AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Applied Psychology

ISSN: 0021-9010

2022, Vol. 107, No. 8, 1243–1260
https://doi.org/10.1037/apl0000927

A Comprehensive Examination of the Cross-Validity of Pareto-Optimal Versus Fixed-Weight Selection Systems in the Biobjective Selection Context

Wilfried De Corte[1], Filip Lievens[2], and Paul R. Sackett[3]
[1] Department of Data-Analysis, Faculty of Psychology, Ghent University
[2] Lee Kong Chian School of Business, Singapore Management University
[3] Department of Psychology, University of Minnesota

The article presents evidence for the cross-validity potential of fixed-weight (FW) versus Pareto-Optimal (PO) selection systems in biobjective selection situations where both the goals of diversity and quality are valued and the importance of the goals is undecided a priori. The article extends previous research by also studying the cross-validity potential of selection systems in the practically most important sample-to-sample cross-validity scenario. We address three research questions: (a) Do different PO systems show comparable levels of relative (i.e., proportional) achievement upon cross-validation? (b) Do PO systems achieve higher levels of relative achievement upon cross-validation than FW selection systems?, and (c) How does the achievement of PO and FW systems, in terms of adverse impact ratios and average performance of the selected applicants, evolve under cross-validation? As a key result, in case of sufficiently large applicant pools (typically 100 applicants or more), PO systems had on average a higher cross-validity potential than the corresponding FW systems. Yet, even for applicant pools as large as 500, FW systems may match the merits of PO systems and we present a straightforward procedure to decide which FW systems may offer a comparable cross-validation potential than the PO systems.

Keywords: cross-validity, adverse impact, personnel selection, Pareto-Optimal, selection design

Supplemental materials: https://doi.org/10.1037/apl0000927.supp

For decades, the diversity/quality trade-off has been a challenge in personnel selection because some of the most valid predictors of job performance, such as cognitive ability tests, show substantial mean differences between majority and minority group applicant populations, resulting in adverse impact (Ployhart & Holtz, 2008; Sackett et al., 2001). The joint role of mean differences and selection ratios in determining adverse impact is well documented (Sackett & Ellingson, 1997). In recent years, the Pareto-Optimal (PO) approach (e.g., De Corte et al., 2007, 2011; Wee et al., 2014) has emerged as a promising proposal for designing selections in the biobjective selection context (Cortina et al., 2017); that is, for settings in which both the goals of diversity and quality are valued but the importance of the goals is undetermined a priori (i.e., a solution is chosen after the range of PO solutions is examined).[1] The PO approach identifies selection systems that are expected to result in the best possible selection diversity (quality) for given levels of quality (diversity). The systems are called PO because the level of diversity (quality) attained by any other feasible system is at best as good as the diversity (quality) level of the PO system, whereas the quality (diversity) level is less than that of the PO system.

Some evidence attests to the increasing use of PO approaches by organizations. First, the original article introducing the PO approach offered a computer program for identifying PO solutions (De et al., 2007). That program has been requested by a wide variety of organizations, in the U.S. and Europe, including private sector firms, government agencies, and consulting firms. Second, recently Rupp et al. (2020) published an article that will undoubtedly help to further disseminate the PO approach among practitioners. This article presented not only flow charts, checklists, and practical guidance on how to use PO but also a very accessible app (i.e., the ParetoR Shiny app) to implement the approach. Third, we conducted a small informal survey among U.S. top employers of industrial and organizational psychologists. Results revealed that some have already used the PO approach in their personnel selection practice for several years. For others, the PO approach is still in a research stage, with its usage restricted to the development of demonstration projects. Besides a concern for ease of application and mathematical sophistication (where the Rupp et al. article should help a lot to remove this concern), the survey also indicated two worries about using the PO approach. At present, the approach focuses on the design of selection systems that are PO with respect to a single

This article was published Online First June 10, 2021.
Wilfried De Corte https://orcid.org/0000-0001-7400-3181
We have no known conflict of interest to disclose.

The computational resources and services used in this work were provided to the first author by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government—Department EWI.

Correspondence concerning this article should be addressed to Wilfried De Corte, Department of Data-Analysis, Faculty of Psychology, Ghent University, H. Dunantlaan 1, Ghent 9000, Belgium. Email: wilfried.decorte@ugent.be

[1] If the importance of the goals is decided, the two objectives can be combined to a single objective using the importance given to the objectives, and the selection is no longer bi—but a single objective. So, biobjective selection implies the boundary condition that the importance of the selection objectives is undetermined a priori.

protected group, whereas practitioners typically deal with situations involving multiple protected groups. Second, there are doubts that PO systems developed for an initial setting will continue to perform well when applied to similar but new settings. In other words, there exist concerns about the cross-validity of PO systems. The present article addresses exactly this cross-validity issue.

At present, little is indeed known about the cross-validity of PO systems, although this information is critically important to assess the practical value of PO selection system design. Research on the cross-validity of biobjective selection systems, as compared to that of single-objective systems, poses particular challenges. For single-objective systems, estimating the achievement (i.e., the criterion-related validity) of the system under cross-validation is straightforward (i.e., estimate the achievement of a predictive model, derived in the initial calibration setting, in the cross-validation setting). However, biobjective systems (i.e., those pursuing both the diversity and the quality achieved by the system) are more challenging, given that achievement is assessed on two dimensions. And, importantly, the biobjective PO approach results in an infinite set of PO systems instead of a single optimal system (i.e., the regression-based system) as is the case in single-objective system design. This requires rethinking cross-validation for biobjective selection systems.

Cross-validation research in the biobjective selection context is usefully aided by also examining what we term "relative" (or "proportional") achievement.[2] Conceptually, we can express achievement on both quality and diversity dimensions as a proportion of the achievement that can possibly be attained upon cross-validation on each dimension. This puts quality and diversity on comparable metrics. It also permits combining (e.g., averaging) the relative achievement on the quality and diversity dimensions into an overall relative achievement index. This relative achievement metric is useful for a variety of purposes. One is the important research question of whether PO systems hold up comparably upon cross-validation across the full range of possible PO systems. PO systems range from the quality-maximizing system at one extreme to the diversity-maximizing system at the other. So, in the initial calibration setting, the achievement on both the quality and diversity dimensions can vary dramatically across PO systems. In terms of the absolute level of achievement, one would expect, for example, that a PO system high on the quality dimension in the calibration setting would show a higher level of quality in the cross-validation setting than would a PO system with a lower level of quality in the calibration setting, with a reversed pattern found on the diversity dimension. At first glance, one might say that the two differ in how well they cross-validate, given these differences in absolute achievement. However, it might prove the case that in terms of relative achievement, both show the same proportion of achievement relative to the achievement that is possible in the cross-validation setting.

A second type of question aided by relative achievement indices involves comparisons between PO systems and other predictor weighting schemes, such as fixed weights (FW). Using standard metrics (average job performance of the selected applicants [AJP], for the quality and adverse impact ratio [AIR], for the diversity objective), comparisons are of necessity limited. As the dimensions of interest are in different noncomparable metrics, results are limited to rough categories: PO beats FW on both dimensions, FW beats PO on both dimensions, or PO and FW each win on one dimension. While this is of some value, using the relative achievement metric has the benefit that it puts both dimensions on a common metric and permits combining the two for an overall quantitative comparison.

The present article addresses the following three research questions: (a) Do different systems across the calibration PO front show comparable relative achievement upon cross-validation? (b) Do calibration PO systems show higher relative achievement upon cross-validation than FW systems?, and (c) How does the achievement of the calibration PO systems in their original metrics (i.e., AJP and AIR) evolve when the systems are applied in cross-validation conditions? Each of these questions addresses a key important issue in evaluating the real incremental contribution of PO as compared to ad hoc and FW systems when designing biobjective selection systems. The answer to the first question may offer an extra criterion for choosing between different PO systems; the relevance of the entire practice of PO selection systems (vis-à-vis FW systems) depends on the answer to the second question, and the answer to the last question addresses the actual as compared to the promised diversity/quality trade-off when implementing PO systems.

We answer the research questions with respect to the three basic cross-validation contexts. The first context involves developing a selection system on a full population and investigating the degree of cross-validation to a subsequent sample drawn from that population (hereafter, population-to-sample). The second deals with developing a selection system-based on a sample and investigating the degree of cross-validation to the population from which that sample was drawn (hereafter, sample-to-population). The third involves developing a selection system on a sample and investigating cross-validation to a subsequent sample drawn from the same population (hereafter, sample-to-sample).

We offer important contributions beyond earlier related research (cf. De Corte et al., 2020; Song et al., 2017). We are the first to address the sample-to-sample cross-validation context, which represents the most realistic setting (e.g., Cattin, 1980) because selection systems are typically developed using sample-based predictor/criterion information and are always applied to finite applicant pools. We extend existing findings on the cross-validity of single-objective selection designs (e.g., Bobko et al., 2007; Schmidt, 1971; Van Iddekinge & Ployhart, 2008) to the considerably more complicated case of biobjective single- and multistage selection design. Crucially, the results that we obtain do not only generalize answers to old questions about the way the calibration achievement of different selection designs evolves under cross-validation but also answer new questions (e.g., about the relative achievement of different PO systems upon cross-validation) that have no counterpart in single-objective selection design.

Overview of Previous Research

The cross-validity of single stage, single-objective selection systems, and more generally the cross-validity of linear predictor composites has been a topic of active research for almost a century. As indicated by the very name "cross-validity," this research essentially studied the criterion-related validity of linear predictor

[2] Relative achievement and the corresponding measures are discussed in detail in the section "Assessing the Cross-Validity of Bi-Objective Selection Systems."

composites in both the sample-to-population and the sample-to-sample cross-validity scenarios (e.g., Cattin, 1980). Validity shrinkage formulas have been developed for both scenarios (e.g., Browne, 1975; Wherry, 1931), whereas simulation and analytic studies examined the conditions under which regression composites outperform FW composites (e.g., Bobko et al., 2007; Cattin, 1978; Einhorn & Hogarth, 1975; Raju et al., 1999). This large body of research led to two results that are particularly relevant for our research questions. The first is that samples of no less than 100 are typically required to expect at least equal cross-validity of regression and FW composites. Our simulation studies, therefore, include only applicant samples of at least 100 because there is reason to expect that the smallest applicant samples for which biobjective optimized PO systems may cross-validate equal or better than FW systems are at least equal to the samples sizes at which single-objective optimized regression composites cross-validate better than FW composites. The second result relates to the difference in population validity between the regression and the FW composites: The smaller this difference, the larger the required sample sizes for regression composites to cross-validate better than FW composites (e.g., Cattin, 1978) and we examine whether this result is also valid in the biobjective context.

Compared to the long-standing literature on the cross-validity of regression and FW composites, the research on the cross-validity of PO versus FW systems has just started. Only two recent articles (i.e., De Corte et al., 2020; Song et al., 2017) focus on the issue. Song et al. (2017) gauged the sample-to-population cross-validity[3] potential of the FW and PO systems as the difference between the average diversity/quality trade-off of the systems in the calibration and the validation conditions. They referred to this difference as the Pareto shrinkage, with the term diversity (quality) shrinkage used for the difference in average calibration and validation diversity (quality) trade-off value of the systems. Larger calibration applicant samples led to decreased diversity and quality shrinkage and, for equal-sized calibration applicant samples, the PO systems with larger population diversity[4] (quality) trade-off values showed larger diversity (quality) shrinkage as compared to the PO systems with smaller population diversity (quality) trade-off values. Nonetheless, the diversity/validity trade-off upon cross-validation of the FW was dominated by the trade-off of some of the studied PO systems (i.e., either the validity or the diversity value of the FW system was smaller than the corresponding value of the PO system, whereas the other trade-off value was at most equal to that of the PO system) for all but the smallest studied calibration applicant sample size (cf. Song et al., 2017, p. 1647).

