© Qianqi Song

DIVERSITY SHRINKAGE OF PARETO-OPTIMAL SOLUTIONS IN HIRING PRACTICE: SIMULATION, SHRINKAGE FORMULA, AND A REGULARIZATION TECHNIQUE

BY

QIANQI SONG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Daniel A. Newman, Co-Chair
Professor James Rounds, Co-Chair
Professor Fritz Drasgow
Associate Professor Victoria Stodden
Assistant Professor Daniel Briley

ABSTRACT

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-optimal 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 without changing the job performance outcome in personnel selection. The current dissertation investigates topics related to a key limitation of the technique—the possibility of shrinkage, especially *diversity shrinkage*, in the Pareto-optimal solutions. The dissertation consists of three studies. Study 1 attempts to study diversity shrinkage and job performance validity shrinkage using Monte-Carlo simulation. Using Monte Carlo simulation, sample size and predictor combinations are varied and cross-validated Pareto-optimal solutions are obtained. Study 2 derives approximate mathematical formulae to directly correct for job performance validity shrinkage and diversity shrinkage when using Pareto-optimal weights. These shrinkage formulae for Pareto-optimal weighting are evaluated using simulation. Finally, Study 3 attempts to develop a Pareto-optimal weighting algorithm that achieves both optimization and regularization (similar to ridge regression, LASSO regression, or elastic nets; in the context of Pareto-optimal weighting with two criteria). An R package is developed to estimate Pareto-optimal solutions in personnel selection (i.e., ParetoR package), which includes: (a) De Corte et al.'s (2007) Pareto-optimization method (i.e., based on the NBI algorithm; used in Study 1), (b) Pareto-optimal shrinkage formula corrections (i.e., as introduced in Study 2), and

(c) a regularized Pareto-optimal method (i.e., as introduced in Study 3). In sum, the current dissertation aims to contribute to the field of diversity selection by investigating job performance validity shrinkage and diversity shrinkage under the Pareto-optimization method to simultaneously optimize both the job performance and diversity of new hires.

ACKNOWLEDGEMENTS

It is with my deepest gratitude that I write down the following words of appreciation. First and foremost, I would like to thank my doctoral committee members, Professor Daniel Briley, Professor Fritz Drasgow, Professor Daniel A. Newman, Professor James Rounds, and Professor Victoria Stodden (in alphabetical order of last names). All have served as my role models and teachers through graduate school, and the completion of this dissertation would not have been possible without their invaluable support. I greatly appreciate the dedication and care that you have offered in the process.

I would like to specially thank my co-advisors, Professor Daniel (Dan) Newman and Professor James (Jim) Rounds. They have always been my mentors, who have provided me with extensive guidance and have helped me grow as a researcher and as a person. Their wisdom, knowledgeableness, and selfless devotion to work and their students have always made me feel greatly fortunate and honored to be their student. I will always keep your teachings in my heart and continue to work hard in my career.

I would like to thank my fellow graduate students, whose support and care have been a priceless part of my journey. I have learnt so much from many of them, and I am grateful to be able to share these treasured years. I would also like to express my wholehearted appreciation to my boyfriend Or Dantsker, who, with his love, understanding, and guidance, has been one of my strongest supports in life and throughout the dissertation completion process. I am ever so grateful to have met you and to have you with me through these memorable years of my life.

With my utmost love and respect, I would like to dedicate this dissertation to the most influential people in my life – my father, Yuxuan Song, and my mother, Qiubo Yang. Your love, wisdom, and courage, will always inspire me throughout my life.

To my parents,

Yuxuan Song and Qiubo Yang

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1
CHAPTER 2: STUDY 1 Diversity Shrinkage: Cross-Validating Pareto-Optimal Weights to Enhance Diversity via Hiring Practices 20
CHAPTER 3: STUDY 2 Development of Pareto-Optimal Shrinkage Formulae 49
CHAPTER 4: STUDY 3 Development of a Regularized Pareto-Optimal Weighting Algorithm 74
CHAPTER 5: GENERAL DISCUSSION AND CONCLUSION 83
TABLES 89
FIGURES 95
REFERENCES 120
APPENDIX A: Multiple Regression Shrinkage Formulae 131
APPENDIX B: Study 1 Supplementary Analyses 132
APPENDIX C: ParetoR R Package and Web Application for Implementing Pareto-Optimal Weighting 134
APPENDIX D: Comparison among Wherry (1931), Olkin & Pratt (1958) and Claudy (1978) Formulae 141
APPENDIX E: Diversity and Validity Shrinkage and Performance of Shrinkage Formulae when Calibration Sample Size is 10,000 144

CHAPTER 1
INTRODUCTION

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 & Ivancevich, 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; and 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.

Diversity and Adverse Impact

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 (*Uniform Guidelines*, 1978)," (Outtz & Newman, 2010). Adverse impact is evidenced by the adverse 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 4/5^{ths} (e.g., 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 (*Uniform Guidelines on Employee Selection Procedures*, 1978). To be more specific, the "4/5^{ths} 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, 2010; Sackett, Schmitt, Ellingson, & Kabin, 2001).

$$AI\ Ratio = \frac{\#\ of\ selected\ minority\ applicants \Big/ total\ \#\ of\ minority\ applicants}{\#\ of\ selected\ majority\ applicants \Big/ total\ \#\ of\ majority\ applicants}$$

Diversity-Performance Trade-Off

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 & McDaniel, 2006]). Hence, organizations often face the seemingly incompatible goals of selecting either for high levels of expected job performance (by maximizing job performance 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-optimal 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 and 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 for dealing 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.

Pareto-Optimal Weighting

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 Lievens, and Sackett (2008) showed how improvements in both criterion validity and the AI ratio could be achieved via a Pareto-optimal-

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).

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 simultaneously: job performance and diversity[1]. 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 (e.g., Figure 1, left panel) 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

[1] Note that the application of Pareto-optimal weighting method does not limit to optimizing job performance and diversity. It could also be used on other trade-off situations where we consider multiple outcomes simultaneously. In the current dissertation, we focus on the trade-off between job performance and diversity in personnel selection.

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.

Normal Boundary Intersection Algorithm. Pareto optimal weighting for multi-criterion optimization (to create Pareto curves; e.g., Figure 1) can be implemented in a variety of ways, one of which is labeled *Normal Boundary Intersection*. Specifically, the Normal Boundary Intersection algorithm (NBI; Das & Dennis, 1998) was used to obtain optimal weights in De Corte et al.'s (2007; see Appendix B) Pareto-optimal implementation to study the diversity-performance tradeoff in personnel selection (see also De Corte, Sackett & Lievens, 2011, Appendix A). An example of NBI is shown in Figure 2. The aim of the algorithm is to find evenly spaced sets of predictor weights that optimize multiple criteria (e.g., two criteria: diversity and job performance) under certain constraints. To implement NBI for personnel selection, the algorithm's objectives and constraints are shown below.

Objectives

Objective 1. $AIratio = \frac{SR_B}{SR_W}$

(adverse impact ratio; diversity outcome of hiring practice)

Objective 2. $R_{perf} = [w_1, \dots, w_k, 1]\boldsymbol{R}[0, \dots, 0, 1]'$

(criterion validity; job performance outcome of hiring practice)

Constraints

Constraint 1. $p_B(SR_B) + (1 - p_B)(SR_W) = SR$

where SR_B is the minority (e.g., Black) selection ratio, SR_W is the majority (e.g., White) selection ratio, SR is the overall selection ratio, p_B is the proportion of minority applicants (e.g., Black applicants) and $1 - p_B$ is the proportion of majority applicants (e.g., White[2] applicants). This constraint illustrates that overall selection ratio [SR] is a function of the selection ratios for the minority group [e.g., selection ratio of Black applicants, SR_B] and the majority group [e.g., selection ratio of White applicants, SR_W.

Constraint 2. $[w_1, \dots, w_k]\boldsymbol{R_u}[w_1, \dots, w_k]' = 1$

where $w_1, \dots, w_k$ are the predictor weights, $\boldsymbol{R_u}$ is the predictor intercorrelation matrix, and $\boldsymbol{R}$ (from Objective 2) is the larger correlation matrix among all the predictors and job performance (e.g., similar to Table 1), where correlations between job performance and each of the *k* predictors are shown in the $(k + 1)^{th}$ column (last column) and $(k + 1)^{th}$ row (last row) of the correlation matrix $\boldsymbol{R}$. Constraint 2 fixes the variance of the composite predictor to be 1. This constraint is placed to ensure a unique solution (De Corte et al., 2007).

Constraint 3. Predictor weights cannot be negative (placed by e.g., De Corte et al.., 2007).

There are three main steps in the NBI algorithm.

[2] For simplicity, we assume that there are only two subgroups in this selection scenario.

Step 1 of NBI algorithm. Find the AI ratio and job performance validity at endpoints (*Shadow F matrix*; Das & Dennis, 1998) and corresponding predictor weights (*Shadow X matrix*).

In this step, sequential least squares programming (SLSQP; an algorithm that finds a set of weights that maximizes the criteria, given the constraints listed above – utilizes a least squares method that is similar to that utilized in least squares regression) is used to (a) obtain the diversity value (i.e., $AIratio_{max}$) and job performance criterion validity (i.e., $R_{perf_{min}}$) when only diversity was treated as criterion (i.e., when only diversity [$AI\ ratio$] was maximized); and (b) diversity value (i.e., $AIratio_{min}$) and job performance criterion validity (i.e., $R_{perf_{max}}$) when only job performance was treated as criterion (i.e., when only R_{perf} was maximized). The diversity and job performance criterion validity corresponding to the endpoints of the Pareto curve are then recorded in a 2×2 matrix called the Shadow F matrix; corresponding predictor weights are recorded in a $k \times 2$ matrix called the Shadow X matrix, where k is the number of predictors. Thus, assume we have specified the number of Pareto points as 21, then the Shadow F matrix and Shadow X matrix can be expressed as follows:

$$Shadow\ F = \begin{pmatrix} AIratio_{max} & AIratio_{min} \\ R_{perf_{min}} & R_{perf_{max}} \end{pmatrix}$$

$$Shadow\ X = \begin{pmatrix} w_{p1.1} & w_{p21.1} \\ w_{p1.2} & w_{p21.2} \\ \vdots & \vdots \\ w_{p1.k} & w_{p21.k} \end{pmatrix}$$

Step 2 of NBI algorithm. Phi trade-off matrix

Criterion values (diversity and job performance) at the two endpoints could further be used to build a Phi trade-off matrix, as shown below. As you can see, the Phi matrix has zeros in the diagonal, and has the range of the AI ratio (i.e., $AIratio_{min} - AIratio_{max}$) and range of job performance criterion validity (i.e., $R_{perf_{min}} - R_{perf_{max}}$) in the off-diagonal. This Phi matrix,

combined with criterion weights at each Pareto point, is used to estimate an optimal set of predictor weights at each Pareto point.

$$\Phi = \begin{pmatrix} 0 & R_{perf_{min}} - R_{perf_{max}} \\ AIratio_{min} - AIratio_{max} & 0 \end{pmatrix}$$

Step 3 of NBI algorithm. Linear interpolation of evenly-spaced solutions between the endpoints

Now that we have estimated the two endpoints (in Step 1), we can estimate the Pareto points between the two endpoints by first creating a line that connects the two endpoints (i.e., the orange line in Figure 2). Then, the NBI algorithm will specify evenly-spaced points along this line, ensuring the number of points along the line (i.e., including the two end points; yellow points in Figure 2) equals the user-specified number of Pareto solutions (e.g., 21 evenly-spaced solutions). Finally, at each Pareto point, the NBI algorithm will project along a perpendicular direction from the initial criterion line (i.e., yellow arrows in Figure 2) until it reaches the border of the solution space (i.e., blue oval in Figure 2). Details of this projection are described below.

To elaborate, in the first step for generating Pareto-optimal weights between the two endpoints, initial criterion weights will be linearly assigned for the two criteria (AI ratio criterion and job performance criterion), based on the user-specified number of desired Pareto points. For example, if we specify the number of Pareto points as 21 (i.e., obtain 21 solutions along the Pareto trade-off curve), then the criterion weight matrix will be a 2×21 matrix:

$$Criterion\ Weight = W_C = \begin{pmatrix} 1 & .95 & .90 & \dots & .05 & 0 \\ 0 & .05 & .10 & \dots & .95 & 1 \end{pmatrix} = [\boldsymbol{w_{C_1}}, \boldsymbol{w_{C_2}}, \dots, \boldsymbol{w_{C_i}} \dots \boldsymbol{w_{C_{21}}}]$$

where criterion weights corresponding to the AI ratio are shown in the first row, and criterion weights corresponding to job performance criterion validity are shown in the second row. The 21 columns (i.e., $\boldsymbol{w_{C_1}}, \boldsymbol{w_{C_2}}, \dots, \boldsymbol{w_{C_i}} \dots \boldsymbol{w_{C_{21}}}$) indicate criterion weights for the 21 Pareto solutions. For example, $\boldsymbol{w_{C_1}} = [1,0]^{\boldsymbol{T}}$ in the first column indicates that at Pareto point #1, job performance

criterion validity will be given a weight of 1, whereas AI ratio will be given a weight of 0. As another example, $\boldsymbol{w_{C_2}} = [.95, .05]^{\boldsymbol{T}}$ in the second column indicates that at Pareto point #2, job performance criterion validity will be given a weight of .95, whereas AI ratio will be given a weight of .05.

Step 4 of NBI algorithm. Projection of evenly-spaced solutions between the endpoints

Following, at each ith Pareto point, the sequential least squares quadratic programming (SLSQP) algorithm was again used to find optimal set of weights (i.e., $[w_{pi.1}, w_{pi.2}, \ldots, w_{pi.k}]$) that minimizes the loss function (converted from Das & Dennis, 1996, Equation 1; De Corte, Lievens and Sackett, 2007, Appendix B, Equation 4):

$$\left[\begin{pmatrix} R_{perf_{pi}} \\ AIratio_{pi} \end{pmatrix} - \begin{pmatrix} R_{perf_{max}} \\ AIratio_{max} \end{pmatrix} - \Phi \boldsymbol{w_{C_i}}^T\right] \times \left[\Phi \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right]^{-1} = \left[\begin{pmatrix} R_{perf_{pi}} \\ AIratio_{pi} \end{pmatrix} - \begin{pmatrix} R_{perf_{max}} \\ AIratio_{max} \end{pmatrix} - \begin{pmatrix} 0 & R_{perf_{min}} - R_{perf_{max}} \\ AIratio_{min} - AIratio_{max} & 0 \end{pmatrix} \boldsymbol{w_{C_i}}^T\right] \times \begin{pmatrix} R_{perf_{min}} - R_{perf_{max}} \\ AIratio_{min} - AIratio_{max} \end{pmatrix}^{-1},$$

where $\boldsymbol{w_{C_i}}$ is a vector of criterion weights that were linearly interpolated between the two endpoints (see Figure 2) in Step 3. This process will iterate through all Pareto points (e.g., 21 points) until we obtain optimal weights for each Pareto point.

Validity and Diversity Shrinkage

Shrinkage and Cross-Validation of 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 commonly used because it provides the best linear unbiased estimates of the predictor weights for predicting a criterion. However, we must consider shrinkage when predictor weights obtained from one sample are used to predict criterion in another sample or used to generalize to the population (Larson, 1931; Wherry, 1931). Because the weights provided by model-fitting procedures 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; the model is overfitted (Yin & Fan, 2001) and shrinkage will likely occur. 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 (ρ^2)* and the *squared cross-validated multiple correlation coefficient ($\rho_c{}^2$)* 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 applying the obtained predictor weights to an independent, validation sample (obtained from the same population as the calibration sample). The obtained result, $\rho_c{}^2$, corrects for overfitting and thus provides a generalizable estimate of the variance accounted by the predictors in the criterion. The statistical approach uses formulae to adjust the squared sample multiple correlation coefficient (R^2) using the sample size (*N*) and the number of predictors (*k*) used to obtain the sample estimate (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 is more cost-effective, given that a second sample is not required (Cascio & Aguinis, 2005).

Shrinkage and Cross-Validation of 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 formulae are based on least-squares assumptions (Dorans & Drasgow, 1980) that do not form the basis for Pareto-optimization. Therefore, in Study 1 we examined the issue of shrinkage in Pareto-optimal weights through empirical approach (i.e., Monte Carlo simulation).

When using Pareto-optimal weighting to reduce adverse impact, the predictor weights simultaneously optimize both job performance and diversity. Because both objectives are optimized by the given set of Pareto-optimal 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-optimal 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 R^2; 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.

Regularization Techniques

One notion that could help us better understand model prediction when generalized to a new sample or population is the so-called *bias-variance trade-off* (e.g., Hastie, Tibshirani &

Friedman, 2009) that is commonly discussed in statistical learning and machine learning. In this context, the terms *bias* and *variance* refer to two sources of error that influence model prediction.

Let us say we are trying to predict values y with predictor values $x_1, x_2, \ldots, x_k$, then the relationship between the two variables can be expressed as:

$$y = f(X) + \epsilon, \epsilon \sim N(1,0)$$

where $f(X)$ is the true function, and error term ϵ follows a normal distribution with mean of zero and variance of one. Our goal is to find a function $\hat{f}(X)$ that can approximate the true function $f(X)$ as much as possible. In other words, we want to minimize mean squared error (MSE), $E\left(y - \hat{f}(X)\right)^2$, for both the current sample and *other samples* drawn from the same population. Mathematically, MSE can be expressed as:

$$E\left[\left(y - \hat{f}(x)\right)^2\right] = Bias\left[\hat{f}(x)\right]^2 + Var\left[\hat{f}(x)\right] + \sigma^2 \quad (1)$$

where $Bias\left[\hat{f}(x)\right]^2$ is the squared bias value of $\hat{f}(x)$, $Var\left[\hat{f}(x)\right]$ is the variance of $\hat{f}(x)$, and σ^2 is the irreducible (random) error (e.g., Hastie et al., 2009, p. 24 Eq. 2.25; James, Witten, Hastie, & Tibshirani, 2013, p. 34, Eq. 2.7). Bias and variance can be further expressed as follows:

$$Bias\left[\hat{f}(x)\right] = E\left[\hat{f}(x) - f(x)\right] \quad (2)$$

$$Var\left[\hat{f}(x)\right] = E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right] = E\left[\hat{f}(x)^2\right] - E\left[\hat{f}(x)\right]^2 \quad (3)$$

Bias (see Equation 2; e.g., Hastie et al., 2009, p. 24 Eq. 2.25) refers to the expected (average) difference between the predicted values from the sample-based prediction equation that is estimated using observed data [$\hat{f}(x)$] and the predicted values based on the true underlying prediction equation [$f(x)$]. Model fitting procedures such as multiple regression aim to minimize bias. *Variance* (see Equation 3; e.g., Hastie et al., 2009, p. 24 Eq. 2.25), on the other hand, illustrates the amount by which $\hat{f}(x)$ change when estimated using different samples (i.e.,

calibration samples) from a population. In short, *bias* is the difference between the mean of $\hat{f}(x)$ and the true value $f(x)$; whereas *variance* is the variance of $\hat{f}(x)$.

When there is high *variance*, models obtained from one sample might not fit as well when applied to another sample. This sample-to-sample variability in model estimates is the cause of shrinkage. In other words, high variance could cause overfitting (i.e., the R^2 from a given sample is too large).

To better understand *bias-variance tradeoff*, let's consider bias and variance of estimates under various model complexity (Figure 3). For the current illustration, we can think of *model complexity* as the number of predictors in the prediction model. In Figure 3, we see that as the model complexity (e.g., number of predictors) increases, the *variance* of the sample-based prediction equation $\hat{f}(x)$ increases monotonically. All else equal, having more predictors results in more uncertainty and thus more sampling variance of $\hat{f}(x)$. In contrast, as the model complexity (e.g., number of predictors) increases, the *bias* of $\hat{f}(x)$ [as an estimate of the true $f(x)$] decreases monotonically (see Figure 3), because increasing the number of predictors tends to predict current sample outcomes better and reduce the biasing influence of omitted variables, for instance. The key insight in Figure 3 is that, if model complexity simultaneously *increases* variance and *decreases* bias, as MSE is a function of the additive sum of bias squared and variance (Equation 1), there should exist an optimal level of model complexity at which the sum of bias squared and variance (i.e., MSE) is minimized. Locating this optimal level of model complexity is the focus of a set of techniques collectively referred to as *regularization*.

Regularization techniques attempt to address shrinkage and the bias-variance trade-off by using a *tuning parameter* that tries to "strike a balance" between "local optimization of prediction and future generalization of prediction" (Oswald & Putka, 2015, 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, regularization techniques are designed to maximize cross-validated R^2 and minimize model overfitting. Regularization techniques include *ridge regression* (e.g., Tikhonov, 1963; Hoerl, 1962), *LASSO regression* (i.e., least absolute shrinkage and selection operator regression; Chen and Donoho, 1994; Chen, Donoho and Saunders, 2001; Tibshirani, 1996), *LARS regression* (i.e., least angle regression; Efron et al., 2004), and *elastic net* (Zou & Hastie, 2005).

At present, however, regularization techniques that can incorporate one or more tuning parameters into a Pareto-optimal weighting algorithm (i.e., for use when simultaneously optimizing on *two* objectives—job performance and diversity) have not yet, to our knowledge, been invented. Development of a new algorithm that achieves some regularization in the context of the Pareto-optimization—multi-objective optimization—technique is the focus of Study 3.

Ridge regression. Ridge regression (e.g., Tikhonov, 1963; Philips, 1962; Hoerl, 1962) achieves regularization by finding regression coefficients that minimize the L_2 penalty term, which is a sum of squared errors (SSE) of prediction in Equation 4.

$$SSE_{L2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \tag{4}$$

where i represents the ith individual in the sample, n is the sample size, j represents the jth predictor, k is the number of predictors, y is the dependent variable, β is the standardized regression coefficient, and λ is the tuning parameter that adjusts for shrinkage. Notice that the first element of the penalty term illustrates bias and variance, which is equivalent to the SSE term from ordinary least squares regression (OLS regression). The ridge penalty term (i.e., L_2

term; Equation 4) also introduces a second element, which involves the tuning parameter (i.e., $\lambda \sum_{j=1}^{k} \beta_j^2$). In practice, ridge regression is similar to OLS regression, except that the tuning parameter gets added to the diagonal of the $X'X$ matrix before taking its inverse:

$$\hat{\beta}_{ridge} = \left(X'X + \lambda I_p\right)^{-1} X'Y$$

This tuning parameter term essentially limits variability of the coefficients across different predictors, which makes the predictor weights slightly more similar to unit weights (and therefore less vulnerable to shrinkage). Thus, by minimizing the ridge regression SSE (Equation 4), we are taking into account both bias and variance (see Equation 1). The tuning parameter λ is usually found via cross-validation, and the resulting set of ridge coefficients attempts to produce the smallest sum of bias and variance (i.e., MSE), striking a balance in the bias-variance trade-off.

LASSO regression. LASSO regression (Chen and Donoho, 1994; Chen, Donoho and Saunders, 2001; Tibshirani, 1996; also briefly described in Oswald & Putka, 2015) methods are used in statistics and machine learning to simultaneously perform both variable selection (choosing which predictors to use in the prediction equation) and regularization. By doing so, it potentially makes the statistical model more parsimonious, and it increases the prediction accuracy when generalizing to a new sample or to the population. LASSO regression uses the L_1 penalty term, as shown in Equation 5.

$$SSE_{L1} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \tag{5}$$

where i represents the ith individual in the sample, n is the sample size, j represents the jth predictor, k is the number of predictors, y is the dependent variable, β is the standardized regression coefficient, and λ is the tuning parameter that adjusts for shrinkage. Notice that the

only difference between the L_2 penalty term (i.e., ridge penalty term; Equation 4) and the L_1 penalty term (i.e., LASSO penalty term; Equation 5) is that the β_j^2 is replaced by $|\beta_j|$ in the last term. One effect of using Equation 5 is that it enables a solution space within which the LASSO coefficients can be zero (i.e., allowing some predictors to be dropped out; Hastie et al., 2009; p. 71), establishing variable selection. Ridge regression, on the other hand, restricts certain coefficients but does not typically allow them to become zero (see Hastie, Tibshirani & Friedman, 2009, section 3.4.3 for more details).

For the sake of completeness, we mention that LARS regression (Efron et al., 2004) is similar to LASSO regression, and it can also be considered as a broader framework that includes LASSO regression (Efron et al., 2004; Hastie et al., 2009; Putka, Beatty & Reeder, 2017). LARS regression uses a step-wise approach that starts with one predictor and slowly increases the coefficients and adding other predictors until it reaches a point determined by the tuning parameter. Thus, LARS accomplishes both regularization and variable selection at the same time.

Elastic net. Elastic net (Zou & Hastie, 2005) combines both ridge regression and LASSO regression by including both the L_1 penalty term and the L_2 penalty term at the same time. The penalty term of elastic net is shown in Equation 6.

$$SSE_{Enet} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\left[(1-\alpha)\sum_{j=1}^{k}\beta_j^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right] \quad (6)$$

where i represents the ith individual in the sample, n is the sample size, j represents the jth predictor, k is the number of predictors, y is the dependent variable, β is the standardized regression coefficient, λ is the tuning parameter that adjusts for shrinkage, and α is a parameter that controls for the relative weighting of the penalty terms in Equation 6. Notice that the

parameter α determines how much relative weight gets assigned to the ridge regression penalty term ($\sum_{j=1}^{k} \beta_j^2$) versus the LASSO regression penalty term ($\sum_{j=1}^{k} |\beta_j|$). At the extremes, when $\alpha = 1$, the elastic net penalty equation (Equation 6) becomes equivalent to the LASSO regression equation (Equation 5); whereas when $\alpha = 0$, the elastic net equation (Equation 6) becomes equivalent to the ridge regression equation (Equation 4). So the elastic net α parameter determines the position of the elastic net solution on a continuum between the ridge regression solution and the LASSO regression solution. Both λ (tuning parameter) and α (which we refer to as the relative weight parameter) are commonly found using cross-validation.

Regularization techniques illustrated in this section are available in various software and computer programs such as R, Python, and MATLAB. Example R packages include "glmnet" (ridge regression, LASSO regression, & elastic net; Friedman, Hastie & Tibshirani, 2010), "lars" (ridge regression, LASSO regression & LARS regression; Hastie & Efron, 2013), "penalized" (ridge regression & LASSO regression; Goeman, 2017), "bigLASSO" (ridge regression & LASSO regression; Zeng & Breheny, 2017), "elasticnet" (LASSO regression & elastic net; Zou & Hastie, 2012). Among these, the "glmnet" R package is capable of conducting all three above mentioned regularization techniques at the same time, depending on the choice of alpha.

CHAPTER 2

STUDY 1

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

The purpose of Study 1 is to demonstrate that diversity shrinkage exists, and to estimate the extent of both diversity shrinkage and validity shrinkage in Pareto-optimal solutions across a range of conditions (sample size, combinations of high- and low-impact predictors, overall selection ratio, proportion of applicants in minority vs. majority group). Although the aforementioned studies (De Corte et al., 2008; Wee et al., 2014) provided evidence that Pareto-optimal weighting could be an effective strategy for explicitly dealing 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 if diversity improvements generalize beyond the sample on which the weights were obtained. That is the purpose of Study 1—to examine the cross-validity of Pareto-optimal solutions.[3] Specifically, by conducting Monte Carlo simulations, we examined the extent to which Pareto-optimal solutions *shrink* (i.e., the extent to which the expected job performance and expected diversity decrease in size when the predictor weights obtained in the original sample are applied to new samples), under different predictor combinations and sample sizes.

[3] 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.

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 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 shrunken Pareto-optimal solutions to 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). 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 examined the degree of validity shrinkage and diversity shrinkage when only one objective was maximized. We then extended the results to multi-criterion optimization, and we also compared cross-validated Pareto-optimal solutions to unit-weighted 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 subtest 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 versus 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 Study 1, 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 (Machinery Repair) as a typical scenario in which the predictor validities and predictor subgroup differences (*d*) were *positively* related, across predictors (i.e., *positive validity-d correspondence*) [e.g., for Machinery Repair jobs, clerical speed tests show the lowest criterion validity and the *lowest* subgroup *d*]. For the sake of comparison, we also chose a second job family (Control & Communication) to represent the scenario where predictor validities and predictor subgroup differences (*d*) were *negatively* related, across predictors (i.e., *negative validity-d correspondence*) [e.g., for control & communication jobs, technical knowledge tests

show the lowest criterion validity and the *highest* subgroup *d*]. To elaborate, *positive validity-d correspondence* scenarios should logically result in smaller diversity improvements (because high-impact predictors receive more weight as performance is maximized), whereas *negative validity-d correspondence* scenarios should result in larger diversity improvements (because high-impact predictors tend not to receive more weight as performance is maximized). Note that even under positive validity-*d* correspondence, the shrunken diversity improvement due to Pareto weighting is still expected to be greater than zero.[4]

In total, we examined 15 simulation conditions [5 sample size × 3 predictor conditions (i.e., Bobko-Roth predictors, plus cognitive subtest predictors for two job families)]. For each condition, we conducted 1,000 replications. For example, for the $N = 40$ condition using the Bobko-Roth predictors, we generated 1,000 calibration samples of size 40. In each replication, predictor weights were generated in a calibration sample, and then applied to an independent validation sample. The resulting cross-validated criterion validity and diversity were obtained. To estimate validity shrinkage, we subtracted the cross-validated criterion validity from the calibration sample criterion validity. Similarly, to estimate diversity shrinkage, we subtracted the cross-validated diversity value from the calibration sample diversity value. Finally, to estimate the diversity improvement possible when using Pareto-optimal weights, we compared the cross-validated diversity values to diversity values obtained when using unit-weights.

Procedure

[4] 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.

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. 1,000 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 *d*s were then calculated for each calibration sample.

We note that race is represented as a binary variable in the current simulation. Thus, in order 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 *d*s 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 *d*s. Specifically, the following steps were used to generate the samples.

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

$$r_{pb} = d/\sqrt{d^2 + 1/(p(1-p))} \tag{7}$$

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):

$$r_{bis} = r_{pb}(\sqrt{p(1-p)}/\lambda) \tag{8}$$

where p is the proportion of minority applicants, $(1 - p)$ is the proportion of majority applicants, and λ 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 $p \times N$ simulated applicants were labeled "2" (indicating White applicant), with the 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 ds 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 & Lyon, 2009). Generating separate distributions for each race has

the drawback that a 2-population mixture distribution is not necessarily normally distributed, and further might produce overall sample correlations across the two groups that are systematically different from the population correlations in Tables 1 and 2 (see Ostroff, 1993).[5]

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 as consistent as possible with previous studies

[5] At a reviewer's request, we also ran the simulations by generating two separate normal distributions (one for each race) that differed only in their means (i.e., same as De Corte et al., 2007). The results of Study 1 were nearly identical under the two procedures for simulating race: (a) the biserial method – in which data are simulated from a single continuous distribution that includes race as a variable vs. (b) the two-distribution method – in which data are simulated from two separate distributions. Results of this comparison are available from the first author. Our reviewer further noted that there is a conceptual point to be made here about one's theoretical definition of race. Under the two-distribution method, race is handled as a natural dichotomy; whereas under our current biserial method, race is treated as a reflection of a continuous underlying distribution of racial identification, racialized experience, and racial image (which gets dichotomized when race is measured, and when one considers adverse impact—consistent with our simulation).

