7-2020

Robustness, sensitivity and sampling variability of Pareto-optimal selection system solutions to address the quality-diversity trade-off

Wilfried DE CORTE
Ghent University

Paul SACKETT
University of Minnesota - Twin Cities

Filip LIEVENS
Singapore Management University, filiplievens@smu.edu.sg

Robustness, Sensitivity and Sampling Variability of Pareto-Optimal Selection System Solutions to Address the Quality-Diversity Trade-off

Wilfried De Corte

Ghent University

Paul Sackett

University of Minnesota, Twin Cities Campus

Filip Lievens

Singapore Management University, Lee Kong Chian School of Business

Author Note

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, 9000 Ghent, Belgium. E-mail: wilfried.decorte@ugent.be

Robustness, Sensitivity and Sampling Variability of Pareto-Optimal Selection System Solutions to Address the Quality-Diversity Trade-off

Abstract

In case that both the goals of selection quality and diversity are important, a selection system is Pareto-optimal (PO) when its implementation is expected to result in an optimal balance between the levels achieved with respect to both these goals. The study addresses the critical issue whether PO systems, as computed from calibration conditions, continue to perform well when applied to a large variety of different validation selection situations. To address the key issue, we introduce two new measures for gauging the achievement of these designs and conduct a large simulation study in which we manipulate 10 factors (related to the selection situation, sensitivity/robustness, and the selection system) that cumulate in a design with 3888 cells and 24 selection systems. Results demonstrate that PO systems are superior to other, non PO systems (including unit weighed system designs) both in terms of the achievement measures as well as in terms of yielding more often a better quality/diversity trade-off. The study also identifies a number of conditions that favor the achievement of PO systems in realistic selection situations.

RUNNING HEAD: Robustness, Sensitivity and Sampling Variability of PO Selection Systems

KEYWORDS: adverse impact, personnel selection, Pareto-optimal, selection design, robustness, sensitivity, sampling variability

Robustness, Sensitivity and Sampling Variability of Pareto-Optimal Selection System Solutions to Address the Quality-Diversity Trade-off

Selection system design requires a number of decisions, including the type and number of predictors that will be used, the selection rule, the sequencing and weighing of the predictors, as well as the between stage retention rates that will be implemented in case of multi-stage selection. To assist making these decisions in cases that value both the quality (i.e., the expected job performance) and the diversity of the selected applicant group, De Corte, Sackett and Lievens (2011) proposed a decision-aid for identifying Pareto-optimal (PO) selection designs. As the concept of Pareto-optimality is relatively new to the psychological literature, it might be confusing to someone who infers that the concept refers to a single optimal solution. Rather, there is a PO solution for every attainable level of diversity, i.e., it is the system that produces the highest level of expected performance among systems producing that specific level of diversity. The result is a set of PO solutions, commonly referred to as a "Pareto front", ranging from a performance-maximizing solution to a diversity maximizing solution. The argument is that PO solutions should be preferred to non-PO solutions. However, the choice among PO solutions is a value judgment, rather than a technical problem, as it depends on the relative value the organization assigns to the performance and diversity objectives.

Recently, Cortina, Aguinis, and DeShon (2017) reviewed key methodological developments in the last century and listed the Pareto-optimal (PO) approach as one of the methodological approaches in the last 10 years that has key applied implications for dealing with subgroup differences in personnel selection (see also the large-scale reviews of Bobko & Roth, 2013; Ryan & Ployhart, 2014). Although we agree that the PO decision-aid offers a sound, psychometrically based contribution to the selection design problem, some critically important issues remain unresolved. At present, the PO approach has been primarily

examined with meta-analytic input data. This is problematic for two key reasons. First, the model and the assumptions that drive the calculation of the expected selection outcomes will virtually never hold perfectly in the actual selection domain. Second, the PO decision-aid uses data on the composition of the applicant pool, as well as on the effect sizes, the validities and the intercorrelations of the predictors that are approximate at best. For example, little is known how results of PO systems are affected if the applicant pool is substantially smaller and different in composition than expected, fewer minority candidates are retained, and/or their scores on the selection procedures are differently distributed than assumed.

Due to these key unresolved issues we do not know the impact of (a) violations from the model assumptions and (b) deviations from the input data on the PO results obtained. We are also in the dark which of these factors might have the most impact on the results. Our purpose is to provide insight into the value of PO selection system design in a large variety of selection conditions. We address robustness (related to the assumption violations), sensitivity (related to the input data deviations), and sampling variability issues.

The structure of this paper is as follows. After providing an overview of previous developments on PO selection design, motivating the key research issues of the paper and summarizing prior research (e.g., Song, Wee & Newman, 2017), we propose new measures and a novel methodology for gauging the achievement of PO selection designs. This methodology is subsequently implemented within a factorial design to study the achievement of various PO designs when applied to a variety of validation settings that all differ from the calibration conditions (i.e., the assumptions and the predictor/criterion effect size and correlation data) used in deriving the PO systems. We also report on the relationship between the achievement level and the key dimensions that differentiate between the calibration conditions and the validation conditions that characterize the selection settings. Finally, we compare the achievement of PO designs in a large variety of selection settings to that of other

design choices (e.g., unit weighed predictor composites).

In our study, we use simulation methods rather than analyzing real data sets or focusing on any one actual intervention implementing the principles of PO selection design. This is a deliberate choice, based on two considerations. First, real data sets, including not only the predictor but also the criterion scores of the entire applicant pool, are seldom if ever available in a real selection context. Second, our objective goes beyond an assessment of the achievement of PO selection designs in a particular context. Instead, our aim is to shed light on the achievement of these systems in a wide variety of selection contexts. Simulation methods fit this aim much better than one single selection case.

Assessing the Achievement of PO Selection Systems

PO Selection Systems: A Brief Tutorial

For selection applications where the goals of selection quality and diversity are both of importance, De Corte, Sackett and Lievens (2011) proposed a psychometrically based decision-aid for rational selection design that results in selection systems that offer an optimal balance (i.e., a PO trade-off) between the two valued goals. The decision-aid conceives the shaping of a selection process as a series of mutually dependent decisions that define the resulting *selection systems* as particular sets of concrete choices with respect to the (1) predictor subset, (2) selection rule, (3) predictor staging, (4) predictor sequencing, (5) predictor weighing, and (6) between stage retention rates that will be implemented during the selection process.

To derive the PO selection systems, the decision-aid proceeds in two steps. The first step, the *inventory stage*, consists of identifying the set of selection systems that are feasible within constraints that govern the planned selection process. Constraints may include limits on selection costs and limits on the number of stages in a selection system, among others. In the second step, *the computational stage*, the decision-aid computes from this set the subset of

selection systems that are PO with respect to the selection diversity and quality goals.

These computations are based on the model proposed by De Corte, Lievens and Sackett (2007) for gauging the quality and diversity outcome value of selection systems, and the formulae invoked by the model depend on essentially three assumptions: (a) the joint distribution of the selection predictors and the job performance criterion is multivariate normal in both the majority and the minority applicant population[1], (b) the initial applicant pool is a mixture of *infinite size* of both applicant populations, and (c) a top down selection rule (without applicant drop out) applies. In addition, the model calculations require data on the validity, the intercorrelation and the effect size of subgroup differences of the available predictors as well as data on the final selection rate and the majority/minority composition of the total applicant pool. Henceforth, the assumptions, together with the input data used in the calculations, are referred to as the set of *calibration conditions* from which the results of the decision-aid are derived and the symbol C_c will be used to denote the set.

To illustrate, consider the example situation, henceforth referred to as situation S_0, where the first, inventory stage results in considering the following five predictors for selecting with a .20 selection rate in an applicant pool consisting of 80 percent candidates from the majority and 20 percent applicants from the minority population: (1) a cognitive ability (CA) test, (2) a structured interview (SI), (3) a conscientiousness (CO) measure, (4) a biographical inventory (BI), and (5) an integrity test (IN). In the inventory stage it is further decided that only three different selection scenarios are feasible: (a) a single stage scenario in which the final accept/reject decision is based on a weighed composite of the CA, CO, BI and IN predictors; (b) a two stage scenario where the candidates are first screened on the basis of a weighed composite of CA, CO and BI, and the remaining candidates (anywhere between 35 and 60 percent of the initial number of applicants) are selected using a weighed composite of the SI and IN predictors; and (c) a three stage scenario where the intermediate retention

decisions involve top-down selection on a CA and IN composite (retaining anywhere between 60 and 75 percent of the candidates) and a CO and BI composite (retaining 35 to 45 percent of the initial candidates), for the first and the second stage respectively, and the SI predictor is used in the final selection stage. Finally, suppose that the inventory stage also leads to the decision that the predictors may have weights between 0 and 1 when forming the predictor composites; a decision which implies that several additional scenarios such as, for example, a two stage scenario using only the CA predictor and the SI predictor are also feasible.

Given the above detailed situation S_0, and using data estimates on the applicant group composition, the predictors and the criterion (the example uses the predictor/criterion data values displayed in Table 1, Selection Environment 3), the decision-aid next proceeds by computing, over all feasible selection systems, the subset of systems that are PO with respect to the selection diversity and quality goals. Panel A of Figure 1 portrays the results of this second step, using the expected job performance of the selected applicants (expressed in standard score units) and the selection ratio in the minority applicant group as gauges for the selection quality and diversity goal respectively. The upper bold line in Panel A represents the set of PO goal trade-offs (i.e., the PO trade-off curve or Pareto front), whereas the area enclosed by the upper and lower (orange) lines depicts the entire gamut of achievable quality/diversity trade-offs. The figure in Panel A also represents a number of particular PO trade-off points (i.e., the points P1 to P4) on the PO trade-off curve. It is of key importance to note that these PO points not only correspond to a particular value for the quality/diversity trade-off, but are each also associated with a particular selection system. For example, PO trade-off point number 2 (point P2 on the figure) is associated with a two stage selection system in which the first stage selection, retaining 60 percent of the candidates, is based on a weighed composite of the CA and CO predictors (with weights equal to 0.707 and 0.687); whereas a composite of the SI and the IN predictors (with weights equal to 0.677 and .750) is

used in the final selection stage.