De Corte et al. (2020) also used simulation methods, focusing on the population-to-sample cross-validity potential of FW and PO systems using both the diversity/quality trade-off and a newly introduced relative achievement measure for the assessment. They were able to calculate the relative quality achievement upon cross-validation of the systems, but available methods were unable to assess the relative validation diversity achievement. Due to this computational problem, De Corte et al. presented results only with respect to the relative validation quality achievement of the systems. Online material also provided some initial results on the sample-to-population cross-validity of PO systems: For the population-to-sample cross-validity case, PO systems had a higher relative quality achievement upon cross-validation than corresponding FW systems, even at the smallest studied validation applicant pool size of 80. They also observed a monotonically increasing relationship between the population quality trade-off value and the relative validation quality achievement value of PO systems. Online material confirmed the Song et al. (2017) finding that PO systems with a higher population diversity (quality) value showed larger diversity (quality) shrinkage than PO systems with a lower population diversity (quality) value. Yet, larger (smaller) diversity (quality) shrinkage did not correspond to lower (higher) relative validation diversity (quality) achievement, but rather to higher (lower) relative validation diversity (quality) achievement.

In sum, these two prior studies reported rather favorable results on the cross-validity of PO systems. Yet, the three key issues on the cross-validity of PO and FW systems that we discussed in the intro (e.g., cross-validation in the typical sample-to-sample context) have remained largely unresolved.

Method

Simulation Design and Procedure

We used simulation methods in a factorial design, where the cells of the design correspond to crossing three factors. The first two factors relate to the size of the calibration and cross-validation applicant pool, each having five levels with values of 100, 200, 500, 1000, and infinite, respectively. Together, the two factors represent various instances of the three cross-validation scenarios. For example, the combination of any level of the first factor (except for the last) with the last level of the second factor reflects *sample-to-population* cross-validation. Any combination of one of the finite valued levels of the first and second factor reflects *sample-to-sample* cross-validation. Finally, any combination of the last level of the first factor with a finite valued level of the second factor reflects *population-to-sample* cross-validation.

The third factor in the design, with two levels, represents the number of stages (either one or two) in the studied selection situation where five predictors (see Table 1) are available in the selection process. The predictor/criterion data in Table 1 derive from the seminal work of Schmitt et al. (1997), Bobko et al. (1999) and Roth et al. (2011) and were also used by both Song et al. (2017) and De Corte et al. (2020). For the single-stage selection setting, the set of feasible selection systems corresponds to the set of all systems that assign a nonnegative weight to the predictors. In the two-stage setting, the set of feasible selection systems consists of all systems using a nonnegative weighed composite of the cognitive ability, the conscientiousness, and the biodata predictor in stage 1 and a nonnegative weighed composite of the structured interview and the integrity predictors in stage 2, implementing a retention rate between .25 and .40 after the first stage and a final selection rate equal to 0.1.

Within each cell of the design we repeatedly applied the computational cycle detailed in Appendix B. The average (across the repetitions) results for the measures described in the next section are subsequently used to answer our research questions.

[3] Song et al. (2017) studied the sample to population cross-validity with a validation sample of 10,000 as a proxy for the validation applicant population.

[4] The population diversity/quality trade-off value of a selection system refers to the diversity/quality trade-off value of the system when applied to the applicant population.

Table 1
Predictor/Criterion Population Data for the Studied Selection Situation

Predictor	$d^{\#}$	1	2	3	4	5
1. Cognitive ability	0.72					
2. Structured interview	0.32	.31				
3. Conscientiousness	−0.09	.03	.13			
4. Biodata	0.39	.37	.16	.51		
5. Integrity	0.04	.02	−.02	.34	.25	
Criterion						
1. Performance	0.38	.52	.48	.22	.32	.20

Note. $^{\#}d$ corresponds to the standardized mean difference between the majority and the minority applicant populations. The data in Table 1 correspond to results presented in Bobko et al. (1999); Roth et al. (2011); Schmitt et al. (1997), and Song et al. (2017).

Assessing the Cross-Validity of Biobjective Selection Systems

Measure

The research questions of the article relate to either the absolute achievement (i.e., the diversity/quality and, more specifically, the AIR/AJP trade-off) or the relative (i.e., proportional) achievement of different biobjective selection systems. The AIR/AJP trade-off achieved by a selection system is used to assess the achievement of a system when addressing how the trade-offs of PO and FW systems evolve under cross-validation. To distinguish calibration and validation conditions, we use the terms *calibration diversity/quality* and *validation diversity/quality trade-off*, respectively.

The AIR/AJP trade-off measure is unfit to study the cross-validation potential of different PO systems and offers only limited help for comparing the cross-validation potential of PO and FW systems. For biobjective selections, the trade-off is a bivalued outcome with component values that involve *incommensurable* dimensions (i.e., AJP and AIR are on different, inconvertible metrics). Aggregating the AJP and AIR values to a single-valued assessment of the level of achievement of the systems is therefore only possible when the importance of the goals is determined in which case the selection is no longer biobjective (see footnote 1). The incommensurability problem further implies that the outcomes of different biobjective systems can only be compared, and are therefore called *comparable*, if the component differences of the systems are not in a different direction (e.g., in comparing PO and FW systems, one wins on one dimension and the other wins on the other dimensions). Only if one approach wins on both dimensions or wins on one dimension and scores equal on the other can one make an overall evaluative statement as to which approach is superior in that setting. Unfortunately, previous research by De Corte et al. (2020) and Song et al. (2017) shows that the trade-offs of different PO systems and FW systems are often incomparable. Even when the systems have comparable trade-offs, the extent that one of the systems has a better trade-off as compared to the other cannot be assessed.

Using diversity and quality shrinkage measures (see Song et al., 2017) instead of the validation trade-off is no real alternative. Aggregating the diversity and quality shrinkage values to a Pareto shrinkage value (cf. Figure 5 in Song et al., 2017) also collides with the incommensurability of the diversity and quality scales, which explains why Song et al. (2017) present no results with respect to overall Pareto shrinkage but only with respect to diversity and quality shrinkage *separately*. Also, prior cross-validity research on single-objective selection designs documented that shrinkage and level of achievement are quite different things. Whereas FW composites, as compared to regression composites, almost invariably show substantially less validity shrinkage this does not imply that FW composites have a higher or even an equal cross-validity value (i.e., show an equal or better level of achievement upon cross-validation), especially with growing sample sizes.

To address the limitations of the trade-off and shrinkage measures, we adopt a set of three new measures: the *relative diversity achievement*, the *relative quality achievement*, and the *global relative achievement* measure. The new measures, first proposed in De Corte et al. (2020), avoid the problem of aggregating the incommensurable diversity and quality dimensions by rescaling these dimensions to the same, dimensionless scale with an identical effective range using the well-known zero-one linear normalization technique from the field of multiobjective programming (Tamiz & Jones, 1997). More specifically, the validation diversity and quality of a selection system are first rescaled as the proportion of the maximum possible (over the set of all feasible selection systems) diversity and quality gain achievable in the validation sample at the quality (diversity) level of the system. The resulting proportions constitute the relative diversity and the relative quality achievement value of the system. In sharing the same 0–1 anchored scale, the two relative achievement values can be meaningfully aggregated to a single number, using equal weights for both values to reflect the condition that the importance of the objectives is undetermined. The resulting global relative achievement measure therefore captures the relative achievement of the systems independent from the importance of the diversity and quality objectives as is consistent with the biobjective selection situation (cf. footnote 1).

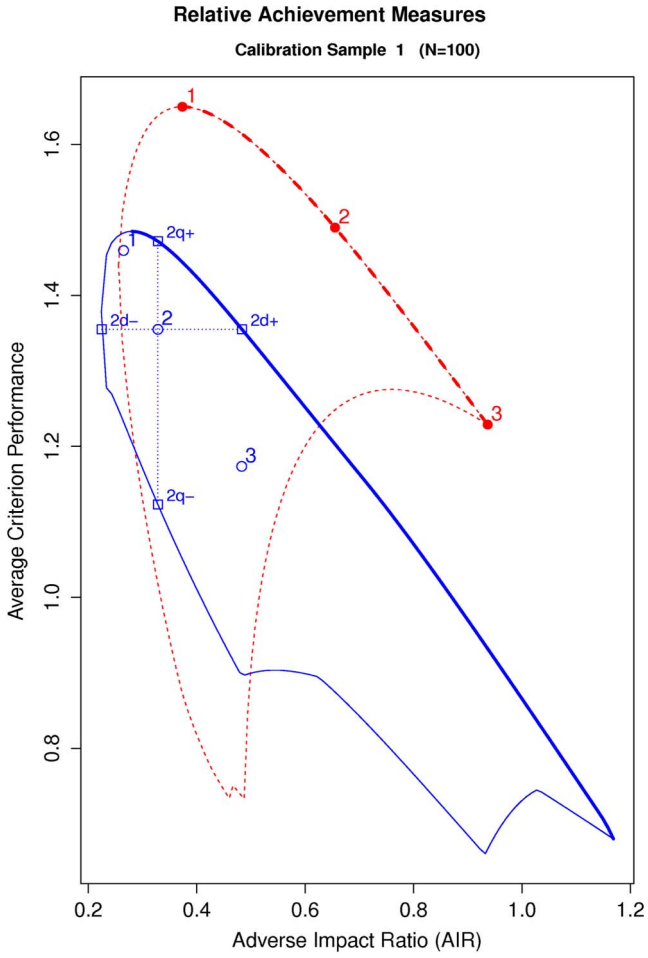
Except for cross-validation conditions involving very small samples (see below), the global relative achievement measure is almost always applicable and leads to a quantification instead of a mere ordinal assessment of the relative achievement of the systems. The terms *relative calibration diversity (quality) achievement*, *relative validation diversity (quality) achievement*, and *global relative calibration (validation) achievement* are used as needed for clarification.

Illustration

Figure 1 illustrates the calculation of the new measures. The illustration pertains to the single-stage selection setting detailed in Table 1 where the set of feasible selection systems corresponds to all systems that assign a nonnegative weight to the predictors. The figure represents the gamut of attainable (across the set of feasible systems) diversity/quality trade-offs in both the calibration and the validation conditions. The dashed/dotted red curves enclose the calibration gamut, whereas the solid blue curves enclose the validation gamut. Both gamuts are computed for the same selection rate of .10 and an identical .20/.80 minority/majority applicant representation, but use different predictor/criterion correlation and effect size data in the calculations. The Table 1 population correlation and effect size values are used for the validation gamut, whereas the calibration gamut is based on correlation and effect size data

Figure 1

Computation of the Relative Validation Diversity and Quality Achievement of PO System 2



Note. PO = Pareto-Optimal. The red dashed and blue solid curves enclose the gamut of achievable diversity/quality trade-offs in the calibration ($N = 100$) and the validation condition ($N =$ Infinity), respectively. The red points identify three systems that are PO in the calibration condition and the corresponding blue points indicate the validation trade-off of these systems. The blue points 2q+ and 2q− (2d+ and 2d−) show the maximal and minimal quality (diversity) value attainable (across the feasible systems) in the validation condition at the validation diversity (quality) level of PO system 2. The relative validation quality (diversity) achievement of PO system 2 equals the ratio of (a) the distance between the blue points 2 and 2q− (2d−) and (b) the distance between the blue points 2q+ (2d+) and 2q− (2d−). See the online article for the color version of this figure.

derived from a sample of size 100. The gamuts therefore exemplify the sample-to-population cross-validation situation.