(De Corte et al., 2007; Wee et al., 2014), we set *SR* at .15 and the number of Pareto points at 21. We used p = .167 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 $p \approx 1/8$ used by De Corte et al. (2008, p. 187).[6] 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.[7] Similarly, the calibration sample diversity (denoted $AI\ Ratio_{cal}$; which is monotonically related to the calibration sample race point-biserial multiple correlation, denoted $R_{race.cal}$) is given by the correlation between the weighted predictor composite scores and the binary race variable. Specifically, the AI ratio can be directly obtained from $R_{race.cal}$ using the following formulae from Newman, Jacobs, and Bartram (2007, p. 1404), given that *SR* and *p* are known for each condition.

$$d = r/\sqrt{(1-r^2)p(1-p)}$$

$$x_{cut} = z_{cut}\sqrt{1+d^2[p(1-p)]} - dp$$

$$AI\ ratio = \frac{SR_{Black}}{SR_{White}} \approx \frac{(1.64x_{cut} + \sqrt{.76x^2{}_{cut} + 4})}{\left(\sqrt{e^{(d^2+2x_{cut}d)}}\right)[1.64(x_{cut}+d) + \sqrt{.76(x_{cut}+d)^2 + 4}]}$$

[6] In response to a reviewer comment, we conducted supplementary simulations to evaluate whether the Pareto curves and shrunken Pareto curves change when there are changes in (a) the selection ratio, and (b) the proportion of applicants from the minority group. These supplementary simulation results are presented in Appendix B. Results suggest that: (a) as the selection ratio decreases, available solutions will provide lower AI ratios, and diversity shrinkage will increase, and (b) as the proportion of applicants from the minority group decreases, available solutions will provide slightly higher AI ratios, and diversity shrinkage will also increase. As such, results of the current simulation might overstate the magnitude of diversity shrinkage compared to conditions with higher selection ratios and higher proportions of minority applicants, and might understate diversity shrinkage compared to conditions with lower selection ratios and lower proportions of minority applicants.

[7] Similar to Ree, Earles, and Teachout (1994), we provide estimates in the *R* rather than the R^2 metric for criterion-related validity.

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 pre-employment 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 calibration sample point-biserial multiple correlations (i.e., $R_{race.cal}$), 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 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 ($AI\ Ratio_{unit}$) 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 ($\bar{R}_{perf.cal}$) at the Pareto point where criterion validity is maximized—i.e., 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 ($\overline{AI\ Ratio}_{cal}$) 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 $\bar{R}_{race.cal}$ to $\overline{AI\ Ratio}_{cal}$.[8] 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 ($\bar{R}_{perf.unit}$) and diversity ($\overline{AI\ Ratio}_{unit}$), 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

[8] We note that our procedure of averaging the multiple correlation (R_{race}) 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.

samples. The unit-weighted solution was obtained by plotting the mean unit-weighted criterion validity ($\bar{R}_{perf.unit}$) against the mean unit-weighted diversity ($\overline{AI\ Ratio}_{unit}$).

8. Calculate validity shrinkage and diversity shrinkage. For each point along the Pareto-optimal curve, the amount of validity shrinkage (ΔR) 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 (ΔAI 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 4 through 6. The Pareto curves for the Bobko-Roth predictors are presented in Figure 4, and the curves for the cognitive subtest predictors, for each of the two job families, are presented in Figures 5 and 6.[9] To understand how to interpret these figures, look at the example graph in Figure 7. In Figure 7, the Pareto curves for the calibration samples (i.e., solid lines) and the Pareto curves for the validation samples (i.e., dotted lines) are shown. The Pareto curves for the validation samples (dotted lines in Figure 7) are the *shrunken* Pareto curves. That is, each shrunken 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 7, 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

[9] Results for the other five ASVAB job families from Wee et al. (2014) may be obtained from the first author.

Pareto solution, or each Pareto point) to a lower expected value of both criterion validity and diversity. In Figure 7, 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 shrinkage 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 7) 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 7) 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 7's illustration, validity shrinkage and diversity shrinkage for Point 4 are two component vectors of Pareto shrinkage. In Figures 4 through 6, 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 4 through 6 showed consistent and intuitive findings regarding sample size: validity shrinkage and diversity shrinkage both decreased as sample size increased. For the Bobko-Roth (Figure 4) and cognitive subtest (Figures 5 and 6) 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 the Bobko-Roth (Figure 4) and cognitive subtest (Figures 5 and 6) 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, due to a nonlinear relationship between shrinkage and sample size.[10] 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 (Figure 4) and the cognitive subtest (Figures 5 and 6) 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),

[10] 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 $(n - 1)/(n - k - 1)$—i.e., 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.

validity shrinkage (ΔR), diversity value (i.e., AI ratio), and diversity shrinkage (ΔAI 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 (ΔR) 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 (ΔAI 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 (Table 3) and cognitive subtest (Table 4) predictors, the results indicated greater shrinkage on the maximized objective. Specifically, diversity shrinkage was larger when diversity was maximized (i.e., ΔAI ratio values in the bottom halves of Tables 3 and 4), as compared to when criterion validity was maximized (i.e., ΔAI 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 ΔAI ratio = 1.24 when diversity was maximized, but only ΔAI ratio = 0.00 when criterion validity was maximized. For the cognitive subtest predictors (Table 4, when N = 40), diversity shrinkage was ΔAI ratio = 0.05 (Machinery Repair) and = 0.03 (Control & Communication) when diversity was maximized, but was smaller at ΔAI ratio = 0.01 (Machinery Repair) and = 0.00 (Control & Communication) when criterion validity was maximized.

Likewise, validity shrinkage was larger when criterion validity was maximized (i.e., ΔR values in the top halves of Tables 3 and 4), as compared to when diversity was maximized (i.e.,

ΔR 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 $\Delta R = -.01$ (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 4 through 6. 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 4 and Table 3) than for the cognitive subtest predictors (Figures 5 and 6, 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 4 (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 5 and Figure 6 (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 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 (due to 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 4/5^{ths} 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, due to 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 4), 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 out 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 4 through 6) to the standard 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 4 through 6) 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 4 vs. Figure 5) and the particular job family (e.g., Figure 5 vs. Figure 6) considered. Based on our simulation results, the Bobko-Roth predictors (Figure 4) 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 (Figure 5). 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 sub-optimal, 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., ΔAI 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 8, left panel) and the cognitive subtest predictors (i.e., Figure 8, right panel).[11] For the Bobko-Roth predictors (Figure 8, left panel), the solid line represents the unshrunken diversity improvement (i.e., diversity

[11] Results in Figure 8 right panel pertain to the Machinery Repair job family; results for other job families may be obtained from the first author.

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 8 (left panel), at smaller sample sizes (e.g., $N = 40$), the unshrunken diversity improvement value (ΔAI ratio = 0.28) over-estimates the actually-attainable shrunken/cross-validated diversity improvement value (ΔAI ratio = -0.01). In contrast, at larger sample sizes (e.g., $N = 500$), the unshrunken diversity improvement (ΔAI ratio = 0.08) and the cross-validated diversity improvement (ΔAI ratio = 0.06) do not differ substantially. A similar pattern of results is observed for the cognitive subtest predictors (bottom two lines in Figure 8, right panel).

In Figure 8, 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 8 (both left and right panels) indicate diversity improvement calculated as the difference between the AI ratio of Pareto weights at $R = .50$ versus the AI ratio of unit weights [in the calibration sample (solid lines) and in the validation sample (dotted lines)]. As such, the bottom two lines in each panel indicate diversity improvement at no cost in terms of job performance (compared to unit weights), and the top two lines in each panel indicate diversity improvement at a modest cost in terms of job performance (i.e., dropping from unit-weighted R to $R = .50$). As expected, larger diversity improvements can be obtained if small decrements in job performance are accepted. These

diversity improvements (the topmost dotted lines in Figure 8) can be obtained despite diversity shrinkage.

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

Study 1, which examined 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 ΔAI 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., Goldstein et al., 2010; Hough, et al., 2001), foster more favorable diversity recruiting (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 Study 1 extended previous research on cross-validation/shrinkage to the multi-criterion 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.[12] 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 formulae) 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 12). 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 formulae (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

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

expect larger diversity shrinkage. This is consistent with what we see in Study 1 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, Study 1 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 face through the use of Pareto-optimal weighting strategies. Conversely, these results also suggest that organizations with much smaller applicant pools ($N << 100$; e.g., start-up firms) might generally benefit from using unit-weighting strategies. One promising area for future research here would be to investigate the possibility that the sample-size tipping point between the recommended use of Pareto weights versus unit weights could be influenced by enhancing calibration sample size via meta-analysis (which could be accomplished using nonlocal calibration samples/validity studies and meta-analysis with local Bayes estimation; Brannick, 2001; Newman et al., 2007). In other words, perhaps even organizations with small applicant pools could augment their calibration sample sizes by using meta-analysis, in order to make Pareto weighting and diversity improvement more feasible.

Additionally, previous research indicates it is generally desirable to combine cognitive predictors (which typically exhibit high criterion validities and high impact) with noncognitive

predictors such as structured interviews or personality tests (which typically exhibit moderate criterion validities and low impact; Hough et al., 2001; Ryan & Ployhart, 2014). Our study suggests that shrinkage affects incremental diversity differently from how it affects incremental validity. As discussed above, whereas validity shrinkage *decreases* with increasing criterion validity (footnote 12), diversity shrinkage *increases* with increasing diversity (because increasing diversity here implies that the predictors do not relate to race, $R^2_{race} = 0$). Thus, caution is required when generalizing the incremental diversity gains that may be obtained from low impact predictors (i.e., low impact predictors imply more diversity shrinkage), especially if the diversity gains were estimated in small or even medium-sized samples (e.g., 500 or fewer applicants, in the case of Bobko-Roth predictors).

Across different job families, there were moderate variations in the job performance and diversity outcomes possible, even when the same set of predictors was used. This result held even after accounting for shrinkage, and thus suggests that practitioners should consider tailoring their selection systems to specific job functions or families, when diversity is a valued outcome. This is consistent with previous calls for developing more construct-oriented approaches to personnel selection (e.g., Schneider, Hough, & Dunnette, 1996), if the applicant pool is sufficiently large.

In Study 1, the Control and Communication job family revealed that large diversity improvements were possible (and unit weights were particularly suboptimal), in part because the predictors exhibited *negative validity-d correspondence* (i.e., the higher-impact predictors had lower criterion validities) for this job family. In contrast, the Machinery Repair job family exhibited *positive validity-d correspondence* (i.e., the higher-impact predictors had higher

criterion validities, similar to the Bobko-Roth scenario), which meant that shrunken diversity improvement was smaller—while still greater than zero.

Limitations and Directions for Future Research

In Study 1, 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_{perf} and R_{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 formulae were based on least squares assumptions that do not hold for Pareto-optimization. However, researchers could consider the extent to which current formulae can be adapted to provide reasonable approximations, even for Pareto-optimal solutions. Alternatively, statistical formulae for Pareto-optimal shrinkage could be specifically developed. In Study 2, we explore development of shrinkage formulae for Pareto-optimal weighting (multi-objective weighting method).

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. Some might wonder 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).

CHAPTER 3

STUDY 2

Development of Pareto-Optimal Shrinkage Formulae

(Approximation of Diversity Shrinkage from Pareto-Optimal Weights)

Study 1 demonstrated that diversity shrinkage exists, and is primarily affected by sample size and the particular combination of high- versus low-impact predictors. The purpose of Study 2 is to attempt to develop some closed-form equations to approximately correct for validity shrinkage and diversity shrinkage when using Pareto-optimal weighting. Shrinkage formulae for Pareto-optimal weighting are derived, and then tested using Monte-Carlo simulation. The proposed method calculates the shrunken Pareto surface by using an interpolation between shrinkage-adjusted values for the endpoints of the Pareto surface.

In particular, Study 2 will attempt to make the following two contributions. First, we will develop a set of formulae to correct for validity shrinkage and diversity shrinkage under Pareto-optimal weighting. Second, we will evaluate the performance of the shrinkage formulae using Monte-Carlo simulation, by comparing formula-adjusted shrunken R^2 estimates for both diversity and job performance against shrunken R^2 outcomes from the simulation.

Method

Development of Pareto-Optimal Shrinkage Correction Formulae

We derive our shrinkage correction formulae for Pareto-optimal weighting based on two main steps. First, we use traditional shrinkage formulae (Browne, 1975; Wherry, 1931) to calculate shrunken R^2 at the two endpoints of the Pareto curve (i.e., at the leftmost endpoint, where job performance is maximized and diversity is ignored; and at the rightmost endpoint, where diversity is maximized and job performance is ignored). Second, we interpolate shrunken

R^2 values between these two endpoints, using a shrunken-to-unshrunken range ratio (described below). We base the derivation on following four main premises.

Premise #1. Classic shrinkage correction. When correcting for shrinkage in OLS regression with a single criterion variable, *the cross-validated squared multiple correlation* ($\hat{\rho}_c{}^2$; i.e., the estimated multiple R^2 in a *new sample*) can be estimated using a formula from Browne (1975; this formula was developed based on the Lord-Nicholson shrinkage formula [Lord, 1950; Nicholson, 1960]):

Browne's shrinkage formula for $\hat{\rho}_c{}^2$ (Browne, 1975, p. 85)

$$\hat{\rho}_c{}^2 = \frac{(N-k-3)\hat{\rho}^4 + \hat{\rho}^2}{(N-2k-2)\hat{\rho}^2 + k} \tag{9}$$

where $\hat{\rho}_c{}^2$ is the corrected cross-validated squared correlation coefficient; $\hat{\rho}^2$ is the corrected population squared multiple correlation coefficient obtained from a formula [such as Wherry's (1931) shrinkage formula[13]: see Equation 10 below, *N* is the calibration sample size, and *k* is the number of predictors.

In this formula, $\hat{\rho}^2$ is the *population squared multiple correlation coefficient*, and comes from Wherry's (1931) shrinkage formula.

Wherry's (1931) shrinkage formula for ρ^2 (see Cattin, 1980, p. 409)

$$\hat{\rho}^2 = 1 - \frac{N-1}{N-k-1}(1-R^2) \tag{10}$$

[13] We conducted additional simulation to compare the performance of Wherry (1931), Olkin & Pratt (1958) and Claudy (1978) formulae. Based on RMSE and Bias results, we selected Wherry formula to be used as part of the Pareto-optimal shrinkage formulae method. For details of the simulation, see Appendix D.

where $\hat{\rho}^2$ is an estimate of 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.

Shrinkage correction at each endpoint of the Pareto curve, using classic formulae.

The Pareto curve has two endpoints: the endpoint where job performance is maximized, and the endpoint where diversity is maximized. Pareto-optimal shrinkage formulae must correct for shrinkage in 2 dimensions: (a) job performance validity shrinkage, and (b) diversity shrinkage. When correcting for job performance validity shrinkage, the first step is to apply the classic shrinkage formulae to job performance (multiple R_{perf}) at the endpoint of the Pareto curve where performance is maximized (e.g., solution point #2 in Figure 1).

When correcting for diversity shrinkage, the first step is to apply the classic shrinkage formulae to diversity (multiple R_{race}) at the endpoint of the Pareto curve where diversity is maximized (e.g., solution point #21 in Figure 1).

Browne's (1975) formula is applicable to correct for validity shrinkage of a certain criterion at the endpoint where that criterion is maximized. That is, at the endpoint where job performance was maximized, job performance criterion shrinkage can be corrected using the Browne formula (with Wherry formula inputs). Similarly, at endpoint where diversity was maximized, diversity shrinkage can be corrected using the Browne formula (with Wherry formula inputs).

Premise #2. At the endpoint where a certain criterion was least maximized, there is no validity shrinkage for that criterion. Study 1 showed that there is no shrinkage on the non-maximized criterion. Thus, the cross-validated R^2 value should be approximately equal to the calibration

value (i.e., unshrunken), at the endpoint of the Pareto curve where the criterion is not being maximized. For example, there is no job performance criterion validity shrinkage at the endpoint where diversity was maximized (i.e., job performance was least maximized). At this endpoint of the Pareto curve, the cross-validated job performance criterion validity should be approximately equal to the calibration sample job performance criterion validity. Similarly, there is no diversity shrinkage at the endpoint where job performance was maximized (i.e., where diversity was least maximized). At this other endpoint of the Pareto curve, the cross-validated diversity R^2 should be approximately equal to the calibration sample diversity R^2.

Premise #3. The *relative criterion weights* (i.e., relative amount of value placed on each criterion at each point along the Pareto curve) in the validation sample stay the same as the relative criterion weights in the calibration (unshrunken) sample. For example, at the endpoint where job performance was maximized, the criterion weight for job performance criterion validity is 1 and the criterion weight for diversity is 0 for both calibration (unshrunken; e.g., light blue point in Figure 9) and cross-validated (shrunken; e.g., purple point in Figure 9) criterion outcomes. Similarly, at the endpoint where diversity was maximized, the criterion weight for diversity is 1 and the criterion weight for job performance is 0 for both calibration (unshrunken; e.g., light green point in Figure 9) and cross-validated (shrunken; e.g., dark green point in Figure 9) criterion outcomes. Further, at the Pareto point where the relative criterion weights for job performance criterion validity and diversity are both .5, job performance and diversity were similarly maximized. At this point on the Pareto curve, the relative criterion weights are .5 for both the calibration sample (unshrunken; e.g., orange point in Figure 9) and cross-validated (shrunken; e.g., red point in Figure 9) criterion outcomes.

Premise #4. Position of each Pareto point as a proportion of the full Pareto range is the same for the shrunken (adjusted) Pareto curve as for the unshrunken (calibration) Pareto curve. As seen in Figure 9, for each Pareto point, the vertical distance from that Pareto point to the endpoint of the Pareto curve can be represented as a proportion of the total range of the Pareto curve (total Pareto range = vertical distance between the endpoints of the Pareto curve). This quantity is labeled the *proportion of Pareto range*, and can be calculated for (a) job performance outcomes in the calibration sample Pareto curve, and (b) job performance outcomes in the shrunken (adjusted) Pareto curve. Alternatively, in the case of calculating diversity outcomes using the horizontal proportion of Pareto range, the proportion can be calculated for: (a) diversity outcomes in the calibration sample Pareto curve, and (b) diversity outcomes in the shrunken (adjusted) Pareto curve.

Premise 4 can perhaps most clearly be expressed as an equation, showing that for each criterion, the *proportion of Pareto range* is the same between calibration and shrunken Pareto curves. This is depicted in Figure 9. The difference between maximum job performance criterion validity (i.e., $R_{Cal.max}$; light blue point) and job performance criterion validity at certain Pareto point (i.e., $R_{Cal.p}$; certain point along the solid red line) on the calibration sample Pareto curve (e.g., light blue bracket in Figure 9), can be represented as a proportion of the total Pareto range [i.e., the difference between maximum job performance criterion validity (i.e., $R_{Cal.max}$; light blue point) and minimum job performance criterion validity (i.e., $R_{Cal.min}$; light green point) on the calibration sample Pareto curve (e.g., dark blue bracket in Figure 9). This is called the proportion of Pareto range.

In order to derive a closed-form shrinkage formula that can be used across the entire Pareto curve, we impose the assumption/constraint that the *proportion of Pareto range* for each point on the shrunken Pareto curve will be equal to the *proportion of Pareto range* for each point on the (unshrunken) calibration sample Pareto curve:

$$\frac{R_{Cal.max}-R_{Cal.P}}{R_{Cal.max}-R_{Cal.min}}=\frac{R_{Adj.max}-R_{Adj.P}}{R_{Adj.max}-R_{Adj.min}} \quad (11)$$

The proportion of Pareto range for the shrunken Pareto curve is set to be equal to the proportion of Pareto range for the unshrunken Pareto curve (i.e., in Figure 9, the proportion of light blue bracket to dark blue bracket is equal to the proportion of light green bracket and dark green bracket). The proportion of Pareto range for the shrunken Pareto curve is the ratio of the difference between maximum job performance criterion validity (i.e., $R_{Adj.max}$) and job performance criterion validity at certain Pareto point (i.e., $R_{Adj.p}$) for the validation sample (shrunken Pareto curve; e.g., light blue bracket in Figure 9) to the difference between maximum job performance criterion validity (i.e., $R_{Adj.max}$) and minimum job performance criterion validity (i.e., $R_{Adj.min}$) for the validation sample (shrunken Pareto curve; e.g., dark blue bracket in Figure 9). For example, in Figure 9, the ratio of the light blue bracket to dark blue bracket on the right hand side is equal to the ratio of the light green bracket to dark green bracket on the left hand side. The relationship is expressed in Equation 11 and Figure 10.

Pareto-Optimal Shrinkage Formulae

The Pareto-optimal shrinkage formulae is shown in Equation 12 and 13. The formulae inputs include: calibration sample size (i.e., N_{Cal}); number of predictors (i.e., k); calibration criterion value at current Pareto point (i.e., $R_{Cal.p}$); calibration criterion value at endpoint where the criterion was maximized (i.e., $R_{Cal.max}$); and calibration criterion value at the endpoint where

the criterion was least maximized (i.e., $R_{Cal.min}$; or, the endpoint where the other criterion was maximized).

$$R_{Adj.p} = R_{Adj.max} - \left(R_{Cal.max} - R_{Cal.p}\right) \times \frac{R_{Adj.max} - R_{Adj.min}}{R_{Cal.max} - R_{Cal.min}} \quad (12)$$

Where $R_{Adj.max}$ can be estimated using traditional shrinkage formulae (see Eq. 14 & 15):

$$R_{Adj.max} = f(R^2 = (R_{Cal.max})^2, N = N_{Cal}, k = k) \quad (13)$$

where

- N_{Cal}: Calibration sample size
- k: Number of predictors
- $R_{Cal.p}$: Calibration sample criterion value at current Pareto point
- $R_{Adj.p}$: Shrinkage-formula-adjusted Pareto-optimal criterion value at current Pareto point
- $R_{Cal.max}$: Calibration sample criterion value at endpoint where the criterion was maximized
- $R_{Cal.min}$: Calibration sample criterion value at the endpoint where the criterion was least maximized
- $R_{Adj.max}$: Calibration sample criterion at the endpoint where the criterion was least maximized, adjusted for shrinkage using Wherry and Browne formula.

Step a. Wherry's shrinkage formula for ρ^2 (Wherry, 1931)

$$\hat{R}^2{}_{Adj.max} = 1 - \frac{N-1}{N-p-1}(1 - R^2) \quad (14)$$

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

Step b. Browne's shrinkage formula for $\rho_c{}^2$

$$R^2_{Adj.max} = \frac{(N-k-3)\hat{R}^4_{Adj.max} + \hat{R}^2_{Adj.max}}{(N-2k-2)\hat{R}^2_{Adj.max} + k} \quad (15)$$

where $R^2_{Adj.max}$ is the corrected cross-validated squared correlation coefficient; $\hat{R}^2_{Adj.max}$is the corrected population squared multiple correlation coefficient obtained from a formula such as Wherry's (1931) shrinkage formula, *N* is the calibration sample size, and *k* is the number of predictors.

The absolute value of $R_{Adj.p}$ is the square-root of $R^2_{Adj.p}$, and the direction of the value (i.e., positive or negative) is defined by the direction of $R_{Cal.p}$. That is, the relationship between $R_{Adj.p}$ and $R^2_{Adj.p}$ are as follows:

$$R_{Adj.p} = \begin{cases} \sqrt{R^2_{Adj.p}} & (R_{Cal.p} \geq 0) \\ -\sqrt{R^2_{Adj.p}}^2 & (R_{Cal.p} < 0) \end{cases} \quad (16)$$

When applied to the personnel selection scenario, the two Pareto-optimal shrinkage formulae for job performance and diversity, which can be applied to estimate the entire shrunken Pareto surface, are as follows:

Job Performance:

$$R_{perf_{Adj.P}} = R_{perf_{Adj.max}} - \left(R_{perf_{Cal.max}} - R_{perf_{Cal.P}}\right) \times \frac{R_{perf_{Adj.max}} - R_{perf_{Adj.min}}}{R_{perf_{Cal.max}} - R_{perf_{Cal.min}}} \quad (17)$$

Diversity:

$$R_{race_{Adj.P}} = R_{race_{Adj.max}} - \left(R_{race_{Cal.max}} - R_{race_{Cal.P}}\right) \times \frac{R_{race_{Adj.max}} - R_{race_{Adj.min}}}{R_{race_{Cal.max}} - R_{race_{Cal.min}}} \quad (18)$$

where R_{perf} (job performance criterion validity) is the correlation between predictor weighted composite score and job performance score; R_{race} (race bivariate correlation) is the correlation

between predictor weighted composite score and race dummy variable (e.g., 0 indicates minority applicants, 1 indicates majority applicants).

We can also further estimate the shrunken diversity AI ratio (i.e., by converting the multiple *R* for race into an AI ratio) by using the following equations from Newman, Jacobs, and Bartram (2007):

$$d = r/\sqrt{(1-r^2)p(1-p)} \tag{19}$$

$$x_{cut} = z_{cut}\sqrt{1+d^2[p(1-p)]} - dp \tag{20}$$

$$AI\ ratio = \frac{SR_{Black}}{SR_{White}} \approx \frac{(1.64x_{cut} + \sqrt{.76x^2{}_{cut} + 4})}{\left(\sqrt{e^{(d^2+2x_{cut}d)}}\right)[1.64(x_{cut}+d) + \sqrt{.76(x_{cut}+d)^2+4}]} \tag{21}$$

Example Application of Pareto-Optimal Shrinkage Formulae

To demonstrate how to apply the above Pareto-optimal shrinkage formula, let us look at an example that uses the Pareto-optimal shrinkage formula to: (a) estimate shrunken job performance criterion validity, and (b) estimate shrunken AI ratio.

Estimate Shrunken Job Performance Criterion Validity. Let us say we want to estimate the shrunken job performance criterion validity for calibration sample of size 40, with the Bobko-Roth predictors (see Table 1), for the Pareto-optimal weights that correspond to the 11th Pareto point. Here, calibration sample size $N = 40$ and number of predictors $k = 5$. We also need the following values: $R_{perf_{Cal.p}}$, $R_{perf_{Cal.min}}$, and $R_{perf_{Cal.max}}$ from the calibration sample (unshrunken) Pareto-optimal solution. The Pareto-optimal solutions can be obtained from packages and programs such as the ParetoR package (Song et al., 2017) or the TROFSS program (De Corte et al., 2007). The Pareto-optimal solution for this example scenario is shown in Table 6. We have following set of input values:

- Calibration sample size: $N = 40$

- Number of predictors: $k = 5$
- Proportion of minority applicants: $p = 1/6$; $p = \frac{number\ of\ minority\ applicants}{total\ number\ of\ applicants} = 1/6$
- Selection ratio: $sr = 0.15$; $sr = \frac{number\ of\ applicants\ selected}{total\ number\ of\ applicants} = .15.$
- Calibration sample criterion validity at Pareto point 11: $R_{perf_{Cal.p11}} = .51$
- Calibration sample criterion validity where diversity was maximized: $R_{perf_{Cal.min}} = .28$
- Calibration sample criterion validity where performance was maximized:$R_{perf_{Cal.max}} = .69$

Next, using Wherry formula (Equation 14) and Browne formula (Equation 15), we estimate shrunken criterion validity at the point where job performance was maximized:

Wherry formula (refer to Equation 14):

$$\hat{R}^2_{perf_{Adj.max}} = 1 - \frac{N-1}{N-k-1}\left(1 - R^2_{perf_{Cal.\max}}\right)$$

$$= 1 - \frac{40-1}{40-5-1}(1 - (.69)^2)$$

$$= .41$$

Browne formula (refer to Equation 15)

$$R^2_{perf_{Adj.max}}$$

$$= \frac{(N-k-3) \times \left(\hat{R}^2_{perf_{Adj.max}}\right)^2 + \hat{R}^2_{perf_{Adj.max}}}{(N-2k-2) \times \hat{R}^2_{perf_{Adj.max}} + k}$$

$$= \frac{(40-5-3) \times (.41)^2 + .41}{(40-2\times 5-2) \times .41 + 5}$$

$$= .35$$

Now, we can calculate $R_{perf_{Adj.p}}$ from $R^2_{perf_{Adj.max}}$ through Equation 16:

$$R_{perf_{Adj.max}} = \begin{cases} \sqrt{R^2_{perf_{Adj.max}}} & (R_{perf_{Cal.p}} \geq 0) \\ -\sqrt{R^2_{perf_{Adj.max}}} & (R_{perf_{Cal.p}} < 0) \end{cases} = \begin{cases} \sqrt{.36} & (.69 \geq 0) \\ -\sqrt{.36} & (.69 < 0) \end{cases} = .60$$

Finally, plug $R_{perf_{Cal.p}}$, $R_{perf_{Cal.max}}$, $R_{perf_{Cal.min}}$, $R_{perf_{Adj.max}}$ into Equation 17:

$$R_{perf_{Adj.P}}$$

$$= R_{perf_{Adj.max}} - \left(R_{perf_{Cal.max}} - R_{perf_{Cal.P}}\right) \times \frac{R_{perf_{Adj.max}} - R_{perf_{Adj.min}}}{R_{perf_{Cal.max}} - R_{perf_{Cal.min}}}$$

$$= 0.60 - (0.69 - 0.51) \times \frac{0.60 - 0.28}{0.69 - 0.28}$$

$$= 0.46$$

Estimate Shrunken AI Ratio. Let us say we want to estimate the shrunken diversity (AI ratio) for a calibration sample of size 40, with the Bobko-Roth predictors, for the Pareto-optimal weights that correspond to the 11th Pareto point. Here, calibration sample size $N = 40$, and number of predictors $k = 5$. The set of input values is as follows:

- Calibration sample size: $N = 40$
- Number of predictors: $k = 5$
- Proportion of minority applicants: $p = 1/6$
- Selection ratio: $sr = .15$
- Calibration sample diversity (race) at Pareto point 11: $R_{race_{Cal.P11}} = .01$
- Calibration sample diversity (race) where diversity was maximized: $R_{race_{Cal.max}} = -.11$
- Calibration sample diversity (race) where job performance was maximized:$R_{race_{Cal.min}} = .33$

Using the Wherry formula (Equation 10), we can estimate:

$$\hat{R}^2_{race_{Adj.max}}$$

$$= 1 - \frac{N-1}{N-k-1}\left(1 - R^2_{race_{Cal.\max}}\right)$$

$$= 1 - \frac{40-3}{40-5-1}\left(1 - (-.11)^2\right)$$

$$= -.13$$

$$\rightarrow 0\ (-.13 < 0)$$

Further, using the Browne formula (Equation 9), we can estimate

$$R^2_{race_{Adj.max}}$$

$$= \frac{(N-k-3) \times \left(\hat{R}^2_{race_{Adj.max}}\right)^2 + \hat{R}^2_{race_{Adj.max}}}{(N-2k-2) \times \hat{R}^2_{race_{Adj.max}} + k}$$

$$= \frac{(40-5-3) \times (0)^2 + 0}{(40-2\times 5-2) \times 0 + 5}$$

$$= 0$$

Then, $R_{race_{Adj.P}}$ can be estimated by plugging $R_{race_{Cal.p}}$, $R_{race_{Cal.max}}$, $R_{race_{Cal.min}}$, $R_{race_{Adj.max}}$ into Equation 18:

$$R_{race_{Adj.P}}$$

$$= R_{race_{Adj.max}} - \left(R_{race_{Cal.max}} - R_{race_{Cal.P}}\right) \times \frac{R_{race_{Adj.max}} - R_{race_{Adj.min}}}{R_{race_{Cal.max}} - R_{race_{Cal.min}}}$$

$$= 0 - (-.11 - .01) \times \frac{0 - .33}{-.11 - .33}$$

$$= .09$$

We can also further estimate shrunken AI ratio through Equations 19 to 21.

$$z_{cut} = -1.04$$

where z_{cut} is the standard normal cut score across all applicants given selection ratio of 0.15 (i.e., value of normal distribution probability density function at percentile of 1 - 0.15 = 0.85).

$$d = \frac{R_{race_{Adj.p}}}{\sqrt{\left(1 - R^2_{race_{Adj.p}}\right) p(1-p)}} = \frac{0.09}{\sqrt{(1 - 0.09^2) \times 0.167 \times (1 - 0.167)}} = 0.24$$

$$x_{cut} = z_{cut}\sqrt{1 + d^2[p(1-p)]} - dp$$

$$= (-1.04) \times \sqrt{1 + 0.24^2 \times [0.167 \times (1 - 0.167)]} - 0.24 \times 0.167$$

$$= -1.08$$

$$AI\ ratio \approx \frac{\left(1.64x_{cut} + \sqrt{.76x^2{}_{cut} + 4}\right)}{\left(\sqrt{e^{(d^2 + 2x_{cut}d)}}\right)\left[1.64(x_{cut} + d) + \sqrt{.76(x_{cut} + d)^2 + 4}\right]}$$

$$= \frac{\left(1.64 \times (-1.08) + \sqrt{.76 \times (-1.08)^2 + 4}\right)}{\left(\sqrt{e^{(0.24^2 + 2 \times (-1.08) \times (0.24))}}\right) \times \left[1.64 \times (-1.10 + 0.24) + \sqrt{.76 \times (-1.08 + 0.24)^2 + 4}\right]}$$

$$= 0.73$$

Simulation

A simulation study was conducted to evaluate the performance of the Pareto-optimal shrinkage formulae (Equations 12, 14, and 15). The shrinkage formulae were evaluated using a combination of cognitive and non-cognitive predictors (i.e., cognitive ability, conscientiousness, integrity, structured interview, biodata; see Table 1). We obtained predictor correlations, job performance validity, and subgroup *d* of the five predictors via meta-analysis (Bobko & Roth, 2013; Roth, Switzer, Van Iddeking, & Oh, 2011; Van Iddeking, Roth, Raymark, & Odle-Dusseau, 2012; McKay & McDaniel, 2006), and we call the predictor combination "Bobko-Roth predictors". Five calibration sample sizes were studied: *n* = 40, 100, 200, 500, and 1000. Each condition was replicated 1000 times.