Besides the PO trade-offs on the upper curve, Panel A of Figure 1 also displays five additional sets of trade-offs that all have the same diversity value as one of the PO trade-offs on the upper curve, but are inferior in terms of the quality value. Exploring these will be a major component of this paper: we will generate selection systems that are inferior to PO systems when assumptions are met, and then examine the achievement of these PO and inferior systems when assumptions are violated.

The trade-offs on the lower curve (labeled with the letter Z) represent the worst possible trade-offs, whereas the trade-offs labeled with the letter U refer to trade-offs associated with selection systems in which any predictor that is assigned a non-zero weight in a selection system is given a weight of one. These fixed weight systems are henceforth referred to as unit weighed systems and they reflect the practice of using unit weighting to either the totality or a subset of the available predictors. So, unit weighed systems do not necessarily assign a weight of one to each predictor in the composite, but for each composite at least one of the predictors has a weight of one. The unit weighed system U1, for example, refers to a three stage selection system with weights one and zero for the first stage predictors CA and IN respectively, weight one for both the second stage predictors CO and BI, and weight one to the third stage predictor SI. Finally, the trade-offs of the remaining three sets[2] (i.e., the sets T1 to T4, F1 to F4 and S1 to S4) correspond to feasible selection systems that are characterized by a quality trade-off value of a given fixed percentage as compared to the quality value achieved by the PO system that shows the same diversity trade-off value. These reflect twenty-five (T), fifty (F), and seventy-five (S) percent of the quality achieved via the PO system.

In summary, the application of the decision-aid determines which of the feasible selection systems are PO and which are non-PO. Also, only PO systems should be

implemented because all other systems (e.g., the U systems or the T, F and S systems in Figure 1) result in a quality/diversity trade-off that can be bettered by a PO system (e.g., the system U2 is bettered by both P2 and P3). The decision-aid does not indicate which PO system is to be preferred. As noted by De Corte et al. (2011, p. 913), the final decision in favor of a particular PO system calls for "a value judgment on the particular kind of balance between selection quality and work force diversity one is aiming at".

Key Research Issues

The results of the decision-aid are all dependent on the validity of the calibration conditions C_c. Yet, there is no doubt that these conditions will rarely, if ever, correspond to the unknown conditions that characterize the real selection situation. So, although input predictor, criterion, and applicant data values might come from a prior local validity study or from generalized validity evidence (e.g., transporting validity from a closely related setting or meta-analytic findings), they might at best approximate the values that will be found in the actual selection situation of interest. In addition, recruitment efforts may in real situations result in a size and a composition of the applicant pool such that different retention rates and a different selection rate than the rates initially used in deriving the PO systems must be applied to obtain the required number of selected candidates. Finally, it will almost surely be the case that the majority and minority candidates in the applicant pool will not represent samples from a multinormal distribution with mean and correlation structure values as assumed under C_c, but rather come from a possibly nonnormal distribution with a different mean and correlation structure.

So, the PO selection systems identified by the decision-aid correspond to calibration conditions C_c that at best approximate the typically unknown conditions that characterize the actual selection application. Denoting the actual prevailing conditions, henceforth also referred to as the *validation conditions*, as C_v, the key issue then becomes how the

achievement of the PO selection systems, as computed under the calibration conditions C_c, evolves when these systems are implemented for a selection application where the validation conditions C_v apply. Also, it is equally important to assess the two major types of circumstances that may impact the achievement level of the PO selection systems: (1) the nature of the selection environment and the calibration conditions C_c the systems are computed from, and (2) the features that differentiate between the calibration conditions C_c and the actually prevailing validation conditions C_v. Finally, it is also worthy to consider whether the achievement in the validation conditions varies across the range of PO systems and to study the possibly different impact on the achievement in the validation condition of PO as compared to non PO systems such as the unit weighed systems.

As a consequence, the first key research issue of the paper focuses not only on the achievement level of PO systems when applied in a large variety of validation settings, but also on the relative impact on the achievement level of (1) the nature of the selection environment and the calibration conditions the systems are computed from, and (2) the features that differentiate between the calibration and the validation conditions. As a second research issue, the paper compares the achievement, across various validation circumstances, of PO selection systems to the achievement of non PO systems, including unit weighted systems. Together both of these issues speak not only to the robustness for violations of the distributional assumptions and the sensitivity to variability in the input data of PO selection system design, but also to the relative level of robustness and sensitivity of these systems relative to other non PO systems. Finally, by considering actually realized applicant pools as finite sized samples obtained under the validation conditions C_v, we address the issue of sampling variability in the PO systems achievement as well.

Previous Related Research

In general, prior related research focused on comparing the quality/diversity trade-off of

PO and unit weighed selection systems, as computed from calibration conditions C_c, to the trade-off achieved by the systems when applied in validation settings C_v that differ only in terms of the predictor/criterion correlation and effect size data values (e.g., De Corte et al., 2011; Song et al., 2017; and Wee et al., 2014). Also, only single stage selection systems were investigated. In a first study, De Corte et al. (2011) used the calibration trade-off value of PO and unit weighed selection systems to link each of these systems to a corresponding set of so-called "dominated" selection systems; that is to a set of systems that under C_c result in quality and diversity trade-off values that are at best equal to the trade-off achieved by the former system. The achievement of the systems in the validation setting C_v was subsequently assessed as the proportion of times that the trade-off of these systems continued to dominate the trade-off of their corresponding dominated systems when they were all applied under C_v. In a second sample to population cross validation study De Corte et al. compared (a) the average value on the quality objective of sample based, calibration PO and unit weighed systems when implemented in the population, validation setting to (b) the corresponding average that is optimally achievable in the validation setting at the identical value for the diversity objective. Wee, Newman and Joseph (2014) studied the gain in the diversity objective when using a PO selection system instead of the unit weighed system, assigning a weight of one to each predictor, across a large number of (population) validation settings, each corresponding to a different set of values for the predictor/criterion effect size and correlation data. Finally, Song et al. (2017) also adopted a sample to population cross validation approach to study the quality (equated to the validity of the predictor composite) and diversity (indexed by the adverse impact ratio, AIR) shrinkage when PO systems computed from sample calibration data are applied to approximate population settings.

By and large, the previous studies reported rather favorable results on the achievement of PO selection systems when applied in new, validation settings. De Corte et al. (2011) found

that in these new settings both PO and unit weighed systems continue to outperform their dominated systems to a fairly similar degree. Also, evaluated at identical diversity levels, PO systems, computed from calibration sample predictor/criterion data, maintained on average a high quality level relative to the average quality achieved by the corresponding PO systems as derived from the population validation data. Wee et al. (2014) concluded that the average gain in the diversity objective, when using the PO system with the same quality level as the unit weighed system, remains substantial across different validation settings. The average gain was also quite stable across different levels of sampling variability in the validation predictor/criterion data. Finally, Song et al. (2017) observed that validity shrinkage in the validation setting is fairly negligible when the PO systems are computed from calibration data obtained from samples of at least 100. Diversity shrinkage is more pronounced for samples of the same size, however, especially when some of the selection predictors show small effect sizes as is illustrated in Figure 2 of Song et al. by the considerably larger shrinkage along the diversity axis as compared to the shrinkage along the validity axis. The shrinkage also relates to the type of PO system: PO systems that give priority to the quality objective are more prone to validity shrinkage, whereas PO systems that favor the diversity objective show more diversity shrinkage. Finally, even accounting for the shrinkage observed for PO systems computed from small sample predictor/criterion data, these systems still offered potential for diversity/validity improvements over unit weighted selection systems.

Although previous studies suggests that PO systems may compare favorably to other selection system designs, further research is highly needed for several reasons. First, the previous studies do not cover the robustness issue and are all limited to the situation where the initial (calibration) and the new (validation) setting differ only in terms of the predictor/criterion correlation and effect size data, thereby neglecting the common instance where the calibration and the validation setting also differ in terms of the selection rate and

the size and composition of the applicant pool. Second, thus far only single stage selection systems have been investigated. Third, and even more importantly, all previous results are tentative at best because they relate to situations where the validation setting involves the applicant *population* instead of samples of limited size from this population[3]. Yet, as argued by Cattin (1980), personnel selection researchers and practitioners are essentially interested in how well selection systems derived in the calibration condition will perform in new, *limited sized validation sample conditions*. Finally, all previous research fails to address the question whether the quality/diversity trade-off of the PO systems achieved in the validation condition continues to compare favorably *to the diversity/quality trade-offs that are at all possible in the validation situation.*[4]

Figure 2 is particularly helpful to explain the latter issue. At the same time, the figure illustrates the main difference between the approach of Song et al. (2017) and the one adopted in the present paper to evaluate the validation potential of calibration based PO systems. The figure relates to a single stage selection situation (using a weighed composite with nonnegative weights of the five predictors of Selection Environment 3 detailed in Table 1) with a .15 selection rate and a 167 proportion of minority candidates in the applicant population. The vertical and horizontal axis of the figure correspond to the quality objective (operationalized as the predictor composite validity) and the diversity objective (measured as the minority applicant selection rate), respectively. The points $\overline{1}_c$, $\overline{2}_c$, …, $\overline{10}_c$ on the figure indicate the average quality/diversity trade-off achieved by 10 PO systems in the calibration condition across a total of 10,000 sample calculations, where the computation of the PO system trade-off values in each replication is based on the predictor/criterion correlation and effect size values as obtained for samples of size 40 from the total applicant population. The other points, with labels $\overline{1}_v$, $\overline{2}_v$, …, $\overline{10}_v$, correspond to the average trade-off achieved by the calibration PO systems when implemented in the validation context (i.e., with respect to the

population predictor/criterion correlation and effect size data of Environment 3 in Table 1). To evaluate the validation potential of the PO systems Song et al. focus on the diversity (validity) shrinkage of the PO systems, defined as the difference between the calibration and validation average diversity (quality) trade-off value of the systems. The dashed lines in the figure exemplify this diversity (quality) shrinkage for the PO systems 1 (which strongly favors the quality objective), 8 and 10 (which strongly favors the diversity objective). Observe that the diversity (quality) shrinkage of the PO systems 1, 8 and 10 in Figure 2 clearly substantiates the conclusion of Song et al. that PO systems, that favor more strongly the diversity (quality) objective, show larger diversity (quality) shrinkage.