The figure also represents the calibration and validation PO front (i.e., the set of PO trade-offs in the calibration and validation condition) using the red bold dashed and the blue bold solid curve segments respectively, with three PO trade-offs, indicated by the red-filled circle points 1 to 3, highlighted on the calibration PO front. The PO trade-offs 1 and 3 correspond to the quality and diversity-maximizing calibration PO systems respectively whereas PO trade-off 2 represents a balanced diversity/quality PO system. The validation trade-off of the PO systems is indicated by the blue-hollow circle points 1–3 respectively. Note that none of the calibration PO systems is PO in the validation condition and that for each of the three pairs (1, 2), (1, 3), and (2, 3) the validation quality value of the first PO system in the pair exceeds the second system's corresponding value, whereas the reverse is true for the validation diversity value. It is thus not possible to decide whether the three PO systems cross-validate differently by looking only at the AIR/AJP trade-off values of the systems.

Finally, the figure also depicts the quantities used in obtaining the global relative validation achievement value for PO system 2. Thus, the blue-hollow square points 2q− and 2q+ (2d− and 2d+) show the minimum and maximum (over the set of feasible selection systems) attainable quality (diversity) value at the validation diversity (quality) level of PO system 2 in the validation condition. The difference between the quality (diversity) value of the points 2q+ (2d+) and 2q− (2d−) indicates the maximum gain (across all feasible systems) in quality (diversity) attainable at the diversity (quality) value of System 2 in the validation condition. From the validation diversity/quality trade-off value of System 2, equal to 0.33/1.35, and the minimum and maximum possible diversity value attainable at the validation quality value of 1.35 of the system, equal to 0.48 (cf. point 2d+) and 0.22 (cf. point 2d−) respectively, the relative validation diversity value of System 2 is obtained as (0.33–0.22)/(0.48–0.22) = .40. Similarly, but this time using the quality values of the points 2q+ and 2q−, the relative validation quality achievement of System 2 is equal to (1.35–1.12)/(1.47–1.12) = .67. Averaging both values results in the global relative validation achievement of System 2 of .535 = (.40 + .67)/2.

Further Comments on the Relative Validation Achievement Measure

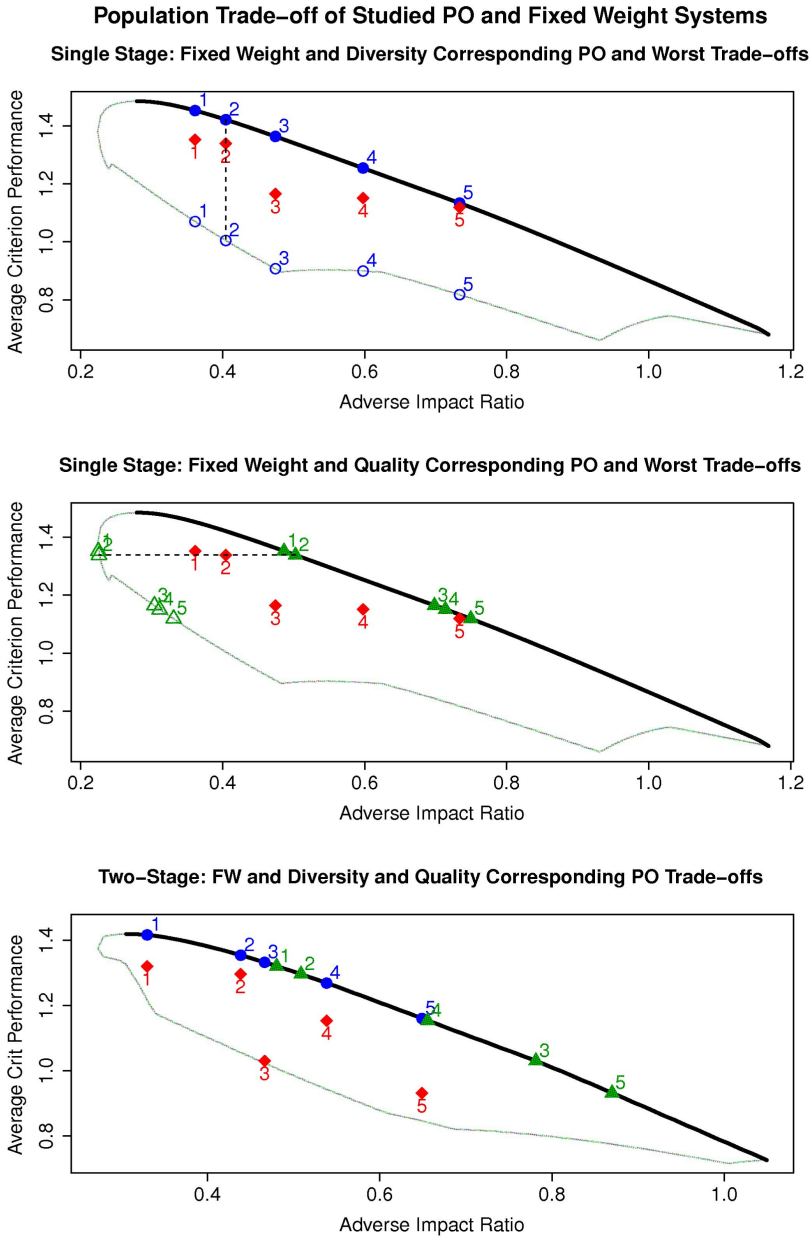
It is of key importance to distinguish the single-valued global relative validation achievement measure, which quantifies the *global level of achievement* of a system in a given validation condition, from the double valued validation diversity/quality trade-off of the system expressing the level of AIR and AJP achieved upon cross-validation. As the importance of the selection objectives is undetermined, the new measure assigns the same value to systems that vary in relative validation diversity and quality achievement but have the same total on the two components. However, note that this only means that the two systems show the same global level of achievement; it does not imply equal achievement (i.e., equal validation diversity/quality trade-offs).

The central purpose of the global relative achievement measure is to provide the means for addressing our research questions about how different PO systems perform upon cross-validation and how PO and FW systems compare upon cross-validation. We already noted that the measure achieves this purpose by avoiding the incommensurability problem. Here, we point to an additional consideration that further explains the format of the constituent relative diversity and quality achievement measures. When applied in any particular validation sample, the diversity (quality) value of a selection system stems from two sources, the first related to the system, which is the effect of interest, and the second related to the particularities of the validation sample. However, the impact of the latter source harbors a confound in that the effect varies across the quality (diversity) dimension and, hence, is different for different

PO and FW systems. In particular, the effect depends on the range of diversity (quality) values achievable (across the set of feasible systems) at each possible quality (diversity) level (cf., the irregular shape of the gamut of attainable trade-offs in the Figures 1 and 2). The rescaling components (i.e., the maximum/minimum possible diversity/quality values) used in the relative diversity and quality achievement measures serve to remove this sample specific confound and are therefore also sample specific. Any other choice for these component

Figure 2

Population Quality/Diversity Trade-Off of the PO and FW Selection Systems in the Single- and Two-Stage 5-Predictor Situation



Note. PO = Pareto-Optimal; FW = Fixed weight. The upper two panels refer to the single-stage 5-predictor situation, whereas the bottom panel refers to the two-stage predictor situation. In the panels, the red diamond points depict the population diversity/quality trade-off of the FW Systems detailed in Table 2. The blue-filled circle (green-filled triangle) points represent the population diversity/quality trade-off of the diversity (quality) corresponding PO systems. The curved solid lines enclose the gamut of attainable diversity/quality trade-off; whereas the bold solid line segments represent the front of PO trade-offs. The upper (middle) panel illustrates the computation of the relative quality (diversity) achievement of FW system 2. See the online article for the color version of this figure.

values (e.g., using a constant value of zero for the minimum possible AIR) would be inappropriate. Also, note that the rescaling components are well-defined and free of any arbitrariness because they all pertain to the validation condition (sample) and are obtained across the set of feasible selection systems that must be defined as a first step in deriving PO systems (cf. De Corte et al., 2007, 2011).

Applying the relative achievement measures is not without challenges, however. First, the relative diversity (quality) achievement measure is undefined when the denominator of the measure equals zero, but the problem is essentially limited to settings involving small applicant samples (100 or less) with a low minority representation (.20 or less) and a small selection ratio (.10 or less). For larger sample sizes zero denominator values quickly become increasingly rare so that the limitation is not really critical, especially because the practice of PO design systems is not intended for such small sample situations (cf. De Corte et al., 2011). Second, the calculation of the relative achievement values of the systems in a validation sample poses a computational challenge. Whereas De Corte et al. (2020) partly addressed this computational problem, we solved it via a modified version of the hybrid ant colony optimization algorithm (Schlueter et al., 2009). Details of the new algorithm and an additional study investigating its accuracy are reported in Appendix A. Our study is, therefore, the first to exploit the full potential of the relative achievement measures in studying the cross-validity of PO and FW systems.

Computational Cycle

To address our research questions a computational cycle consisting of four steps was repeated 2000 times in each cell of the design. Appendix B details the steps of the computational cycle. Here, we note that each cycle starts from calibration and validation applicant pool data sampled with a fixed .2/.8 minority/majority proportion from the predictors/criterion population distribution with mean and correlation structure given in Table 1. At the end, each cycle returns the calibration and validation trade-off, the relative validation diversity and quality achievement as well as the global relative validation achievement of 5 FW and 10 PO systems. The five FW systems vary only across the two selection settings (i.e., single- and two-stage selection) and are chosen so that the population diversity/quality trade-off of the systems varied substantially across the diversity/quality dimension. The 10 PO systems consist of 5 pairs, each of which corresponds to one of the FW systems. The first PO system within a pair is the *diversity equivalent* PO system, whereas the other is the *quality equivalent* system. The diversity equivalent PO system has the same population diversity level as the FW system it corresponds to, whereas the quality equivalent PO system shares the quality level of the corresponding FW systems. Table 2 details the studied FW systems and Figure 2 depicts the population diversity/quality trade-offs of the FW and PO systems in both the single- and two-stage selection situation.

Table 2
Studied Fixed-Weight (FW) Selection Systems

5-Predictor selection situation				
		Two-stage		
FW system	Single-stage	Stage 1	Stage 2	RRate
1	1 + 2 + 3 + 4	1 + 3	4	.40
2	1 + 2 + 3 + 4 + 5	1 + 2 + 3	4 + 5	.25
3	1 + 3 + 4 + 5	1 + 3	5	.36
4	2 + 3 + 4 + 5	2 + 3	4	.25
5	2 + 3 + 5	1 + 2 + 3	5	.40

Note. Predictors: 1 = Cognitive Ability; 2 = Structured Interview; 3 = Conscientiousness; 4 = Biodata (BI); 5 = Integrity (IN). RRate: Retention rate after stage 1; Final selection rate is 0.1. Each row of the table identifies a particular FW system used in either the single- or two-stage setting. For example, the first row indicates that in the single-stage setting FW system 1 corresponds to using the unit weighed composite of the predictors 1, 2, and 3; whereas in the two-stage setting the system corresponds to using the unit weighed composite of the predictors 1 and 3 in stage 1 and predictor 4 in stage 2.