The simulation, which includes eight steps, was conducted using the ParetoR package (Song et al., 2017) and the R programming language (R version 3.2.0; R Core Team, 2015).

Step 1[14]*. Generate calibration samples*. We simulated 1,000 samples for each of five sample size conditions (*n* = 40, 100, 200, 500, and 1000) from the population correlation matrix (see Table 1).

Step 2. Generate validation samples. Next, we simulated an additional 1,000 samples for each sample size condition, from the same population correlation matrix (Table 1).

Step 3. For each replication, calculate calibration sample tradeoff frontier of 21 Pareto solutions. Each of the 21 Pareto solutions includes a set of Pareto-optimal predictor weights, an estimate of criterion validity (correlation between weighted predictor composite and job performance), and an estimate of diversity (correlation between weighted predictor composite and race, which is later transformed to AI ratio—see Step 7 below).

Step 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.

Step 5. For each replication, apply Pareto-optimal shrinkage formulae (Equation 12, 14, and 15) to the calibration sample, to obtain formula-based estimates of the shrunken criterion validity and shrunken diversity. The Pareto shrinkage formulae are used as follows:

1) Obtain R_{perf} and R_{race}, at two endpoints: where performance was maximized and where diversity was maximized.

[14]Steps 1 to 4 are same as Song, Wee and Newman (2017; p. 1641 – 1643). For more details of Steps 1 to 4, please refer to Song et al. (2017).

a. R_{perf}: bivariate correlation of predictor weighted composite score and job performance

b. R_{race}: bivariate correlation of predictor weighted composite score and diversity variable (e.g., race dummy variable where White applicants are coded as 1 and Black applicants are coded as 0)

c. $R_{perf_{Cal}}$: R_{perf} of calibration sample

d. $R_{perf_{Val}}$: R_{perf} of validation sample

e. $R_{race_{Cal}}$: R_{race} of calibration sample

f. $R_{race_{Val}}$: R_{race} of validation sample

g. $R_{perf.max}$: Pareto-optimal output for job performance (R_{perf}) at the endpoint where *job performance* was *maximized*

h. $R_{perf.min}$: Pareto-optimal output for job performance (R_{perf}) at the endpoint where *job performance* was *least maximized*

i. $R_{perf.p}$: Pareto-optimal output for job performance (R_{perf}) at the current Pareto point

j. $R_{race_{max}}$: output for *diversity* (R_{race}: the correlation between the predictor composite and the race variable) at the endpoint where *diversity* was *maximized*; (i.e., where the correlation between race variable and job performance was least maximized)

k. $R_{race_{min}}$: output for *diversity* (R_{race}) at the endpoint where *diversity* was minimized (i.e., where the difference between racial groups is maximized, *diversity* outcomes are *least maximized*)

l. $R_{race.p}$: output for diversity (R_{race}) at the current Pareto point

2) Obtain criterion weights

Estimate *relative criterion weights* for performance (w_{perf}) and diversity (w_{race}) using the following formulae:

$$w_{race} = \frac{R_{perf_{Cal.max}} - R_{perf_{Cal.p}}}{R_{perf_{Cal.max}} - R_{perf_{Cal.min}}} = \frac{R_{perf_{Cal.p}} - R_{perf_{Cal.max}}}{R_{perf_{Cal.min}} - R_{perf_{Cal.max}}} \quad (22)$$

Notice that w_{race} (the criterion relative weight for *diversity*) is defined in terms of the Pareto range for *job performance*. This is because at the Pareto endpoint where job performance in maximized (where $R_{perf_{Cal.p}} = R_{perf_{Cal.max}}$), that is where the diversity criterion is weighted zero. Similarly, w_{perf} (the criterion relative weight for *job performance*) is defined in terms of the Pareto range for *diversity* (i.e., w_{perf} can be calculated by using a version of Equation 22, in which job performance and diversity subscripts are switched).

3) Obtain shrunken R values for job performance and diversity, at both of the endpoints of the Pareto curve, via Wherry (1931) shrinkage formula and Browne (1975) shrinkage formulae

- $R_{perf_{Adj.max}}$: $R_{perf_{Cal.max}}$ (i.e., calibration job performance criterion validity at endpoint where *job performance* was *maximized*) after correcting for shrinkage using Wherry and Browne formulae
- $R_{race_{Adj.max}}$: $R_{race_{Cal.max}}$ (i.e., calibration diversity at endpoint where *diversity* was *maximized*) after correcting for shrinkage using Wherry (1931) shrinkage formula and Browne (1975) shrinkage

    3a) Wherry's (1931) shrinkage formula for ρ^2 (Equation 10).

    3b) Browne's shrinkage formula for $\rho_c{}^2$ (Equation 9).

4) Obtain shrunken R^2 values of Pareto points between the endpoints

Job performance:

$$R_{perf_{Adj.p}} = (w_{race} \times \left(R_{perf_{Adj.min}} - R_{perf_{Adj.max}}\right) + R_{perf_{Adj.max}}) \quad (23)$$

Diversity:

$$R_{race_{Adj.p}} = (w_{perf} \times \left(R_{race_{Adj.min}} - R_{race_{Adj.max}}\right) + R_{race_{Adj.max}}) \quad (24)$$

Step 6. Calculate RMSE and Bias

For each Pareto point (total of 21 Pareto points), we calculated the root-mean-squared-error (RMSE) and bias for formula-adjusted shrunken job performance validity ($R_{perf_{Adj.P}}$) and formula-adjusted shrunken race bivariate correlation ($R_{race_{Adj.p}}$):

$$RMSE\left(R_{perf_{Adj.P}}\right) = \sqrt{\frac{\sum_{j=1}^{1000}\left(R_{perf_{Adj.p}} - R_{perf_{Val.p}}\right)^2}{1000}} \quad (25)$$

$$RMSE\left(R_{race_{Adj.p}}\right) = \sqrt{\frac{\sum_{j=1}^{1000}\left(R_{race_{Adj.p}} - R_{race_{Val.p}}\right)^2}{1000}} \quad (26)$$

$$Bias\left(R_{perf_{Adj.P}}\right) = \frac{\sum_{j=1}^{1000}\left(R_{perf_{Adj.p}} - R_{perf_{Val.p}}\right)}{1000} mean \quad (27)$$

$$Bias\left(R_{race_{Adj.p}}\right) = \frac{\sum_{j=1}^{1000}\left(R_{perf_{Adj.p}} - R_{perf_{Val.p}}\right)}{1000} \quad (28)$$

RMSE and bias were calculated for each trial, and then aggregated (average) over 1,000 trials.

Step 7. Aggregate R_{perf} and R_{race} results over 1,000 trials and obtain $\bar{R}_{perf}$, $\bar{R}_{race}$, and $\overline{AIratio}$.

First, for each of the Pareto-optimal points (e.g., 21 points), estimate the mean of R_{perf} and R_{race} results over five calibration sample size conditions (calibration sample size = 40, 100, 200, 500, 1000) and 1,000 trials. For example, if the number of Pareto-optimal solutions is 21, then as a result of the aggregation, for each calibration sample size (e.g., 40), we will have 21

mean R_{perf} (i.e., $\bar{R}_{perf_1}, \bar{R}_{perf_2}, \ldots, \bar{R}_{perf_{21}}$) and 21 mean R_{race} (i.e., $\bar{R}_{race_1}, \bar{R}_{race_2}, \ldots, \bar{R}_{race_{21}}$). In all, after the procedure, we will obtain 5 (number of calibration sample size conditions) × 21 (number of Pareto-optional solutions) $\bar{R}_{perf}$ and $\bar{R}_{race}$ values.

Next, estimate $\overline{AIratio}$ from $\bar{R}_{race}$ using the following formulae from Newman, Jacobs, and Bartram (2007, p. 1404), given that *SR* (selection ratio) and *p* (proportion of minority applicants in the applicant pool) are known for each condition. In all, 5 (number of calibration sample size conditions) × 21 (number of Pareto-optional solutions) $\overline{AIratio}$ values will be obtained after the procedure.

Step 8. Plot the Pareto-optimal trade-off curves and shrinkage formula results.

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. Specifically, the tradeoff curves for calibration samples were obtained by plotting the mean calibration criterion validity ($R_{perf_{Cal}}$) against mean calibration diversity ($R_{race_{Cal}}$). The tradeoff curves for validation samples were obtained by plotting the mean validation (cross-validated) criterion validity ($R_{perf_{Val}}$) against mean validation (cross-validated) diversity ($R_{race_{Val}}$). The shrinkage formula results were obtained by plotting the criterion validity obtained from the shrinkage formula ($R_{perf_{Adj}}$) and the diversity obtained from shrinkage formula ($R_{race_{Adj}}$).

Sensitivity Analysis

To test whether the results were due to the specific choice of values for selection ratio and proportion of minority applicants, we conducted a sensitivity analysis. Specifically, we

conducted additional simulation with calibration sample of size 40 under the following conditions: 1) proportion of minority applicants = .10 (i.e., 1/10), .167, (i.e., 1/6), .25 (i.e., 1/4); and 2) selection ratio = .10, .15, .30. A total of $3 \times 3 = 9$ additional conditions were evaluated. We chose to further examine the condition when the calibration sample size is 40, as previous results showed that across all conditions, validity and diversity shrinkage were largest when the calibration sample size was small. The Bobko-Roth predictor correlations were used for the sensitivity analysis and were based on meta-analytic values (Table 1). Consistent with the previous simulation, each condition was replicated for 1,000 trials, and the results from each replication were aggregated by each condition.

Results

Shrinkage Formula Results

Results of the simulation are shown in Figures 11a to 11e. In Figures 11a to 11e, square-dotted black lines represent the calibration sample Pareto curve (unshrunken Pareto curve)), circle-dotted blue lines represent the cross-validated Pareto curve (shrunken Pareto curve), and triangle-dotted red lines represent the calibration sample Pareto curve adjusted with the Pareto-optimal shrinkage formula (adjusted Pareto curve). We can see that in each figure, the formula-adjusted Pareto curves (i.e., triangle-dotted red lines) are much closer to the cross-validated Pareto curves (i.e., circle-dotted blue lines) than are the calibration sample Pareto curves (i.e., square-dotted black lines). In other words, using the Pareto shrinkage formula appears to improve the accuracy with which the cross-validated Pareto curve of estimated (see results below for estimates of RMSE and bias).

Nonetheless, the adjusted Pareto formula is not perfect. For example, as seen in Figure 11a, the formula-adjusted Pareto curve is generally slightly lower on the vertical axis (i.e., it

estimates slightly lower job performance validity) compared to the shrunken Pareto curve, and is to the right on the horizontal axis (i.e., it estimates slightly higher diversity outcomes, closer to an AI ratio of 1.0) compared to the shrunken Pareto curve. This indicates that in this (Bobko-Roth) sample, *the Pareto-optimal shrinkage formula is capable of adjusting for validity and diversity shrinkage in Pareto-optimal solutions, but it generally slightly over-adjusts for job performance validity shrinkage and slightly under-adjusts for diversity shrinkage*.

Results of RMSE for the formula-based estimations are shown in Figures 12a and 12b, whereas bias are shown in Figures 12c and 12d, respectively. RMSE (which is the square-root of MSE—sum of variance and bias) illustrates the cross-validation error. As mentioned earlier, bias is the difference between estimated value and true value, in this case the difference between formula-adjusted shrunken criterion value (e.g., job performance criterion validity, AI ratio) and the shrunken criterion value obtained by simulation (true shrunken criterion value). The aim of the shrinkage formulae is to estimate shrunken criterion value in the validation sample, or obtain estimates that have smallest bias and RMSE.

Figures 12a through 12d compare the RMSE and bias of calibration sample (unshrunken) Pareto-optimal solution and the formula-adjusted Pareto-optimal solution. In Figure 12a, we see that the RMSE (cross-validation error) are very similar between unshrunken (calibration sample) and formula-adjusted Pareto solutions (especially when calibration sample size is large), and is primarily a function of sample size. That is, as calibration sample size increases, RMSE decreases. This is consistent with what we observed in Study 1. Figure 12c illustrates the *bias* in job performance validity. Results suggest that: (a) unshrunken (calibration sample) Pareto-optimal solutions suffer from a sizeable positive bias (overestimation) of criterion validity, that grows to the extent that job performance is being maximized (i.e., the positive bias is larger on

the left-hand side of the figure, where Multiple *R* is overestimated by around +.07), (b) formula-adjusted Pareto solutions do not suffer from positive bias of criterion validity [if anything, the shrinkage adjustment formula produces a small negative bias (underestimation) of Multiple *R*, around -.02, and this small negative bias does not depend upon whether job performance is being maximized], and (c) all estimation bias approach zero as sample size increases toward *N* = 500. Note that previous studies (e.g., Drasgow, Dorans, & Tucker, 1978; Cattin, 1980) found that regression shrinkage formulae such as the Browne formula also demonstrate negative bias (i.e., estimated value is smaller than cross-validated value). Thus, the negative bias observed in the current study is consistent with previous findings based on regression shrinkage formula.

Figure 12b shows RMSE for both unshrunken and formula-adjusted Pareto-optimal solutions. Results suggest that: (a) unshrunken (calibration sample) solutions suffer from a large amount of error (RMSE), especially when diversity is being maximized (right-hand side of Figure 12b), (b) formula-adjusted Pareto solutions suffer from a much smaller degree of error, which is fairly constant across different Pareto-optimal points, and (c) unshrunken Pareto-optimal solution do not approach the formula-adjusted Pareto curve (in terms of RMSE) until the sample size is near *N* = 1,000. In Figure 12d, we see that the formula-adjusted Pareto-optimal solutions are relatively unbiased, compared to the unshrunken Pareto solutions (which are negatively biased, especially to the extent that diversity is being maximized and sample size is small. The Pareto-optimal shrinkage formulae tend to slightly underestimate race bivariate correlations, and thus overestimate AI ratios (i.e., they give conservative estimate, suggesting that estimated diversity returns will be harsher than what they actually turn out to be in reality), even when the sample size is large.

In conclusion, we found that the formula-adjusted Pareto-optimal solutions yield smaller RMSE and bias than unshrunken (calibration sample) Pareto-optimal solution. Pareto-optimal shrinkage formula provides job performance validity and AI ratio estimates that are closer to the validation sample estimates than calibration sample. Larger adjustments are seen when calibration sample size is small. In general, the formula-adjusted job performance validity and formula-adjusted race bivariate correlation were smaller than validation sample job performance validity and validation sample race bivariate correlation, providing conservative estimates of the validation sample criterion value.

Sensitivity Analysis

To evaluate performance of the Pareto-optimal shrinkage formula under various proportion of minority applicant and selection ratio conditions, we also further conducted a sensitivity analysis under the following conditions:

- Proportion of minority applicants (Prop)
    - Prop = 0.25 (i.e., 1/4)
    - Prop = 0.17 (i.e., 1/6) (original condition)
    - Prop = 0.10 (i.e., 1/10)
- Selection ratio (SR)
    - SR = 0.30
    - SR = 0.15 (original condition)
    - SR = 0.10

Results are shown in Figure 13. The results suggest that the shrinkage correction formulae consistently does good job in estimating cross-validated solutions, across various

proportions of minority representation in the applicant pool and various selection ratio conditions.

Discussion

General Findings

In the current study, we developed approximate formulae to correct for validity shrinkage and diversity shrinkage when using Pareto-optimal weighting to optimize on two criteria simultaneously. Simulation results showed that the formulae can be used to easily estimate shrunken Pareto-optimal solutions (i.e., when Pareto-optimal weights are applied to new samples). It provides relatively accurate estimates (as compared to unshrunken calibration sample estimates) of shrunken validity when job performance was maximized, and relatively accurate estimates (as compared to calibration sample estimates) of shrunken diversity (AI ratio) when diversity is maximized. The current study marks the first attempt of which we are aware to develop shrinkage methods for multi-objective optimization in personnel selection. We also supported that the application of classic shrinkage formulae (Browne, 1975; Wherry, 1931) to multi-objective weighting, allowing for future expansion of the formulae to be applied to broader applications, such as algorithms where more than two criterion are maximized.

Results showed that the performance of shrinkage formulae varies for different calibration sample sizes. The pattern can be seen in Figures 11a through 11e. When calibration sample size was small (i.e., $N = 40$, $N = 100$, and $N = 200$), the formula-adjusted Pareto curve was more accurate estimate of the cross-validated (shrunken) Pareto curve, in comparison to the original calibration-sample Pareto curve. This trend does not hold, however, when calibration sample size gets very large. When calibration sample size is greater than 500 (see Figures 11d and 11e), the formula-adjusted Pareto curve becomes less accurate than the original calibration

sample Pareto curve. As such, *we recommend that when using a set of predictors similar to the Bobko-Roth predictors, the shrinkage formulae should only be applied when the corresponding calibration sample size is less than N = 500*.

The formulae developed in the current study allow researchers and practitioners to predict validity and diversity when generalized to another sample. This allows for easy estimation of shrinkage-adjusted job performance validity and diversity (e.g., AI ratio) without the need for simulation, thus promoting the application of the Pareto-optimal method to enhance diversity in hiring to a broader audience. Practitioners (e.g., hiring design teams, hiring managers) now could make an approximation and evaluation of the result of a hiring design before implementing the predictor weighting strategy to make decisions on job applicants. We also include in the current paper an example adaptation and tutorial for researchers and practitioners to easily adapt the method with step-by-step demonstrations.

We have also included the shrinkage formulae as part of a Pareto-optimal R package (the ParetoR package). Similar to features provided by common software and packages for OLS regression, the newly added Pareto-optimal shrinkage formulae feature provides users the adjusted validity and diversity results for application of Pareto-optimal weights to personnel selection. This allows for one-stop estimation of Pareto-optimal solutions. For details of the update, please refer to Appendix C.

Limitations and Future Research

One limitation of the current formula is that at the endpoint where job performance was maximized, the formula does not provide a completely unbiased estimate of cross-validated results, and there is space for further improvement. This might be due to the fact that the Wherry (1931) and Browne (1975) formulae, which are included in the Pareto-optimal shrinkage

formulae method, provide negative bias when estimating shrunken R^2 (e.g., Drasgow, Dorans, & Tucker, 1978; Cattin, 1980).

One promising area of future research is to develop a Pareto-optimal algorithm that can develop Pareto weights which are less susceptible to shrinkage to begin with. For example, in the field of machine learning and statistical learning, regularization techniques address shrinkage by using a *tuning parameter* that tries to "strike a balance" between, "local optimization of prediction and future generalization of prediction" (Oswald & Putka, 2015, p. 52). Regularization techniques include ridge regression (e.g., Tikhonov, 1963; Hoerl, 1962), LASSO regression (Chen and Donoho, 1994; Chen, Donoho and Saunders, 2001; Tibshirani, 1996; also briefly described in Oswald & Putka, 2015), and elastic nets (Zou & Hastie, 2005). We could combine the logic of regularization with the Pareto-optimal method to develop a Pareto-optimal weighting algorithm that achieves both optimization and regularization (similar to ridge regression, LASSO regression, or elastic nets; in the context of Pareto weighting with two criteria). With development of such a method, researchers and practitioners could expect that the predictor weights obtained from an initial sample (i.e., calibration sample) could better generalize to other samples. This is the aim of Study 3.

CHAPTER 4

STUDY 3

Development of a Regularized Pareto-Optimal Weighting Algorithm

Study 1 demonstrated that diversity shrinkage exists, and showed how both diversity shrinkage and job performance validity shrinkage depend upon the extent to which a particular criterion is being maximized. Study 2 attempted to develop and evaluate closed-form equations for shrinkage correction (both diversity shrinkage correction formulae and job performance shrinkage correction formulae). Finally, Study 3 attempts to develop a Pareto-optimal weighting algorithm using an elastic net-like method, which is designed to achieve both diversity-performance optimization and regularization.

In Study 3, we will ultimately compare three methods for estimating cross-validated diversity and job performance outcomes using Pareto-optimal weighting: (a) calibration sample *multiple R* estimates obtained using the Pareto-optimal weighting method introduced by De Corte et al. (2007), (b) formula-adjusted adjusted calibration sample *multiple R* estimates [corrected using the shrinkage formulae derived in Study 2], and (c) calibration sample *multiple R* estimates obtained using the regularized Pareto-optimal weighting method that will be discussed in this study. Using Monte-Carlo simulation, all three methods will be evaluated on both the RMSE and bias with which they estimate the shrunken *R* (from the validation sample), for both diversity and job performance outcomes.

Method

Development of Regularized Pareto-Optimal Method

As a recap, the NBI algorithm first finds endpoints using a linear-regression-like algorithm – sequential least squares quadratic programming (SLSQP[15]). Then, based on the endpoints, it generates an equally spaced set of solutions between the two endpoints. Study 1 showed that the largest shrinkage took place at two endpoints (i.e., diversity shrinkage is largest at endpoint where diversity was maximized; job performance criterion validity shrinkage is largest at the endpoint where job performance was maximized). Relatedly, Study 2 showed that if we correct for the endpoints, we will be able to nonlinearly interpolate an approximation of the cross-validated Pareto-optimal solution. Next, the following simple approach will be used to develop our Pareto-optimal regularization method. First (Step 1), we will employ regularization methods (e.g., elastic net) to solve the endpoints of the Pareto curve in a manner that minimizes cross-validation error (i.e., Equation 1) at the endpoints. Next, Step 2 and Step 3 of the Pareto-optimal regularization method will be the same as the NBI method. Finally, a regularization method (e.g., elastic net) will again be used on a projection of evenly-spaced sub-solutions between the endpoints.

Step 1 of Pareto-optimal regularization method: Employ regularization methods to solve the endpoints of the Pareto-optimal curve.

In this step, a regularization method (specifically, the elastic net-like method) is used to obtain (a) the diversity value (i.e., $AIratio_{max}$) and job performance criterion validity (i.e., $R_{perf_{min}}$) at the endpoint of the Pareto curve where only diversity was treated as a criterion (i.e., when only $AIratio$ was maximized); and (b) the diversity value (i.e., $AIratio_{min}$) and job performance criterion validity (i.e., $R_{perf_{max}}$) at the endpoint of the Pareto curve where only job performance was treated as a criterion (i.e., when only R_{perf} was maximized).

[15] Sequential least squares programming (SLSQP) was implemented in R using the 'nloptr' package: Johnson, S. G., The NLopt nonlinear-optimization package, http://ab-initio.mit.edu/nlopt

As the elastic net technique incorporates both ridge regression and LASSO regression techniques [i.e., when $\alpha = 1$, the elastic net penalty equation (Equation 6) becomes equivalent to the LASSO regression equation; whereas when $\alpha = 0$, it becomes equivalent to the ridge regression equation], we base our regularized Pareto curve method on the elastic net idea to first regularize the endpoints of the Pareto curve. We do this by adding the elastic net penalty term, $\lambda\left[(1-\alpha)\sum_{j=1}^{k}\beta_j^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right]$, to the criterion set to be minimized. Thus, the criterion to be minimized by the SLSQP algorithm now can be expressed by the following:

$$\boldsymbol{w_{C_i}}^{\boldsymbol{T}} \times \begin{pmatrix} AIratio_{pi} \\ R_{perf_{pi}} \end{pmatrix} + \lambda\left[(1-\alpha)\sum_{j=1}^{k}\beta_j^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right]$$

as compared to the criterion minimized by SLSQP in the NBI algorithm (adapted by De Corte et al., 2017):

$$\boldsymbol{w_{C_i}}^{\boldsymbol{T}} \times \begin{pmatrix} AIratio_{pi} \\ R_{perf_{pi}} \end{pmatrix}$$

where $\boldsymbol{w_{C_i}}$ is a vector of criterion weights that were linearly interpolated between the two endpoints (see Figure 2), $AIratio_{pi}$ is the $AI\ ratio$ at Pareto point i, $R_{perf_{pi}}$ is the R_{perf} at Pareto point i, β is the standardized regression coefficient, λ is the tuning parameter that adjusts for shrinkage, and α is a parameter that controls for the relative weighting of the penalty terms in Equation 6.

Cross-validation was used to select the λ parameter and α parameter[16]. This was done by first setting up a grid, or potential solutions. For example, potential solutions for λ include:

[16] We note that sometimes, α parameter is chosen based on theoretical needs. For example, if we intend for variable selection, which allows for some of the predictor weights to equal zero, then we can set $\alpha = 1$. Otherwise, if we intend to find solutions with non-zero but small predictor weights, we can set $\alpha = 0$. If want to find solution for which both ridge and LASSO characteristics are equally taken into account, then we can set $\alpha = 0.5$. However, we attempt to

$[10^{-1.00}, 10^{-0.95}, \ldots, 10^{-0.05}, 10^{0}, 10^{0.05}, \ldots, 10^{1.95}, 10^{2.00}]$, and potential solutions for α include: $[0, 0.05, 0.10, \ldots, 0.95, 1.00]$. We then pair-up all possible λ and α values to form potential solutions for λ and α pairs. Following, cross-validation will be used to find a λ and α pair that minimizes cross-validation error (i.e., MSE for the k training samples). Then, the optimal pair of λ and α were used for the elastic net penalty term (Equation 6), and predictor weights were estimated with the aim of minimizing the elastic net penalty term.

Step 2 of Pareto-optimal regularization method. Phi trade-off matrix (same as Step 2 of NBI method, or the Pareto-optimal method introduced by De Corte et al., 2007)

Criterion values (diversity and job performance) at the two endpoints could further be used to build a Phi trade-off matrix, as shown below. As you can see, the Phi matrix has zeros in the diagonal, and has the range of the AI ratio (i.e., $AIratio_{min} - AIratio_{max}$) and range of job performance criterion validity (i.e., $R_{perf_{min}} - R_{perf_{max}}$) in the off-diagonal. This Phi matrix, combined with criterion weights at each Pareto point, is used to estimate an optimal set of predictor weights at each Pareto point.

$$\Phi = \begin{pmatrix} 0 & R_{perf_{min}} - R_{perf_{max}} \\ AIratio_{min} - AIratio_{max} & 0 \end{pmatrix}$$

Step 3 of Pareto-optimal regularization method. Linear interpolation of evenly-spaced solutions between the endpoints (same as Step 3 of NBI algorithm, or the Pareto-optimal method introduced by De Corte et al., 2007)

Now that we have estimated the two endpoints (in Step 1), we can estimate the Pareto-optimal points between the two endpoints by first creating a line that connects the two endpoints (i.e., the orange line in Figure 2). Then, the algorithm will specify evenly-spaced points along

---

take a data-driven and exploratory approach and decide on the value of α parameter based on cross-validation results.

this line, ensuring the number of points along the line (i.e., including the two end points; yellow points in Figure 2) equals the user-specified number of Pareto solutions (e.g., 21 evenly-spaced solutions). Finally, at each Pareto point, the algorithm will project along a perpendicular direction from the initial criterion line (i.e., yellow arrows in Figure 2) until it reaches the border of the solution space (i.e., blue oval in Figure 2). For details of this procedure, see Step 3 of the NBI method, as illustrated in Chapter 1: Introduction.

Step 4 of Pareto-optimal regularization method. Projection of evenly-spaced sub-solutions between the endpoints using regularization method

Following, at each ith Pareto point, the SLSQP algorithm was again used to find optimal set of weights (i.e., $[w_{pi.1}, w_{pi.2}, \dots, w_{pi.k}]$) that minimizes the loss function:

$$\left[\begin{pmatrix} R_{perf_{pi}} \\ AIratio_{pi} \end{pmatrix} - \begin{pmatrix} R_{perf_{max}} \\ AIratio_{max} \end{pmatrix} - \Phi \boldsymbol{w_{c_i}}^T\right] \times \left[\Phi \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right]^{-1} + \lambda \left[(1-\alpha)\sum_{j=1}^{k}\beta_j^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right]$$

$$= \left[\begin{pmatrix} R_{perf_{pi}} \\ AIratio_{pi} \end{pmatrix} - \begin{pmatrix} R_{perf_{max}} \\ AIratio_{max} \end{pmatrix} - \begin{pmatrix} 0 & R_{perf_{min}} - R_{perf_{max}} \\ AIratio_{min} - AIratio_{max} & 0 \end{pmatrix} \boldsymbol{w_{c_i}}^T\right]$$

$$\times \begin{pmatrix} R_{perf_{min}} - R_{perf_{max}} \\ AIratio_{min} - AIratio_{max} \end{pmatrix}^{-1} + \lambda \left[(1-\alpha)\sum_{j=1}^{k}\beta_j^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right]$$

where $\boldsymbol{w_{c_i}}$ is a vector of criterion weights that were linearly interpolated between the two endpoints (see Figure 2), β is the standardized regression coefficient, λ is the tuning parameter that adjusts for shrinkage, and α is a parameter that controls for the relative weighting of the penalty terms in Equation 6. This process will iterate through all Pareto points (e.g., 21 points)

until we obtain optimal weights for each Pareto point. Here, similar to Step 1, the elastic penalty term $\lambda\left[(1-\alpha)\sum_{j=1}^{k}\beta_j^2 + \alpha\sum_{j=1}^{k}|\beta_j|\right]$ was added to the loss function to allow for regularization.

An R function – *ParetoElnet* – was developed based on ParetoR function in the ParetoR package to implement the above procedures for the Pareto-optimal regularization method. We updated the ParetoR package (Song et al., 2017) to include implementation of the ParetoElnet function. For more information of the updated ParetoR package and the *ParetoElnet* function, see Appendix C.