Whereas Song et al. (2017) focus on diversity (quality) shrinkage to assess the validation potential of the PO systems, the present approach proposes comparing the validation diversity/quality trade-off of the PO systems to the set of trade-offs that are at all possible in the validation condition. The area enclosed by the solid line contour in Figure 2 represents the latter gamut of possible diversity/quality trade-offs, and the average diversity/quality trade-offs achieved by the PO systems in the validation condition (i.e., the trade-offs $\overline{1}_v$, $\overline{2}_v$, …, $\overline{10}_v$) are all within the gamut. Surprisingly however, and although the PO systems that favor more strongly the diversity (quality) objective show the largest diversity (quality) shrinkage, these systems perform rather inversely when compared relative to the maximum and minimum possible diversity (quality) level that can be achieved at their corresponding quality (diversity) level. As an example, consider the diversity shrinkage of the calibration PO systems 1 (giving maximum priority to the quality objective) and 10 (giving maximum priority to the diversity objective) and compare this shrinkage to the relative position of the average validation trade-off of the systems with respect to the points P1 and W1 (for system 1) and the points P10 and W10, respectively. Clearly, calibration PO system 10 shows a larger diversity shrinkage (equal to the difference between the diversity value of

the average calibration trade-off $\overline{10}_c$, .21, and the diversity value of the average validation trade-off $\overline{10}_v$, .12, resulting in a value of .09) than system 1 (with a diversity shrinkage equal to .08 - .06 = .02). Yet, compared to system 1, the .12 diversity value of the average validation trade-off of system 10 corresponds to a higher proportional achievement level relative to the best and worst possible diversity level that can be achieved in the validation context at the same quality level. (i.e., a proportional achievement level for system 10 equal to $\frac{.12-.06}{.14-.06} = .75$, with .06 and .14 the worst and the best possible diversity value in the validation condition at the system 10 quality value of .38, cf. the points W10 and P10 in Panel A; and a proportional achievement level for system 1 equal to $\frac{.06-.05}{.08-.05} = .33$ for system 1). Alternatively PO system 1 has a larger quality shrinkage than system 10 despite the fact that its validation quality value shows a substantially higher proportional achievement level as compared to system 10. So, focusing on the diversity/quality trade-offs that are at all possible in the validation condition (present approach), instead of using shrinkage (Song et al. approach) as a measure for gauging the achievement of PO systems in the validation condition, may very well lead to quite different conclusions about the validation potential of calibration PO and other selection systems.

The next section further develops the basic idea underlying the present approach for gauging the validation potential of calibration based selection systems. These developments lead to new measures for quantifying the achievement of PO and other selection system designs when applied in validation settings involving applicant groups of both limited and unlimited size. The new measures also enable a straightforward comparison of the achievement level of the various systems in these settings.

Measuring the Achievement of Selection Systems in the Calibration Condition

We first consider measuring the achievement level of selection systems as obtained in the calibration stage. In this stage, the systems are computed using the model proposed by De

Corte et al. (2007) implying that the achievement level of the systems expresses the achievement as obtained with respect to an infinitely sized applicant pool; that is with respect to the total applicant *population*. Panel A of Figure 1 represents such a situation. Suppose now that we aim for a measure, with values ranging between 0 and 1, to assess the achievement of the selection systems P1 to P4 and Z1 to Z4 depicted in the panel. In that case, the obvious choice is to assign a value of 1 to the systems P1,…, P4 and a value of 0 to the systems Z1,…, Z4 because the former systems show the maximum possible quality at the corresponding diversity level, whereas the latter systems have the worst possible quality at the same diversity level. Given these values, it is then straightforward to assign achievement values to the other systems (e.g., U1, F2, and so on) reflecting the percentage of the possible improvement over the Z system that is obtained with the PO system. More specifically, the achievement of these other systems can be expressed as a proportion relating (a) the difference in quality value of the system and the quality value of the worst possible system with the same diversity value to (b) the difference in quality value of the best and the worst possible system with the same diversity value. In the extreme rare event that the latter difference equals zero we adopt the convention that the system has a performance value of one.

As an illustration of the proposed achievement measure, consider the system U2. The system shows a quality/diversity trade-off of 1.086/0.126, whereas the best and worst possible systems with the same diversity value (i.e., P2 and Z2) have a quality value of 1.256 and 0.788 respectively. With these values, the achievement of system U2 is then equated to $\frac{1.086-0.788}{1.256-0.788} = .64$. In other words, system U2 obtains 64% of the gain over the worst possible system that could be obtained with the best (i.e., PO) system.

In what follows, the above described measure for the achievement of a selection system will be referred to as the *calibration quality achievement* of the system. So, the calibration

quality achievement of a selection system indicates the proportional achievement, on the quality objective, of the system at its corresponding diversity level as computed under the calibration conditions C_c.

As the natural companion of the former gauge, we also introduce the *calibration diversity achievement measure*. Similar to the calibration quality achievement measure, the calibration diversity achievement measure indicates the proportional achievement in the calibration condition, but this time with respect to the diversity objective, of a selection system at its corresponding quality level. Using system S1 of panel A (with a diversity/quality trade-off value of .10/1.18) as an example, it can be seen that the systems with labels W1 and B1 have the same quality value (i.e, 1.18) as the system S1, with W1 showing the worst possible diversity value (i.e., .07) and B1 the best possible diversity value (i.e., .15). The calibration diversity achievement therefor equals (.10-.07)/(.15-.07)=.38.

Observe that the calibration diversity and the calibration quality achievement measure are undefined if the denominator in the corresponding proportion equals zero. As illustrated in Panel A of Figure 1, this will typically be the case for only four selection systems: the systems NB, NE, NO and P1. System P1, for example, shows a zero difference between the lowest and the highest attainable diversity trade-off value at its quality trade-off level, but in this case as well as for the other three systems it is obvious to equate the corresponding calibration diversity (quality) achievement measure to one. It is also important to note that both new measures result in dimensionless quantities that share the same metric. Although the quantities still relate to one specific selection objective, they do no longer share the metric of the objective. In particular a value of, for example, .75 on the *calibration quality achievement* (*calibration diversity achievement*) measure does not mean that the system has a value of .75 for the quality (diversity) objective but that its achievement on the quality (diversity) objective is at 75 percent of the gain over the worst possible system that could be obtained

with the best system at the same diversity (quality) level of the system. Also, because both new measures share the same metric, it is admissible to combine their values to one aggregate achievement measure by taking the average of the two measure values.

Measuring the Achievement of Selection Systems in the Validation Condition

In selection practice, one is less interested in the achievement of selection systems when applied to an entire population but rather in the expected achievement of the systems when applied to a future, finite applicant pool because real world applicant pools are always of finite size. We therefor focus on measuring the achievement of selection systems in validation conditions involving either a single finite sized applicant pool or a population of finite sized pools. Panel B of Figure 1 (the subsection "Computing the Validation Achievement of Selection Systems" details the procedure for obtaining the results depicted in the panel) illustrates the development of the achievement measures in the first case. The panel depicts the trade-offs achieved by the calibration selection systems of Panel A when applied to a given applicant pool of size 250 with an equal number of minority and majority applicants using an overall selection rate of 0.3. Panel B also shows the gamut of trade-offs that can be achieved in the validation applicant pool.

Comparing both panels of Figure 1 illustrates how the gamut of achievable quality/diversity trade-offs and the trade-offs of the PO and the non PO selection systems as obtained under the calibration condition C_c may change substantially for the validation applicant pool. First, the upper and lower boundary of the gamut of achievable trade-offs consists of only a limited number of points because the validation condition involves a finite sized applicant pool such that only certain values for the minority selection rate are possible. Similarly, and for the same reason, the gamut no longer corresponds to the area enclosed by the boundary points, but reduces to the collection of vertical dashed lines connecting the corresponding upper and lower boundary points (cf. the vertical orange dashed lines in Panel

B). Also, within each vertical line, only a finite number (quickly increasing with the size of the applicant pool) of different quality values is achievable. Finally, observe the changes in the trade-off achieved by the selection systems in the calibration condition (cf. Panel A) versus the validation applicant pool. Consider, for example PO system P2. Under C_c, the system has a quality/diversity trade-off of 1.26/0.13 (cf. Panel A), whereas the same system results in a trade-off of 1.09/.16 when applied to the validation pool. Also, none of the systems that are PO under C_c remain PO in the validation pool because each one is dominated by a feasible system that has the same diversity, but a higher quality value (cf. the systems corresponding to the trade-offs B1,…, B4 in Panel B).

Despite the differences between the calibration and the validation conditions, the principle used to measure the achievement of the selection systems in the calibration condition can also be invoked to gauge the achievement of these systems when applied to the validation applicant sample. To distinguish the resulting measures for the applicant sample in the validation context from the corresponding measures in the calibration condition, they are henceforth referred as the *sample validation diversity achievement* and the *sample validation quality achievement* respectively. Thus, given the trade-offs achieved in the validation pool of, for example, P2, B2, W2, W2D and B2D (i.e., 1.09/.16, 1.20/.16, .51/.16, 1.09/.12 and 1.09/.23 for P2, B2, W2, W2D and B2D, respectively; cf. Panel B), the sample validation quality achievement of system P2 can now be equated to (1.09-.51)/(1.20-.51)=0.84; whereas the sample validation diversity achievement of the system equals (.16-.12)/(.23-.12)=.36.

If the validation conditions refer to a population of finite sized applicant pools, the sample validation achievement value of the selection systems will vary across the set of all possible applicant samples that are consistent with the validation conditions C_v. To account for this sampling variability the *validation diversity (quality) achievement* of a selection system under such more general validation conditions C_v is henceforth defined as the

expected sample validation diversity (quality) achievement across all possible applicant samples according to C_v.