Results

Research Question 1: Do Different PO Systems Show Comparable Levels of Relative Validation Achievement?

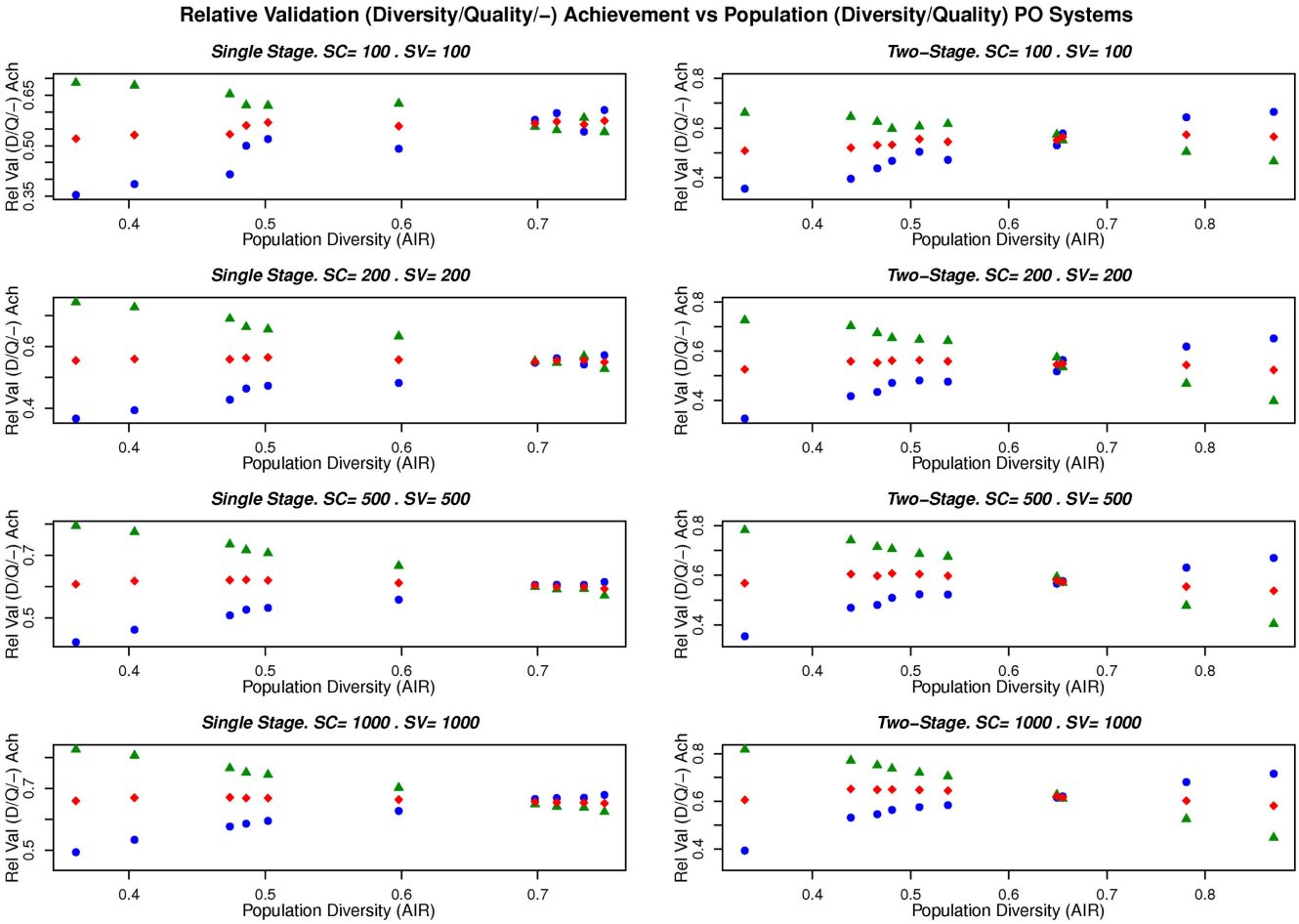
We started by analyzing the relationship between the population diversity trade-off value and the average (across the 2000 repetitions) relative validation (diversity/quality/-) achievement of the PO systems studied in the selection environments for the different calibration/validation sample-size conditions. The panels in Figure 3 display the results of the analyses for both selection settings and a representative subset of the sample-to-sample cross-validation conditions. Per panel, the scatter plot[5] of the green triangle, the blue circle, and the red diamond points shows the relationship between the population diversity trade-off and the relative validation quality, the relative validation diversity, and the global relative validation achievement of the PO systems, respectively. In general, the plots confirm that PO systems with a higher (lower) population diversity (quality) trade-off value show a higher (lower) relative validation diversity (quality) achievement value. The plots also indicate that all different PO systems show approximately the same global relative validation achievement value. Finally, we inspected the plots corresponding to the population-to-sample and sample-to-population cross-validation conditions: Again, all different PO systems show a fairly equal global relative validation achievement in each selection environment.

The above results imply that the question of the cross-validity potential of PO systems can be addressed without making a distinction among different PO systems: The average (across the 10 PO systems and the 2000 repetitions) global relative validation achievement of the systems provides an adequate summary of their cross-validation potential. The upper part of Table 3 summarizes these average global relative validation achievement values, and the middle and lower part of the table detail the averages obtained for the population-to-sample and the sample-to-population cross-validation conditions.

[5] The scatter plots for the calibration/validation sample size of 100 are based on some 60% of the replications because the relative validation diversity achievement is undefined in the remaining 40% of the cases.

Figure 3

Relationship Between the Population Diversity and the Relative Validation Diversity, the Relative Validation Quality and the Global Relative Validation Achievement of the PO Systems



Note. FW = Fixed weight. In the panels, the population diversity corresponds to the adverse impact ratio (AIR). The blue circle points indicate the relative validation diversity achievement, the green triangle points correspond to the relative validation quality achievement and the red diamond points indicate the global relative validation achievement of the PO systems. Each panel corresponds to a particular sample-to-sample cross-validation scenario, with SC and SV referring to the size of the calibration and validation sample, respectively. See the online article for the color version of this figure.

Not surprisingly, the PO systems show both a substantially higher sample-to-population and a slightly higher[6] population-to-sample cross-validity compared to the corresponding sample-to-sample cross-validity. The cross-validity values also increase for higher calibration and/or validation sample sizes, but the sample-to-sample and the population-to-sample cross-validity values remain rather modest even in the largest calibration and/or validation size conditions. The global relative validation achievement of these systems hovers on average between .54 and .66 in the sample-to-sample cross-validation conditions.

We also studied the variability of the global relative validation achievement values of the PO systems across the repetitions. Results showed that the different PO systems have a virtually equal sampling variability. Table 3 reports the average value (across PO systems and repetitions) of the standard deviation of the global relative validation achievement between brackets. As the maximum possible standard deviation is .29 (i.e., the standard deviation of the uniform distribution on the 0–1 interval), Table 3 indicates that the cross-validity of the PO systems varies considerably (from near zero to near one) across the repetitions for all cross-validation cases, thereby extending De Corte et al.'s (2020) findings to the sample-to-sample cross-validity case. However, the variability of the global relative achievement decreases with larger calibration and/or validation sample sizes.

[6] The similarity of the population-to-sample and the sample-to-sample results can be explained by first noting that the PO systems in step 3 of the computational cycle are computed using the procedure presented in De Corte et al. (2006), implying that the PO systems are not PO with respect to the sample but with respect to the population that corresponds to the sample. Second, repeating the computational cycle many times assures that the average of the latter population statistics and the corresponding averages of the relative achievement values of the PO systems converge to the initial population statistics and the corresponding population to sample relative achievement values respectively.

Table 3
Sample-to-Sample, Population-to-Sample and Sample-to-Population Cross-Validity of PO and Fixed-Weight (FW) Systems

	Sample-to-sample cross-validity							
	CS = VS = 100		CS = VS = 200		CS = VS = 500		CS = VS = 1000	
Situation	PO	FW	PO	FW	PO	FW	PO	FW
Single-stage	.55 (.26)	.54 (.26)	.56 (.20)	.52 (.19)	.61 (.15)	.53 (.14)	.66 (.13)	.55 (.12)
Two-stage	.54 (.26)	.49 (.27)	.55 (.21)	.45 (.21)	.58 (.16)	.44 (.16)	.63 (.14)	.43 (.14)
	Population-to-sample cross-validity							
	CS = Inf; VS = 100		CS = Inf; VS = 200		CS = Inf; VS = 500		CS = Inf; VS = 1000	
Situation	PO	FW	PO	FW	PO	FW	PO	FW
Single-stage	.60 (.26)	.55 (.26)	.60 (.20)	.52 (.19)	.63 (.15)	.53 (.14)	.68 (.13)	.55 (.12)
Two-stage	.57 (.26)	.49 (.27)	.57 (.21)	.46 (.21)	.59 (.15)	.44 (.14)	.63 (.13)	.43 (.12)
	Sample-to-population cross-validity							
	CS = 100; VS = Inf		CS = 200; VS = Inf		CS = 500; VS = Inf		CS = 1000; VS = Inf	
Situation	PO	FW	PO	FW	PO	FW	PO	FW
Single-stage	.79 (.17)	.70 (.15)	.86 (.12)	.70 (.15)	.94 (.06)	.70 (.15)	.97 (.03)	.70 (.15)
Two-stage	.79 (.20)	.40 (.25)	.87 (.14)	.40 (.25)	.94 (.07)	.40 (.25)	.96 (.04)	.40 (.25)

Note. CS = Calibration sample size; VS = Validation sample size; Inf = Infinity; PO = Pareto-Optimal; FW = Fixed weight. In each cell, the first value corresponds to the average (across the repetitions and across the different PO or FW systems) global relative validation achievement and the second value, between parenthesis, to the standard deviation of the global relative validation achievement.

Research Question 2: Do PO Systems Achieve Higher Levels of Relative Achievement Upon Cross-Validation Than FW Systems?

Below we discuss the results obtained using the relative achievement measures, postponing the presentation of the findings related to the AJP/AIR trade-off until the next section because of the obvious relevance of the latter findings when addressing the third research question on how the calibration trade-off of the FW and PO systems evolves under cross-validation. In addition to the average global relative validation achievement values of the PO systems, Table 3 also summarizes the corresponding average values of the FW systems (across the five systems and the repetitions) in the different cross-validation conditions. The comparison of both sets of values reveals that the average global relative validation achievement of the PO systems exceeds the corresponding average of the FW systems for sample sizes of at least 100.[7] The difference in cross-validation potential also grows for larger sample sizes as PO systems cross-validate better for larger sample-size conditions, whereas the FW systems show no such effect. The difference in average cross-validity potential is small to modest, however, except for the sample-to-population cross-validity.

The above result only partly answers the question about the cross-validity potential of PO and FW systems, however. Whereas the PO systems all show approximately the same global relative validation achievement, this is not the case for FW systems. The global relative validation achievement of the FW systems correlates substantially with their *population global relative achievement* (i.e., the global relative achievement in the population). Hence, a complete comparison of PO and FW systems must focus not only on the average global relative validation achievement of the two, but also on the difference between the individual FW and their two corresponding PO systems.