Evaluation of the Pareto-Optimal Regularization Method

A simulation study was conducted to evaluate the performance of the Pareto-optimal regularization method. Similar to Study 1 and Study 2, the Pareto-optimal regularization method was evaluated using the Bobko-Roth cognitive and non-cognitive predictor correlation matrix as the population matrix (Table 1). Five calibration sample sizes were studied: *n* = 40, 100, 200, 500, and 1000. The validation sample size was 10,000, and each condition was replicated 1000 times. For each sample size condition, diversity shrinkage and job performance criterion validity shrinkage were compared between: (a) De Corte's (2007) Pareto-optimal method (as introduced in De Corte et al., 2007), (b) Pareto-optimal shrinkage formulae (as introduced in Study 1), and (c) regularized Pareto-optimal method (as introduced in the current study). RMSE and bias were calculated to evaluate the accuracy between calibration sample solution and cross-validated solution, at each point along the Pareto curve. The simulation was conducted using the ParetoR R package (Song et al., 2017), the newly developed ParetoElnet function, and the R programming language (R version 3.2.0; R Core Team, 2015).

Results

Comparisons among (a) (unshrunken) Pareto-optimal curve obtained using the original Pareto-optimal curve method (i.e., the algorithm introduced by De Corte et al., 2007); (b) the formula-based (adjusted) Pareto-optimal curve; and (c) the Pareto-optimal curve obtained using elastic net (regularized) Pareto-optimal solutions; are shown in Figures 14a to 14e. Job performance validity shrinkage (i.e., the vertical difference between calibration and validation sample's Pareto-optimal curves) and diversity shrinkage (i.e., the horizontal difference between calibration and validation sample's Pareto-optimal curves) were smaller for the regularized Pareto-optimal solution (difference between solid and dotted lines with red triangles) as compared to the original unshrunken Pareto-optimal results (difference between solid and dotted lines with black circles).

The validation sample Pareto-optimal curve obtained using regularized Pareto-optimal method (i.e., dotted line with red triangles) was almost overlapping with the validation sample Pareto-optimal curve obtained using original Pareto-optimal method (i.e., dotted line with black circles; see Figure 14a). This means that (a) we could decrease job performance validity shrinkage and diversity shrinkage when using the regularized Pareto-optimal method, instead of the original (unshrunken) Pareto-optimal method, and (b) the cross-validated (validation sample; or shrunken) Pareto solutions found using regularized predictor weights are almost identical to those obtained using the original (unshrunken) Pareto-optimal method predictor weights. That is, by using the regularized Pareto-optimal method (in comparison to the original unshrunken Pareto method), we could expect to obtain the same shrunken Pareto curve in a new sample, but with less overestimation of the Pareto-optimal solutions in the calibration sample.

To study shrinkage across the whole Pareto-optimal curve, we also present the RMSE and bias across different Pareto points in Figures 12a to 12d. The RMSE of job performance

validity was similar to the original Pareto-optimal method solutions and formula-based adjusted solutions, with the RMSE for the regularized Pareto-optimal method being the smallest (although by a negligible amount; Figure 12a). As shown in Figure 12b, the RMSE of diversity validity for regularized Pareto-optimal method was: (a) larger than the RMSE for the formula-adjusted Pareto curves (Study 2), but smaller than the RMSE for the original unshrunken Pareto curves, and (b) largest at the Pareto point where diversity was maximized. As shown in Figure 12c, bias of the job performance validity estimation under the regularized Pareto technique was smaller (less positive) than the positive bias of the unshrunken Pareto curve, and less negative than the small negative bias of the formula-adjusted Pareto curve. As such, in some cases the regularized Pareto technique may be a happy medium between the unshrunken and the formula-adjusted Pareto curves. In Figure 12d (bias in the estimation of diversity outcomes), we again see that the regularized Pareto technique is less biased than the unshrunken Pareto technique, but more biased than the formula-adjusted Pareto technique.

Discussion

In the current study, we developed and evaluated a new multi-objective optimization algorithm that provides Pareto-optimal weights with minimum job performance validity shrinkage and diversity shrinkage. The technique allows for practitioners to obtain a set of predictor weights (i.e., Pareto-optimal weights) that could simultaneously optimize two criteria and provide similar selection outcomes (e.g., diversity and expected job performance) across multiple samples. This is especially important in scenarios where calibration sample (e.g., current employee data) size is small and where multiple samples (e.g., across different time points and settings) were selected for one job role specification. Such examples include seasonal internship hiring, where candidates were selected for similar or same internship positions every

year. To allow for broader application of the method, we have updated the ParetoR package (Song, Wee & Newman, 2017) to implement the regularized Pareto-optimal method (see Appendix C) and to implement the shrinkage correction formula (Study 2). Finally, as regularization techniques originate from the literatures in statistical learning and machine learning, the current study is a multi-disciplinary attempt to expand personnel selection methods by bridging between I/O psychology and other disciplines to study diversity issues in hiring. The regularization technique (i.e., an analog to elastic net) originates from the literatures on statistical learning and machine learning (James, Witten, Hastie, & Tibshirani, 2013). Thus, Study 3 is also a multi-disciplinary attempt to expand personnel selection methods by bridging between I/O psychology and other disciplines to study diversity issues in hiring.

When looking at Figures 14a through 14e, we can compare the regularized Pareto curve against the calibration sample Pareto curve and (shrunken) validation sample Pareto curve. First, the regularized Pareto curve is less biased (i.e., smaller overestimation) than the calibration sample Pareto curve, in terms of its ability to estimate the (shrunken) validation sample Pareto curve. So the regularization technique helps to partly remove some of the overly-optimistic results of the calibration sample Pareto solution. Second, when the regularized Pareto curve is cross-validated to a new sample, it is shown that the regularization technique is roughly equivalent to the original De Corte et al. Pareto technique, in terms of the cross-validated (shrunken) Pareto curve that ultimately characterizes the selection practitioner's choice set in a new sample. In other words, regularization prevents some of the over-promising that occurs with the De Corte Pareto curve, but delivers no real difference from the De Corte Pareto curve in job performance or diversity outcomes after cross-validation to a new sample (Figures 14a-14e).

CHAPTER 5

GENERAL DISCUSSION AND CONCLUSION

The current dissertation sought to contribute to the literature on adverse impact, personnel selection, and diversity in hiring by investigating validity shrinkage and diversity shrinkage under the Pareto-optimization method (De Corte et al., 2007) to simultaneously optimize both the job performance and diversity of new hires.

Study 1 examined the extent of shrinkage in Pareto-optimal solutions. We found evidence for the magnitude of job performance validity shrinkage and diversity shrinkage across the entire Pareto-optimal trade-off curve. Specifically, job performance validity shrinkage was greater when the job 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 cognitive and noncognitive 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.

In Study 2, we developed and evaluated a set of shrinkage formulae for the Pareto-optimal weighting technique, similar to the shrinkage formulae used for ordinary least-squares (OLS) regression (e.g., Wherry, 1931; Browne, 1975); that estimate adjusted R^2. An evaluation conducted using Monte-Carlo simulation showed that the Pareto curve shrinkage formulae could

well estimate validation sample Pareto-optimal results (i.e., the cross-validated, or shrunken Pareto curve). The formulae have potentially important implications for practice, as it allows for practitioners to evaluate how well a selection decision based on Pareto-optimal weights on a current employee sample (i.e., calibration sample) could generalize to an applicant pool (i.e., validation sample). It also extends the application of Browne (1975) and Wherry (1931) formulae (the formulae used to correct for shrinkage in OLS regression) beyond OLS regression (where only one criterion is optimized), to conditions where two or more criterion are optimized at the same time. To our knowledge, Study 2 is the first attempt to develop a set of shrinkage formulae to estimate shrunken validity and diversity under multiple-criterion optimization in personnel selection.

In Study 3, we developed and evaluated a new multiple-criterion optimization algorithm that provides Pareto-optimal weights with minimum shrinkage. The technique allows for practitioners to obtain a set of predictor weights (i.e., Pareto-optimal weights) that could simultaneously optimize two criteria, but with Pareto solutions that are less vulnerable to shrinkage (the technique seems especially useful under small sample size conditions; e.g., $N <$ 100). This is especially important in a scenario where the calibration sample size (e.g., current employee data) is small.

We have chosen to study calibration sample sizes of 40, 100, 200, 500, and 1,000; to capture the reasonable range of validation sample sizes used in past studies (see Newman, Jacobs, & Bartrm, 2007, p. 1397, for a review; also see more general review by Shen, Kiger, Davies, Rasch, Simon, & Ones, 2011). However, a committee member had pointed out that in practice, organizations who uses complex selection methods such as the Pareto-optimal methods might tend to be large organizations who regularly work with large sample sizes such as a

calibration sample size of 10,000. Thus, using the same methods as Studies 1 and 2, we further studied the diversity and validity shrinkage and performance of shrinkage formulae when calibration sample size is 10,000. Results of this additional analysis are shown in Appendix E. In general, when calibration sample size is 10,000, there is negligible diversity shrinkage and validity shrinkage. The Pareto-optimal shrinkage formulae slightly underestimates the (shrunken) validation sample Pareto-optimal solution, providing a conservative estimate of the shrunken Pareto-optimal solution.

We also have some limitations. Across Studies 1 through 3, we based simulations primarily on meta-analytic population correlation estimates for one set of widely-used predictors[17]: the Bobko-Roth predictors (i.e., combination of cognitive and noncognitive predictors; biodata, cognitive ability test, conscientiousness test, structured interview, integrity test). The Bobko-Roth predictors were chosen as they are some of the most commonly used predictors in personnel selection (Bobko & Roth, 2013), and we based our simulation studies on the most up-to-date meta-analytic values. Nevertheless we want to note that our current treatment of validity and diversity shrinkage, and the Pareto-optimal regularization technique, are evaluated almost exclusively in the context of the Bobko-Roth predictor combination. As briefly mentioned in the Discussion section of Study 1, characteristics of predictor combinations, such as predictor intercorrelations, could influence shrinkage. Thus, it would be interesting in the future to further investigate the influence of predictor combinations on shrinkage. Such studies could include simulation studies that systematically varies predictor characteristics (e.g., predictor intercorrelation, job performance validity, subgroup d).

[17] Study 1 also investigated a set of cognitive subtest predictors (verbal ability, mathematical ability, technical knowledge, clerical speed; based on ASVAB data).

We investigated job performance validity shrinkage and diversity shrinkage when predictor weights were applied to one validation sample. This is applicable to hiring scenarios when all selections for a certain group job opening were done in one setting and based on one applicant pool. In practice, however, we often conduct selection in multiple settings and based on multiple applicant pools. For example, internship selection usually happens every year, even multiple times in a year. In each rounds of selection, we try to select for the same internship positions, preferably using the same selection system and predictor weights across rounds. Thus, instead of generalizability of predictor weights from calibration sample to one validation sample, we often are also interested in the generalizability to multiple samples. Future study could further investigate the job performance validity shrinkage and diversity shrinkage when applied to multiple samples with sampling variations.

Related question had also raised the attention of scholars in bioinformatics, statistics and computer science (e.g., Bernau et al., 2014; Molinaro et al., 2005). Recently, Bernau et al. (2014) had proposed a "leave-one-dateset-in" *cross-study validation* method to take into account sampling variation, by using multiple independent datasets, when evaluating generalizability of algorithms. Specifically, each data set is used turn as training data (calibration), and all other data as testing data (validation sample). The cross-validation outputs are then analyzed for evidence of study outliers and for estimation of cross-study validation accuracy. Further, Riester et al. (2014) proposed "leave-one-dataset-out" cross-study validation where one sample was used as testing data and rest of the samples used as training sample. Bernau et al. (2014) found that summary statistics of cross validation (CV) and cross-study validation (CSV) are not highly correlated, which indicates that cross-study prediction (evaluated using CSV) and within-study prediction (evaluated using CV) are "two related but distinct problems" (Bernau et al., 2014; p.

i111). Cross-study validation could help provide "more realistic expectations of future prediction model" (Bernau et al., 2014; p. i112) by taking into account sampling variation in assessment of generalizability.

In the regularized Pareto-optimal method that we developed in the current study, we used the cross-validation method to evaluate how the regularized Pareto-optimal solution obtained from one sample from the population (i.e., calibration sample) could generalize to one other sample from the population (i.e., validation sample). In future studies, it would be interesting to implement the cross-study validation method to evaluate generalizability to the regularized Pareto-optimal method. Further, we could also consider using the cross-study validation method to find optimal alpha and lambda parameters for the regularized Pareto-optimal method. Given results from previous studies (e.g., Bernau et al., 2014; Riester et al., 2014), we expect that the updated algorithm using cross-study validation will have superior multi-sample generalizability when applied to multiple samples with sampling variation.

Another perspective for thinking about sampling variation is the comparison between "local" and "meta-analytic" validity data (Newman et al., 2007). In practice, we often base our calibration on the employee data from the current organization (i.e., local data). Sometimes, we might also be able to get access to related data collected in other organizations or samples (i.e., meta-analytic data). Local data has the advantage that it's most related to the organization, but it is influenced by sampling variation. On the other hand, meta-analytic data is less influenced by sampling variation (due to large sample size), but also less related to the local organization. Future studies could investigate combination of local and meta-analytic data based on Bayesian logic (see Brannick, 2001; Newman, Jacobs and Bartram, 2007). The research could be beneficial for organizations to find predictor weights that could be more generalizable across

samples compared to weights obtained using local data alone. In a more extreme practice scenario, certain firms might not have the ability to collect calibration samples (e.g., when designing a selection system for a brand new job position that did not exist in the organization before). In such case, organizations might choose to refer to meta-analytic data or similar data from other organizations to develop Pareto-optimal weights.

Last, but not least, we want to note that the implementation of the Pareto-optimal method (multiple objective optimization technique) is not limited to optimizing job performance and diversity in hiring scenarios. Similar to how it was used in the field of engineering to optimize energy use and performance (e.g., Shirazi et al., 2014), Pareto-optimal methods could be applied to other research topics. The optimization techniques could especially be applicable in scenarios when there is a tradeoff (i.e., when it is difficult to maximize or minimize two or more objectives).

In sum, the current dissertation (a) showed that diversity shrinkage exists (and should not be ignored) when applying Pareto-optimal weighting techniques to enhance diversity in hiring; (b) demonstrated that the extent of the shrinkage depends on calibration sample size such that smaller calibration sample size yields larger shrinkage; (c) developed shrinkage formulae to correct for shrinkage in Pareto-optimal weighting; and (d) introduced a regularized Pareto-optimal weighting algorithm that accounts for shrinkage when applying Pareto-optimal solutions to a new sample. This work has implications for research and practice involving organizational diversity in hiring, and expands methodology to optimize on multiple criteria (rather than single criterion) simultaneously.

TABLES

Table 1

Bobko-Roth Predictor Inter-correlations, Black-White Subgroup ds*, and Criterion Validities*

Variable	1	2	3	4	5
Roth et al., 2011 predictors					
1. Biodata	1.00				
2. Cognitive Ability	.37	1.00			
3. Conscientiousness	.51	.03	1.00		
4. Structured Interview	.16	.31	.13	1.00	
5. Integrity Test	.25	.02	.34	-.02	1.00
Criterion Validities	.32	.52	.22	.48	.20
Black-White Subgroup *d*	.39	.72	-.09	.39	.04

Note. Meta-analytic predictor inter-correlation matrix for all five predictors was obtained from Roth, Switzer, Van Iddekinge and Oh (2011; pp. 918-919, Table 6). Criterion validities for biodata, cognitive ability, Conscientiousness, and structured interview were obtained from Roth et al. (2011; p. 915, Table 5). We used an updated estimate of integrity test validity provided by Van Iddekinge, Roth, Raymark and Odle-Dusseau (2012; p. 511, Table 2; ρ = .18), which we then corrected for direct range restriction using u = .90 (Van Iddekinge et al., 2012, p. 516). In all, the validity and predictor inter-correlation input values were based on job applicant information and were corrected for range restriction and criterion unreliability only. Subgroup *d* values for predictors were obtained from Bobko and Roth (2013; pp. 94-96, Tables 1 and 2). Job performance Black-White subgroup *d* was .38 (McKay & McDaniel, 2006; p. 544, Table 2).

Table 2

Cognitive Subtests Predictor Inter-correlations, 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 inter-correlation matrix, criterion validities and subgroup *d*s 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).

Table 3

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

	Endpoint where criterion validity is maximized							
	Criterion Validity (*R*)				Diversity (AI ratio)			
Calibration Sample Size	Unit-Weighted Validity	Calibration Sample Validity	Cross-Validated Validity	Validity Shrinkage (Δ*R*)	Unit-Weighted AI ratio	Calibration Sample AI ratio	Cross-Validated AI ratio	Diversity Shrinkage (ΔAI ratio)
N = 40	.58	.69	.63	.06	0.63	0.48	0.48	0.00
N = 100	.58	.67	.65	.03	0.63	0.47	0.47	0.00
N = 200	.58	.67	.65	.02	0.63	0.47	0.47	0.00
N = 500	.58	.66	.66	.01	0.63	0.47	0.46	0.00
N = 1000	.58	.66	.66	.00	0.63	0.46	0.46	0.00
	Endpoint where diversity is maximized							
	Criterion Validity (*R*)				Diversity (AI ratio)			
Calibration Sample Size	Unit-Weighted Validity	Calibration Sample Validity	Cross-Validated Validity	Validity Shrinkage (Δ*R*)	Unit-Weighted AI ratio	Calibration Sample AI ratio	Cross-Validated AI ratio	Diversity Shrinkage (ΔAI ratio)
N = 40	.58	.25	.27	-.02	0.63	2.15	0.91	1.24
N = 100	.58	.22	.23	-.01	0.63	1.46	1.03	0.43
N = 200	.58	.22	.21	.00	0.63	1.25	1.07	0.18
N = 500	.58	.21	.21	.00	0.63	1.20	1.12	0.08
N = 1000	.58	.21	.21	.00	0.63	1.17	1.14	0.03

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.

Table 4

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

	Endpoint where criterion validity is maximized							
	Criterion Validity (*R*)				Diversity (AI ratio)			
Calibration Sample Size	Unit-Weighted Validity	Calibration Sample Validity	Cross-Validated Validity	Validity Shrinkage (Δ*R*)	Unit-Weighted AI ratio	Calibration Sample AI ratio	Cross-Validated AI ratio	Diversity Shrinkage (ΔAI ratio)
N = 40	.60/.52	.64/.57	.59/.53	.04/.04	0.35/0.35	0.32/0.46	0.32/0.46	0.01/0.00
N = 100	.60/.52	.62/.57	.61/.55	.02/.02	0.35/0.35	0.31/0.48	0.31/0.48	0.00/0.00
N = 200	.60/.52	.62/.56	.61/.55	.00/.01	0.35/0.35	0.30/0.48	0.30/0.48	0.00/0.00
N = 500	.60/.52	.62/.56	.61/.55	.00/.00	0.35/0.35	0.30/0.48	0.30/0.48	0.00/0.00
N = 1000	.60/.52	.62/.56	.62/.55	.00/.00	0.35/0.35	0.30/0.48	0.30/0.48	0.00/0.00
	Endpoint where diversity is maximized							
	Criterion Validity (*R*)				Diversity (AI ratio)			
Calibration Sample Size	Unit-Weighted Validity	Calibration Sample Validity	Cross-Validated Validity	Validity Shrinkage (Δ*R*)	Unit-Weighted AI ratio	Calibration Sample AI ratio	Cross-Validated AI ratio	Diversity Shrinkage (ΔAI ratio)
N = 40	.60/.52	.37/.42	.38/.43	-.01/-.01	0.35/0.35	0.89/0.89	0.83/0.86	0.05/0.03
N = 100	.60/.52	.36/.43	.36/.43	.00/.00	0.35.0.35	0.92/0.92	0.91/0.92	0.01/0.00
N = 200	.60/.52	.35/.43	.36/.43	.00/.00	0.35/0.35	0.91/0.92	0.92/0.92	0.00/0.00
N = 500	.60/.52	.36/.43	.36/.43	.00/.00	0.35/0.35	0.92/0.92	0.92/0.92	0.01/0.00
N = 1000	.60/.52	.36/.43	.36/.43	.00/.00	0.35/0.35	0.91/0.92	0.92/0.92	0.00/0.00

Note. Validity shrinkage (Δ*R*) refers to the mean criterion validity difference between calibration sample and validation sample values, and indicates job performance shrinkage. Diversity shrinkage (ΔAI ratio) refers to the mean diversity (i.e., AI ratio) difference between calibration sample and validation sample values. In each cell, results for the Machinery Repair job family are shown on the left side of "/" whereas results for the Control & Communication job family are shown on the right side of "/".

Table 5

Summary of Basic Findings

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 4)
    b. In the current simulation, diversity shrinkage was typically negligible for cognitive subtest predictors, whenever sample size was at or above 100. (Figures 5 and 6)

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^2increases (this can be clearly seen in classic shrinkage formulae).
    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 $N \geq 100$), and
    b. Pareto weights can typically yield greater diversity outcomes compared to unit weights, at no cost in terms of job performance. (Figures 4, 5, and 6)
    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 4 and 8)

Table 6

Results of Simulation Study to Evaluate Pareto-Optimal Shrinkage Formula (Calibration Sample Size = 40)

Pareto Point	$R_{perf_{Cal}}$	$R_{perf_{Val}}$	$R_{perf_{Adj}}$	$R_{race_{Cal}}$	$R_{race_{Val}}$	$R_{race_{Adj}}$	$AIratio_{race_{Cal}}$	$AIratio_{race_{Val}}$	$AIratio_{race_{Adj}}$
1	**.28**	.28	.28	**-.11**	.09	**.00**	2.06	0.74	1.00
2	.30	.30	.30	-.10	.10	.01	1.92	0.72	0.97
3	.33	.32	.32	-.09	.11	.02	1.76	0.70	0.94
4	.35	.34	.33	-.08	.12	.02	1.63	0.68	0.92
5	.37	.36	.35	-.07	.13	.03	1.50	0.66	0.89
6	.40	.38	.37	-.06	.14	.04	1.37	0.64	0.87
7	.42	.40	.39	-.05	.15	.04	1.29	0.62	0.85
8	.45	.42	.41	-.04	.16	.05	1.20	0.60	0.82
9	.47	.44	.43	-.03	.17	.06	1.13	0.58	0.80
10	.49	.46	.44	-.01	.18	.08	1.05	0.56	0.77
11	**.51**	**.48**	**.46**	**.01**	**.19**	**.09**	0.97	0.54	**0.73**
12	.53	.49	.47	.02	.20	.10	0.91	0.52	0.72
13	.55	.51	.49	.04	.22	.11	0.85	0.50	0.69
14	.57	.53	.51	.06	.23	.13	0.79	0.48	0.66
15	.59	.54	.52	.09	.24	.15	0.74	0.46	0.61
16	.61	.56	.54	.12	.25	.17	0.68	0.44	0.57
17	.63	.57	.55	.15	.27	.20	0.62	0.42	0.53
18	.65	.59	.57	.18	.28	.22	0.57	0.39	0.50
19	.67	.60	.58	.21	.30	.24	0.50	0.37	0.46
20	.68	.61	.59	.26	.31	.28	0.43	0.34	0.40
21	**.69**	.62	**.60**	**.33**	.33	.33	0.31	0.31	0.31

FIGURES

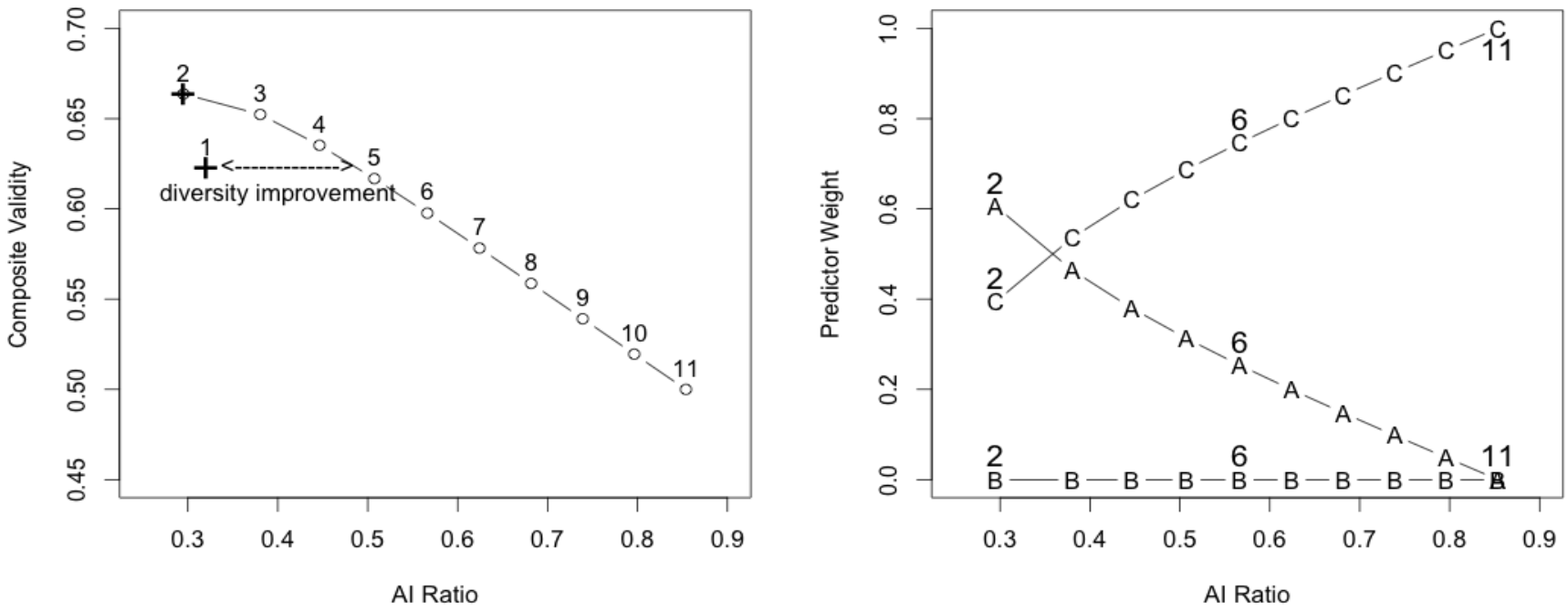


Figure 1. Left panel: Point 1 is the unit weight solution, Point 2 is the regression weight solution (maximizing validity/job performance), and Points 2 through 11 correspond to 10 Pareto weight solutions (i.e., the Pareto-optimal trade-off surface). The horizontal dashed line indicates the *diversity improvement* obtained by using Pareto-optimal weights rather than unit weights. The example scenario uses 3 predictors (A, B, and C) and an overall measure of job performance (Y). *d* values are .9, .7, and .1, respectively; test validities are .6, .5, and .5, respectively, and inter-correlations are r_{AB} = .8, r_{BC} = .6, and r_{AC} = .4.
Right panel: Predictor weighting scheme that corresponds to the Pareto-optimal trade-off surface. Lines marked 'A---A', 'B---B', and 'C---C' correspond to weights for Predictors A, B, and C, respectively. Reprinted from "More than g: Selection quality and adverse impact implications of considering second-stratum cognitive abilities," by S. Wee et al., 2014, Journal of Applied Psychology, 99, p. 552. Copyright 2013 by American Psychological Association.

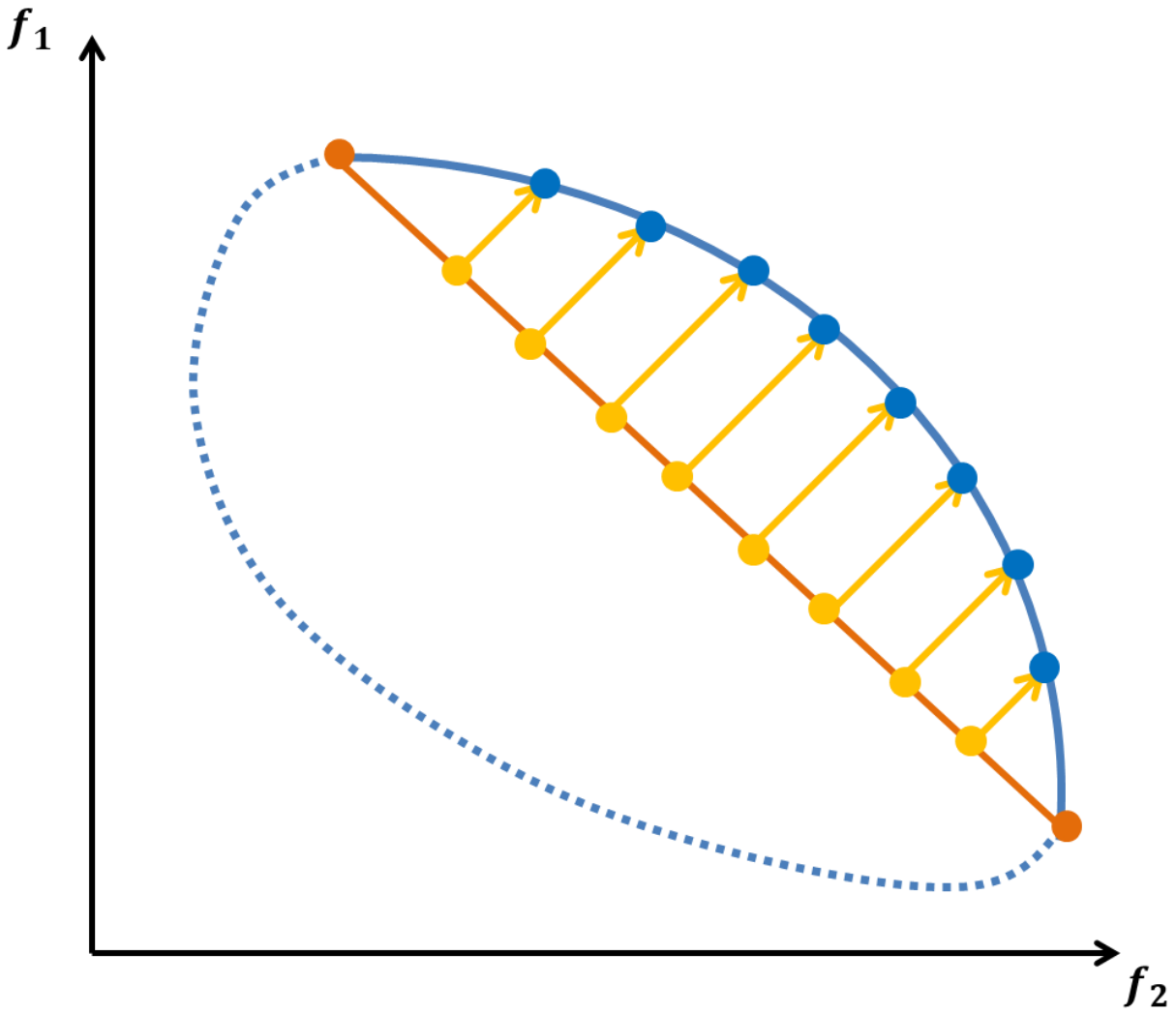


Figure 2. Illustration of the Normal Boundary Intersection surface. f_1 and f_2 represent two criteria that we want to optimize (e.g., job performance and diversity). Blue oval (which includes both dotted and solid border) represents all possible solution space. Orange points represent two extreme endpoints. Orange point on the left is solution where f_1 is maximized, and orange point on the right is solution where f_2 is maximized. Solid blue curve represent Pareto-optimal solutions, where both f_1 and f_2 are optimized. Orange and blue points together represent example Pareto solutions. Normal Boundary Intersection (NBI) algorithm first finds two endpoints (i.e., two orange points) and form a straight line between the two endpoints (i.e., orange line segment). Then, the algorithm finds evenly spaced points (i.e., yellow points) along the orange line. The number of evenly spaced points depends on user-specified number of Pareto-points. In the example here, we have specified 10 Pareto-points, and thus we have two orange endpoints and eight yellow points on the orange line, between the two endpoints. Following, the NBI algorithm projects solutions in each yellow point along perpendicular direction as the orange line to find the corresponding solution (i.e., blue point) on the border of all possible solution space (i.e., solid blue line). The resulting solutions (i.e., blue points), along with solutions at endpoint (i.e., two orange points) are the Pareto-optimal solutions.