As is the case for the calibration achievement measures, the validation achievement measures are undefined if the denominator in the corresponding proportion equals zero. Although the condition will generally not hold for the validation quality achievement measure, the same is not true for the validation diversity measure, especially if the validation conditions relate to a selection with a small selection rate applied to a small applicant pool. In that case, the number of possible values for the selection diversity trade-off, as gauged by either the minority selection rate or the AIR, is (very) small and the worst and the best possible diversity value for a given quality level are often identical[5]. So, although both validation achievement measures are conceptually on an equal footing, the validation quality achievement measure has, compared to the validation diversity measure, the net advantage that it is almost never undefined.

Compared to previously proposed gauges, the novel measures of validation achievement have two distinct advantages. First, the measures offer an adequate, intuitively appealing and easily interpretable quantification of the validation achievement level of a selection system. In essence, the measures tell by means of a proportion how well a system is expected to perform on the quality (diversity) objective in a new setting C_v as compared to the best and the worst possible selection system designs that, under C_v, have the same diversity (quality) value as the system. Also, because the measures are dimensionless and in the same metric they can be combined to a single aggregate validation achievement measure. Second (and except for the earlier discussed limitation for the validation diversity achievement measure), the measures are generally applicable because they can be used to evaluate the validation achievement of both PO and other selection systems with respect to any applicant pool corresponding to any set of validation conditions C_v and therefore enable comparing the validation achievement of

any one selection system with that of any other system either under the same conditions C_v or across different conditions.

Finally, observe that the values on the new validation achievement measures as well as differences between these values can easily be converted to corresponding quantities that are of immediate relevance to practitioners. For example, consider again Panel B of Figure 1. The panel shows that the systems U4, P4, W4, and B4 result in the same minority selection rate of .21. Yet, each of these has quite a different value for the quality objective (i.e., quality values of .36, .82, .93 and 1.15 for W4, U4, P4 and B4, respectively), resulting in sample validation quality achievement values of 0, .58, .72 and 1 respectively. Obviously, the P4 system outperforms the U4 system and the difference in sample validation quality achievement, equal to .14, can be translated to a difference of .93 - .82 = .11 standard units in the expected job performance of the selected applicants.

Computing the Validation Achievement of Selection Systems

We developed two suites of programs and accompanying shell scripts to compute the validation achievement in the validation conditions $\mathrm{C_v}$ of selection systems as derived under the calibration conditions $\mathrm{C_c}$. The first suite is restricted to the study of single stage selection systems with respect to validation conditions involving an infinite sized applicant pool as in the Song et al. (2017) study and the suite is executable on a personal computer. The suite also calculates the shrinkage in the quality and diversity objective of the systems in the validation conditions. The program solves a series of nonlinear optimization problems similar to the ones described in De Corte et al. (2011), using a classic, gradient based sequential quadratic programming algorithm. The suite, including documentation about its usage, can be downloaded from http://users.ugent.be/~wdecorte/software.html and the online material accompanying the paper presents an application studying the robustness and sensitivity of both the shrinkage and the validation achievement of various single stage selection systems

under population validation conditions. These results assist the discussion on the main results reported in the paper.

In contrast to the limited capabilities of the first suite of programs, the second suite addresses both single and multi stage selection systems[6] under general validation conditions related to a finite sized applicant pool. The previous section shows that in that case the computation of the validation diversity (quality) achievement of a selection system in validation conditions C_v requires generating a large number of applicant samples according to C_v, computing the sample validation diversity (quality) achievement of the system for each sample, and taking the average of the resulting achievement values. Because these computations are extremely demanding, the second suite can only be executed on a high performance computing facility.

To generate the applicant samples in the second suite, we use the procedure described by Ruscio and Kaczetow (2008) because it can deal with virtually any type of joint distribution (including real data distributions) of the predictor/criterion in the majority and the minority applicant populations and the procedure can, therefor accommodate a very broad range of C_v conditions. Next, to compute the validation achievement values of the systems in each of the generated applicant samples, we wrote a mixed C and Fortran 77 program. The program repeatedly applies the evolutionary multi objective optimization (EMOO) algorithm as implemented in the NSGA-2 program developed by Deb, Pratap, Agarwal, and Meyarivan (2002) to calculate the maximum and minimum quality (diversity) value that can be achieved (over all feasible selection systems) at the diversity (quality) level obtained by the systems in the sample. We adopted the latter EMOO algorithm because with finite sized validation applicant pools the calculation of both the maximum and minimum achievable quality (diversity) involves the global, constrained optimization of a nonlinear, nonanalytic function (corresponding to either the quality or the diversity of the system) where one of the equality

constraints (related to the diversity or the quality in case the quality or the diversity is maximized or minimized) is also nonanalytic. These optimization problems can not be solved with classical, gradient based methods, such as invoked by the decision-aid of De Corte et al. (2011), leaving no other option than to use a general meta heuristic approach instead. To assist the EMOO algorithm, its execution is preceded by an extensive grid search to generate an initial population of problem variable values that meet the nonanalytic equality constraint.

Although the above procedure succeeds in computing the minimum and maximum achievable quality at the diversity level obtained by a selection system in the validation sample, the procedure is unreliable when solving for the corresponding diversity optimizations at the quality level obtained by the system[7]. Using a different metaheuristic approach (i.e., ant colony optimization, Dorigo and Stutzle, 2004) instead of the evolutionary based approach does not solve the problem. Apparently, the problems with the present procedure to maximize/minimize diversity under the quality equality constraint is caused by the fact that in finite applicant pools the number of possible values for the quality objective (gauged by the average job performance of the selected applicants) at a given diversity level is much larger than the corresponding number of possible values for the diversity objective (gauged by either the minority selection rate or the AIR) at a given quality level, making it much harder to implement the equality constraint with respect to the quality objective as compared to the implementation of the constraint with respect to the diversity objective.

Despite these computational problems, we decided to adopt the general approach for the remainder of the paper, even though this means that only results about the validation quality achievement of the systems will be reported. The decision is motivated by the fact that only the general approach can shed light on the achievement of both single and multi stage selection systems when applied in realistic validation conditions, that is in conditions involving finite sized applicant pools. Also, using the findings from the study reported in the

online supplement it is possible to at least indicate how the results about the validation diversity achievement of different selection systems are expected to evolve. Finally note that, even without the computational problems, the integration of the validity diversity achievement measure in the present study could still be somewhat problematic because the measure is often undefined for small applicant pool validation conditions (cf. the section "Measuring the Achievement of Selection Systems in the Validation Condition").

Studying the Robustness and Sensitivity of Selection Systems

We use simulation methods within a design structured by 10 factors to address the key research questions about the validation achievement of PO and other selection systems when these systems are applied to a large variety of validation selection settings. The design adopts the framework of sensitivity analysis (Saltelli, Tarantola, Campolongo & Ratto, 2004). This framework aims to assess the effect of different sources of uncertainty (variability or error) in the input data of a model on the model output, often using simulation and regression or ANOVA methods within a (preferably) factorial design to identify the most prominent sources of uncertainty or variability. The framework is therefore ideally suited to address the key research questions of the paper. In addition, the present design also permits studying issues concerning the population to sample and the sample to sample cross-validation (cf. Cattin, 1980) of PO selection system designs.

The first three factors of our design, henceforth referred to as the *selection situation factors*, capture the impact of the nature of the selection environment and the initial calibration conditions C_c the systems are computed from. A second set of five factors, addressing the *sensitivity and robustness* issues, relates to the major features that differentiate between the initial conditions C_c and the actually prevailing validation conditions C_v. Finally, the remaining two factors, labeled as the *selection system factors*, structure the characteristics of the analyzed selection systems.

Selection Situation Factors

The first two selection situation factors relate to the overall selection rate under C_c (with three levels: .1, .2 and .4), and the proportional representation under C_c of the majority applicants in the candidate population (with two levels: a .8 and a .5 majority applicant representation) respectively. We included these factors in the design to investigate whether PO and other selection systems, as derived for different combinations of selectivity rate and majority/minority mixture proportion values under C_c, show different levels of robustness and sensitivity. The choice of the actual levels of both factors was driven by a double concern: sufficient variation in the level values to capture an eventual effect of the factors and maintaining a reasonable degree of realism.

The final selection situation factor, labeled as the selection environment factor, has three levels that each refer to a quite different selection setting. Table 1 and 2 detail the three environments. The first table identifies the available predictors in each environment and summarizes the predictor/criterion mean and intercorrelation data used to compute the different selection systems under C_c within the environment. In turn, Table 2 describes the contextual and other relevant constraints that demarcate the set of feasible selection systems for each environment.

We choose these three selection environments because we first and foremost wanted to assess the validation achievement of PO and non PO systems over a wide variety of selection settings, even though this implied considering environments that differ not only in terms of the type and number of the predictors and, hence, in the predictor/criterion data, but also vary with respect to the staging of the predictors (i.e., single vs two stage selection and mixed single, two and three stage selection) and the nature of the feasible selection designs. At the present early stage of research on the robustness and sensitivity of PO selection system design, we decided in favor of including a wide variety of factors, representing all major types

of circumstances that may impact on the validation achievement of the selection systems, rather than focusing too much on a single, albeit important aspect such as the nature of the selection environment. Given the multitude of ways in which selection environments may differ from each other (e.g., with respect to the number and type of predictors, the distribution of the predictor/criterion effect sizes, the factorial structure of the predictor battery and the nature of the set of feasible selection designs), a detailed analysis of the impact of this factor is best postponed until more is known about the other circumstances that are critical to the robustness and sensitivity of PO and other systems. For now, the decision to first consider the full scale of possibly important factors implied choosing between a set of fairly homogeneous levels for the selection environment factor that differ in only one aspect such as, for example, the staging of the selection process, and a set of heterogeneous levels. We decided in favor of the latter option because it enables a more general and informative answer about the validation achievement of PO as compared to other selection systems, even though this choice may entail some difficulties with the interpretation of the effect of the factor.