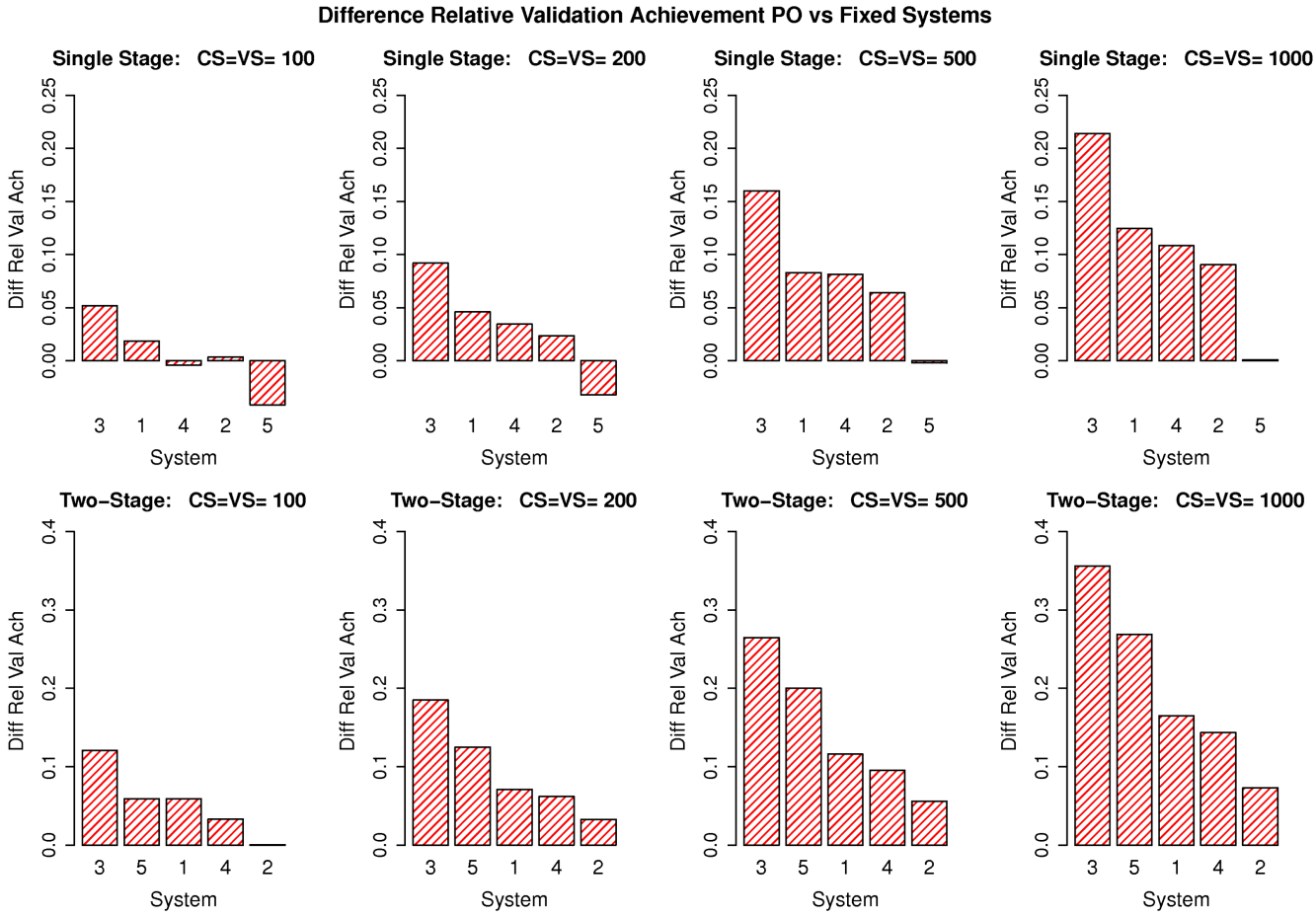
Figures 4 and 5 compare the individual FW and the corresponding diversity equivalent and quality equivalent PO systems. The bar plots in Figure 4 represent the difference in the average global relative validation achievement of the PO and FW systems, whereas Figure 5 addresses the proportion (across repetitions) with which FW systems result in a lower (higher) global relative validation achievement value. In Figure 5, the blue solid lines (red dashed lines) connect the proportion of times that the global relative validation achievement of the two companion PO systems (FW system) exceeds the achievement of the corresponding FW system (two companion PO systems). In the panels of both figures, the FW systems are ordered from left to right according to increasing population global relative achievement explaining why the order of the FW systems differs from one selection setting to another (i.e., the single-stage vs. the two-stage setting).

In general, the Figure 4 plots reveal an inverse relationship between the population relative achievement of the FW systems and the average difference in global relative validation achievement between the system and its corresponding two PO systems. The Figure 5 plots show a similar inverse relationship, but this time with respect to the likelihood that the companion PO systems result in a higher global relative validation achievement than the FW system. Although not shown in a separate figure, the plots for the studied sample-to-population and population-to-sample cross-validation condition further confirm the conclusion that the difference in cross-validation potential of the FW and corresponding PO systems, in favor of the PO systems, decreases for FW systems with increasing population global relative achievement. Also, FW systems with a very high global population relative achievement cross-validate as well or even better than corresponding PO systems in medium-to-large calibration/validation sample-size conditions (see results of FW system number 5 in the upper row panels of Figures 4 and 5).

[7] However, note that the results for the $N = 100$ condition are less reliable because they are based on only about 60% of the replications due to undefined relative achievement values.

Figure 4

Difference in Global Relative Validation Achievement Between Corresponding Sets of PO and FW Systems



Note. PO = Pareto-Optimal; FW = Fixed weight. Each panel corresponds to a particular sample-to-sample cross-validation condition, with CS and VS indicating the size of the calibration and validation sample respectively. The bars show the difference in global relative validation achievement for corresponding sets of PO and FW systems. In each panel, the systems are ordered in increasing population relative achievement along the horizontal axis and the numbering on the horizontal axis corresponds to the numbering of the systems in Figure 2. See the online article for the color version of this figure.

Research Question 3: How Does the Achievement of PO and FW Systems Evolve Under Cross-Validation?

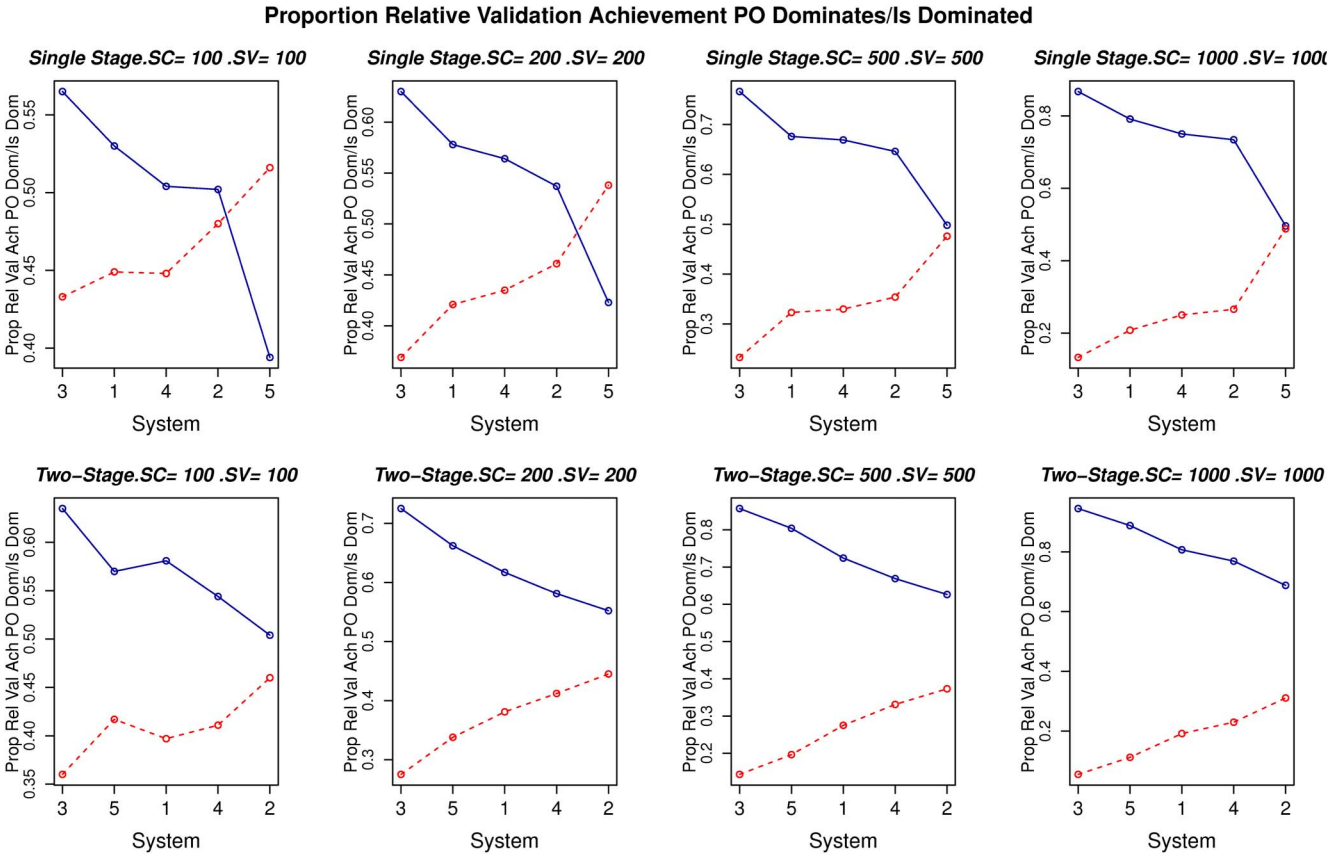
Apart from knowing that PO systems may promise a better global relative validation achievement than FW systems, it is also important to assess how the diversity/quality trade-off of the PO and FW systems behaves under cross-validation. We first address the issue in terms of the average (across the repetitions) calibration and validation trade-off of the different systems. Figure 6 displays these averages for a selected set of sample-to-sample cross-validation scenarios in the studied selection settings. In each panel, the blue-filled circle points and the green-filled triangle points correspond to the average calibration trade-off of the diversity equivalent and the quality equivalent PO systems respectively, whereas the blue-hollow circle points and the green-hollow triangle points indicate the corresponding average validation trade-off of the systems. In turn, the red-hollow diamond points show the average validation trade-off of the FW systems. Finally, the solid and dashed lines in each panel represent the *interpolated front* of the average calibration and validation trade-offs of the PO systems in the cross-validation scenarios.

Figure 6 shows that the shrinkage results obtained by Song et al. (2017) for sample-to-population cross-validation generalize to sample-to-sample cross-validation. The average calibration trade-off of the PO systems typically dominates the corresponding validation trade-off and the domination is more pronounced with smaller samples. The claim that PO systems with a high (low) population diversity (quality) trade-off value show more (less) diversity (quality) than quality (diversity) shrinkage also receives support, especially in single-stage selection. Finally, ancillary analyses, similar to those reported in Figure 6, confirmed the above findings for the sample-to-population but not for the population-to-sample scenarios.[8]

[8] This is explained by the fact that equal deviations in sample predictor composite effect sizes result in quite different AIR deviations when the sample predictor composite effect size is smaller as compared to when it is larger than the population predictor composite effect size.