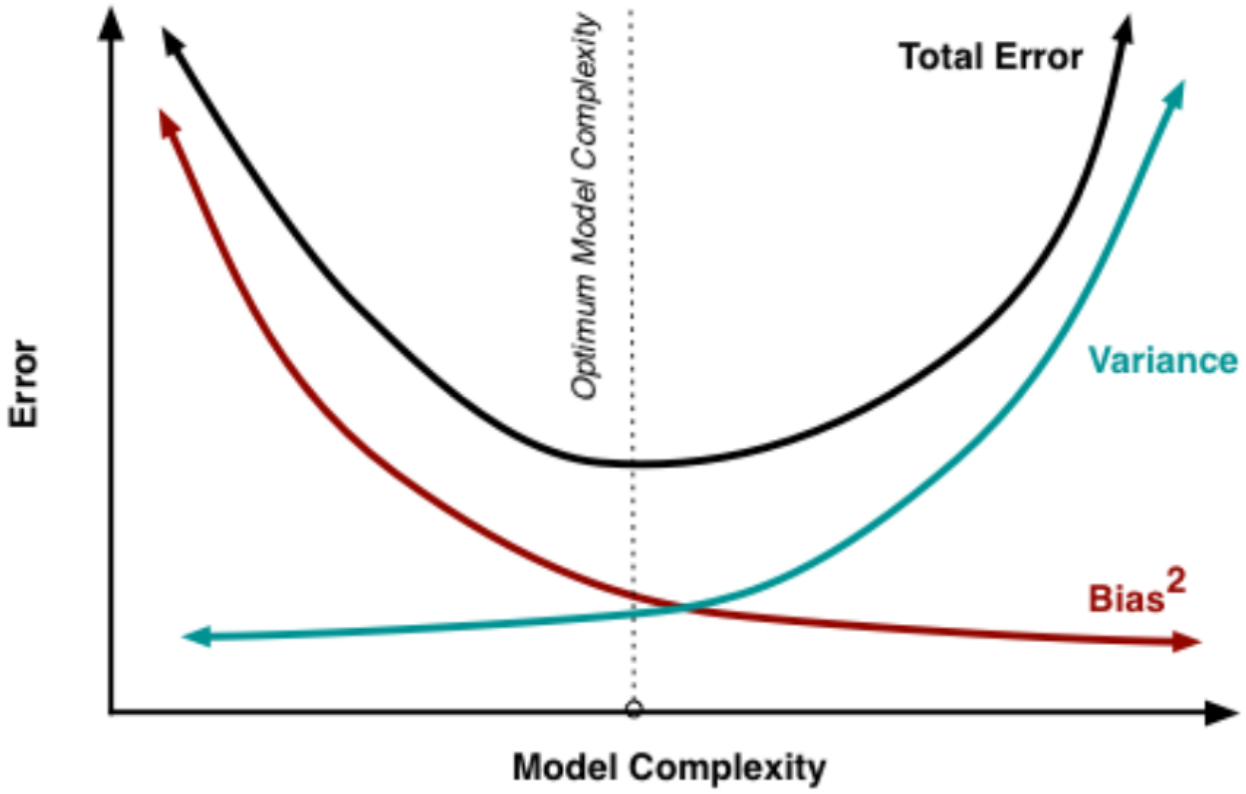


Figure 3. Relationship between model complexity and bias-variance trade-off. Adapted from Fortmann-Roe, S. (2012). Understanding the bias-variance tradeoff. http://www.citeulike.org/group/15400/article/13849984

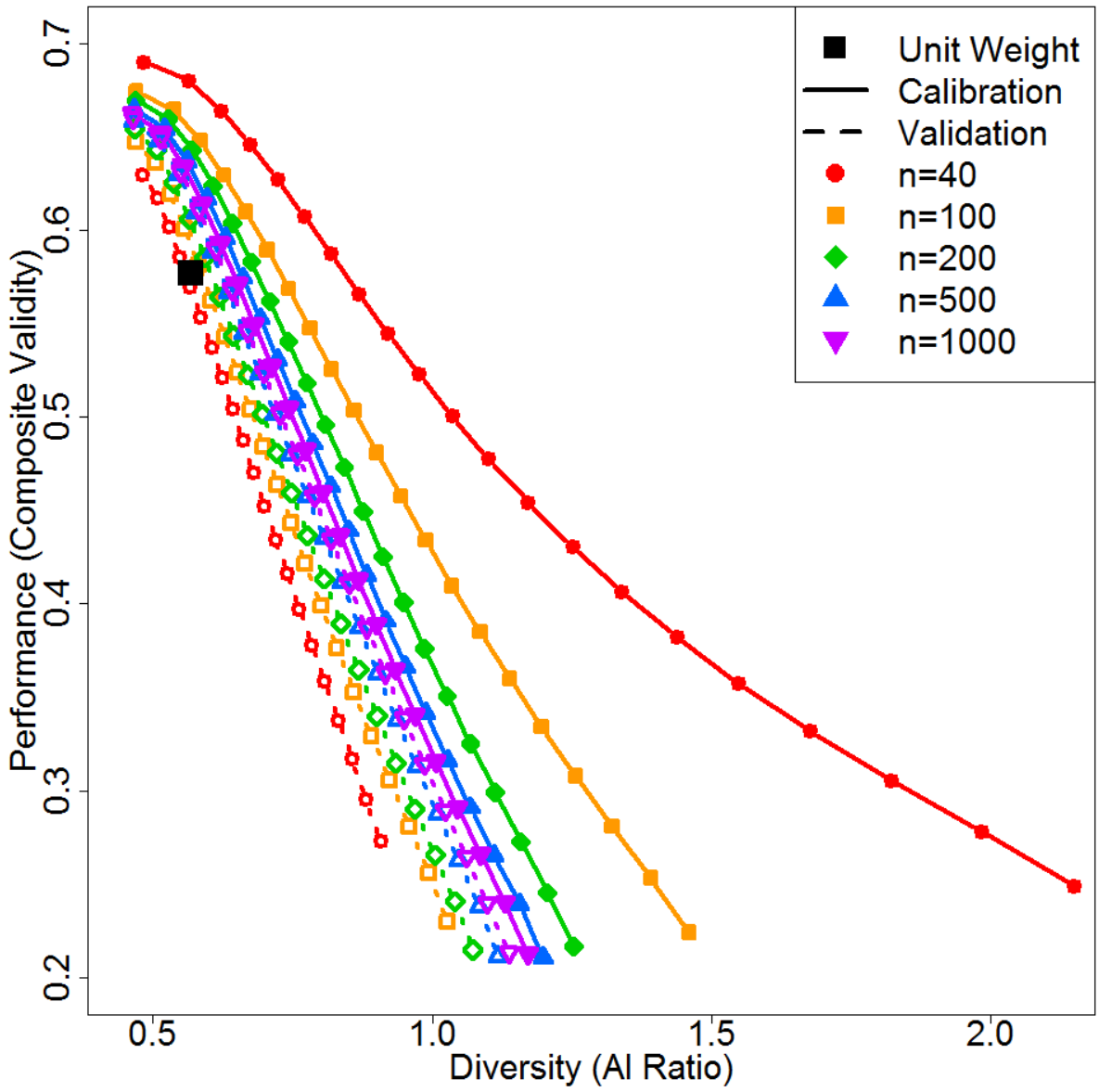


Figure 4. Validity and diversity shrinkage (original/calibration and shrunken Pareto curves) for Bobko-Roth predictors using meta-analytically derived values presented in Table 1.

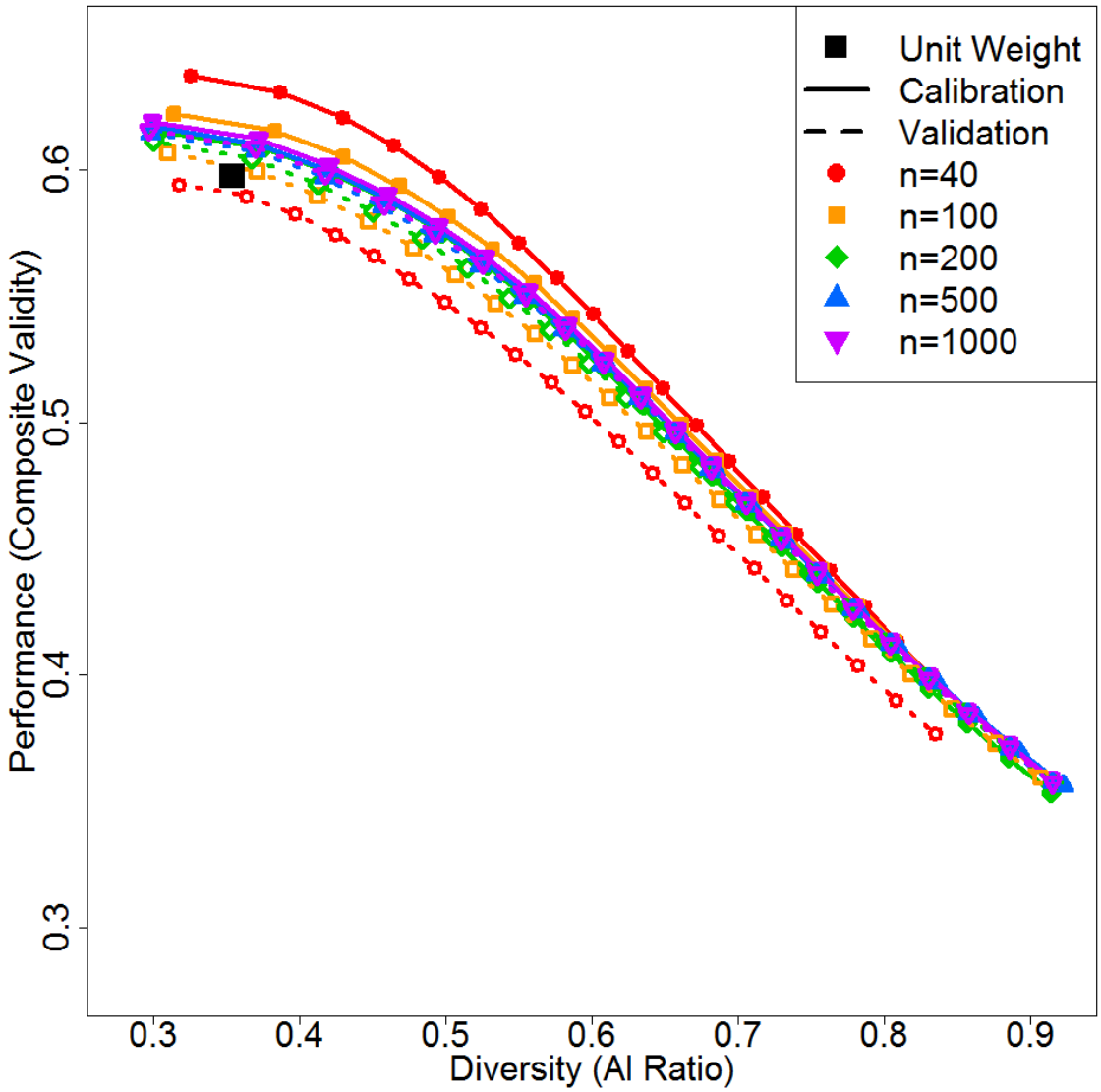


Figure 5. Validity and diversity shrinkage (original/calibration and shrunken Pareto curves) for Machinery Repair job family using cognitive subtest predictors (Wee et al., 2014).

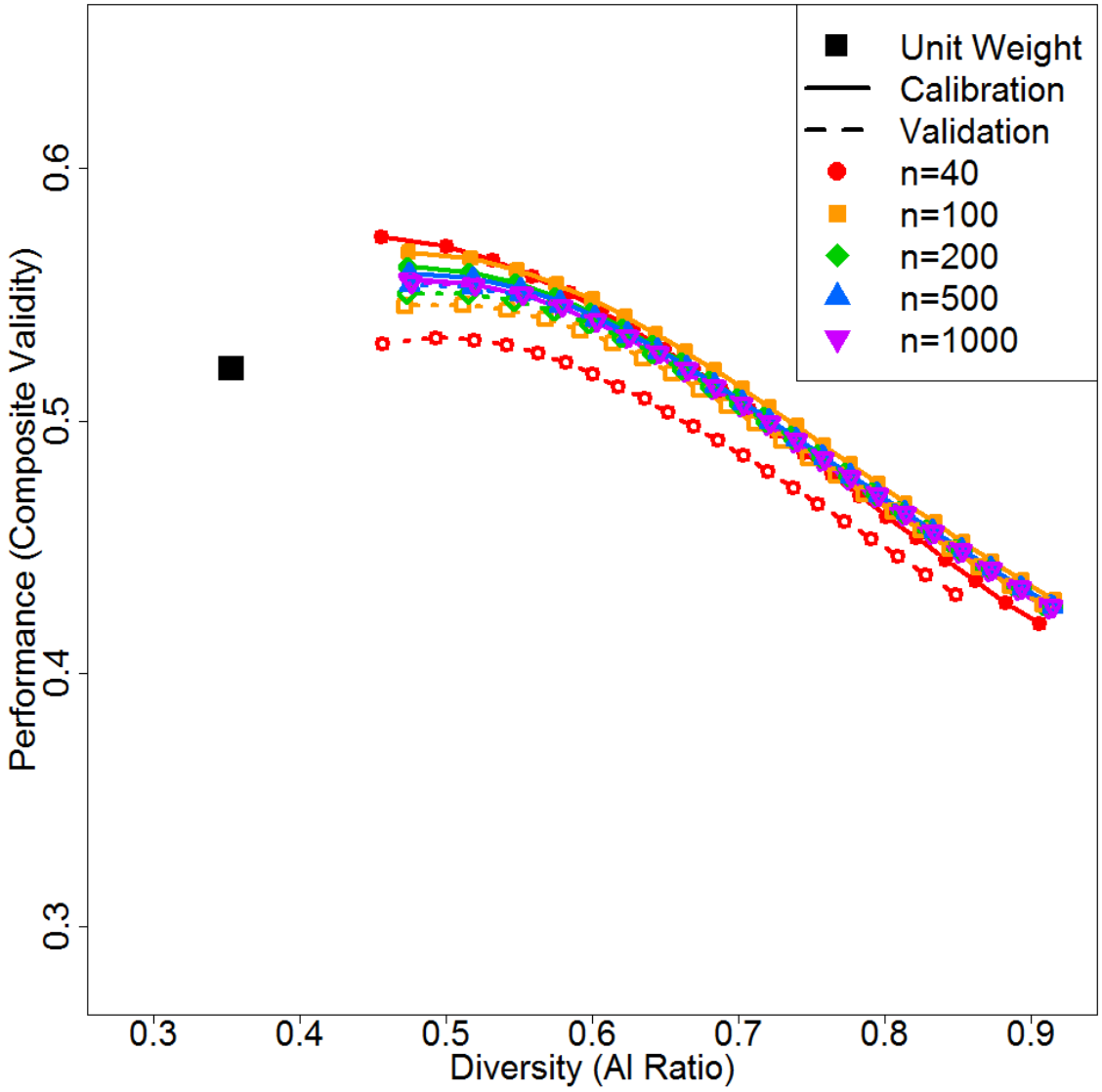


Figure 6. Validity and diversity shrinkage (original/calibration and shrunken Pareto curves) for Control and Communication job family using cognitive subtest predictors (Wee et al., 2014).

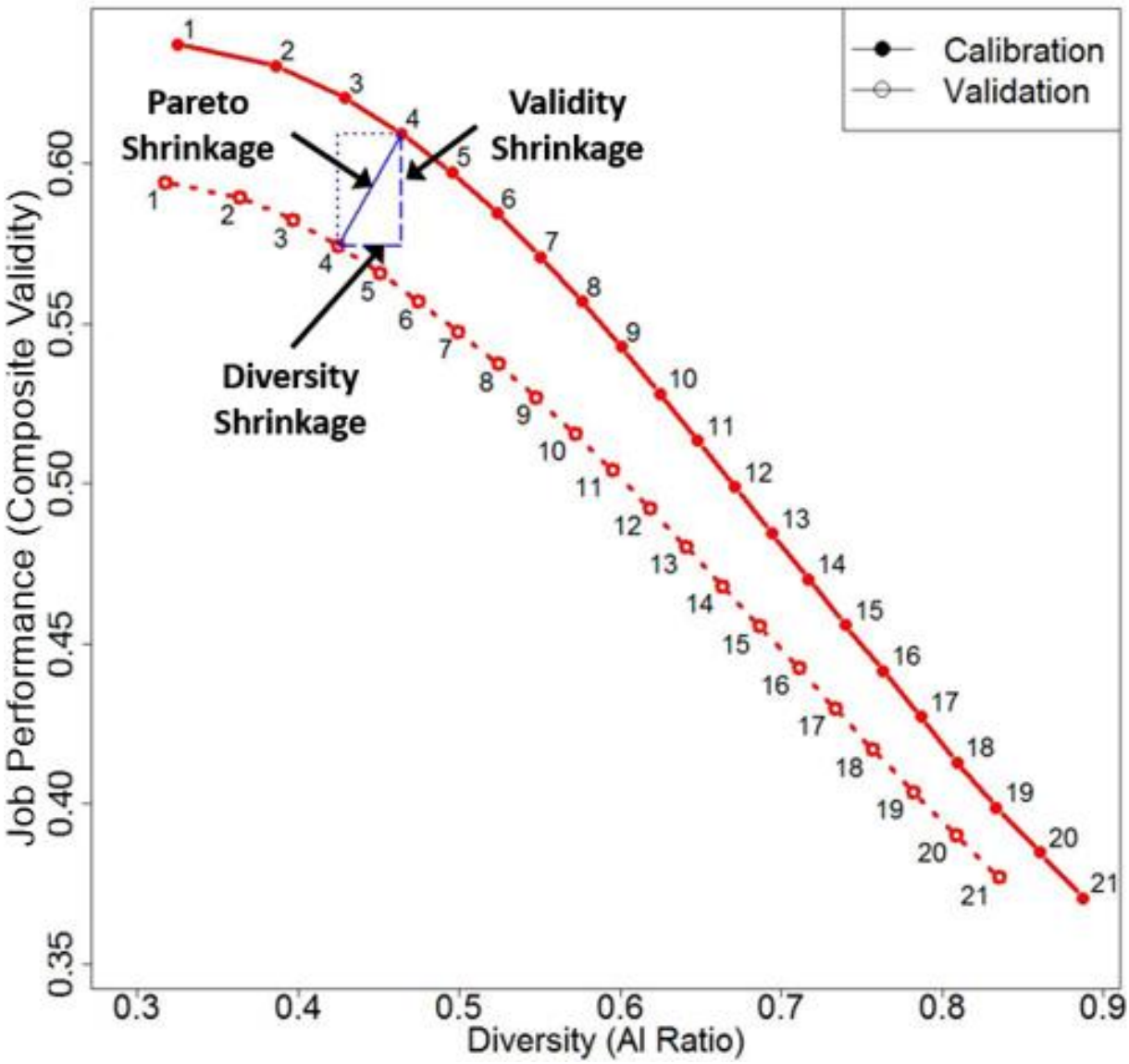


Figure 7. Example scenario demonstrating decomposition of Pareto shrinkage into diversity shrinkage and validity shrinkage. Pareto shrinkage describes the shift of the calibration sample Pareto curve (i.e., each of the 21 Pareto solutions on the solid line) to a lower expected value for both criterion validity and diversity (i.e., each of the 21 Pareto solutions on the dotted line; obtained in the validation sample). Pareto shrinkage can be further decomposed into independent components of validity shrinkage (along the vertical axis) and diversity shrinkage (along the horizontal axis).

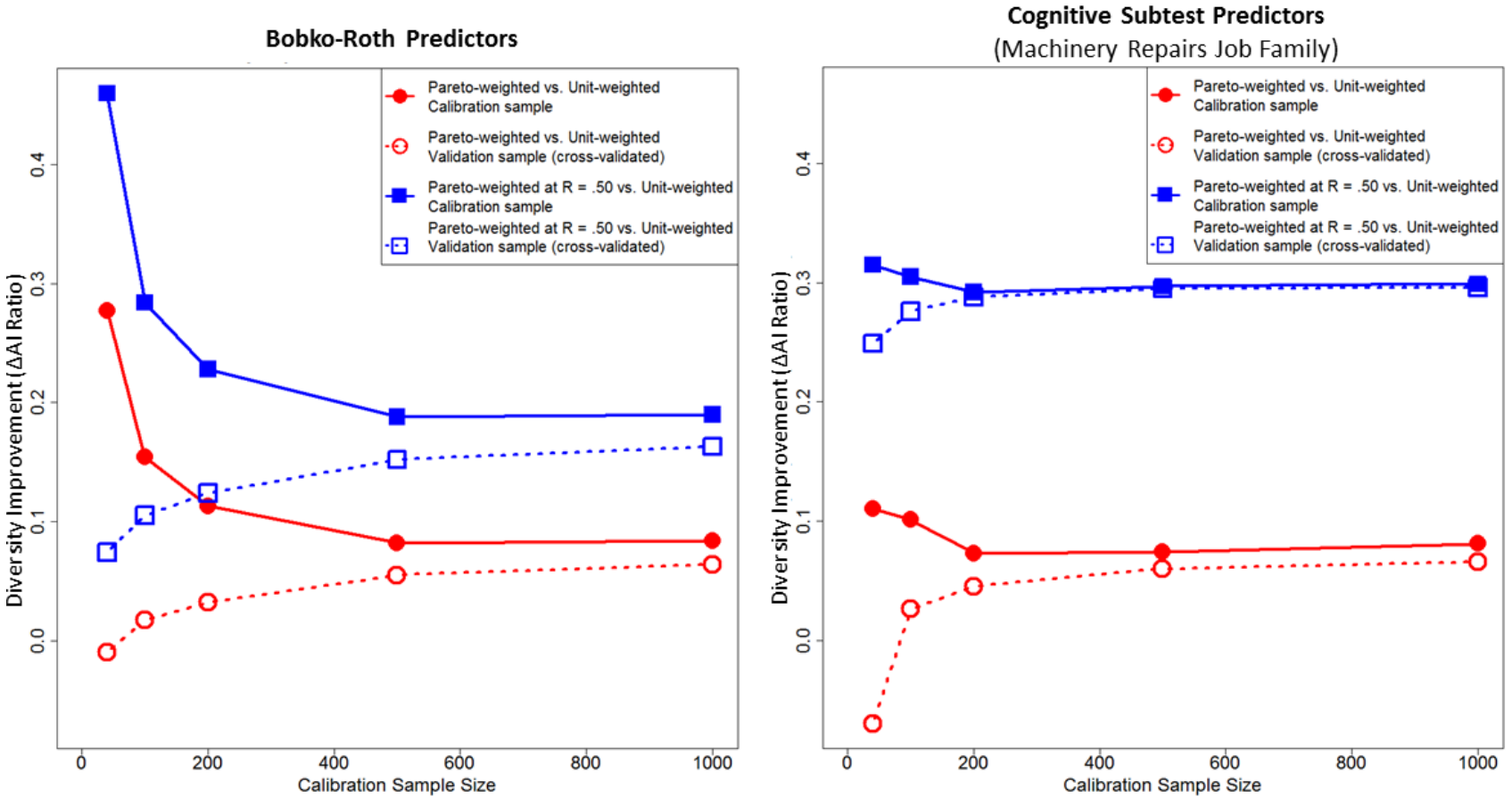


Figure 8. Diversity improvement (ΔAI ratio) by calibration sample size, for the Bobko-Roth predictors (left panel) and the cognitive subtest predictors in the Machinery Repair job family (right panel). Red lines indicate diversity improvement calculated as the difference between AI ratio of Pareto Weights at Unit-Weighted *R* vs. 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* = .50 vs. 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* = .50).

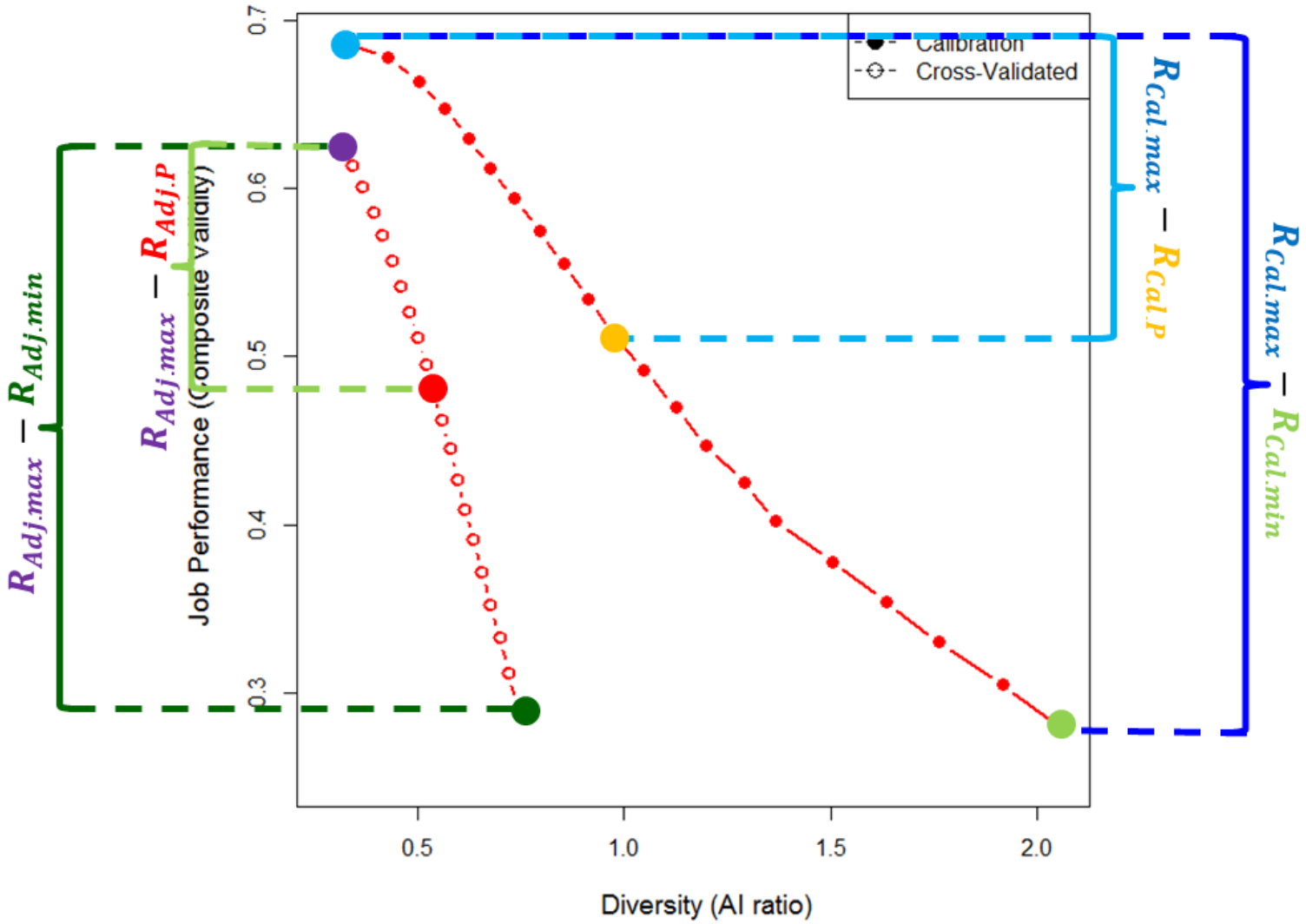


Figure 9. Illustration of the relationship between shrunken and unshrunken ratio in Pareto-optimal shrinkage formula

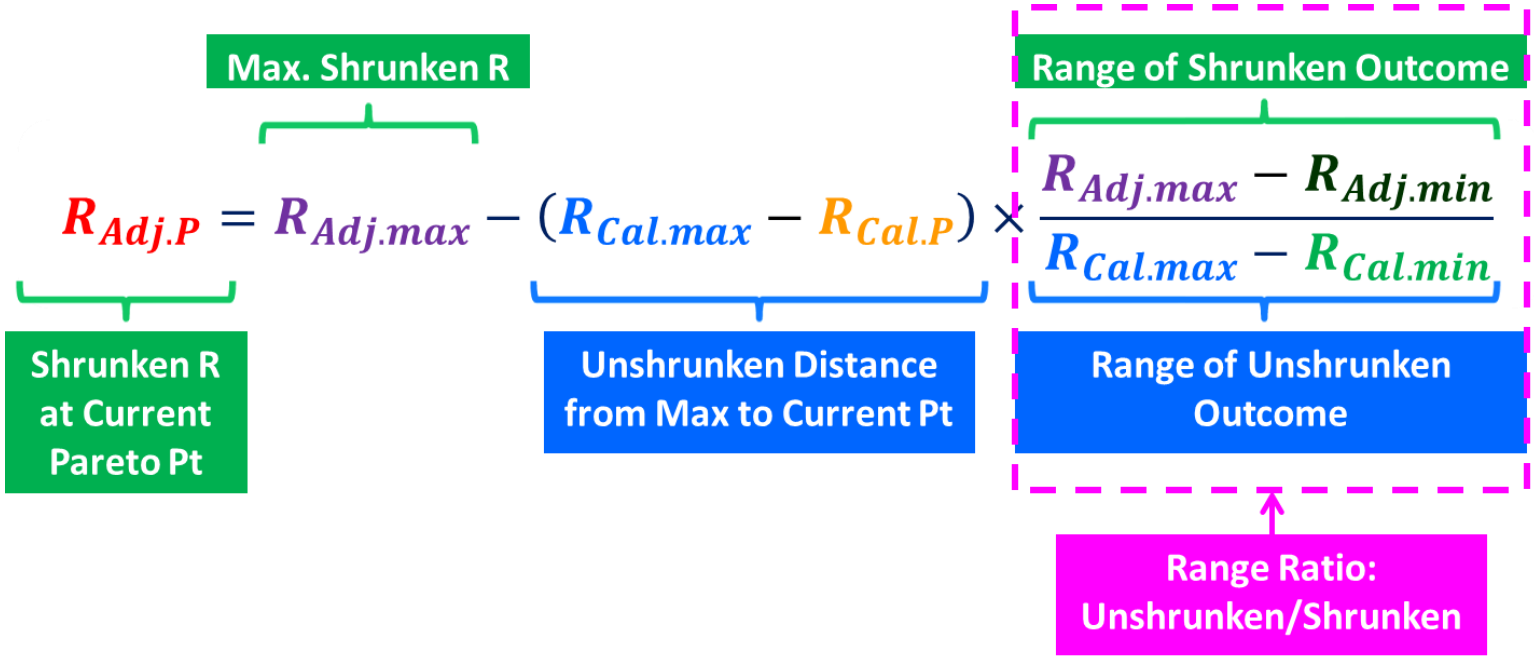


Figure 10. The relationship between shrunken and unshrunken ratio in Pareto-optimal shrinkage formula

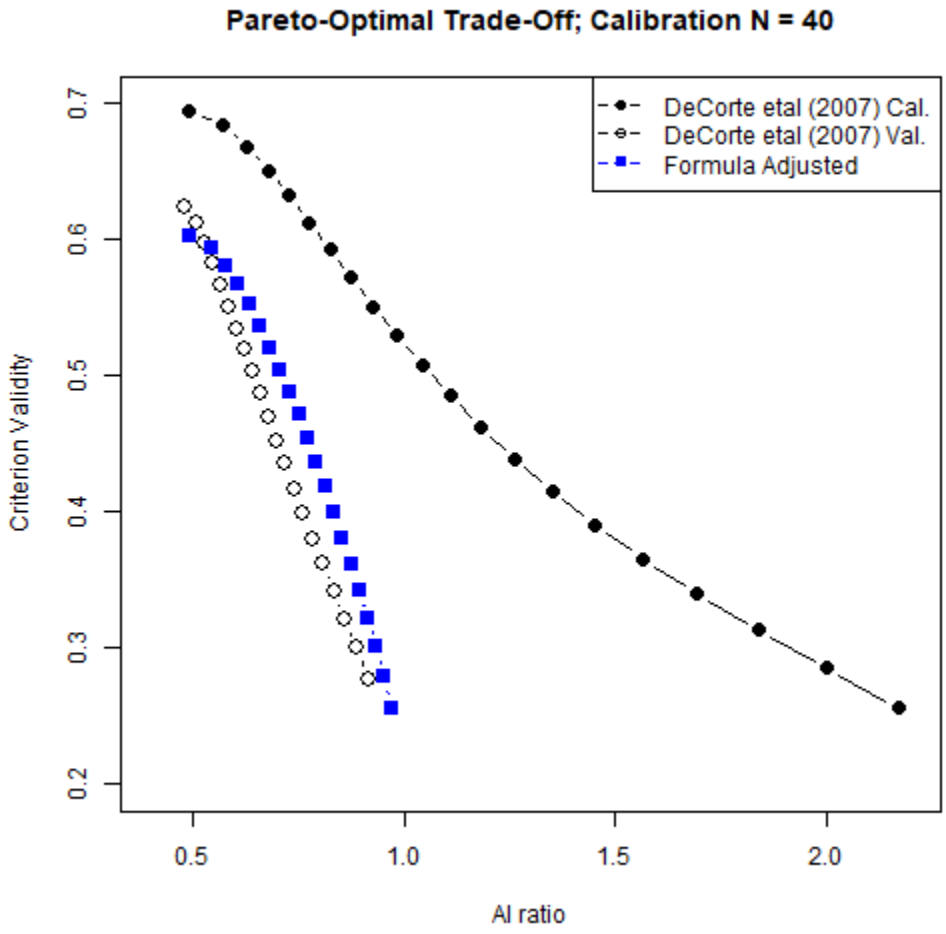


Figure 11a. Comparison among Pareto-optimal trade-off curves calibration sample, cross-validated and shrinkage adjusted criterion validity when calibration sample size is 40.