With three levels for the selection rate and the selection environment, and two levels for the proportional minority/majority representation, the crossing of the three selection situation factors results in a total of 18 different studied selection situations. These 18 situations are at best exemplary for the broad range of situations encountered in practice, but we believe that the situations are sufficiently heterogeneous to assure that the study provides at least guiding evidence on the validation achievement of PO selection system designs.

Factors Differentiating Between the Calibration and the Validation Conditions

The design also includes five factors to capture the ways in which the validation conditions of the selection application, C_v, may deviate from calibration conditions, C_c. The first factor targets the robustness issue because it relates to the nature of the distribution (under C_v) of the predictor/criterion scores in the majority/minority applicant populations

from which the applicant pool is sampled in the validation conditions. The other four factors address the sensitivity issue. More specifically, these factors focus on the differences in the data values under C_v as compared to C_c of (1) the proportional representation of the majority and minority applicants in the applicant pool, (2) the overall selection rate, (3) the size of the applicant pool, and (4) the mean and the intercorrelation of the predictors/criterion in the majority/minority populations the applicant pool is sampled from. Observe that the latter factor relates to differences of the mean and intercorrelation data at the population and not at the sample level.

Although the above five factors permit a fairly exhaustive investigation of the robustness and sensitivity issues, it is again noted that the choice of the number and the nature of the factor levels reflects a balance between the concerns of feasibility and adequate coverage. Thus, proportional representation under C_v of the minority/majority applicants has only two levels: either the same or different to the one under C_c (i.e., if different, the proportion majority applicants under C_v equals .8 (.5) when the corresponding proportion is .5 (.8) under C_c). In turn, the selection rate under C_v is limited to three levels, with level one indicating an identical selection rate and the levels two and three corresponding to the case where the selection rate under C_v is 1.5 and 0.5 times the selection rate under C_c. The size of the applicant pool has four levels, covering the range from rather small (80) to medium (250) to large (800) and very large (2500) applicant pools. Next, the difference of the mean and correlation structure of the joint predictor/criterion distribution under C_c versus C_v, has three levels with level one corresponding to the situation where the mean and the correlation values of the predictors/criterion in the majority/minority applicant populations are identical under C_c and C_v. This situation permits studying population to sample cross-validation issues. The levels two and three represent increasing degrees of difference between the mean and correlation values under C_v vs C_c where the random distorted correlation matrices under C_v

are constructed according to the procedure described by Hardin, Garcia and Golan (2013) and random sampling from a rectangular distribution is used to generate the distorted mean values. Under level two (three), the noise added to the initial C_c correlation structure is at 30 (60) percent of the maximum possible value (to ensure that the distorted matrix is still positive semi-definite) and the mean (i.e., effect size) values are sampled from the rectangular distribution centered on the initial value and having a range equal to .30 (.60). So, the levels two and three of the factor represent the condition of mildly distorted and substantially distorted mean/correlation data respectively. Considering the mean and correlation structure under C_c as a sample structure that can occur under C_v, both conditions enable addressing, albeit in a limited form, sample to sample cross-validation issues.

Finally, the factor about the normality vs non-normality of the joint predictors/criterion score distribution in the parent majority and minority population from which the applicant pools are sampled from under C_v, has three levels. Level one corresponds to sampling from the multinormal distribution, whereas the levels two and three indicate sampling from moderately and severely nonnormal distributions (i.e., generalized lambda distributions, Chalabi, Scott & Wuertz, 2012) respectively. More specifically, the marginal distributions of the predictors/criterion scores have skew and kurtosis of .75 (2.0) and 4 (9) under level 2 (3) of the factor. In this way, both levels reflect the characteristics of predictor/criterion score distributions as often encountered in real samples (cf. Micceri, 1989; Blanca et al., 2013). Also, the heavily skewed criterion score distribution under level three accords with recent arguments by O'Boyle and Aguinis (2012) that actually observed job performance scores follow a Pareto distribution; however, see Beck, Beatty, and Sackett (2014) for a contrary view.

Selection System Factors

The selection systems studied within each cell of the design correspond to the crossing

of two selection system factors: the selection system type factor with six levels and the selection system relative diversity factor with four levels. More specifically, the first five levels of the selection system type factor differentiate the studied selection systems in terms of the calibration quality achievement level attained under C_c. Level one selection systems are PO under C_c, and therefor show the best possible calibration quality achievement value (i.e., a value of one or 100 percent), whereas the level two, three, four and five systems have, under C_c, a 0, 25, 50 and 75 percent calibration quality achievement value respectively. Panel A of Figure 1 illustrates the different types of selection systems: the systems corresponding to the trade-offs P1 to P4 represent level one type of selection systems, the systems corresponding to the trade-offs Z1 to Z4 are level two type systems, and so on. The sixth level of the selection system type factor refers to selection systems in which unit weighed composites are used to perform the selection. In Panel A of Figure 1, these systems correspond to the trade-offs U1 to U4.

The inclusion of the selection system type factor permits addressing the differential robustness and sensitivity of different types of selection systems. Also, given the particular levels chosen for the factor it is possible to study whether PO selection systems, as derived under C_c, continue to outperform other selection system types and, in particular, unit weighed selection systems when these systems are applied in a large variety of validation settings. Note that the study does not include regression weighed selection systems as an additional level for the selection system type factor because these systems may, depending on the predictor/criterion correlation structure, assign negative weights to the predictors in forming the predictor composites. As a consequence the regression weighed systems may violate the constraint on the feasible selection systems, imposed in all three studied selection environments, that only non-negative weights are permissible in forming the predictor composites.

The four levels of the selection system relative diversity factor refer to increasing degrees of diversity achieved by the systems under C_c. The actual values of the four diversity levels vary across the 18 different studied selection situations, however, because these situations differ in terms of the selection environment, the selection rate and the majority/minority applicant composition such that it is impossible to construct selection systems that show identical diversity values across the situations. The relative diversity factor is therefore nested within the crossing of the three selection situation factors. Also, within each situation, the diversity level values were chosen according to two criteria. First, the values must be attainable by at least one of the unit weighed selection systems that are feasible in the situation. Second, the level values should span as evenly as possible the major part of the range of diversity values achievable between the diversity level associated with the highest quality PO system (under C_c) and the diversity corresponding to the least quality PO system (under C_c). The diversity trade-off values corresponding to the PO systems P1 to P4 in Panel A of Figure 1 illustrate the resulting four factor levels for the selection situation S_0 described in the section "PO Selection Systems: a Brief Tutorial".

We added the relative diversity factor to the design because both the measures of calibration and validation quality achievement are defined with reference to the diversity level of the system and it is therefore important to assess whether the robustness/sensitivity of PO selection systems varies, depending on the diversity trade-off value of the systems. Also, Song et al. (2017) found that validity (diversity) shrinkage in PO systems is more pronounced to the extent that the PO system gives priority to the validity (diversity) objective. If this finding would also apply to the validation quality achievement of PO systems, then the low diversity PO systems (i.e., the systems that give a high priority to the quality objective) will show a lower level of validation quality achievement than the high diversity PO systems.

Finally, note that the relative diversity factor harbors an ambiguity with respect to the

unit weighed selection systems. Whereas variable weight systems may show different quality trade-off values for the same diversity trade-off value by adjusting the predictor weights in the composites, this is not the case with unit weighed systems. For these systems the diversity trade-off value corresponds to a unique quality trade-off value and, hence, to a unique value. So, choosing the unit weighed selection systems within each selection situation according to the diversity level value also fixes the calibration quality achievement value of the systems. As a consequence, the levels of the relative diversity factor confound diversity and calibration quality achievement in case (but only in case) of the unit weighed systems, and this confound will have to be taken into account when comparing the validation achievement of unit weighed and PO selection systems.

Overview and Implementation of the Study Design

Table 3 provides a summary of the 10 factors of the design. The design corresponds to the full crossing of nine of the factors, whereas the relative diversity level of the selection systems factor is nested within the crossing of the three selection situation factors. Given the number of levels of the eight factors that are used to provide a fairly exhaustive coverage of the different selection situations and the ways in which real settings deviate from the idealized conditions C_c, the design has a total of 3888 cells, with 24 selection systems (corresponding to the crossing of the two selection system factors) studied in each cell.

The implementation of the design proceeded in two stages, using throughout the minority selection rate and the average score on the job performance criterion as gauges for the diversity and the quality objective respectively. In the first stage a modified version of the COPOSS program (De Corte et al., 2011; De Corte, 2011) is used to identify the 24 selection systems under C_c for each of the 18 different selection situations obtained from the crossing of the three selection situation factors. The second, simulation stage involved the computation of the validation quality achievement of these systems when applied in the validation

conditions C_v corresponding to each of the 3888 cells of the design.

The actual execution of the second stage consisted of two steps. In the first step, the procedure of Ruscio and Kaczetow (2008) was used to generate 500 applicant predictor/criterion data samples within each of the 3888 cells according to the situational features and the C_v conditions that are specific for the cell. In the second step, the above described procedure for assessing the validation quality achievement of the selection systems was applied to each data sample within each cell of the design, resulting in the sample validation quality achievement value of the selection systems for the particular sample.

Obviously, given the size of the design and the numerical complexity, especially of the step to determine the validation quality achievement of the different systems for each sample within each of the cells of the design, the implementation of the study required massive computational resources as can only be delivered by a High Performance Computing facility. In particular, all computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI. All results were subsequently transferred to the SAS/Stat software environment for further analysis, using the means, tabulate and ANOVA procedure to provide answers to the key research questions of the paper.

Validation Quality Achievement of PO vs Other Selection Systems

First, we focus on the validation quality achievement of PO selection systems and, on the outcomes related to the sensitivity and the robustness of these systems. Next, we compare the sensitivity and robustness of the validation quality achievement of PO systems to that of other types of selection systems. Given the categorical measurement level of the studied factors, all results are obtained using appropriate ANOVA models. When the dependent variable in the models is a (function of a) proportion, we first applied the logit transformation to the dependent before executing the ANOVA analyses[8]. For each analysis we report the

percentage of variance explained by the models (i.e., the effect size measure η^2 expressed as a percentage), without correcting for the number of terms in the model because the latter, corrected percentages are virtually identical to the uncorrected percentages because of the large number of observations (i.e., at least 500) within the cells of the design. The effect size measure η^2 (expressed as a percentage) is also used to report the size of the effect of the important terms in the ANOVA models. As the present study is the first to explore the conditions that may affect the actual performance of PO selection designs, all conditions that show at least a small effect size value (i.e., explain at least one percent of the variance, cf. Cohen, 1988) are reported. Finally, note that the effect sizes of the terms in each ANOVA model sum to the effect size of the entire model because the models apply to an orthogonal design.