Figure 5

Probability That the Global Relative Validation Achievement of PO (FW) Systems Exceeds the Global Relative Validation Achievement of the Corresponding FW (PO) Systems



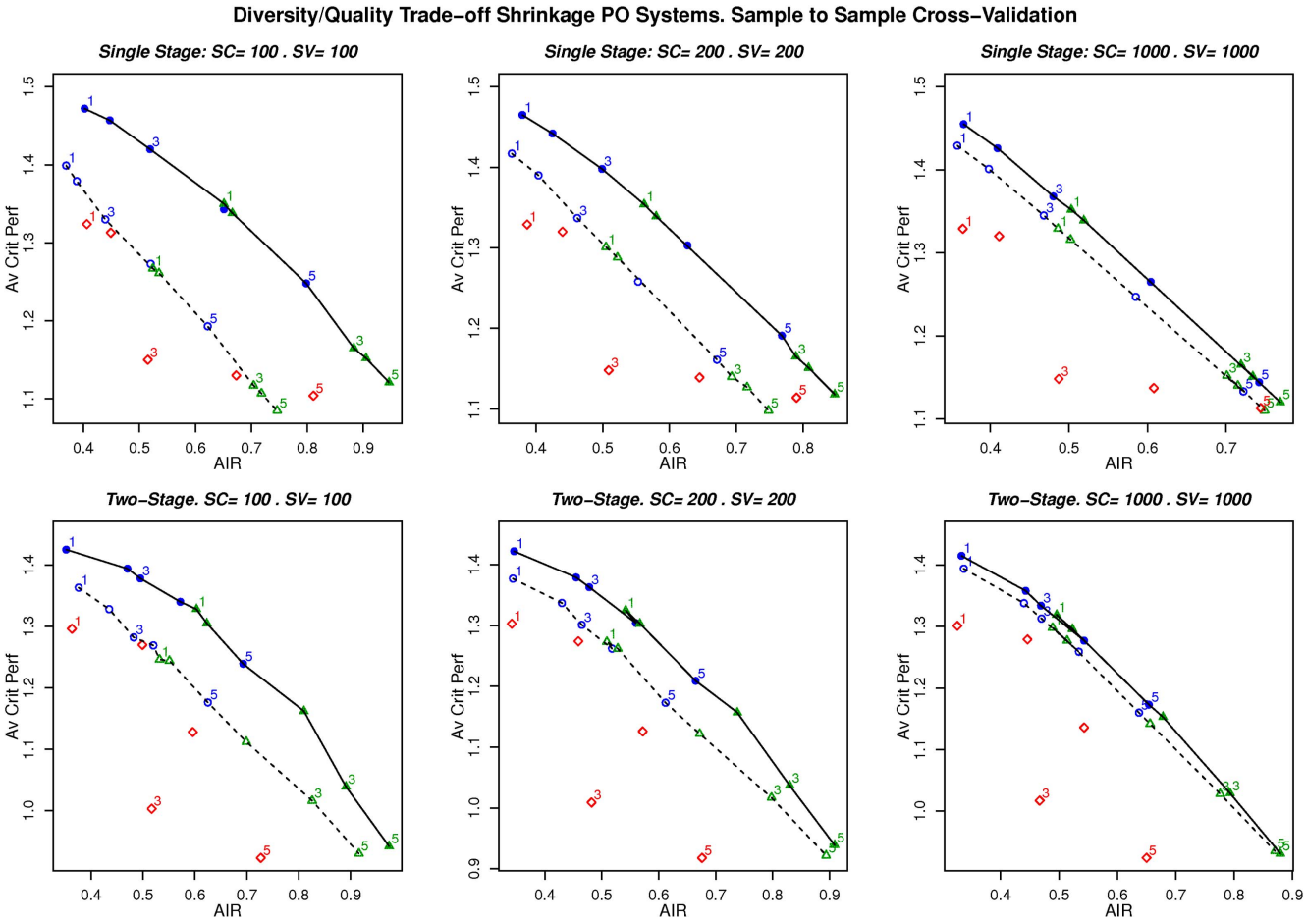
Note. PO = Pareto-Optimal; FW = Fixed weight. Each panel corresponds to a particular sample-to-sample cross-validation condition, with CS and VS indicating the size of the calibration and validation sample, respectively. The blue circle points, connected by a blue solid line (red circle points connected by a red dashed line) indicate the probability that the global relative validation achievement of PO (FW) systems exceeds the global relative validation achievement of the corresponding FW (PO) systems. The numbering on the horizontal axis corresponds to the numbering of the systems in Figure 2 and the systems are ordered in increasing population relative achievement along the horizontal axis. See the online article for the color version of this figure.

Although Figure 6 suggests considerable diversity and quality shrinkage of the calibration PO trade-offs under cross-validation, especially in the smallest applicant sample sizes of 100 and 200, the actual practical importance of the shrinkage remains rather insignificant. To illustrate this, consider, for example, the shrinkage results obtained in the single-stage selection situation for the diversity equivalent PO system number 3 in the 100, 200, and 1000 applicant pool size condition (cf. the blue-filled and the blue-hollow circle point 3 in the upper panels of Figure 6). The average (across repetitions) calibration diversity/quality (i.e., the AJP/AIR) trade-off of the PO system is 1.42/.52, 1.40/.50, and 1.37/.48 in the 100, 200, and 1000 sample-size conditions respectively; whereas the corresponding average validation trade-offs are 1.33/.44, 1.34/.46, and 1.35/.47. So, even in the smallest sample-size condition of 100, the drop in AJP is only 6% (i.e., 100 (1.42–1.33)/1.42 = 6). For the same sample condition, the corresponding drop in AIR, when translated to the corresponding drop in a number of minority hires, is equally unimpressive. Given the .2/.8 proportional representation of the minority/majority candidates and the .1 total selection rate, the average (across repetitions) calibration AIR trade-off value .52 of PO system 3 corresponds to the selection of 1.15 minority applicants; whereas the corresponding average validation diversity trade-off value of .44 of the system translates to the hiring of 1 minority applicant. In practice, this means that the same number of 1 minority applicant is expected to be hired in both the calibration and the validation condition.

Figure 6 also illustrates the potential of using the AJP/AIR trade-off instead of the relative achievement measures to address the research question on the cross-validation potential of PO and FW systems. In particular, the panels in the figure show that the average validation trade-off of the FW systems typically lies below the interpolated front of the average PO validation trade-offs and this is more obvious in the larger calibration/validation sample-size conditions and for FW systems with a low population relative achievement. The result indicates that PO systems with the same average validation diversity (quality) trade-off as an FW system have on average a higher validation quality (diversity) trade-off and that the average trade-off of the FW systems is dominated by a subset of the

Figure 6

Calibration and Validation Diversity)/Quality Trade-Off of the PO Systems for a Representative Set of Sample-to-Sample Cross-Validation Conditions



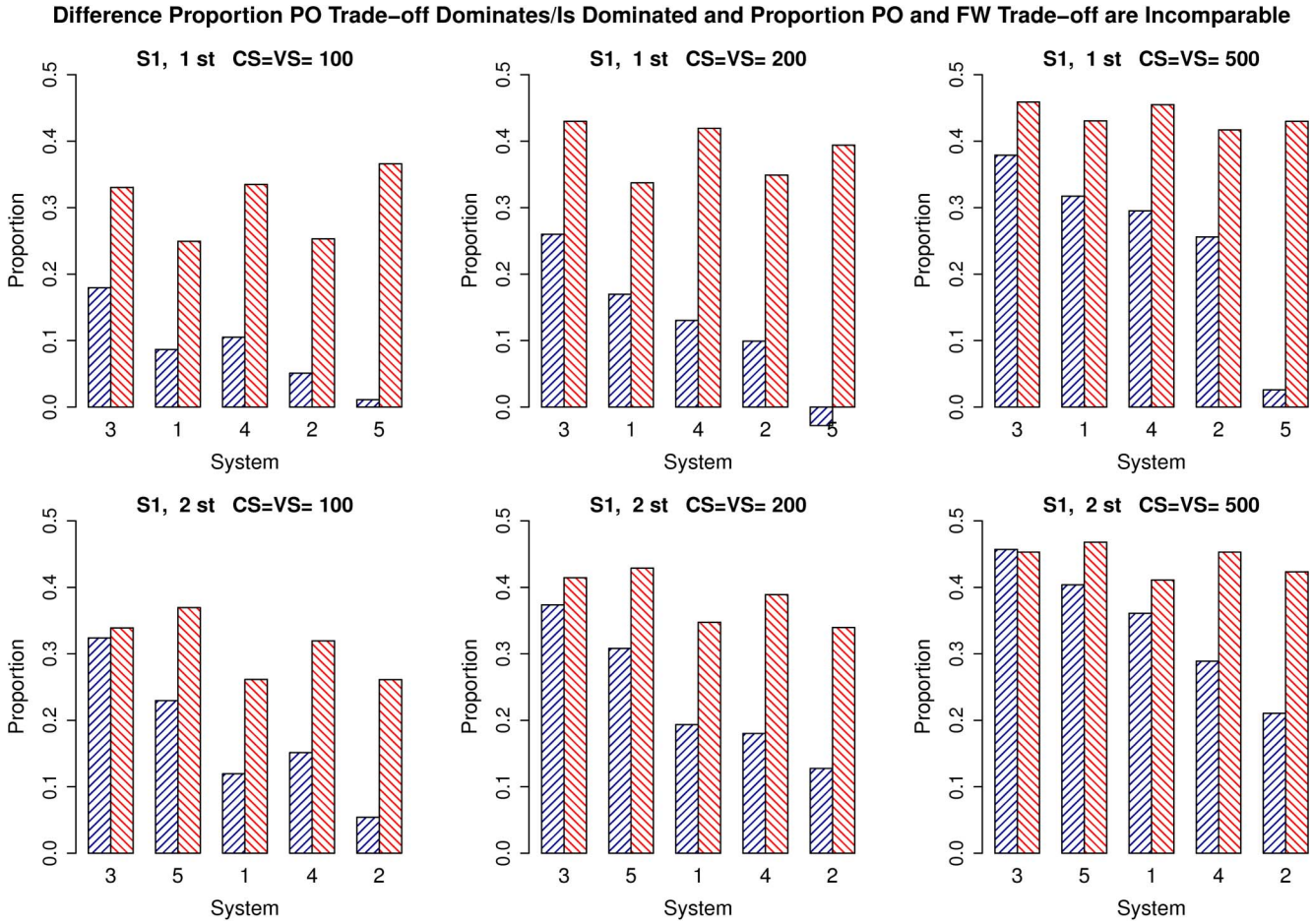
Note. PO = Pareto-Optimal. Each panel corresponds to a particular sample-to-sample cross-validation condition, with CS and VS indicating the size of the calibration and validation sample, respectively. The filled circle and triangle points, connected by solid line segments (hollow circle and triangle points, connected by dashed-line segments) depict the calibration (validation) diversity/quality trade-off of the PO systems. Diversity is gauged by the adverse impact ratio (AIR), whereas quality corresponds to the average job performance of the selected applicants. See the online article for the color version of this figure.

interpolated PO trade-offs. Yet, these findings essentially relate to the average trade-off of the FW systems as compared to *interpolated* average PO trade-offs and not to the average trade-offs of the actually studied PO systems. Neither do they permit a general conclusion about which of the systems, PO versus FW, cross-validate better because the respective trade-offs of the systems are often incomparable (i.e., PO superior on one dimension and FW superior on the other). In fact, across all possible pairs of FW and actually studied PO systems the percentage of incomparable average validation trade-offs lies between 76 and 86 across the different sample-to-sample cross-validation conditions.