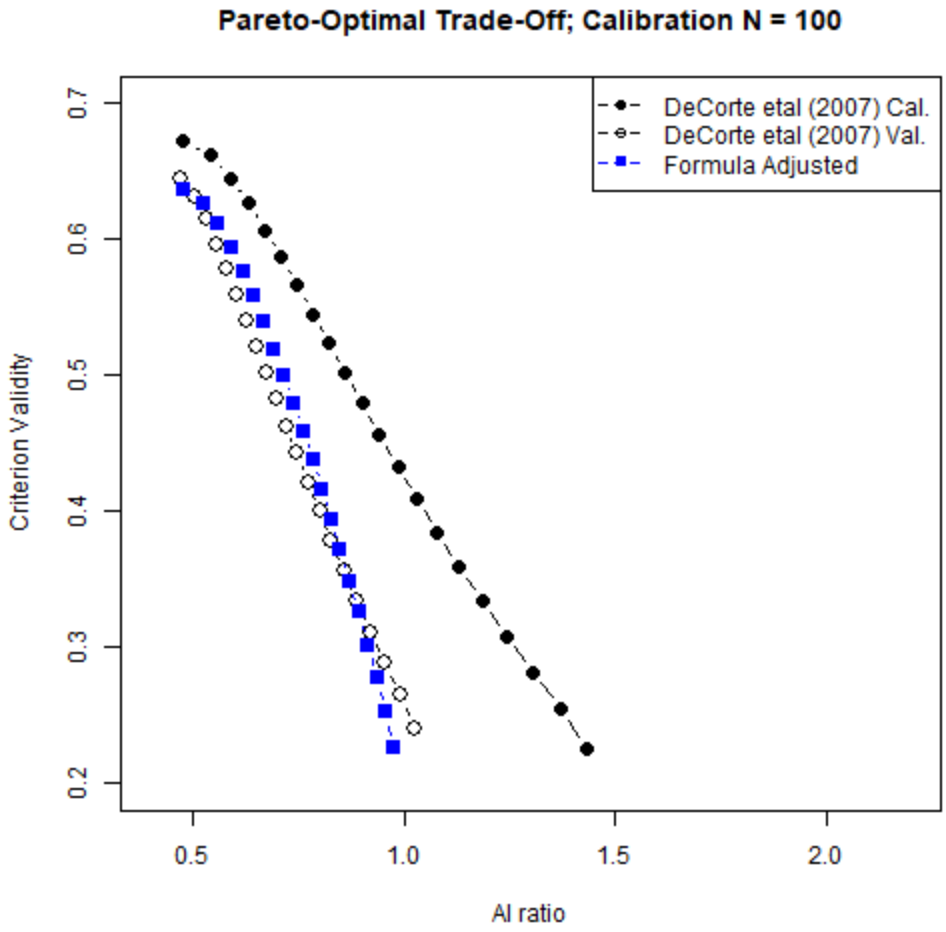


Figure 11b. Comparison among Pareto-optimal trade-off curves calibration sample, cross-validated and shrinkage adjusted criterion validity when calibration sample size is 100.

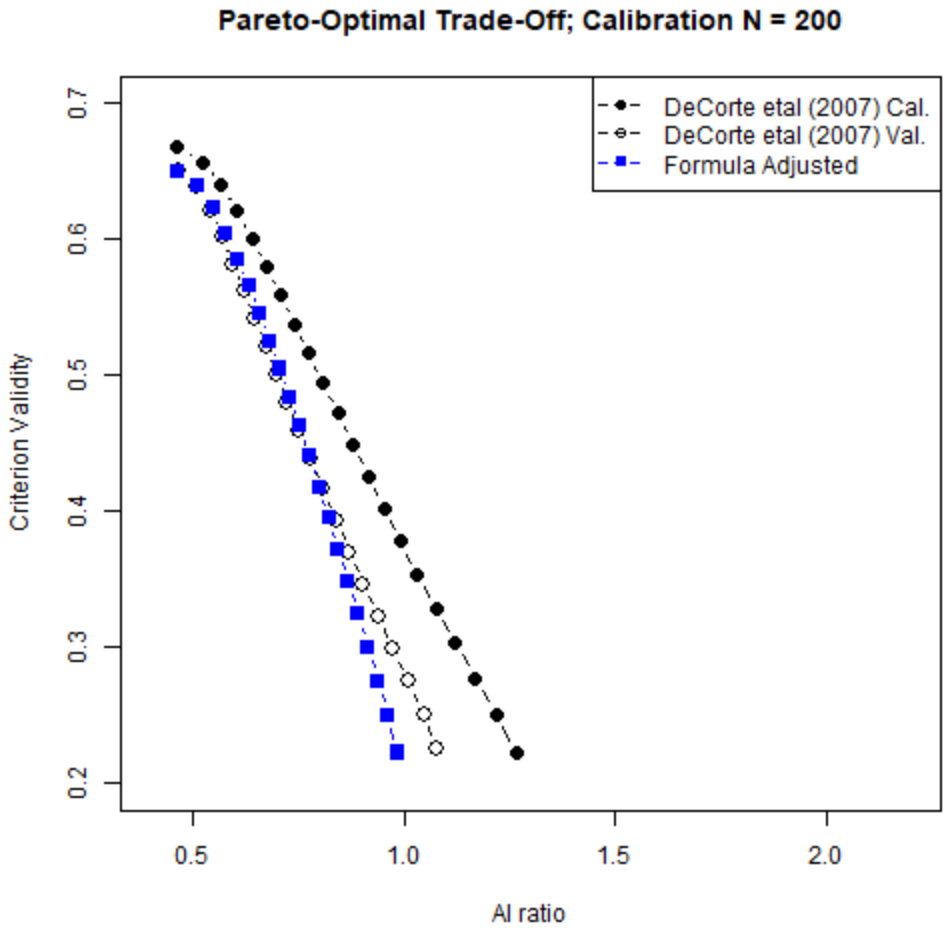


Figure 11c. Comparison among Pareto-optimal trade-off curves calibration sample, cross-validated and shrinkage adjusted criterion validity when calibration sample size is 200.

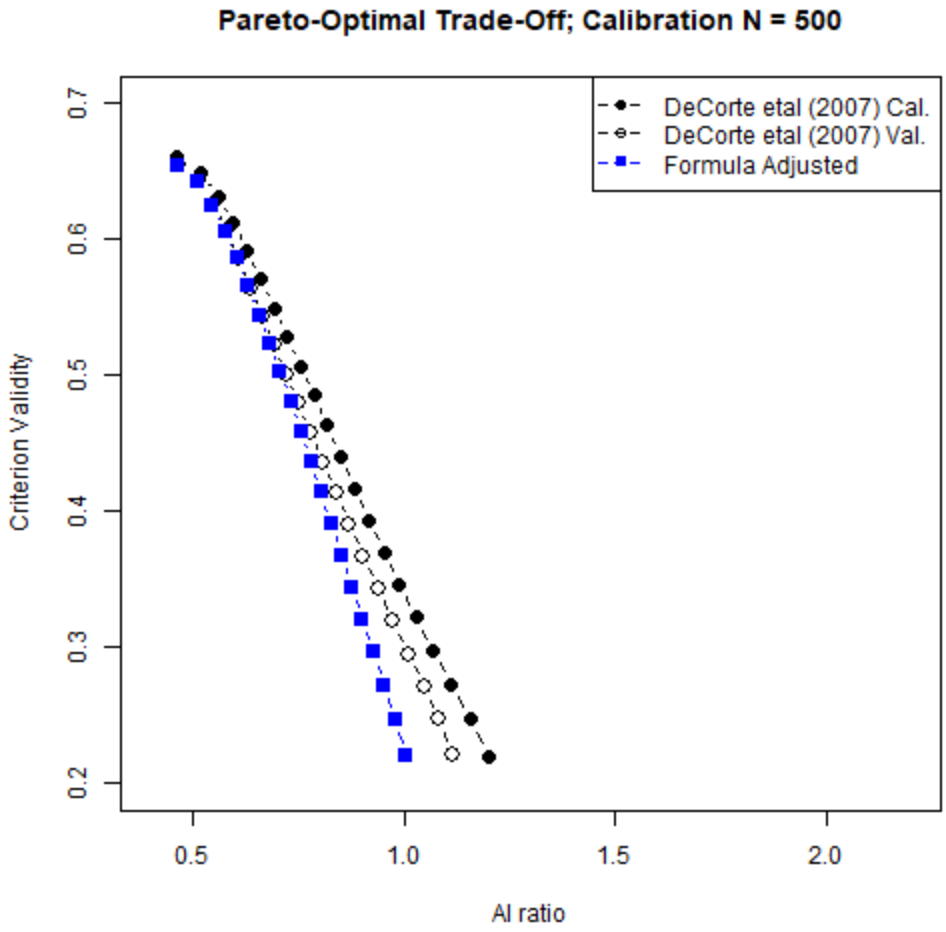


Figure 11d. Comparison among Pareto-optimal trade-off curves calibration sample, cross-validated and shrinkage adjusted criterion validity when calibration sample size is 500.

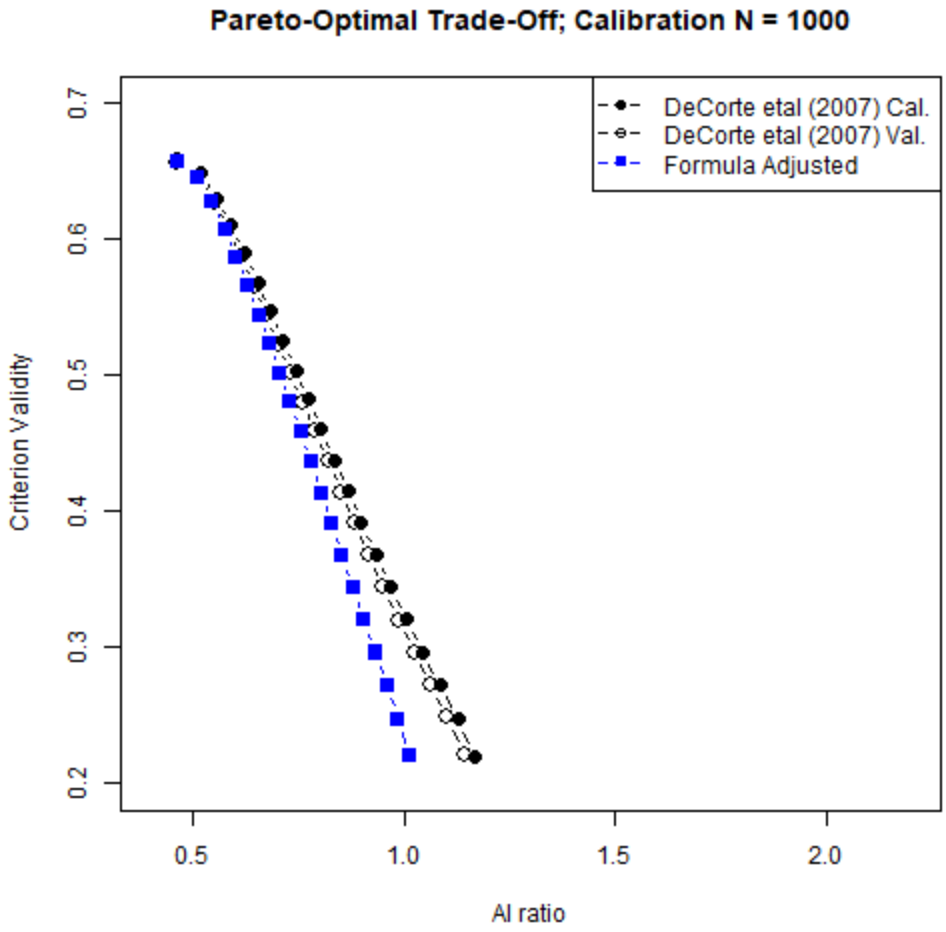


Figure 11e. Comparison among Pareto-optimal trade-off curves calibration sample, cross-validated and shrinkage adjusted criterion validity when calibration sample size is 1000.

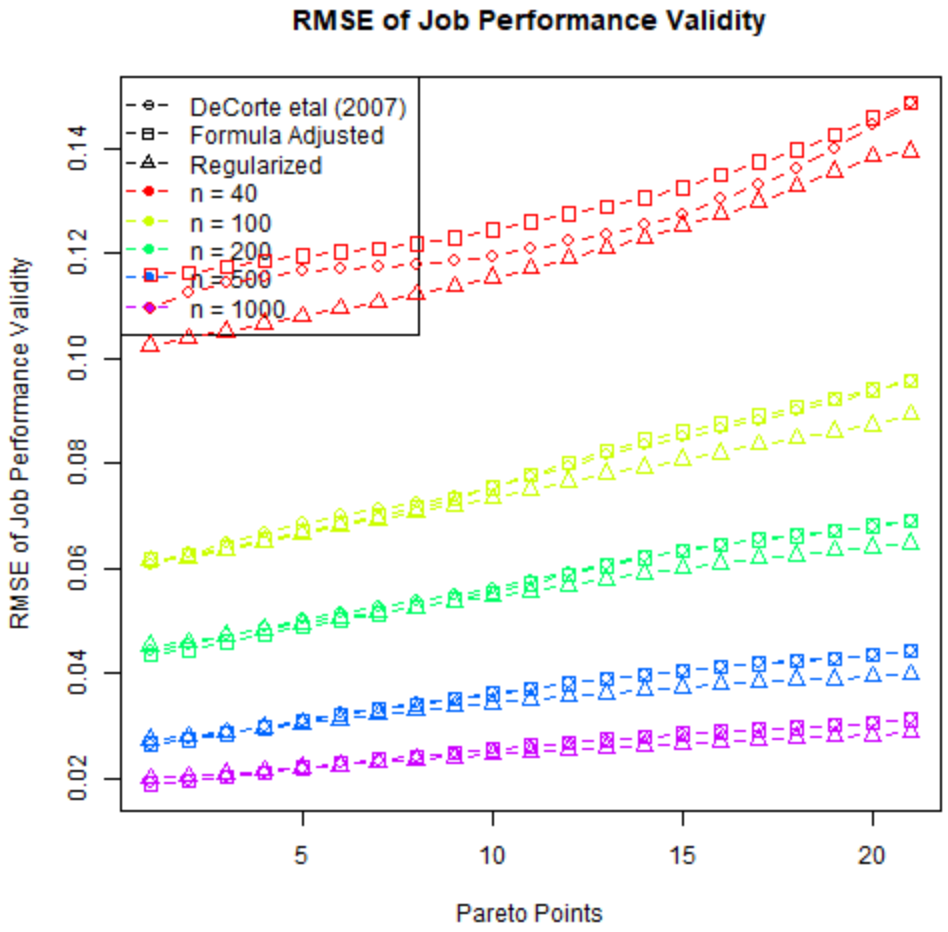


Figure 12a. RMSE of job performance validity for unshrunken, formula-adjusted and regularized Pareto-optimal solution. Pareto point #1 represents the solution where job performance was maximize; Pareto point #21 represents the solution where diversity was maximized. "DeCorte etal (2007)" is the Pareto-optimal method (based on NBI method) that was introduced by De Corte et al., (2017). "Formula Adjusted" is the formula method to correct for validity and diversity shrinkage that was discussed in Study 2. "Regularized" is the regularized Pareto-optimal method that was introduced in Study 3. RMSE of "DeCorte etal (2007)" were estimated as the root-mean-squared difference between calibration sample and validation sample's job performance validity obtained using De Corte et al. (2007) Pareto-optimal method. RMSE of "Formula Adjusted" were estimated as the root-mean-squared difference between the formula-adjusted (Study 2) job performance validity (corrected based on calibration sample solution) and validation sample job performance validity obtained using De Corte et al. (2007) Pareto-optimal method. RMSE of "Regularized" were estimated as the root-mean-squared difference between calibration sample and validation sample's job performance validity obtained using regularized Pareto-optimal method (Study 3). For each point in the Pareto-optimal trade-off curve, RMSE was aggregated (i.e., averaged) over 1000 trials.

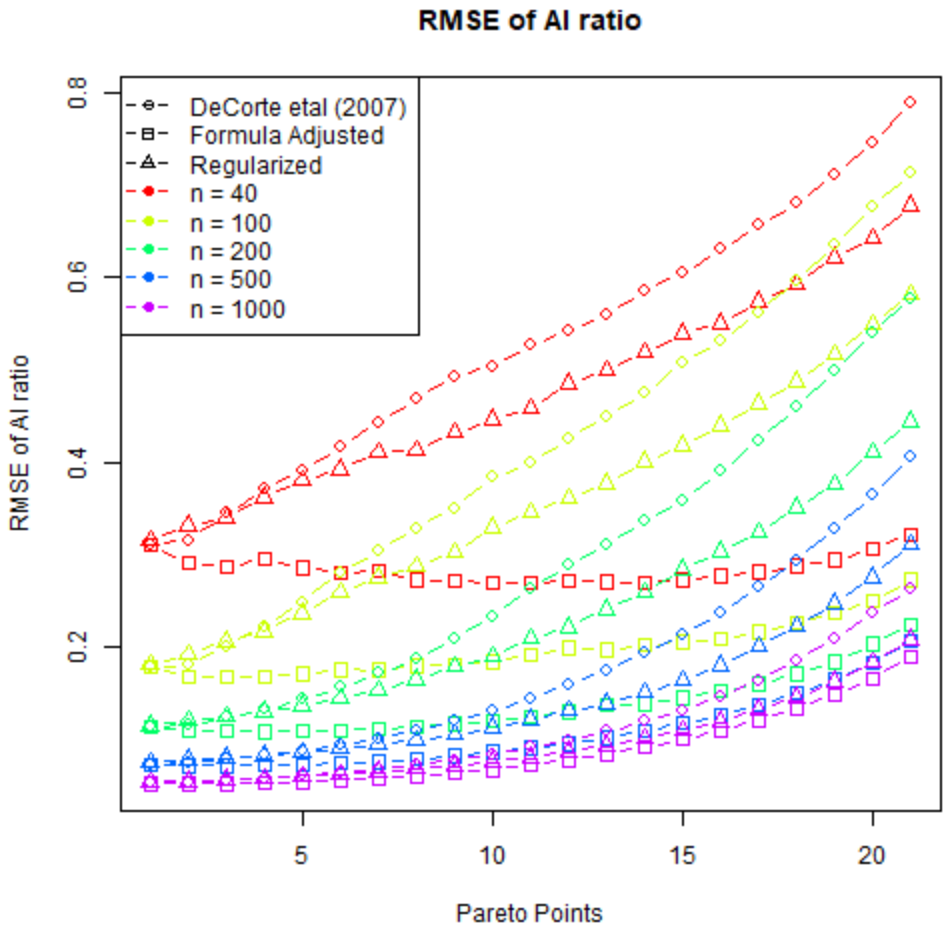


Figure 12b. RMSE of AI ratio for unshrunken, formula-adjusted and regularized Pareto-optimal solution. Pareto point #1 represents the solution where job performance was maximize; Pareto point #21 represents the solution where diversity was maximized. "DeCorte etal (2007)" is the Pareto-optimal method (based on NBI method) that was introduced by De Corte et al., (2017). "Formula Adjusted" is the formula method to correct for validity and diversity shrinkage that was discussed in Study 2. "Regularized" is the regularized Pareto-optimal method that was introduced in Study 3. RMSE of "DeCorte etal (2007)" were estimated as the root-mean-squared difference between calibration sample and validation sample's AI ratio obtained using De Corte et al. (2007) Pareto-optimal method. RMSE of "Formula Adjusted" were estimated as the root-mean-squared difference between the formula-adjusted (Study 2) AI ratio (corrected based on calibration sample solution) and validation sample AI ratio obtained using De Corte et al. (2007) Pareto-optimal method. RMSE of "Regularized" were estimated as the root-mean-squared difference between calibration sample and validation sample's AI ratio obtained using regularized Pareto-optimal method (Study 3). For each point in the Pareto-optimal trade-off curve, RMSE was aggregated (i.e., averaged) over 1000 trials.

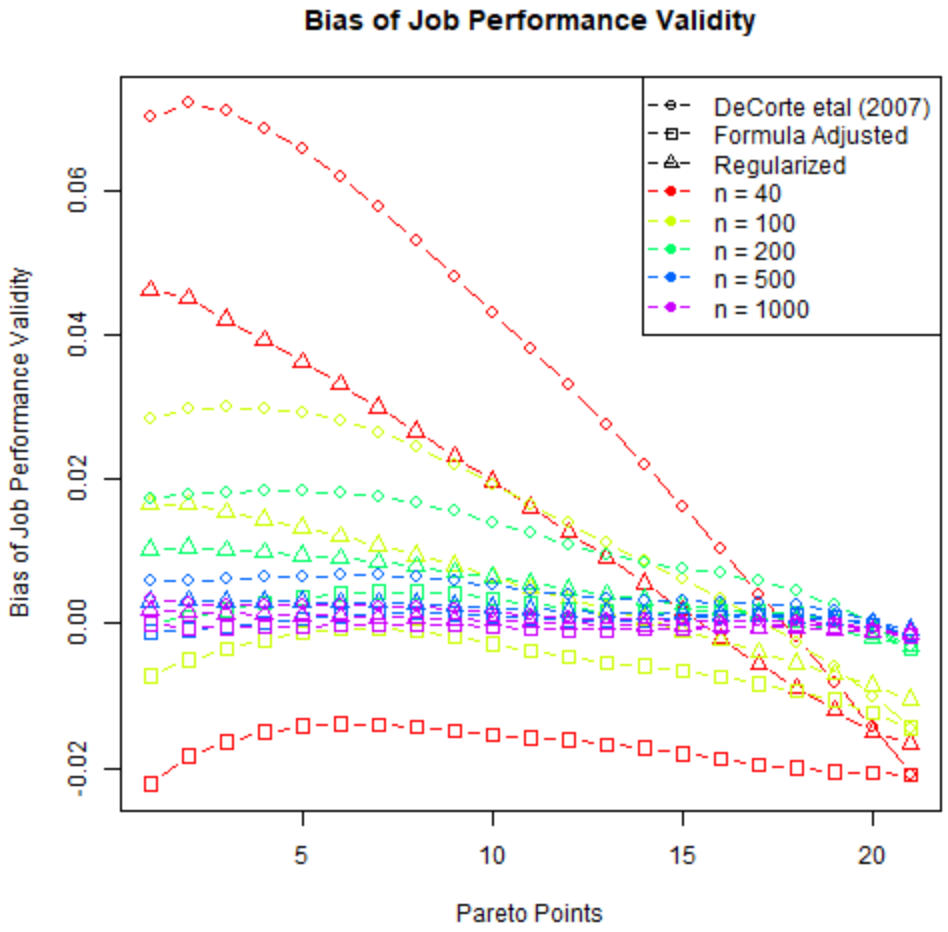


Figure 12c. Bias of job performance validity for unshrunken, formula-adjusted and regularized Pareto-optimal solution. Pareto point #1 represents the solution where job performance was maximize; Pareto point #21 represents the solution where diversity was maximized. "DeCorte etal (2007)" is the Pareto-optimal method (based on NBI method) that was introduced by De Corte et al., (2017). "Formula Adjusted" is the formula method to correct for validity and diversity shrinkage that was discussed in Study 2. "Regularized" is the regularized Pareto-optimal method that was introduced in Study 3. Bias of "DeCorte etal (2007)" were estimated as the mean difference between calibration sample and validation sample's job performance validity obtained using De Corte et al. (2007) Pareto-optimal method. Bias of "Formula Adjusted" were estimated as the mean difference between the formula-adjusted (Study 2) job performance validity (corrected based on calibration sample solution) and validation sample job performance validity obtained using De Corte et al. (2007) Pareto-optimal method. Bias of "Regularized" were estimated as the mean difference between calibration sample and validation sample's job performance validity obtained using regularized Pareto-optimal method (Study 3). For each point in the Pareto-optimal trade-off curve, bias was aggregated (i.e., averaged) over 1000 trials.

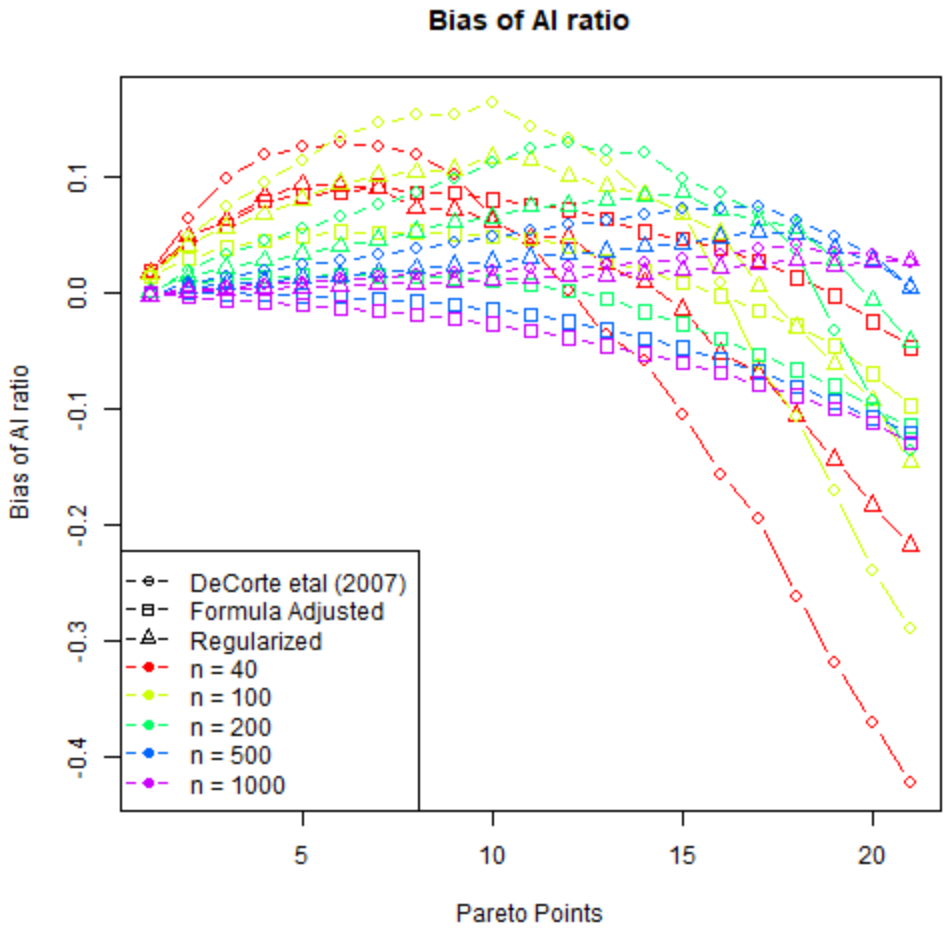


Figure 12d. Bias of AI ratio for unshrunken, formula-adjusted and regularized Pareto-optimal solution. Pareto point #1 represents the solution where job performance was maximize; Pareto point #21 represents the solution where diversity was maximized. "DeCorte etal (2007)" is the Pareto-optimal method (based on NBI method) that was introduced by De Corte et al., (2017). "Formula Adjusted" is the formula method to correct for validity and diversity shrinkage that was discussed in Study 2. "Regularized" is the regularized Pareto-optimal method that was introduced in Study 3. Bias of "DeCorte etal (2007)" were estimated as the mean difference between calibration sample and validation sample's AI ratio obtained using De Corte et al. (2007) Pareto-optimal method. Bias of "Formula Adjusted" were estimated as the mean difference between the formula-adjusted (Study 2) AI ratio (corrected based on calibration sample solution) and validation sample AI ratio obtained using De Corte et al. (2007) Pareto-optimal method. Bias of "Regularized" were estimated as the mean difference between calibration sample and validation sample's AI ratio obtained using regularized Pareto-optimal method (Study 3). For each point in the Pareto-optimal trade-off curve, bias was aggregated (i.e., averaged) over 1000 trials.