Robustness and Sensitivity of PO Selection Systems

To address the first key research question we conducted an ANOVA with the validation quality achievement value of the PO selection systems as the dependent variable. The set of independent variables in the ANOVA comprises all terms in the full model of nine of the ten design factors. The selection system type factor can be dropped because we study only one type of selection system (i.e., PO systems). The model explains 46.4 percent of the total variance and the bulk of the explained variance is due to the main effect of three factors: (1) the diversity of the system under C_c, explaining 19.6 percent of the variance, (2) the selection environment, 14.6 percent, and (3) the size of the applicant pool, 6.5 percent. Table 4 presents the average validation quality achievement value of the PO systems overall and broken down according to the levels of the three factors. The tabled values reveal that, across all studied conditions, the systems that are PO under C_c have an average validation quality achievement value of .661, but the level of achievement varies substantially across the three selection environments and the levels of diversity that characterize the systems. Thus, the validation

quality achievement is highest in selection environment three, whereas PO systems with higher diversity trade-off levels under C_c show a poorer achievement. The higher average validation quality achievement in environment three probably relates to the fact that this environment uses fewer predictors than the other environments. Previous research on shrinkage in regression models and the formulas used to predict the shrinkage (e.g., Cattin, 1980) indicate that the amount of shrinkage is inversely related to the number of predictors in the model; a result that is mirrored by the present finding that the selection environment with the least number of predictors offers the best validation achievement.

In contrast, the result about the lower validation quality achievement of high diversity (and therefore low quality) PO systems defies the expectation as based on the shrinkage results of Song et al. (2017). Whereas Song et al. found that low quality PO systems exhibit less quality shrinkage (i.e., less validity shrinkage as Song et al. use validity for the quality objective) we find that these systems have a lower validation quality achievement than the high quality (low diversity) PO systems. Apparently PO systems that give a higher priority to the diversity objective (and, hence a lower priority to the quality objective) tend to show a smaller validation quality achievement as compared to the lower diversity systems. The finding thereby indicates that quality shrinkage, as proposed by Song et al., could be misinterpreted by users as a gauge for the loss in the quality achievement by a PO system when implemented in validation conditions, at least when smaller quality shrinkage would be considered as an indication of higher quality achievement. Looking back at Figure 2, this finding does not come as a surprise, however, because the figure clearly shows that quality shrinkage and validation quality achievement are rather inverse indicators of the validation potential of selection systems: whereas smaller quality shrinkage might suggest a higher quality achievement, the reverse is the case. This is further substantiated by the results of the study reported in the section “Comparing Shrinkage and Validation Achievement” of the

online material. This study, albeit restricted to validation conditions involving the applicant population, additionally shows that the relationship between the corresponding diversity (quality) shrinkage and validation diversity (quality) achievement measures is not linear and even not entirely monotone.

Although the present research does not permit studying the relation between the relative diversity of the PO systems and validation diversity achievement of the systems in general validation conditions with finite applicant pools, the online material presents at least indicative results on this issue in the case of validation conditions involving applicant populations instead of finite applicant pools. These results show a rather proportional relationship between the relative diversity of a system and its validation diversity achievement, again contrary to the expectation based on the diversity shrinkage results.

From the four factors in the design that aim to study the sensitivity of the PO selection systems for discrepancies between the calibration conditions C_c and the validation conditions C_v only the size of the applicant pool explains at least one percent of the variability in the validation quality achievement values. As expected, the validation quality achievement of the PO systems is higher when the applicant pool is larger. In small applicant pool samples, as compared to large sized samples, the variability of the predictor correlation, validity and effect size values is considerably larger, implying that these values are more often substantially different from the values on which the PO selection systems are based, thereby resulting in a poorer validation quality achievement of the systems.

The effects related to the other sensitivity factors, although statistically significant (as almost all other effects in the ANOVA analysis because of the huge number of cases) explain only a negligible fraction of the total variability. Thus, the validation quality achievement of PO selection systems depends very little on the discrepancies between C_c and C_v as related to the proportional representation of the majority/minority candidates in the applicant pool and

the predictor/criterion mean and correlation structure values in the majority/minority populations, although the level averages of the latter factor show that the average validation quality achievement decreases for larger discrepancies in the predictor/criterion mean and correlation structure.

The ANOVA further indicates that the effect of the factor about the normality vs non-normality of the joint predictors/criterion distribution is also quite small (i.e., less than 1 percent explained variance). Although the validation quality achievement decreases somewhat in settings where the distribution is non-normal, the effect is not entirely consistent across the different environments and the levels of the applicant pool size factor. By and large, the finding implies that the assumption invoked by the decision aid about the multivariate normal distribution of the predictor/criterion scores in the applicant populations is not really critical. Fairly different joint predictor/criterion distributions only marginally affect the validation quality achievement of the PO systems.

Finally, the ANOVA reveals that none of the effects related to the interaction of the selection environment factor with (any combination of) the other factors explains a sizable portion of the total variance, implying that the above discussed effects about selection system diversity level and the size of the actual applicant pool apply in a similar way across the different types of selection environment and therefor are quite general. The result is also of key importance with regard to future studies about the features of the selection environment that impact on the actual performance of PO systems because it suggests that this future research can be conducted using a much more simple design that focuses on only these features without considering any additional factors.

Sampling Variability of the Validation Quality Achievement of PO Systems

With more than 50 percent unexplained variance, the ANOVA also shows that the sampling variability of the validation quality achievement of PO systems is quite large. Figure

3 illustrates this by showing the density plot of the validation quality achievement by selection environment (upper panel), by size of the applicant pool (middle panel) and by the diversity level of the PO systems (lower panel). Within the panels we also represented for each density the .1 (filled square) and the .9 quantile (filled circle) of the density, thereby indicating the interval that contains the 80 percent middle values of the validation quality achievement. Even for the largest applicant pool size, the width of this interval, with .1 and .9 quantile values of .52 and .92, is still quite substantial.

To determine the conditions that affect the sampling variability we applied a second ANOVA, with the within-cell (logit transformed) interquartile range of the validation quality achievement of the PO systems as the dependent variable and the main and the interactions effects (up to the fourth order) of the nine relevant factors of the design as independent variables. The model explains 94.4 percent of the variance. As expected, the effects related to the number of selected applicants provide together the largest contribution (i.e., the size of the applicant pool, 42.2 percent, and the selection rate under C_c and C_v factors with 5.0 and 2.3 percent, respectively), whereas the relative diversity factor (26.5 percent), the discrepancy between the moments of the joint predictor/criterion score distribution under C_c versus C_v (3.9 percent) and the selection environment factor (4.6 percent) are largely responsible for the remaining part of the explained variance. The average values of the dependent variable corresponding to these factors further show that the sampling variability of the validation quality achievement of a PO system is directly proportional to the relative diversity level of the system (i.e., the interquartile range values for the diversity levels 1 to 4 are 0.134, 0.150, 0.174 and 0.213), increases for systems using a larger number of predictors (i.e., the interquartile range values for the environments 1 to 3 are 0.179, 0.170, and 0.155), and decreases in situations with a larger number of (selected) applicants. The variability also increases for bigger differences between the validation predictor/criteria moment data and the

calibration moment data used to derive the PO systems (i.e., interquartile range values of 0.157, 0.164 and 0.182 for the levels one to three of the difference of the mean and correlation structure of the joint predictor/criterion score distribution under C_c versus C_v).

Integrating the results of the previous analyses it can be concluded that the conditions that substantially affect the magnitude and the variability of the validation quality achievement of a PO system are by and large the same. One may expect a higher validation quality achievement, and at the same time be more confident about this expectation (i.e., the sampling variability is smaller) when implementing a low diversity PO system, derived from fairly accurate predictor/criterion data and involving a small number of predictors, in a selection situation with a large number of (selected) applicants. However, the substantial decline from the value of 1 for the calibration quality achievement to the value of .661 for the validation quality achievement of the PO systems may raise concerns about the real practical utility of adopting these designs instead of other, more simple designs to address the selection quality/diversity quandary. To settle this issue, the next sections compare the robustness, the sensitivity and the sampling variability of both PO and other non-PO selection system designs, including the unit weighed designs.

Comparing the Robustness, Sensitivity and Sampling Variability of PO and Sub-PO Systems

We first report the results of the analysis comparing the validation quality achievement of the PO and the 0, 25, 50 and 75 percent calibration quality achievement systems. The analysis again applies an ANOVA model to the (logit transformed) validation quality achievement of the selection systems as the dependent variable, but this time using a slightly restricted model containing the main effects and all possible interactions up to the seventh order of all ten factors in the design. We imposed the restriction to stay within the limitations inherent to the SAS ANOVA procedure. The restriction does not affect the quality of the

analysis, however. Because the design is orthogonal, the sum of squares (and, hence, the proportion of explained variance) associated with the different effects remains the same whatever the set of effects that is included in the model.