The findings reported in Figure 7 further detail the incomparability issue at the level of the individual trade-offs of the FW and the two corresponding PO systems within each repetition instead of at the level of the average (across the 2000 repetitions within each cell of the design) trade-offs across all pairs of FW and PO systems. In the panels of the figure, the red downward shaded bars represent the proportion (across) the trials that the FW systems (ordered from left to right according to increasing population relative achievement value) and the two corresponding PO systems have an incomparable trade-off. These proportions show that the PO and FW systems have incomparable trade-offs in roughly 30%–50% of the trials even when considering only the triads of the FW and the diversity and quality equivalent PO systems. For the remaining cases where the validation trade-offs of the FW and the corresponding PO systems are comparable, the proportions represented by the blue upside shaded bars show that the validation trade-off of the FW systems is more often dominated by the corresponding PO systems than the reverse is the case. The excess domination of the validation trade-off of the PO systems also increases with sample size, but FW systems with a higher population relative achievement result less frequently in dominated validation trade-offs than FW systems with a lower population relative achievement.

Figure 7

Difference Proportion PO Validation Trade-Off Dominates/Is Dominated by the Corresponding FW Trade-Off and Likelihood of Incomparable Trade-Offs



Note. PO = Pareto-optimal; FW = Fixed weight. Each panel corresponds to a particular sample-to-sample cross-validation condition, with CS and VS indicating the size of the calibration and validation sample respectively. The numbering on the horizontal axis corresponds to the numbering of the systems in Figure 2 and the systems are ordered in increasing population relative achievement along the horizontal axis. In each panel, the blue upward shaded bars indicate the difference in proportion that the PO validation trade-off dominates/is dominated by the validation trade-off of the corresponding FW system; whereas the red downward shaded bars show the likelihood that the validation trade-off of the PO and FW systems are incomparable. See the online article for the color version of this figure.

Additional Studies

Our Supplemental Online Material reports additional studies using the 9-predictor selection situation of De Corte et al. (2020). We also examined the cross-validation potential of PO and FW systems for a selection rate of .4 instead of .1. Results echoed and extended our above findings by showing that also with a selection rate of .4 the PO systems continue to maintain an advantage for cross-validation conditions with a validation applicant pool of at least 100, although the advantage is somewhat smaller in the 9-predictor situation.

Discussion

Key Contributions

The article presents the first all-round examination of the cross-validity potential of PO as compared to FW selection systems. As argued above, this examination requires developing and using a new set of measures (i.e., the relative achievement measures) for gauging the cross-validity of biobjective selection systems. Shrinkage and trade-off measures, as applied in previous research on the cross-validity of both single- and biobjective selection systems, fall short to fully capture the cross-validity of biobjective systems because these systems simultaneously pursue two goals (i.e., quality and diversity) that relate to incommensurable dimensions. As a consequence, separate results on the quality and diversity shrinkage of different selection systems under cross-validation cannot be combined to a single number quantifying the extent that these systems cross-validate. In contrast, our study is the first to fully implement the new relative achievement measures and, in particular, the global relative validation achievement measure, that overcomes the incommensurability issue that plagues the study of the cross-validity of biobjective selection systems.

Our study is, therefore, also unprecedented in comparing the cross-validity of different PO systems as well as in comparing the cross-validity of PO and FW systems *for any pair of PO or PO and FW systems*. Whereas previous research on the cross-validity of biobjective systems (e.g., Song et al., 2017) necessarily stops with the separate evaluation of the extent of diversity and quality shrinkage under cross-validation and ends up with two sets of incomparable shrinkage results, our study goes the full mile by virtue of using the global relative validation achievement measure. However, note that the relative achievement measures are not used when studying the way in which the diversity/quality trade-off of selection systems evolves under cross-validation. For this, the present study adopts the same procedure as previous research by studying the expected trade-off of the systems in the different cross-validation conditions.

Besides presenting and using new measures for quantifying the cross-validity potential of biobjective selection systems, the article offers two more unique contributions. Thus, we extend previous results on the sample-to-sample cross-validity of regression and FW systems in the single-objective selection context (e.g., Raju et al., 1999; Schmidt, 1971) to corresponding results for PO and FW systems in the biobjective situation. As selection systems are always initially developed on (aggregated) sample data and subsequently always applied to finite-sized applicant groups, sample-to-sample cross-validity, instead of population-to-sample or sample-to-population cross-validity, is the most informative type of cross-validity for the selection practitioner.

Second, the focus on biobjective selection systems leads to studying and answering research questions that have no counterpart in the single-objective cross-validity literature. In the latter literature, the cross-validity issue boils down to a single research question about how the cross-validity of the regression composite compares to the cross-validity of the FW composite. In contrast, the same issue implies multiple and new research questions in the biobjective selection context as many (in fact, an infinite number of) different PO systems are now possible, instead of only a single regression system. One of these new questions is whether the different PO systems have approximately the same cross-validation potential (our first research question). Other additional, new research questions concern the relation between the cross-validity of FW and PO systems (our second research question). Answering the latter question requires a more elaborate procedure than the procedure used to address the corresponding question in the single-objective selection context. For a start, we now have to compare FW and multiple PO systems. Also, as PO systems vary across the range of feasible diversity and quality trade-off values, it is natural to compare their cross-validity to FW systems that also vary on these dimensions. This implies studying not only the FW system where *all available* predictors are equally weighted (as is the case in the previous cross-validity literature in the single-objective selection context) but also FW systems where only a subset of the available predictors are equally weighted to form the selection predictor composite. As discussed in the next section, examining the considerably more complex comparison of the sample-to-sample cross-validity of FW and PO systems not only leads to new insights that apply specifically to the biobjective selection context, but also to a renewed interest for the traditional cross-validity research in the single-objective situation.

Main Findings and Conclusions

The first main finding is that different PO systems all show approximately the same global relative validation achievement value in each of our conditions. As a key implication, selection practitioners retain all options when choosing between the different PO systems, as these systems cross-validate equally well. Deciding between the PO systems remains a value-based judgment reflecting where organizations stand on trading off diversity and quality. We further discovered that the cross-validation potential of PO systems, as gauged by the relative validation achievement criterion, is quite modest (i.e., in the .54−.68 range) in both the sample-to-sample and population-to-sample cross-validation scenarios, even for validation sample sizes of 1000.

Second, our results confirm and generalize the claim of previous research that PO systems maintain an advantage over FW systems under cross-validation. The results based on the validation trade-off criterion, although limited to pairs of FW and PO systems with a comparable validation trade-off, as well as the more general results obtained when using the relative achievement measures corroborate the claim. In particular, the latter results show that the average global relative validation achievement of the PO systems exceeds the corresponding average of the FW systems for applicant pool sizes of at least 100. However, we also found that FW systems are differently outperformed by the corresponding PO systems. Whereas all PO systems result in more or less the same global relative validation achievement, FW systems show considerable variability, depending on the population relative achievement value of the systems. The higher the population relative achievement value of an FW system, the higher its global relative validation achievement, such that FW systems with a (very) high population relative achievement continue to outperform corresponding PO systems for calibration/validation sample sizes up to 500 and even larger. This result leads to perhaps the study's single most important advice to the selection practitioner when facing the choice between alternative selection system designs. The practitioner should, as a first task, estimate the diversity/quality trade-off and the global relative achievement value of the feasible FW systems using available predictor/criterion effect size and correlation data, whether obtained locally, from meta-analyses, or a combination of both. A program to perform this task can be downloaded from https://users.ugent.be/~wdecorte/software.html. If any of the systems shows a high to very high global relative achievement value (i.e., a value of at least .90), if an applicant pool of at most 1000 is anticipated, and if the FW system shows an acceptable diversity/quality trade-off, then it should probably be preferred above any PO system. The latter condition is of key importance, however. If an FW meets the former two conditions, but results in a diversity/quality trade-off that runs against the organization's position on trading off diversity for quality, then it is still better to decide in favor of a PO system that shows the desired trade-off as this system will cross-validate almost as well as the FW system.

The result that the cross-validity (whatever the type of cross-validity) of the FW systems varies proportionally with the population relative achievement of the FW systems mirrors a finding already established in the earlier studies on the cross-validation potential of single-objective selection systems. Although frequently left unmentioned at least some of these studies (e.g., Cattin, 1978; Einhorn & Hogarth, 1975) explicitly indicate that the tendency for

regression composites to cross-validate better than FW systems relates to the comparative population validity of the regression and the FW systems.

As to the finding that PO systems cross-validate on average equal or better than FW systems for applicant samples as less as 100 we see two possible explanations. First, the average of the ratio between the population global relative achievement of the FW and the corresponding PO system was quite low in our study and was definitely substantially lower than the corresponding ratio between the FW and the regression population validity in many earlier cross-validity studies in the single-objective cross-validity literature. For example, in the study of Raju et al. (1999), the ratio between the FW population validity and the regression composite population validity was .466/.479 = .973, whereas in our study the average of the ratio between the population global relative achievement of the FW and the corresponding PO system was only .704. Second, we studied only PO systems using nonnegatively weighted predictor composites because negative weights imply that the scores of the applicants on some of the valid predictors are counted against the candidates. Because constrained linear composites (as compared to unconstrained ones) have lesser tendency for overfitting (cf., the logic behind ridge, lasso, and elastic net regression) PO systems based on nonnegatively constrained composites are expected to cross-validate better than PO systems using unconstrained predictor composites. As a further consequence, the sample sizes at which constrained PO systems cross-validate on par with FW systems will be lower. Given the above argument, surprisingly, none of the previous research on the cross-validity of regression and FW composites in the single-objective selection context studied and compared FW and nonnegatively constrained regression composites.

The results reported in Table 3 also show that the variability of the relative validation achievement of PO systems across replications only minimally exceeds the corresponding variability of FW systems. Although the variability decreases for larger validation applicant pool conditions, the achievement values of both PO and FW systems span almost the entire maximum possible 0–1 range, even for the largest studied validation sample-size conditions.

Finally, we also examined whether PO systems are expected to result in a better diversity/quality trade-off than FW systems under cross-validation. The expected quality/diversity trade-off of FW systems is typically below the interpolated front of the PO trade-offs, but the result is less clear for smaller validation sample sizes and a larger number of predictors. Although the result is by itself insufficient to justify the general conclusion that PO systems cross-validate better than FW systems because it applies only to PO and FW systems with a comparable (average) validation trade-off it indicates that it is virtually always possible to construct PO systems that cross-validate better than any given FW system. When the trade-offs of the PO and the FW systems are comparable, we finally showed that the likelihood that the trade-off of FW systems is dominated by the trade-off of the corresponding PO systems is higher than the reverse, especially in cross-validation conditions with larger validation applicant pools and a smaller number of predictors.

Limitations and Avenues for Further Research

Our study does not address the possibility that the calibration and validation conditions relate to different populations and that the applicant pools in the calibration and validation conditions come from different populations. The calibration and/or validation applicant pools might also not be random samples from the population. For example, applicant pools might be preselected or prone to self-selection, leading to range restriction and/or different effect sizes for selection predictors in the validation condition. As another limitation, our study left out the possibility of applicant withdrawal or job refusal and that the focal criterion behavior is multidimensional instead of unidimensional. Future research can tackle these issues (e.g., by mimicking the preselection/self-selection process when generating validation pools).

Future research should focus on reducing the considerable loss in relative achievement of PO systems in future applications. Inspired by general linear modeling developments (e.g., Putka et al., 2018), the introduction of regularization techniques in the computation of PO systems might be particularly promising (see Song, 2018). As a final avenue of future research, we suggest the further development of methods for computing selection systems that are PO with respect to several, instead of just one protected group (cf. Song & Tang, 2020, for an initial approach) and the implementation of these methods in the study of the cross-validity of PO systems.