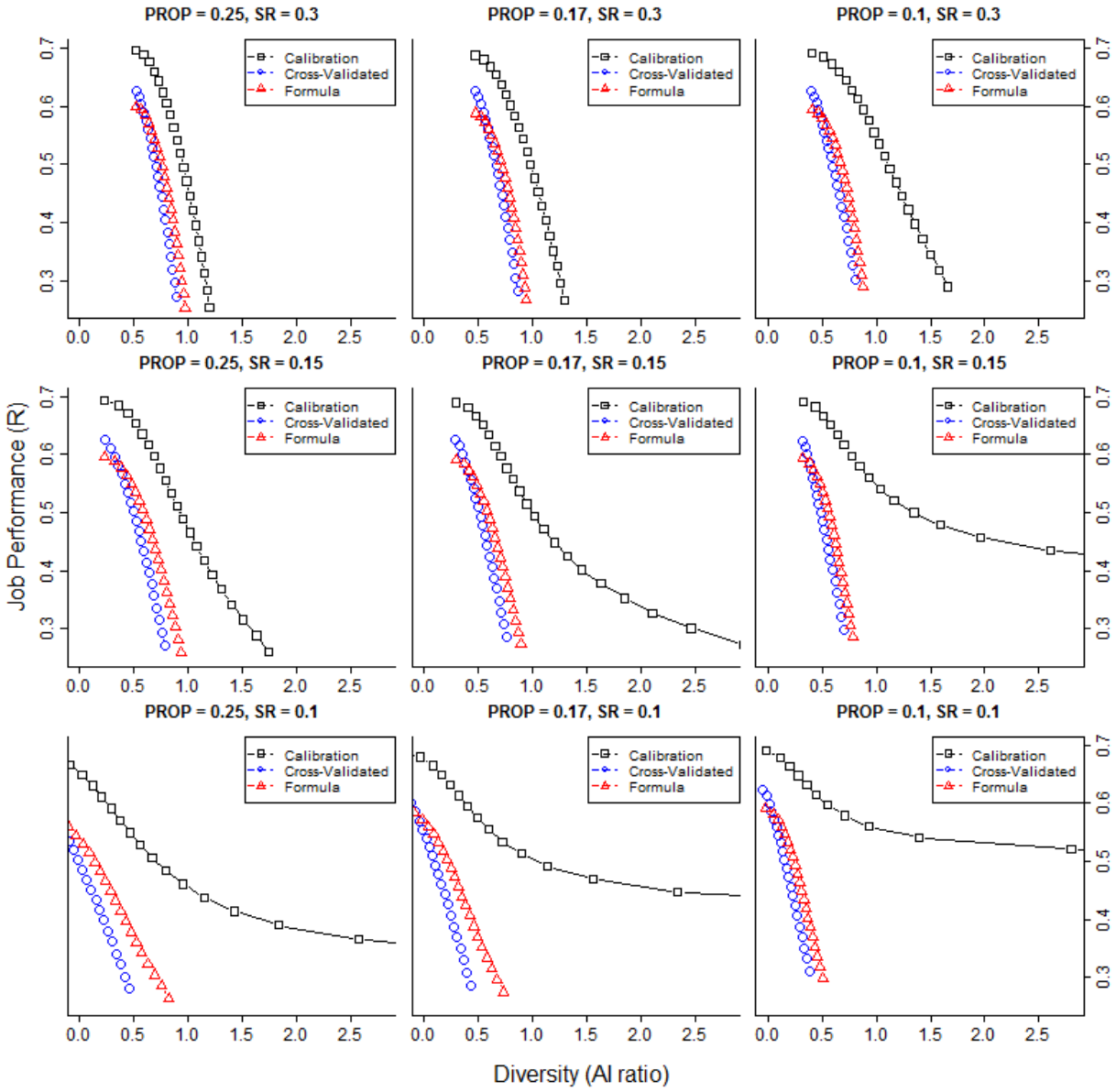


Figure 13. Comparison among different proportion of minorities and selection ratio.

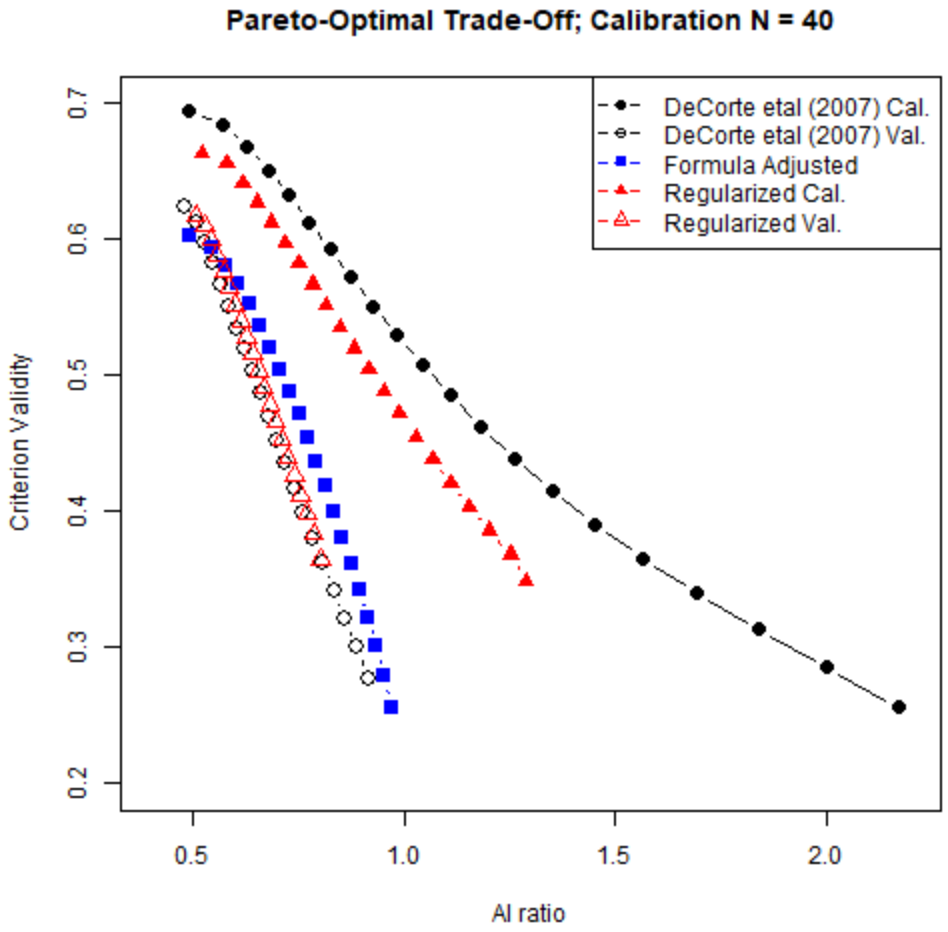


Figure 14a. Comparison among (a) calibration sample Pareto-optimal curve (based on De Corte et al., 2007 method), (b) validation sample Pareto-optimal curve (based on De Corte et al., 2007 method), (c) formula-adjusted Pareto-optimal curve (obtained based on calibration sample), (d) calibration sample regularized Pareto-optimal curve, and (e) validation sample regularized Pareto-optimal curve when calibration sample size = 40. "DeCorte etal (2007) Cal." and "DeCorte etal (2007) Val." are the calibration and validation sample Pareto-optimal curve obtained by using the method that was introduced by De Corte et al., (2007). "Formula Adjusted" is the Pareto-optimal curve after applying shrinkage formulae (discussed in Study 2) to the Pareto-optimal solution obtained based on calibration sample. "Regularized Cal." and "Regularized Val." are the calibration and validation sample regularized Pareto-optimal curve obtained by using the regularized Pareto-optimal method that was introduced in Study 3.

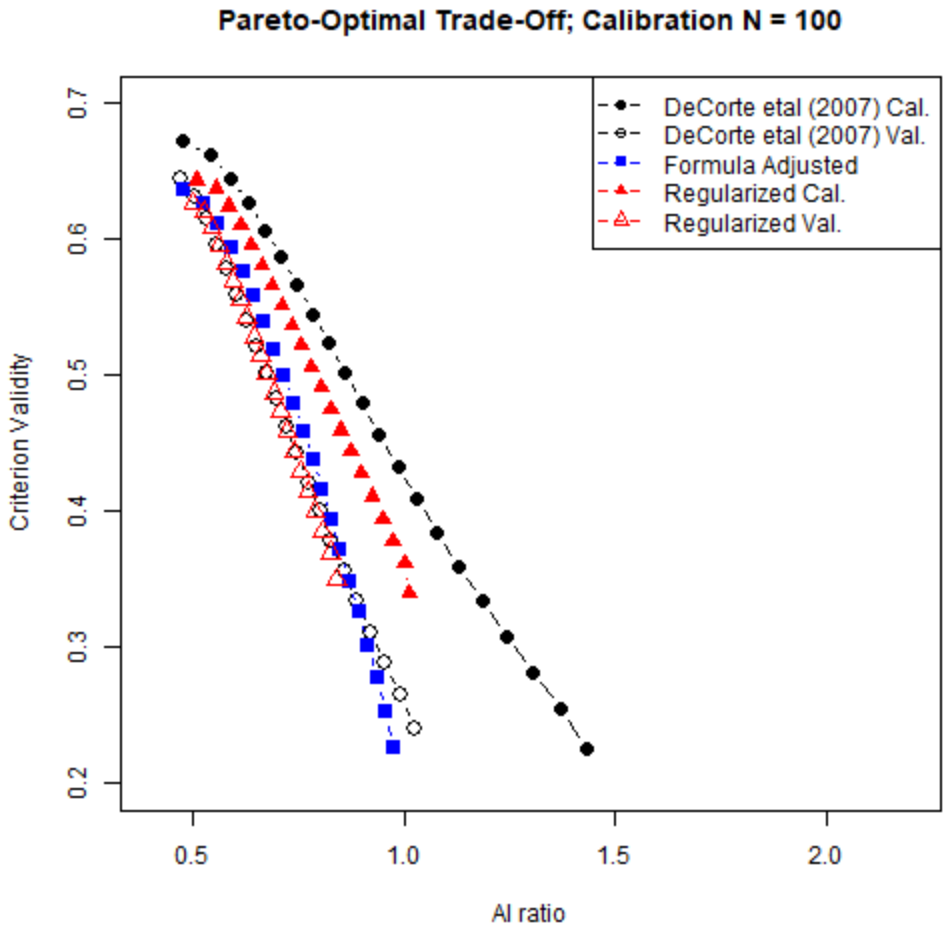


Figure 14b. Comparison among (a) calibration sample Pareto-optimal curve (based on De Corte et al., 2007 method), (b) validation sample Pareto-optimal curve (based on De Corte et al., 2007 method), (c) formula-adjusted Pareto-optimal curve (obtained based on calibration sample), (d) calibration sample regularized Pareto-optimal curve, and (e) validation sample regularized Pareto-optimal curve when calibration sample size = 100. "DeCorte etal (2007) Cal." and "DeCorte etal (2007) Val." are the calibration and validation sample Pareto-optimal curve obtained by using the method that was introduced by De Corte et al., (2007). "Formula Adjusted" is the Pareto-optimal curve after applying shrinkage formulae (discussed in Study 2) to the Pareto-optimal solution obtained based on calibration sample. "Regularized Cal." and "Regularized Val." are the calibration and validation sample regularized Pareto-optimal curve obtained by using the regularized Pareto-optimal method that was introduced in Study 3.

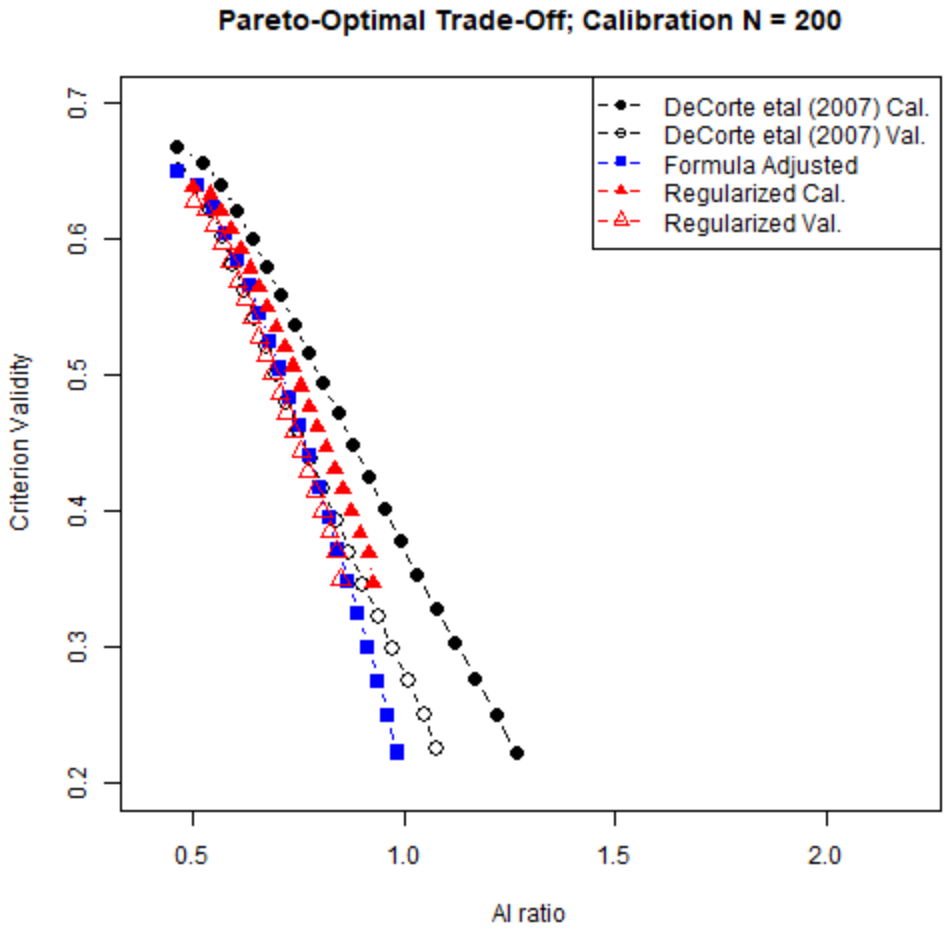


Figure 14c. Comparison among (a) calibration sample Pareto-optimal curve (based on De Corte et al., 2007 method), (b) validation sample Pareto-optimal curve (based on De Corte et al., 2007 method), (c) formula-adjusted Pareto-optimal curve (obtained based on calibration sample), (d) calibration sample regularized Pareto-optimal curve, and (e) validation sample regularized Pareto-optimal curve when calibration sample size = 200. "DeCorte etal (2007) Cal." and "DeCorte etal (2007) Val." are the calibration and validation sample Pareto-optimal curve obtained by using the method that was introduced by De Corte et al., (2007). "Formula Adjusted" is the Pareto-optimal curve after applying shrinkage formulae (discussed in Study 2) to the Pareto-optimal solution obtained based on calibration sample. "Regularized Cal." and "Regularized Val." are the calibration and validation sample regularized Pareto-optimal curve obtained by using the regularized Pareto-optimal method that was introduced in Study 3.

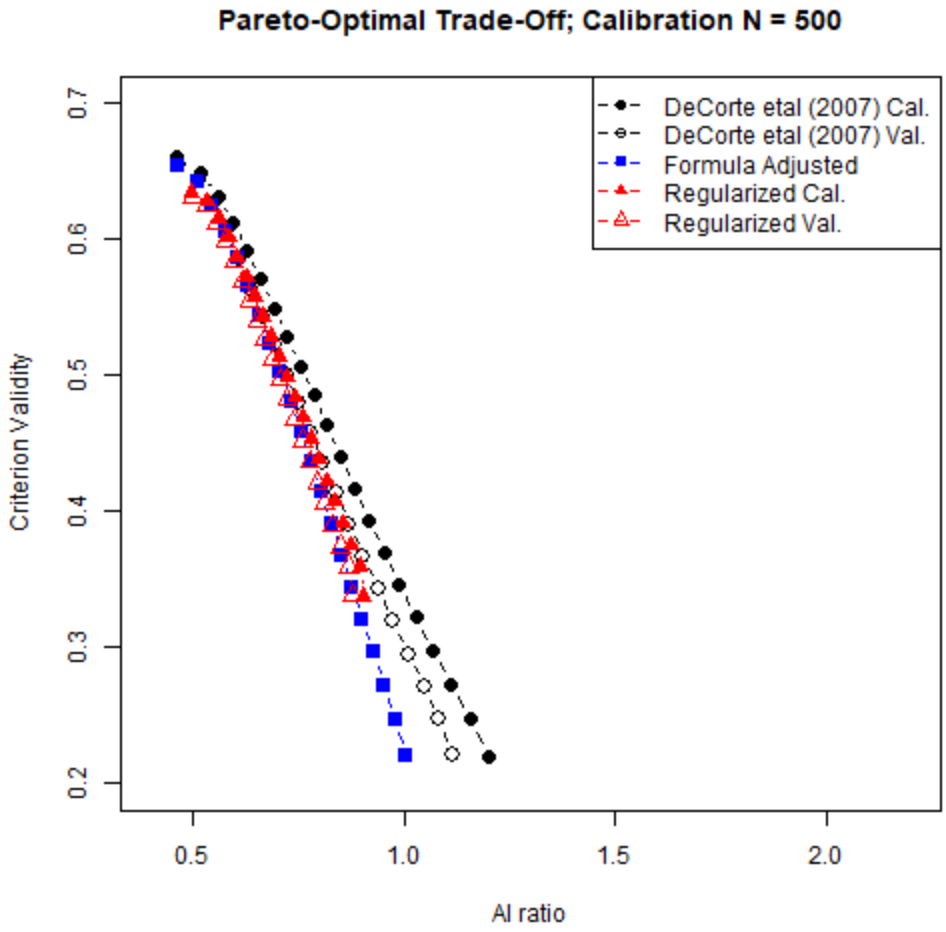


Figure 14d. Comparison among (a) calibration sample Pareto-optimal curve (based on De Corte et al., 2007 method), (b) validation sample Pareto-optimal curve (based on De Corte et al., 2007 method), (c) formula-adjusted Pareto-optimal curve (obtained based on calibration sample), (d) calibration sample regularized Pareto-optimal curve, and (e) validation sample regularized Pareto-optimal curve when calibration sample size = 500. "DeCorte etal (2007) Cal." and "DeCorte etal (2007) Val." are the calibration and validation sample Pareto-optimal curve obtained by using the method that was introduced by De Corte et al., (2007). "Formula Adjusted" is the Pareto-optimal curve after applying shrinkage formulae (discussed in Study 2) to the Pareto-optimal solution obtained based on calibration sample. "Regularized Cal." and "Regularized Val." are the calibration and validation sample regularized Pareto-optimal curve obtained by using the regularized Pareto-optimal method that was introduced in Study 3.

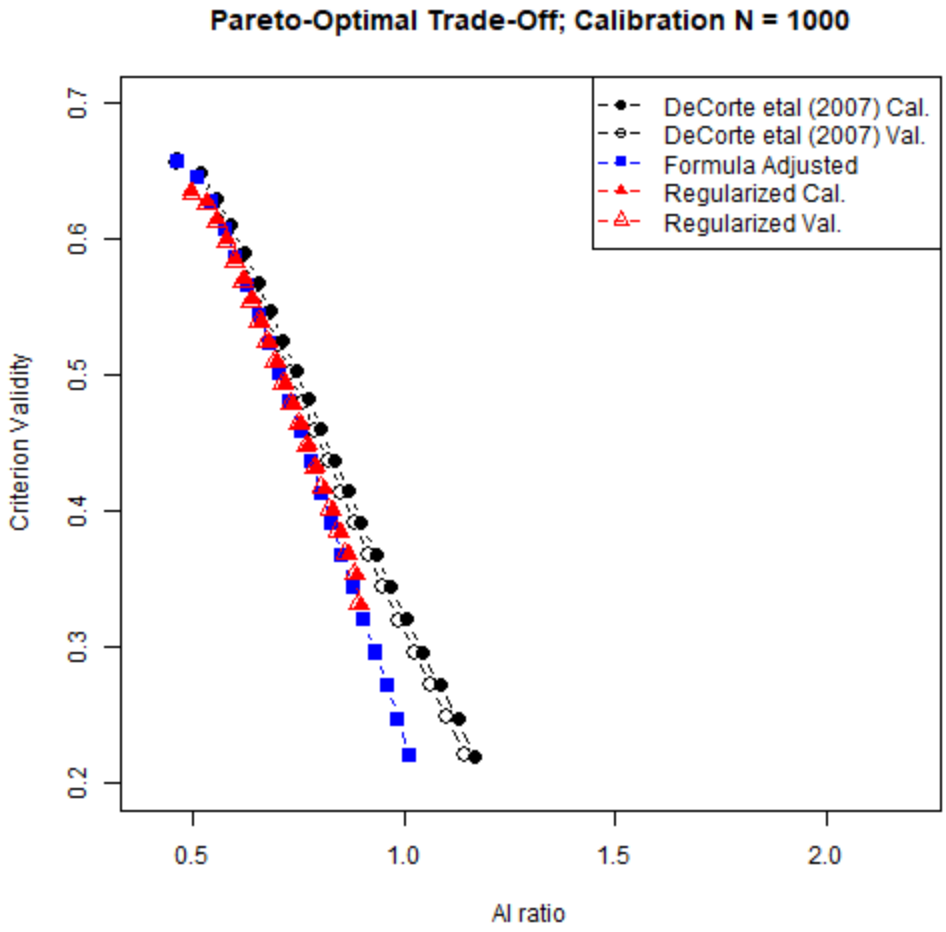


Figure 14e. Comparison among (a) calibration sample Pareto-optimal curve (based on De Corte et al., 2007 method), (b) validation sample Pareto-optimal curve (based on De Corte et al., 2007 method), (c) formula-adjusted Pareto-optimal curve (obtained based on calibration sample), (d) calibration sample regularized Pareto-optimal curve, and (e) validation sample regularized Pareto-optimal curve when calibration sample size = 1000. "DeCorte etal (2007) Cal." and "DeCorte etal (2007) Val." are the calibration and validation sample Pareto-optimal curve obtained by using the method that was introduced by De Corte et al., (2007). "Formula Adjusted" is the Pareto-optimal curve after applying shrinkage formulae (discussed in Study 2) to the Pareto-optimal solution obtained based on calibration sample. "Regularized Cal." and "Regularized Val." are the calibration and validation sample regularized Pareto-optimal curve obtained by using the regularized Pareto-optimal method that was introduced in Study 3.

REFERENCES

Arthur, W., Jr., Edwards, B. D., & Barrett, G. V. (2002). Multiple-choice and constructed response tests of ability: Race-based subgroup performance differences on alternative paper-and-pencil test formats. *Personnel Psychology*, *55*, 985-1008. doi:10.1111/j.1744-6570.2002.tb00138.x

Avery, D. R., & McKay, P. F. (2006). Target practice: An organizational impression management approach to attracting minority and female job applicants. *Personnel Psychology, 59,* 157-187.

Avery, D. R., McKay, P. F., Wilson, D. C., & Tonidandel, S. (2007). Unequal attendance: The relationships between race, organizational diversity cues, and absenteeism. *Personnel Psychology, 60,* 875-902. doi:10.1111/j.1744-6570.2007.00094.x

Bedeian, A. G., Sturman, M. C., & Streiner, D. L. (2009). Decimal dust, significant digits, and the search for stars. *Organizational Research Methods*, 12, 687-694. doi: 10.1177/. 1094428108321153

Bell, M. P., Connerley, M. L., & Cocchiara, F. K. (2009). The case for mandatory diversity education. *Academy of Management Learning and Education, 8,* 597-609. doi:10.5465/amle.2009.47785478

Bernau, C., Riester, M., Boulesteix, A. L., Parmigiani, G., Huttenhower, C., Waldron, L., & Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12), i105-i112.

Bobko, P. (2001). Correlation and regression: *Applications for industrial organizational psychology and management*. Thousand Oaks, CA: Sage Publications.

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*, 561-589.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10,* 689–709. doi:10.1177/1094428106294734

Bobko, P., & Roth, P. L. (2013). Reviewing, categorizing, and analyzing the literature on Black–White mean differences for predictors of job performance: Verifying some perceptions and updating/correcting others. *Personnel Psychology, 66,* 91-126. doi:10.1111/peps.12007

Brannick, M. T. (2001). Implications of empirical Bayes meta-analysis for test validation. *Journal of Applied Psychology, 86,* 468-480. doi:10.1037/0021-9010.86.3.468

Browne, M. W. (1975). A comparison of single sample and cross-validation methods for estimating the mean squared error of prediction in multiple linear regression. *British Journal of Mathematical and Statistical Psychology, 28,* 112-120.

Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *Annual Review of Organizational Psychology and Organizational Behavior, 2,* 47-74. 10.1146/annurev-orgpsych-032414-111427

Cascio, W. F., & Aguinis, H. (2005). Test development and use: New twists on old questions. *Human Resource Management, 44*, 219-235. doi:10.1002/hrm.20068

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology, 65,* 407-414.

Chen, S. S. & Donoho, D. L. (1994). Basis pursuit. In *28th Asilomar Conf. Signals, Systems Computers.*

Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM Review*, 43(1), 129-159.

Claudy, J. G. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, 2, 595-607.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69(3), 161.

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*, 631-657.

De Corte, W. (1999). Weighing job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology*, *84*, 695-702.

De Corte, W. (2006). TROFSS User's Guide.

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,* 1380-1393. doi: 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,* 183–194. doi:10.1111/j.1468-2389.2008.00423.x

De Corte, W., Sackett, P., & Lievens, F. (2011). Designing Pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology, 90,* 907-926.

Dorans, N. J., & Drasgow, F. (1980). A note on cross-validating prediction equations. *Journal of Applied Psychology, 65,* 728-730. doi:10.1037/0021-9010.65.6.728

Drasgow, F., Dorans, N. J., & Tucker, L. R. (1979). Estimators of the squared cross-validity coefficient: A Monte Carlo investigation. *Applied Psychological Measurement*, 3(3), 387-399.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13,* 171-192.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407-499.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43,* 38290-39315.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software,* 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

Gilbert, J. A., & Ivancevich, J. M. (2000). Valuing diversity: A tale of two organizations. *Academy of Management Perspectives, 14,* 93-105. doi:10.5465/AME.2000.2909842

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694-734. doi:10.5465/AMR.1993.9402210155

Goeman, J. J. (2017). Penalized R package, version 0.9-50.

Goldstein, H. W., Scherbaum, C. A., & Yusko, K. P. (2010). Revisiting *g*: Intelligence, adverse impact, and personnel selection. In J. L. Outtz (Ed.), *Adverse impact: Implications for*

organizational staffing and high stakes selection (pp. 95-134). New York, NY: Routledge/Taylor & Francis Group.

Hastie, T. & Efron, B. (2013). Lars R package, version 1.2.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485-585). Springer, New York, NY.

Hattrup, K., Rock, J., & Scalia, C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology, 82,* 656-664.

Hoerl A. E. (1962). Application of ridge analysis to regression problems, *Chemical Engineering Progress*, 1958, 54–59.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9,* 152-194. doi:10.1111/1468-2389.00171

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer.

Jayne M. E. A., & Dipboye, R. L. (2004). Leveraging diversity to improve business performance: Research findings and recommendations for organizations. *Human Resource Management, 43,* 409-424. doi:10.1002/hrm.20033

Joshi, A., & Roh, H. (2009). The role of context in work team diversity research: A meta-analytic review. *Academy of Management Journal, 52,* 599-627. doi: 10.5465/AMJ.2009.41331491

Kochan, T., Bezrukova, K., Ely, R., Jackson, S., Joshi, A., Jehn, K., Leonard, J., Levine, D., & Thomas, D. (2003). The effects of diversity on business performance: report of the diversity research network. *Human Resource Management, 42,* 3-21.

Konrad, A. (2003). Defining the domain of workplace diversity scholarship. *Group and Organization Management, 28,* 4-17.

Larson, S. C. (1931). The shrinkage of the coefficient of multiple correlation. Journal of *Educational Psychology*, *22*, 45-55.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications, Inc.

Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample*. (Research Bulletin No. 50-40) Princeton, NJ: Educational Testing Service.

McKay, P. F., & McDaniel, M. A. (2006). A reexamination of Black-White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91*, 538-554. doi:10.1037/0021-9010.91.3.538

Molinaro,A.M. et al. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21, 3301-3307.

Newman, D. A., Hanges, P. J., & Outtz, J. L. (2007). Racial groups and test fairness, considering history and construct validity. *American Psychologist*, *62*, 1082-1083. doi:10.1037/0003-066X.62.9.1082

Newman, D. A., Jacobs, R. R., & Bartram, D. (2007). Choosing the best method for local validity estimation: Relative accuracy of meta-analysis versus a local study versus Bayes-

analysis. *Journal of Applied Psychology*, *92*, 1394-1413. doi:10.1037/0021-9010.92.5.1394

Newman, D. A., & Lyon, J. S. (2009). Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability, and diversity. *Journal of Applied Psychology*, *94*, 298-317. doi:10.1037/a0013472

Nicholson, G. (1960). Prediction in future samples. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 322-330). Stanford, CA: Stanford University Press.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201-211.

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology*, *78*, 569-582.

Oswald, F. L., & Putka, D. J. (2015). Statistical methods for big data: A scenic tour. In S. Tonidandel, E. B. King, and J. M. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology* (pp.43-63). New York, NY: Routledge.

Outtz, J. L. (Ed.). (2010). *Adverse impact: Implications for organizational staffing and high stakes selection*. New York, NY: Routledge/Taylor & Francis Group.

Outtz, J. L., & Newman, D. A. (2010). A theory of adverse impact. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 53-94). New York, NY: Routledge/Taylor & Francis Group.

Phillips, D. L. (1962). A Technique for the Numerical Solution of Certain Integral Equations of the First Kind. *Journal of the ACM*. 9: 84. doi:10.1145/321105.321114

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,* 153-172. doi:10.1111/j.1744-6570.2008.00109.x

Pulakos, E. D., & Schmitt, N. (1996). An evaluation of two strategies for reducing adverse impact and their effects on criterion-related validity. *Human Performance, 9,* 241-258.

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2017). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, doi: 10.1177/1094428117697041.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology, 79,* 518-524.

Riester, M., Wei, W., Waldron, L., Culhane, A. C., Trippa, L., Oliva, E., Kim, S. H., Michor, F., Huttenhower, C., Parmigiani, G. & Birrer, M. J. (2014). Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *Journal of the National Cancer Institute,* 106(5). doi:10.1093/jnci/dju048

Roberson, Q. M., & Park, H. J. (2007). Examining the link between diversity and firm performance: The effect of diversity reputation and leader racial diversity. *Group and Organization Management, 32,* 548-568.

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,* 297-330. doi:10.1111/j.1744-6570.2001.tb00094.x

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

Ryan, A. M., & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65*, 693-717.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*, 707-721.

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*, 302-318. doi:10.1037/0003-066X.56.4.302

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274. doi:10.1037/0033-2909.124.2.262

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*, 719-730. doi:10.1037/0021-9010.82.5.719

Schneider, R. J., Hough, L. M., & Dunnette, M. D. (1996). Broadsided by broad traits: How to sink science in five dimensions or less. *Journal of Organizational Behavior, 17,* 639-655.

Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., & Ones, D. S. (2011). Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology*, 96(5), 1055.

Shirazi, A., Najafi, B., Aminyavari, M., Rinaldi, F., & Taylor, R. A. (2014). Thermal–economic–environmental analysis and multi-objective optimization of an ice thermal energy storage system for gas turbine cycle inlet air cooling. *Energy*, 69, 212-226.

Song, Q., 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. doi: 10.1037/apl0000240

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society (Series B: Methodological), 58,* 267-288.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*. 4: 1035–1038.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology, 97,* 499-530. doi: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*, 547-563. doi:10.1037/a0035183

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2*, 440-457.

Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education, 69*, 203-224.

Zeng, Y., & Breheny, P. (2017). The bigLASSO Package: A Memory-and Computation-Efficient Solver for LASSO Model Fitting with Big Data in R. *arXiv preprint arXiv:1701.05936*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320. https://arxiv.org/abs/1701.05936.

