasong.shinyapps.io/ParetoR/). The web application (like the ParetoR package) uses only a correlation matrix, selection ratio, proportion of applicants from the minority group, and subgroup $d$ values as input. It then provides a full set of Pareto solutions and their corresponding predictor weights.

**Call for Nominations**

The Publications and Communications (P&C) Board of the American Psychological Association has opened nominations for the editorships of the *Journal of Experimental Psychology: Animal Learning and Cognition*, *Neuropsychology*, and *Psychological Methods* for the years 2020 to 2025. Ralph R. Miller, PhD, Gregory G. Brown, PhD, and Lisa L. Harlow, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2019 to prepare for issues published in 2020. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations are also encouraged.

Search chairs have been appointed as follows:

- *Journal of Experimental Psychology: Animal Learning and Cognition*, Chair: Stevan E. Hobfoll, PhD
- *Neuropsychology*, Chair: Stephen M. Rao, PhD
- *Psychological Methods*, Chair: Mark B. Sobell, PhD

Candidates should be nominated by accessing APA’s EditorQuest site on the Web. Using your browser, go to https://editorquest.apa.org. On the Home menu on the left, find “Guests/Supporters.” Next, click on the link “Submit a Nomination,” enter your nominee’s information, and click “Submit.”

Prepared statements of one page or less in support of a nominee can also be submitted by e-mail to Sarah Wiederkehr, P&C Board Editor Search Liaison, at swiederkehr@apa.org.

Deadline for accepting nominations is Monday, January 8, 2018, after which phase one vetting will begin.

Received June 30, 2016
Revision received May 8, 2017
Accepted May 9, 2017