Conclusion

Our study provides much-needed evidence for claims (cf. De Corte et al., 2020; Song et al., 2017) that PO systems maintain an advantage over FW systems under cross-validation. We conducted an all-round examination of the cross-validation potential of these systems by examining not only sample-to-population, population-to-sample but also sample-to-sample cross-validation scenarios. When selections involve sufficiently large applicant pools (at least 100 applicants), PO systems had on average a higher cross-validity potential than the corresponding FW systems. Yet, even for applicant pools as large as 1000, in limited situations FW systems may match the merits of PO systems and we provide selection specialists with a straightforward procedure to decide whether FW systems may offer a comparable or better cross-validation potential than the PO systems.

References

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*(4), 689–709. https://doi.org/10.1177/1094428106294734

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

Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical & Statistical Psychology*, *28*(1), 79–87. https://doi.org/10.1111/j.2044-8317.1975.tb00550.x

Cattin, P. (1978). A predictive-validity-based procedure for choosing between regression and equal weights. *Organizational Behavior and Human Performance*, *22*(1), 93–102. https://doi.org/10.1016/0030-5073(78)90007-7

Cattin, P. (1980). Estimation of the predictive power of a regression model. *Journal of Applied Psychology*, *65*(4), 407–414. https://doi.org/10.1037/0021-9010.65.4.407

Cortina, J. M., Aguinis, H., & DeShon, R. P. (2017). Twilight of dawn or of evening? A century of research methods in the Journal of Applied

Psychology. *Journal of Applied Psychology*, *102*(3), 274–290. https://doi.org/10.1037/apl0000163

De Corte, W., Lievens, F., & Sackett, P. R. (2006). Predicting adverse impact and mean criterion performance in multistage selection. *Journal of Applied Psychology*, *91*(3), 523–537. https://doi.org/10.1037/0021-9010.91.3.523

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

De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing Pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology*, *96*(7), 907–926. https://doi.org/10.1037/a0023298

De Corte, W., Sackett, P. R., & Lievens, F. (2020). Robustness, sensitivity, and sampling variability of Pareto-Optimal selection systems to address the quality-diversity trade-off. *Organizational Research Methods*, *23*(3), 535–565. https://doi.org/10.1177/1094428118825301

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, *13*(2), 171–192. https://doi.org/10.1016/0030-5073(75)90044-6

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

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689–732. https://doi.org/10.1177/1094428117697041

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement*, *23*(2), 99–115. https://doi.org/10.1177/01466219922031220

Roth, P. L., Switzer, F. S., III, 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*(4), 899–935. https://doi.org/10.1111/j.1744-6570.2011.01231.x

Rupp, D. E., Song, Q. C., & Strah, N. (2020). Addressing the so-called validity-diversity trade-off. Exploring the practicalities and legal defensibility of Pareto-optimization for reducing adverse impact within personnel selection. Industrial and Organizational Psychology: Perspectives on *Science and Practice*, *13*(2), 246–271. https://doi.org/10.1017/iop.2020.19

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology*, *50*(3), 707–721. https://doi.org/10.1111/j.1744-6570.1997.tb00711.x

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

Schlueter, M., Egea, J. A., & Banga, J. R. (2009). Extended ant colony optimization for non-convex mixed integer nonlinear programming. *Computers & Operations Research*, *36*(7), 2217–2229. https://doi.org/10.1016/j.cor.2008.08.015

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, *31*(3), 699–714. https://doi.org/10.1177/001316447103100310

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*(5), 719–730. https://doi.org/10.1037/0021-9010.82.5.719

Song, Q. C. (2018). *Diversity shrinkage of Pareto-optimal solutions in hiring practice: Simulation, shrinkage formula and regularization technique* [Doctoral dissertation]. University of Illinois at Urbana-Champaign.

Song, Q. C., & Tang, C. (2020, April). *Adverse impact reduction for multiple subgroups: A Pareto-optimization approach. In Q. C. Song & S. Wee (Co-chairs), Multi-objective optimization in the workplace: Addressing adverse impact in selection* [Symposium]. 35th Annual Convention of the Society for Industrial and Organizational Psychology, Austin, Texas, United States.

Song, Q. C., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology*, *102*(2), 1636–1657. https://doi.org/10.1037/apl0000240

Tamiz, M., & Jones, D. F. (1997). An example of good modelling practice in goal programming: means for overcoming incommensurability. In R. Caballero, F. Ruiz, & R. Steuer (Eds.), *Advances in multiple objective and goal programming* (pp. 29–37). Springer. https://doi.org/10.1007/978-3-642-46854-4_3

Van Iddekinge, C. H., & Ployhart, R. E. (2008). Developments in the criterion-related validation of selection procedures: A critical review and recommendations for practice. *Personnel Psychology*, *61*(4), 871–925. https://doi.org/10.1111/j.1744-6570.2008.00133.x

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

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, *2*(4), 440–457. https://doi.org/10.1214/aoms/1177732951

(*Appendices follow*)

Appendix A

Computation of the Relative Achievement Measures

We first recapitulate the procedure for assessing the relative achievement of a given selection system when the system is applied in the applicant population (i.e., an applicant pool of infinite size). In that case, the population formula for computing the selection outcomes in general, including multistage, selections proposed by De Corte et al. (2006) can be applied to calculate the diversity value (e.g., the AIR) and the quality value (e.g., the average criterion performance of the selected applicants) of the selection system. Next, standard, gradient-based algorithms (e.g., sequential quadratic programming) can be used to reliably address the nonlinear constrained optimization problems that solve for the maximum and the minimum possible diversity (quality) value at the given quality (diversity) value of the selection system (cf. De Corte et al., 2020). The standard algorithms can in this case be applied because both the objective function (i.e., the function related to the quantity—either the diversity or the quality—that must be optimized) and the nonlinear constraints (related to the condition that the maximum/minimum must be determined for a given value of the other selection objective) are continuous, differentiable functions.

For the same reasons, the standard optimization algorithms can also be used for calculating the relative validation achievement in case of a finite applicant pool, provided that the effect size and the validity of the selection composite are chosen for gauging the selection objectives. With other metrics for the selection objectives and, in particular when using the AIR and the average job performance of the selected applicants as is required for multistage selections, the objective and constraints are no longer analytic functions, leaving only metaheuristic methods to solve for the relative (diversity/quality/-) achievement of the selection systems. Thus, De Corte et al. (2020) apply a method based on the NSGA-2 evolutionary optimization algorithm of Deb et al. (2002). However, the procedure fails to calculate the relative diversity achievement, primarily because of difficulties related to the implementation of the nonlinear equality constraint that the maximum/minimum of the diversity must be determined at the given quality trade-off value of the system. Because the relative achievement of a selection system equals the average of its relative diversity and relative quality achievement, the failure also blocks the computation of the relative achievement of selection systems in finite applicant samples.

We resolved the computational problems related to the assessment of the relative diversity achievement via a modified version of the hybrid ant colony optimization algorithm of Schlueter et al. (2009). The new algorithm first invokes a Latin hypercube sampling strategy to obtain an initial set of feasible (but not optimal) solutions to the maximization/minimization problems implied by the calculation of the relative diversity achievement. The resulting set is subsequently used within a relaxation strategy to solve the maximum/minimum validation diversity value at the validation quality value of the selection system. Instead of implementing the nonlinear validation quality constraint in full precision, a relaxed version of the constraint involving three digits precision is used instead.

In a subsequent study, we tested the accuracy of the new computational procedure. The study exploits the fact that, for single-stage selections, the relative validation achievement of selection systems can be assessed using both the standard gradient-based as well as the nonstandard metaheuristic algorithm when the effect size and validity metrics are used to represent the selection diversity and quality objectives. The results show that the average relative validation achievement values of the systems, obtained by the three-digit precision metaheuristic algorithm, are very similar to the corresponding averages obtained using the gradient-based algorithm.

Appendix B

Computational Cycle

The computational cycle comprises the following four steps: (a) randomly generating a calibration and a validation data sample; (b) computing the diversity/quality trade-off attained by the five fixed-weight systems in the calibration data sample; (c) computing two corresponding sets of calibration PO systems; and (d) applying the FW and calibration PO weight systems to the validation sample and calculating the validation trade-off and relative validation achievement values of the systems. The average, across the repetitions, of the thus obtained trade-off and achievement values is subsequently used as an estimate of the expected trade-off and relative achievement of the studied selection, whereas the standard deviation of the values across the repetitions will occasionally serve to represent the variability.

Step 1: Sampling the Calibration and Validation Applicant Pools

The formula for computing PO selection systems assumes that the predictors and the criterion follow a multivariate normal finite mixture distribution in the total applicant population (cf. De Corte et al., 2006, 2007). The calibration and validation applicant pools therefore correspond to random samples from this mixture distribution with the given size and minority/majority applicant representation. When the calibration (validation) condition of the cell corresponds to the applicant population, the applicant pool is the population.

(*Appendices continue*)

Step 2: Computation of the Calibration Trade-Off of the Five FW systems

The formula in De Corte et al. (2006) is used to calculate the calibration trade-off of the FW systems. The calculations are based on the predictor effect size and predictor/criterion intercorrelation values derived from the calibration sample predictor/criterion score data.

Step 3: Computation of the Two Sets of Corresponding Calibration PO Systems

The calculation of the calibration PO systems is also based on the formula of De Corte et al. (2006) such that standard gradient-based optimization methods can be used to compute the systems. As with the computation of the calibration trade-off of the FW system, the computation of the PO systems also starts from the predictor effect size and predictor/criterion intercorrelation values derived from the calibration sample predictor/criterion score data. The calibration PO systems of the first set have the same diversity level as one of the fixed-weight systems, whereas the calibration PO systems of the second set share the quality level of one of the fixed-weight systems. Because the diversity/quality trade-off value of the fixed-weight systems varies across repetitions, the resulting calibration PO systems also vary from one repetition to the other. An adapted version of the program described in De Corte et al. (2011) is used to perform the calculation of the calibration PO systems. The adapted program adds an equality constraint to the nonlinear programs used to calculate the PO systems. The equality constraint enforces the requirement that the PO system should have the same diversity (quality) value as its corresponding FW system.

Step 4: Computation of the Validation Trade-Off and the Relative Validation Achievement of the FW and Calibration PO Systems

The final step calculates the validation diversity/quality trade-off and the relative validation achievement (i.e., the triple of relative validation diversity, relative validation quality, and global relative validation achievement) of the fixed-weight and corresponding calibration PO systems when applied to the *validation applicant pool data*. For finite applicant pool conditions, the metaheuristic algorithm described in Appendix A is used to perform the calculation of the relative validation achievement, whereas a classic gradient-based algorithm is invoked to obtain the results in case of an infinite applicant pool. To compute the validation trade-off of the systems, the formula of De Corte et al. (2006) is used in case of an infinite validation applicant pool. With finite applicant pools, the validation trade-off can be calculated directly from the validation applicant pool predictor/criterion score data.

At the end of the computational cycle, calibration and validation trade-off values as well as relative validation diversity, relative validation quality, and global relative validation achievement values are available for both the 5 FW and the 10 calibration PO systems. Averaging these values across the 2000 repetitions within each cell of the study design provides an accurate estimate of the trade-off and global relative achievement one may expect to obtain under the selection conditions specified by the cell of the design. This is particularly the case for the average relative achievement values because the standard error of these values is at most equal to .288/sqrt(2000) < .007.

Received January 7, 2020
Revision received April 19, 2021
Accepted April 21, 2021 ■