Zou, H., & Hastie, T. (2012). elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. R package version 1.1.

APPENDIX A

Multiple Regression Shrinkage Formulae

The squared multiple correlation coefficient observed in a calibration sample (R^2) is typically an overestimate of both the population squared multiple correlation (ρ^2) and the cross-validated squared multiple correlation ($\rho_c{}^2$). The *population squared multiple correlation coefficient (ρ^2)* estimates what R^2 would be in the general population (i.e., if the same predictors were used to predict the same criterion, but with potentially different predictor weights, as obtained in the entire population; see Darlington, 1968, p. 173). Commonly used shrinkage formulae for ρ^2include 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).

Olkin and Pratt's shrinkage formula for ρ^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 formulae has been used to describe shrinkage for the *cross-validated squared multiple correlation coefficient ($\rho_c{}^2$)*, which is an estimate of what R^2 would be in a new sample. $\rho_c{}^2$ is an estimate of what R^2 would be if the original calibration-sample regression weights were applied in the entire population; which is equal to the expected value of what R^2 would be in a new (cross-validated) sample drawn from that population. To estimate $\rho_c{}^2$, one must use a different formula that use the population squared multiple correlation coefficient (ρ^2) as input. Browne's (1975) shrinkage formula—which was based on the Lord-Nicholson shrinkage formula (Lord, 1950; Nicholson, 1960)—is commonly used to estimate $\rho_c{}^2$. Yin and Fan (2001) conducted a simulation study to investigate the effectiveness of various shrinkage formulae and found that Olkin and Pratt's (1858) shrinkage formula for ρ^2 and Browne's (1975) shrinkage formula for $\rho_c{}^2$ outperformed other shrinkage formulae.

Browne's shrinkage formula for $\rho_c{}^2$

$$\hat{\rho}_c{}^2 = \frac{(N-k-3)\hat{\rho}^4 + \hat{\rho}^2}{(N-2k-2)\hat{\rho}^2 + k}$$

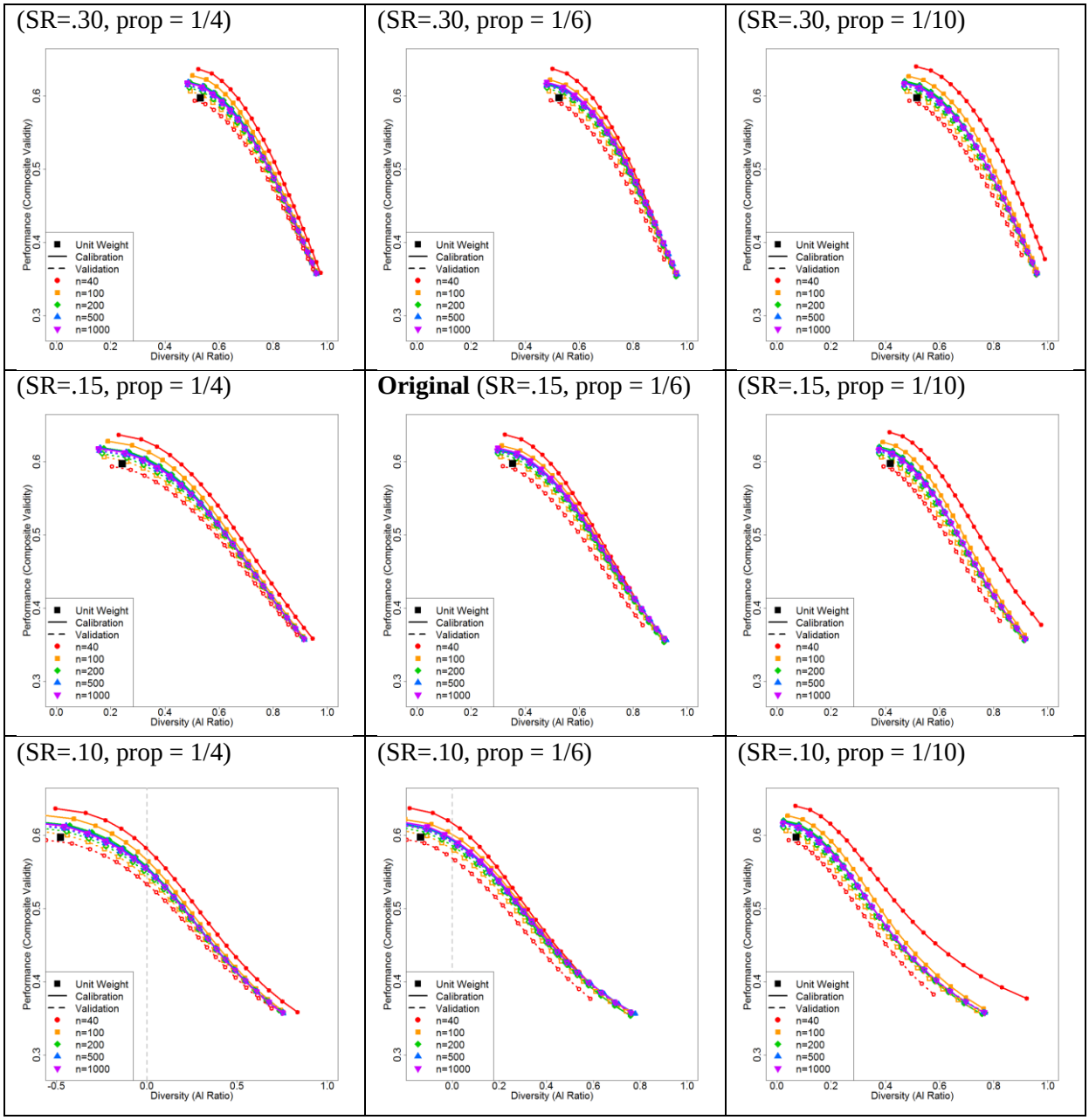
where $\hat{\rho}_c{}^2$ is the corrected cross-validated squared correlation coefficient; $\hat{\rho}^2$ is the corrected population squared multiple correlation coefficient obtained from a formula such as Olkin and Pratt's (1958) shrinkage formula and Wherry's (1931) shrinkage formula, *N* is the calibration sample size, and *k* is the number of predictors.

APPENDIX B

Study 1 Supplementary Analyses

Table 7.

Comparison of Results Based on Different Selection Ratios (SR) and Proportions of Applicants from Minority Group (prop) [Cognitive Subtests: Machinery Repair Job Family]

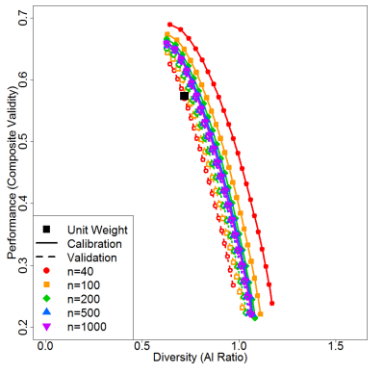


Note. SR = selection ratio (overall), prop = Proportion of applicants from minority group. [For example, prop = 1/6 means that among all the applicants, 1/6 of the applicants were minority applicants (e.g., Black applicants)].

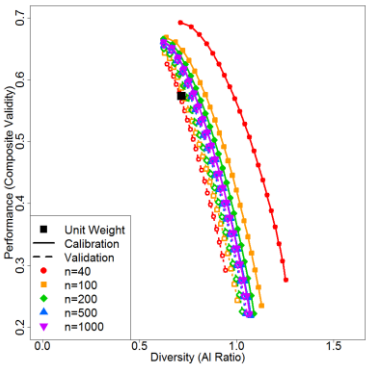
Table 8.

Comparison of Results Based on Different Selection Ratios (SR) and Proportion of Applicants from Minority Group (prop) [Bobko-Roth Predictors]

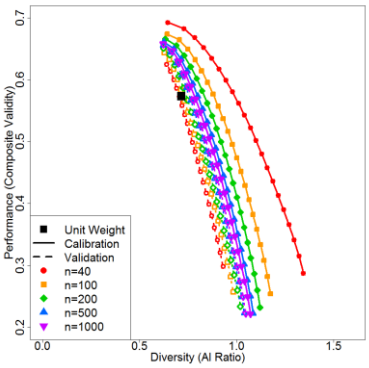
(SR=.30, prop = 1/4)



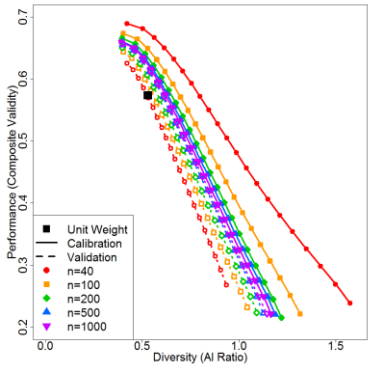
(SR=.30, prop = 1/6)



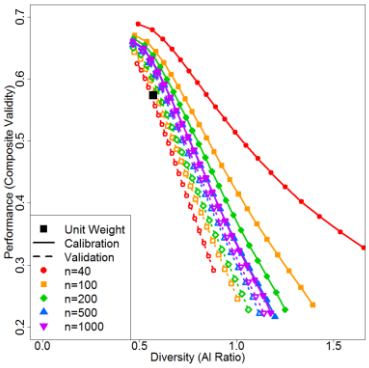
(SR=.30, prop = 1/10)



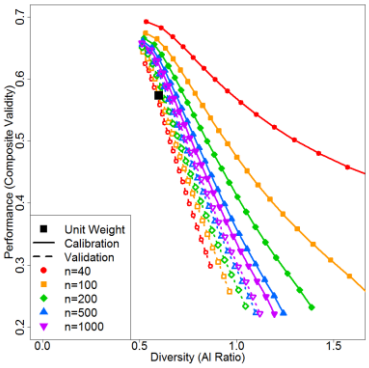
(SR=.15, prop = 1/4)



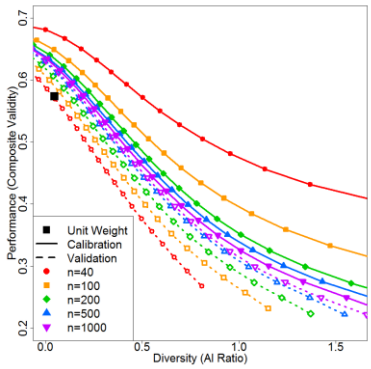
Original (SR=.15, prop = 1/6)



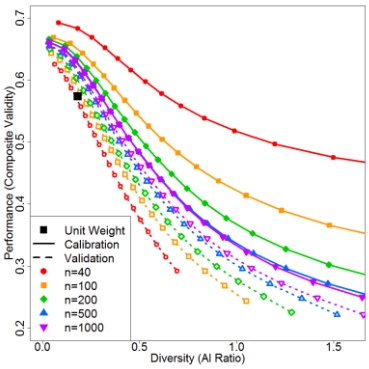
(SR=.15, prop = 1/10)



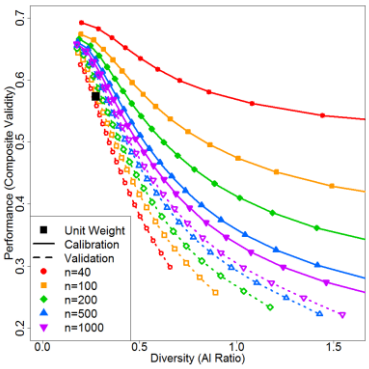
(SR=.10, prop = 1/4)



(SR=.10, prop = 1/6)



(SR=.10, prop = 1/10)



Note. 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)].

APPENDIX C

ParetoR R Package and Web Application for Implementing Pareto-Optimal Weighting

Pareto-Optimization in Diversity Hiring
Developer: Q. Chelsea Song
Contact: qianqisong@gmail.com
Last Update: 03/23/2018

The commands below and more details of the package can be found at:
https://github.com/Diversity-ParetoOptimal/ParetoR

Objective

The current R package provides a set of Pareto-optimal solutions that simultaneously optimize both diversity and criterion validity in a personnel selection scenario [see Song, Wee, & Newman (2017). The current package allows for implementation of (1) Pareto-optimal method that was adapted from De Corte, Lievens & Sackett (2007); (2) Pareto-optimal shrinkage formulae to estimate formula-adjusted shrunken Pareto-optimal solutions (see Study 2 of the current dissertation); (3) regularized Pareto-optimal method (see Study 3 of the current dissertation).

Instructions

Install and Load Package

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")
```

Main Functions

ParetoR function
Pareto-optimal method introduced by De Corte, Lievens & Sackett (2007)

Example Implementation

1. Specify four inputs (example from De Corte, Lievens & Sackett (2007) is given below):

```
# (1) Proportion of minority applicants (prop) = (# of minority applicants)/(total # of applicants)
## Example:
```

```
    prop <- 1/4
 # (2) Selection ratio (sr) = (# of selected applicants)/(total # of applicants)
 ## Example:
    sr <- 0.10
 # (3) Subgroup differences (d): standardized mean differences between minority and majority
subgroups, on each predictor (in applicant pool)
 ## Example:
    d <- c(1.00, 0.23, 0.09, 0.33)
 # (4) Correlation matrix (R) = criterion & predictor inter-correlation 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))
```

2. 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).

ParetoShrinkage function
Estimate shrunken Pareto-optimal solution based on Pareto-optimal shrinkage formulae introduced in Study 2 of the current dissertation

Example Implementation

1. Specify inputs

```
 # (1) Calibration sample size
    Ncal <- 100
 # (2) Number of predictors
    numpred <- 4
 # (3) Number of Pareto-optimal points (i.e., number of sets of predictor weights)
    p_pareto <- 21
 # (4) Vector of calibration sample job performance validity
    load(R_perf_cal)
 # (5) Vector of calibration sample race bivariate correlation [i.e., correlation between race
dummy variable (0-minority, 1-majority) and predictor composite score]
```

```
load(R_race_cal)
```

2. Paste and run the following command in R console or RStudio:

```
# Estimate shrunken Pareto-optimal solution
ParetoShrinkage(Ncal = Ncal, numpred = numpred, p_pareto = p_pareto,
                R_perf_cal = R_perf_cal, R_race_cal = R_race_cal,
                use = "Wherry)
```

Output Description

1. Formula-adjusted Pareto-optimal solution

ParetoElnet function
Regularized Pareto-optimal method introduced in Study 3 of the current dissertation

Example Implementation

1. Specify inputs (example from De Corte, Lievens & Sackett (2007) is given below):

```
# (1) Calibration sample size
## Example
    n_cal = 100
# (2) Subgroup differences (d): standardized mean differences between minority and majority
subgroups (i.e., majority - minority), on each predictor and criterion (in applicant pool)
## Example
    d <- c(1.00, 0.23, 0.09, 0.33, .30)
# (3) Correlation matrix (R) = Criterion predictor inter-correlation matrix
# Format: Predictor_1, ..., Predictor_n, Criterion
## Example
    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) Validation sample size
## Example
    n_val = 10000
# (5) Grid of alpha values to try
## Example
    alpha.grid <- seq(0,1,length=3)
# (6) Grid of lambda values to try
## Example
    lambda.grid <- 10^seq(1,-2,length=11)
# (7) Proportion of minority applicants (prop) = (# of minority applicants)/(total # of applicants)
## Example
```

```
    prop <- 1/4
# (8) Selection ratio (sr) = (# of selected applicants)/(total # of applicants)
## Example
    sr <- 0.10
# (9) Spac = number of Pareto points
## Example
    Spac <- 21
```

2. Paste and run the following command in R console or RStudio:

```
# Fit Regularized Pareto-optimal model with parameter selection via cross-validation
    cv.out = cv.ParetoElnet(n_cal = n_cal, D = D, R = R,
                    n_val = n_val,
                    lambda.grid = lambda.grid, alpha.grid = alpha.grid,
                       prop = prop, sr = sr, Spac = Spac)
```

Output Description

1. Regularized 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).

cv.ParetoElnet function
Regularized Pareto-optimal method introduced in Study 3 of the current dissertation with parameter selection based on cross-validation method

Example Implementation

Example 1: Calibration raw data as input

1. Specify inputs (example from De Corte, Lievens & Sackett (2007) is given below):

```
# (1) Calibration data
# Format: Predictor_1, ..., Predictor_n, Job Performance Validity, Race dummy variable (e.g., 0-minority; 1-majority)
## Example
    load(data_cal)
# (2) Validation sample size
## Example
    n_val = 10000
# (3) Grid of alpha values to try
## Example
    alpha.grid <- seq(0,1,length=3)
# (4) Grid of lambda values to try
## Example
```

```
    lambda.grid <- 10^seq(1,-2,length=11)
# (5) Proportion of minority applicants (prop) = (# of minority applicants)/(total # of applicants)
## Example
    prop <- 1/4
# (6) Selection ratio (sr) = (# of selected applicants)/(total # of applicants)
## Example
    sr <- 0.10
# (7) Spac = number of Pareto points
## Example
    Spac <- 21
```

2. Paste and run the following command in R console or RStudio:

```
# Fit Regularized Pareto-optimal model with parameter selectoin via cross-validation
    cv.out = cv.ParetoElnet(data_cal = data_cal, n_val = 10000,
                            lambda.grid = lambda.grid, alpha.grid = alpha.grid,
                            prop = prop, sr = sr, Spac = Spac)
```

Example 2: Statistics of calibration sample data as input (i.e., calibration sample size, standardized subgroup mean difference, predictor and criterion correlation matrix). Input do not include calibration sample raw data set.

1. Specify inputs (example from De Corte, Lievens & Sackett (2007) is given below):

```
# (1) Calibration sample size
## Example
    n_cal = 100
# (2) Subgroup differences (d): standardized mean differences between minority and majority subgroups (i.e., majority - minority), on each predictor and criterion (in applicant pool)
## Example
    d <- c(1.00, 0.23, 0.09, 0.33, .30)
# (3) Correlation matrix (R) = Criterion predictor inter-correlation matrix
# Format: Predictor_1, ..., Predictor_n, Criterion
## Example
    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) Validation sample size
## Example
    n_val = 10000
# (5) Grid of alpha values to try
## Example
    alpha.grid <- seq(0,1,length=3)
# (6) Grid of lambda values to try
## Example
```

```
    lambda.grid <- 10^seq(1,-2,length=11)
# (7) Proportion of minority applicants (prop) = (# of minority applicants)/(total # of applicants)
## Example
    prop <- 1/4
# (8) Selection ratio (sr) = (# of selected applicants)/(total # of applicants)
## Example
    sr <- 0.10
# (9) Spac = number of Pareto points
## Example
    Spac <- 21
```

2. Paste and run the following command in R console or RStudio:

```
# Fit Regularized Pareto-optimal model with parameter selection via cross-validation
    cv.out = cv.ParetoElnet(n_cal = n_cal, D = D, R = R,
                            n_val = n_val,
                            lambda.grid = lambda.grid, alpha.grid = alpha.grid,
                            prop = prop, sr = sr, Spac = Spac)
```

Output Description

1. Regularized Paret-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) with the model using best parameters selected using cross-validation.
2. Selected model parameters (i.e., alpha parameter and lambda parameter)
3. Plots (i.e., Criterion validity – AI ratio trade-off curve, and Predictor weights across trade-off points).

Note

The program is updated based on the ParetoR package that was introduced in Song et al. (2017). It is partially 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

Song, C., Wee, S., & Newman, D. A. (in press). Diversity shrinkage: Cross-validating Pareto-optimal weights to reduce adverse impact. Journal of Applied Psychology.

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, 631-657.

De Corte, W. (2006). TROFSS User's Guide.

De Corte, W., Lievens, F., & Sackett, P. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. Journal of Applied Psychology, 92, 1380-1393.

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,* 547-563.

Acknowledgements

Great appreciation to Dr. Serena Wee, Dr. Dan Newman, Dr. Wilfred De Corte, and Dr. Victoria Stodden 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 package (https://qchelseasong.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.

APPENDIX D

Comparison among Wherry (1931), Olkin & Pratt (1958) and Claudy (1978) Formulae

In Study 2, we used Monte-Carlo simulation to evaluate three commonly used OLS shrinkage formulae to adjust for shrinkage. Based on evaluation result, Wherry (1931) formula was selected to be used as part of the Pareto-optimal shrinkage formulae.

Method

We examined the performance of three shrinkage formulae when applied as part of a set of Pareto-optimal shrinkage formulae to correct for job performance validity and diversity shrinkage in Pareto-optimal solutions. Specifically, we conducted simulation on Pareto-optimal shrinkage formulae with (1) Wherry (1931) formula, (2) Olkin and Pratt (1958) formula, and (3) Claudy (1978) formula. For each Pareto point (total of 21 Pareto points), we calculated the root-mean-squared-error (RMSE) and bias for formula-adjusted shrunken job performance validity ($R_{perf_{Adj.P}}$) and formula-adjusted shrunken race bivariate correlation ($R_{race_{Adj.p}}$) through the same procedure as Study 2. Specifically:

$$RMSE\left(R_{perf_{Adj.P}}\right) = \sqrt{\frac{\sum_{j=1}^{1000}\left(R_{perf_{Adj.p}} - R_{perf_{Val.p}}\right)^2}{1000}}$$

$$RMSE\left(R_{race_{Adj.p}}\right) = \sqrt{\frac{\sum_{j=1}^{1000}\left(R_{race_{Adj.p}} - R_{race_{Val.p}}\right)^2}{1000}}$$

$$Bias\left(R_{perf_{Adj.P}}\right) = \frac{\sum_{j=1}^{1000}\left(R_{perf_{Adj.p}} - R_{perf_{Val.p}}\right)}{1000} mean$$

$$Bias\left(R_{race_{Adj.p}}\right) = \frac{\sum_{j=1}^{1000}\left(R_{perf_{Adj.p}} - R_{perf_{Val.p}}\right)}{1000}$$

where:

a. R_{perf}: bivariate correlation of predictor weighted composite score and job performance
b. R_{race}: bivariate correlation of predictor weighted composite score and diversity variable (e.g., race dummy variable where White applicants are coded as 1 and Black applicants are coded as 0)
c. $R_{perf_{Adj.p}}$: Formula-adjusted job performance validity (R_{perf}); i.e., calibration sample job performance criterion validity at a Pareto point p, after correcting for shrinkage using 1) Wherry (1931), (2) Olkin and Pratt (1958), or (3) Claudy (1978) shrinkage formula and Browne shrinkage formula

d. $R_{race_{Adj.p}}$: Formula-adjusted bivariate race correlation (R_{race}); i.e., calibration sample diversity at a Pareto point p, after correcting for shrinkage using (1) Wherry (1931), (2) Olkin and Pratt (1958), or (3) Claudy (1978) shrinkage formula and Browne (1975) shrinkage
e. $R_{perf_{Val.p}}$: Validation sample job performance validity (R_{perf}) at a Pareto point p, obtained through simulation
f. $R_{race_{Val.p}}$: Validation sample race bivariate correlation (R_{race}) at a Pareto point p, obtained through simulation

RMSE and bias were calculated for each trial, and then aggregated (average) over 1,000 trials. RMSE (which is the square-root of MSE—sum of variance and bias) illustrates the cross-validation error. Bias is the difference between estimated value and true value, in this case the difference between formula-adjusted shrunken criterion value (e.g., job performance criterion validity, AI ratio) and the shrunken criterion value obtained by simulation (true shrunken criterion value). The aim of the shrinkage formulae is to estimate shrunken criterion value in the validation sample, or obtain estimates that have smallest bias and RMSE.

Results and Discussion

Figure 15 shows the RMSE and bias comparison among the three shrinkage formulae. For both job performance validity and AI ratio, Wherry formula exhibited smallest or second to smallest RMSE and bias across different calibration sample sizes. Olkin and Pratt formula tend to provide highest RMSE for job performance validity and provide largest overestimate of validation sample job performance validity when calibration sample size is 40. Olkin and Pratt formula and Claudy formula also tend to result in larger RMSE and bias for AI ratio as compared to Wherry formula. As a result, we chose to use Wherry shrinkage formula because it is a simpler formula that performs as well as or superior to the two alternatives, in terms of bias and RMSE.

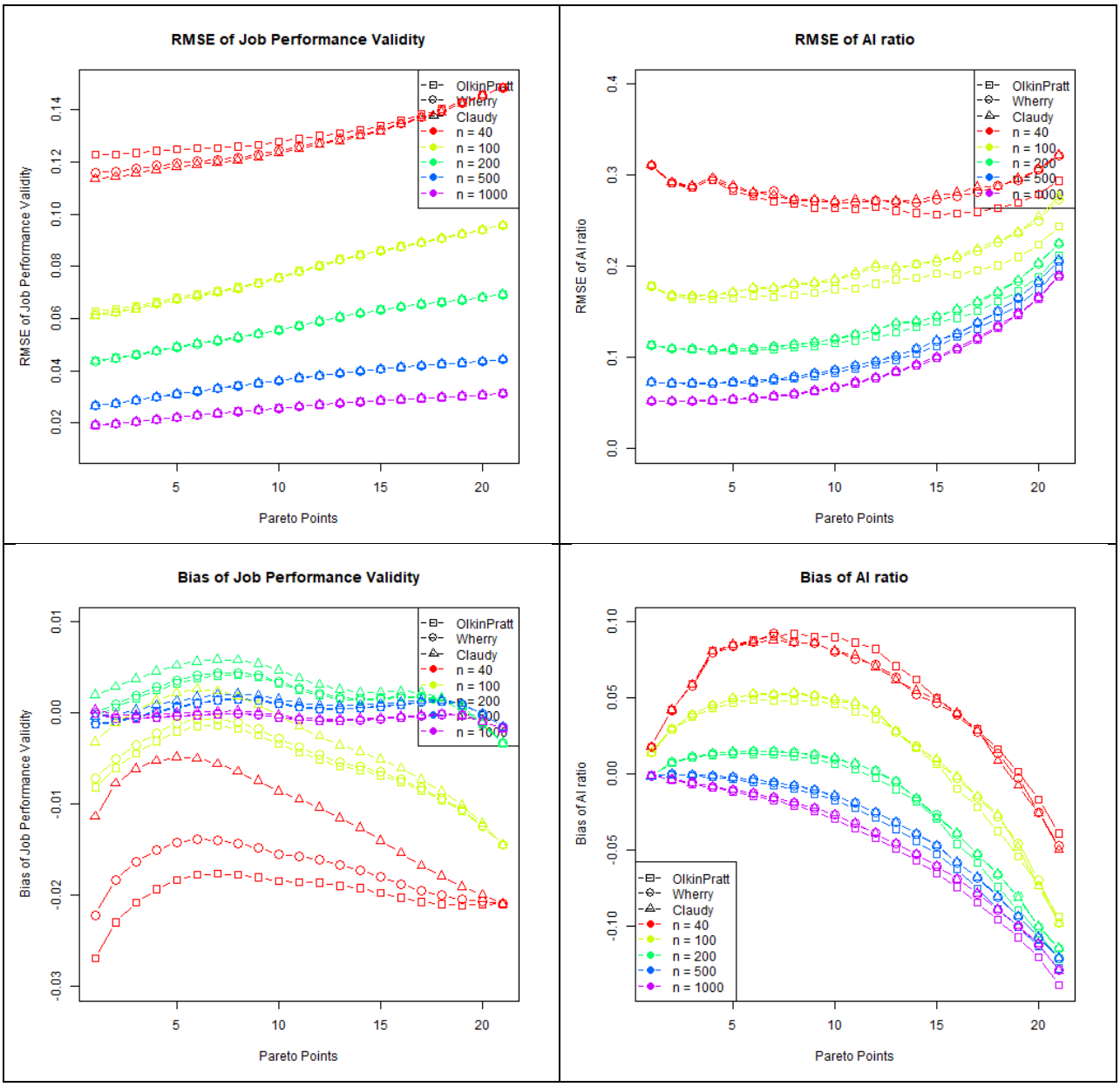

RMSE of Job Performance Validity
RMSE of Job Performance Validity
Pareto Points
RMSE of AI ratio
RMSE of AI ratio
Pareto Points
Bias of Job Performance Validity
Bias of Job Performance Validity
Pareto Points
Bias of AI ratio
Bias of AI ratio
Pareto Points
OlkinPratt
Wherry
Claudy
n = 40
n = 100
n = 200
n = 500
n = 1000

Figure 15.

APPENDIX E

Diversity and Validity Shrinkage and Performance of Shrinkage Formulae when Calibration Sample Size is 10,000

We conducted additional analysis to examine diversity and validity shrinkage and the utility of shrinkage formulae when calibration sample size is huge (N = 10,000). We used the same simulation procedures as Studies 1 and 2. Results are shown in Figure 16. In general, when calibration sample size is 10,000, there is negligible diversity and validity shrinkage. The Pareto-optimal shrinkage formulae underestimate the validation sample (shrunken) Pareto-optimal solution, providing a conservative estimate of the shrunken Pareto-optimal solution.

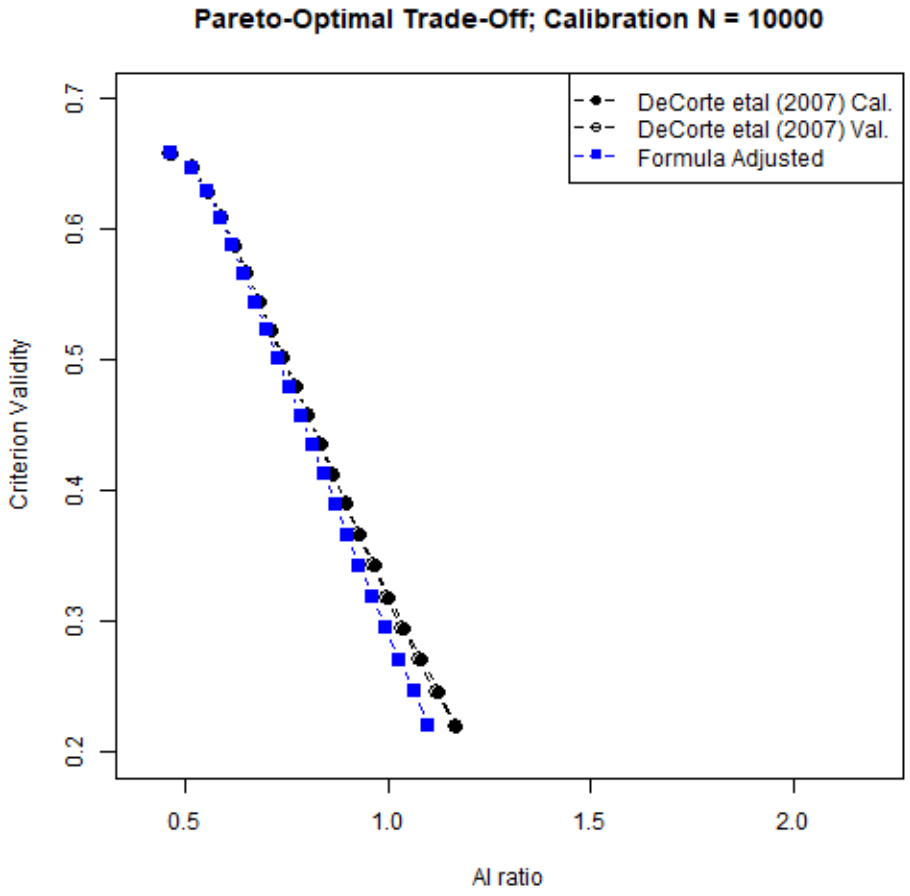


Figure 16. Comparison among Pareto-optimal trade-off curves from calibration sample, cross-validated/validation sample, and shrinkage formulae-adjusted criterion validity when calibration sample size is 10,000.