The ANOVA model explains 52.6 percent of the total variance, with four effects related to the selection system type factor contributing at least one percent: the main effect of selection system type (25.9 percent), and the interaction of selection system type with the selection environment factor (9.1 percent) , the size of the applicant pool (2.8 percent) and the system relative diversity (1.5 percent) respectively. Table 5 summarizes the average validation quality achievement values corresponding to these four effects. The averages support the major conclusion that the order in the validation quality achievement level of the selection systems is maintained when these systems are applied in a large variety of selection settings. Note in particular that the three interaction effects do not invalidate this conclusion. Both overall and within each selection environment, within each size of the total applicant pool, and within each relative diversity level of the systems, the PO systems perform best, followed by the 75 (the S systems), the 50 (the F systems), 25 (the T systems) and the zero percent estimated performance systems (the Z systems), but the degree of separation between the validation quality achievement validation levels of the different selection systems varies significantly across the selection environments, the applicant pool size conditions and the system diversity levels. Also note that the average of validation quality achievement values of the different selection system types across the levels of the applicant pool size factor reflect the expectation that the validation quality achievement of selection systems with a high calibration quality achievement level (i.e., the 75 EP and the PO systems) is directly proportional to the size of the applicant pool, whereas the reverse is the case for the selection systems with a low calibration quality achievement level (i.e., the 0 and the 25 EP systems). Higher variability in the predictor/criterion data because of smaller applicant pool size should

more often benefit the validation quality achievement of systems with a low calibration quality achievement and have the opposite effect for high calibration quality achievement systems.

With more than 47 percent unexplained variance, the ANOVA again indicates substantial sampling variability. Figure 3 further illustrates the issue by displaying the density of the validation quality achievement of the five different selection system types, both overall and by the different selection environments. To study whether the different selection system types are more or less susceptible to sampling variability we performed a follow up ANOVA with the within cell (logit transformed) interquartile range of the validation quality achievement of the systems as dependent and the main effects of all ten design factors as well as the corresponding interactions (up to the fourth order) as independents. The ANOVA explains 94.7 percent of the variance with several effects related to the selection system type factor contributing at least one percent. Briefly summarized, the average interquartile range values corresponding to these effects reveal that the sampling variability is inversely related to the validation quality achievement level of the systems, that the trend is more pronounced for lower relative diversity systems, but weaker for larger sizes of the applicant pool. Yet, despite this variation, the important practical finding remains that PO systems apparently show a smaller within cell sampling variability than the non PO systems.

Comparing the validation quality achievement of PO and Sub-PO Systems at the Same, Single Application Level

Although PO systems maintain the highest validation quality achievement level, without showing more sampling variability, the substantial overlap of the density plots in Figure 3 suggests that PO systems may with some frequency result in a lower validation quality achievement than the sub-PO systems. To gather more precise information about this possibility, we recorded for each sample within each cell of the design the proportion with

which the validation quality achievement of PO systems is at least equal to that achieved by the corresponding 0, 25, 50 and 75 percent calibration achievement systems. Averaged across all samples and cells, these proportions equal .87, .84, .79 and .70, respectively, implying that the overall odds that PO systems outperform (i.e., have a higher validation quality achievement value) 0, 25, 50 and 75 percent systems at the single application level are 6.69, 5.25, 3.76 and 2.33 to one, respectively. These odds clearly show that PO selection systems not only maintain the highest validation quality achievement level, but also are much more likely to perform better than the sub PO systems when applied to the same single selection application.

From a practical perspective, the comparison in terms of robustness, sensitivity and sampling variability between the PO and the sub-PO systems showed that selection practitioners may expect a substantially better and a less variable validation quality achievement when implementing a PO instead of a sub-PO system. In the next section we study whether PO systems also maintain an advantage when compared to simpler unit weighed designs.

Comparing the Robustness, Sensitivity, Sampling Variability and Validation Quality Achievement at the Same, Single Application Level of PO and Unit Weighed Systems

The ANOVA analysis to explore the comparative robustness and sensitivity of PO and unit weighed selection systems, using the full model of all ten factors, explains 48.5 percent of the variance of the validation quality achievement of the systems. Only four of the effects related to the selection system type factor contribute at least one percent to the explained variance: the main effect of system type (13.8 percent), and the first order interaction of selection system type with the selection environment (1.8 percent), the applicant pool size (1.5 percent) and selection system diversity (3.6 percent). Table 5 summarizes the average validation quality achievement values associated with these effects. Except for the latter

selection system type by selection system diversity interaction effect, the averages related to the other effects essentially repeat the findings reported in the previous section, albeit this time with respect to the unit weighed systems: PO systems show a higher validation quality achievement than unit weighed systems, the difference grows for larger pool sizes and varies across selection environments.

The interpretation of the selection system type by selection system relative diversity interaction effect is less straightforward, however, because the relative diversity level of the unit weighed systems is inevitably confounded with the level of calibration quality achievement of these systems. Whereas PO systems have, by definition, 100 percent calibration quality achievement, the unit weighed systems have a calibration quality achievement that varies across the levels of the relative diversity factor. The selection system type by selection system relative diversity interaction may therefor very well reflect this difference in calibration quality achievement rather than indicate that the unit weighed systems have a different validation quality achievement pattern across the levels of the relative diversity factor as compared to that of the PO systems

The study comparing the PO and unit weighed systems also included the above detailed analyses focusing on (a) the susceptibility to within cell sampling variability of the two systems, and (b) the likelihood that the PO systems show a better validation quality achievement than the unit weighed systems at the same, single application level. By and large both analyses result in essentially the same findings as the corresponding studies comparing between PO and non PO systems. Thus, the first additional analysis reveals that PO systems show substantially less sampling variability than the unit weighed systems (cf. Figure 3) and that the difference in sampling variability between the two systems varies across selection environments and across the levels of the relative diversity and the applicant pool size factors. However, the average interquartile range values associated with these interactions never

indicate that PO systems have a larger within cell sampling variability. The variability in the calibration quality achievement of the unit weighed systems, as compared to the corresponding fixed 100 percent achievement of the PO systems, probably explains why the latter systems exhibit a smaller within cell sampling variability of the validation quality achievement values.

In turn, the second additional analysis results in an overall proportion of .75 that PO systems have a better validation quality achievement than unit weighed systems when applied in the same setting. The results of this and the previous analyses therefore warrant the conclusion that PO systems not only outperform variable weight sub-PO systems (i.e., systems using variable weights for the predictors in forming the predictor composites), but also fixed, and in particular, unit weight systems.

Comparing the Validation Trade-off Achieved by the Different Selection Systems

Thus far, all analyses and results focus on the new validation quality achievement measure as the criterion for evaluating the merits under general validation conditions of PO, sub-PO and unit weighed systems. Yet, despite the advantage of using this measure instead of other possible gauges it remains true that selection practitioners will often also be interested in the merits of the different selection systems as operationalized by the diversity/quality trade-off value achieved by the systems under such general validation conditions. In particular, they may wonder whether PO systems, when applied in validation conditions, are expected to result in a better diversity/quality trade-off (i.e., a trade-off where the value of the PO system on one of the objectives is higher than the corresponding value of the non-PO system, whereas the value of the PO system on the other objective is at least as high as the corresponding value of the non-PO system) as compared to the one achieved by a non-PO system in the same setting.

To clarify whether or not this is the case, we conducted a final analysis. For each of the

total of 7,776,000 studied sample selections we registered whether the trade-off achieved by the PO system, as compared to the trade-off achieved by the corresponding 0, 25, 50 and 75 percent calibration achievement systems is better, worse or incomparable. We found that the percentages with which the PO systems result in a better trade-off than the 0, 25, 50 and 75 percent systems equals 50, 48, 47 and 43, respectively, whereas the corresponding percentages with which the PO systems results in a poorer trade-off is equal to 6, 8, 10 and 15. In the remaining 44, 44, 43 and 42 percent of the comparisons with the 0, 25, 50 and 75 percent systems the trade-offs were incomparable in that neither trade-off is better than the other trade-off. Compared to the unit weighed systems, the PO systems result in a better (worse) trade-off in 48 (12) percent of all cases, with 40 percent incomparable results. Note that the substantial percentage of incomparable outcomes once again illustrates that an evaluation and comparison of both PO and non PO systems solely on the basis of the trade-off that these systems achieve under validation conditions is quite unsatisfactory and that achievement measures such as proposed in the paper are necessary to achieve this purpose.

Discussion

When learning about PO selection design, and the decision-aid for deriving these designs in particular, selection experts and practitioners may question whether PO systems will live up to expectations when implemented in a large variety of validation selection situations. These doubts can never be resolved conclusively because every future selection application harbors a number of inherent uncertainties. That said, the present paper offers a theoretical as well as a practical contribution that together succeed in generating rather convincing evidence to decide on the issue of the validation achievement of PO selection system design. From a theoretical perspective the paper introduces two new gauges for expressing the achievement of PO and other selection systems when applied under almost any type of validation condition. Compared to previous approaches, the validation quality and the

validation diversity achievement measures permit an adequate, unbiased and intuitively appealing assessment and comparison of the validation achievement of both PO and non PO selection system designs

From a practical perspective, the paper presents two novel procedures for computing the validation achievement of any selection system design as applied to virtually any selection situation, involving either finite applicant pool or infinite applicant population conditions. These procedures prove reliable, except for the evaluation of the validation diversity achievement in validation situations related to finite applicant pools. Also, the procedure for studying validation achievement with respect to applicant population validation conditions is made available to other researchers and practitioners. We encourage others to use the procedure because the analyses reported in the online material indicate that there is no real alternative short cut procedure that can provide a more easily obtainable estimation of the validation achievement of selection systems, even in the case of applicant population validation conditions. Finally, it is shown how the new procedures can be integrated within a factorial design to provide answers not only about the validation achievement of PO selection system design in a wide variety of validation conditions, but also, and even more importantly, about the major key issues addressed in the paper: do PO selection systems result in a higher validation achievement than non PO systems and is this higher achievement consistent across a large variety of validation conditions?

Are PO selection systems to be preferred to non PO systems?

Given our results, the answer to the above question is strongly in favor of a “yes”. In particular, we found an overall difference in validation quality achievement between the PO and the corresponding unit weighed systems of .16 (cf. Table 5). Using the procedure outlined in the section “Measuring the Achievement of Selection Systems in the Validation Condition”, this difference corresponds to an overall difference in expected job performance

of .10 standard units. Although this may not seem impressive at first, this is the same difference one may expect to obtain when switching from a predictor with a rather low validity of .30 to a predictor with a substantially higher validity of .42 when performing a selection with a .20 selection rate. Also, the gain of .10 standard units in average job performance did not come at the expense of a lower minority selection rate because both the PO and the unit weighed systems showed a virtually identical overall value (i.e., .166 versus .164 for the PO and the unit weighed systems) for this selection rate.

The validation quality achievement of PO systems also consistently and substantially exceeds that of other non PO systems. Furthermore, and although the sampling variability of the validation quality achievement level may be quite large for both PO and non PO systems, the odds, that PO systems have a higher validation quality achievement than the sub PO or the unit weighed systems when they are all applied to the same setting, are well above two to one in virtually all studied validation conditions. Finally, we found no evidence confirming that unit weighed systems are more robust and/or less sensitive than variable weight systems and PO selection systems in particular.

Observe that the present results substantially extend previous findings about the merits of PO selection systems. Whereas all former findings relate to the behavior of these systems in validation contexts involving the total applicant population, the present results inform about the achievement in (small, medium sized, etc.) sample validation conditions that are the real center of interest in an applied setting like personnel selection. In addition, the new measures used to assess the merits of the different selection systems avoid the deficiencies associated with the previously used methods.

Summing up, the message of the present analyses should be clear. If both the goals of selection quality and diversity are of importance, at least approximate data on the predictor/criterion characteristics are available, and provided that the design of the selection

process is not entirely fixed by the constraints of the selection situation, practitioners have good reasons to implement a PO selection system design. Any selection design boils down to a decision that is to be made and according to the results of the present study, PO selection designs represent the decision with the most favorable expected consequences. When probed under a large variety of validation conditions, these designs show the best validation quality achievement. In addition, when applied to the same single selection, the odds that PO systems attain a higher validation quality achievement level than the non PO systems are favorably, exceeding two to one in almost all applications. Finally, compared to non PO systems, the implementation of PO systems results more often in a quality/diversity trade-off that is better than the trade-off achieved by these other systems.

Limitations of the Study and Avenues for Further Research

Let us start by repeating that neither the present nor any following study can produce conclusive and final answers about the robustness and sensitivity of PO selection systems. Although the design of the study aimed for a comprehensive inclusion of the factors that may affect the robustness/sensitivity of these systems, using representative levels for these factors, certain possibly important factors may have been omitted and other more pertinent specifications for the levels of the factors (e.g., smaller applicant pool sizes and higher selection rates than those considered to reflect current labor shortages) may have been missed. Thus, future studies could be designed to provide more detailed answers about the features of the selection environment that either favor or impede the robustness/sensitivity of PO systems. Although we varied the nature of the selection environment and, in particular, the number of available predictors in the different environments, other features, related to, for example, the specific blend of available predictors (e.g., the set of predictors is quite homogeneous, with all predictors assessing either the same construct or different lower order variations of the same construct, or more heterogeneous, focusing on a mixture of different

constructs), the application context and the demarcation of the set of feasible selection systems, were not really considered systematically. To this end, a number of smaller scale studies, focusing on only one or a limited number of these environment features (e.g., comparing single stage versus multi stage environments, keeping all other features constant) may be more appropriate. The present results, showing that none of the interaction effects of the selection environment factor with the other studied factors explains a sizable portion of the PO system variability, indicates that such smaller scale studies may indeed be adequate.

As a second limitation, the study does not present results on the validation diversity achievement of the selection systems primarily because of as yet unresolved technical issues in the computation of the measure in validation conditions with finite applicant pool sizes, but we also note that even without these issues, the measure is still somewhat problematic for studying the validation achievement in small applicant pool validation conditions because its value will often be undefined in these conditions. Also, the online material reports results about both the quality and diversity achievement in case of validation conditions related to applicant populations. These results confirm the finding that lower diversity PO systems (and, hence, higher quality PO systems) tend to show a higher validation quality achievement as compared to the higher diversity PO systems. In addition, they suggest a rather opposite trend with respect to the validation diversity achievement of the systems: lower diversity PO systems have a lower validation diversity achievement. These results once again underscore that focusing on the level of diversity and quality shrinkage between the calibration and the validation condition as proposed by Song et al. (2017) may result in users forming a rather poor picture of the true merits under validation of the selection systems. The online supplement study implements side by side both the Song et al. shrinkage and the novel validation achievement calculations showing that higher rather than lower levels of quality (diversity) shrinkage correspond to higher validation quality (diversity) achievement.

Whereas the present study confirms this for validation conditions involving finite applicant pools with respective to the quality dimension, future studies should consider the diversity dimension as well, provided that the difficulties regarding the computation of the validation diversity achievement in finite applicant pool validation conditions can be resolved.

As a further possibility for future study it would be interesting to assess how the validation achievement of PO systems is affected by common realities such as the refusal of job offers or the drop-out of candidates during the selection process. Several variants of job refusal/drop-out could be considered in the validation condition, paying attention to, among others, the differential effect of random versus systematic forms of candidate self-selection where job refusal/drop-out is related to the quality of the candidates.

Adding the sample to sample cross-validation approach more explicitly to the study design constitutes a final avenue for future research. To achieve this purpose only one major extension is required. Instead of computing the PO and other systems only once, the systems (and the corresponding calibration trade-offs) should be computed with respect to a large number of (calibration) sample based data as done by Song et al. (2017). The sample to sample cross-validity research question can then be addressed by invoking the above outlined procedure for computing the validation achievement of the systems when applied to a large number of finitely sized validation samples and by averaging the thus obtained achievement values.

Conclusion

The paper reports a massive simulation study investigating whether PO systems live up to expectations when implemented under a large variety of validation selection conditions. Although by no means conclusive, the obtained results nevertheless converge to the conclusion that PO systems, as derived by the psychometric approach proposed by De Corte et al. (2007, 2011) are indeed expected to outperform other non PO systems, including unit

weighed systems. The results therefore add substantial weight to the advice that selection practitioners and researcher should consider applying PO selection designs whenever possible. Otherwise they may face complaints and even legal actions because plaintiffs can argue quite convincingly, not only on the formal grounds implied by the formulas of the psychometric approach, but also based on the results of the present study, that a better design was indeed possible.

References

Beck, J. A., Beatty, A. S., & Sackett, P. R. (2014). On the distribution of job performance: The role of measurement characteristics in observed departures from normality. *Personnel Psychology, 67,* 531-566. http://dx.doi.org/10.1111/peps.12060 .

Blanca, M. J., Arnau, J., Lopez-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9,* 78-84. http://dx.doi.org/10.1027/1614-2241/a000057 .

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

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

Chalabi, Y., Scott, D. & Wuertz, D. . (2012). The Generalized Lambda Distribution as an Alternative to Model Financial Returns. Retrieved from https://www.rmetrics.org/sites/default/files/glambda_0.pdf on 30/11/2017.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York: Academic Press.

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,* 274-290. http://dx.doi.org/10.1037/apl0000163 .

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software, 34,* 1-24.

Day, N. E. (1969). Estimating components of a mixture of normal distributions. *Biometrika, 56,* 463-474. http://dx.doi.org/10.1093/biomet/56.3.463

Deb, K., Pratap. A, Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transaction on Evolutionary Computation, 6*, 181-197. http://dx.doi.org/10.1109/4235.996017 .

De Corte, W., Lievens, F., & Sackett, P. R. (2006). Predicting adverse impact and mean criterion performance in multistage selection. *Journal of Applied Psychology, 91*, 523-537. http://dx.doi.org/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*, 1380-1393. http://dx.doi.org/1037/0021-9010.92.5.1380 .

De Corte, W. (2011). COPOSS user's guide. Retrieved from http://users.ugent.be/~wdecorte/software.html.

De Corte, W., Sackett, P. R., & Lievens, F. (2011). Designing Pareto-optimal selection systems: Formalizing the decisions required for selection system development. Selecting predictor subsets considering validity and adverse impact. *Journal of Applied Psychology, 96*, 907-926. http://dx.doi.org/10.1037/a0023298 .

Dorigo, M. & Stutzle, T. (2004). *Ant colony optimization*. MIT Press, Cambridge, MA.

Hardin, J., Garcia, S.R., &Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics, 7*, 1733-1762. http://dx.doi.org/10.1214/13-AOAS638 .

Johnson, J. W., Abrahams, N., & Held, J. D. (2004, April). *A procedure for selecting applicants considering validity and adverse impact.* Poster presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, ILL.

Kehoe, J. F. (2008). Commentary on Pareto-Optimality as a rationale for adverse impact reduction: what would organizations do? *International Journal of Selection and Assessment, 16*, 195-200. http://dx.doi.org/10.1111/j.1468-2389.2008.00424.x

Micceri, T. (1989). The unicorn, the normal curve and other improbable creatures. *Psychological Bulletin, 105*, 156-166. http://dx.doi.org/10.1037/0033-2909.105.1.156

O'Boyle, E., & Aguinis, H. (2012). The best and the rest: revisiting the norm of normality of individual performance. *Personnel Psychology, 65*, 79-119. http://dx.doi.org/10.1111/j.1744-6570.2011.01239.x

Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. . *Multivariate Behavioral Research, 43*, 355-381. http://dx.doi.org/10.1080/00273170802285693

Ryan, A., & Ployhart, R. E. (2014). A century of selection. *Annual Review of Psychology, 65*, 693-717. http://dx.doi.org/10.1146/annurev-psych-010213-115134

Sackett, P., De Corte, W., & Lievens, F. (2008). Pareto-optimal predictor composite formation: A complementary approach to alleviating the selection quality/adverse impact dilemma. *International Journal of Selection and Assessment, 16*, 206-209. http://dx.doi.org/10.1111/j.1468-2389.2008.00426.x

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice. A guide to assessing scientific models.* Chichester: John Wiley & Sons.

Smithson, M, & Verkuilen, J. (2006). A better lemon squeezer? Maximum-Likelihood regression with Beta-distributed dependent variables. *Psychological Methods, 11*, 54-71. http://dx.doi.org/10.1037/1082-989X.11.1.54

Song, Q., Wee, S., & Newman, D. A. (2017). Diversity shrinkage: Cross-validating Pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology*. Advance online publication. http://dx.doi.org/10.1037/apl0000240 .

USA Department of Labor, Manpower Administration (1970). *General Aptitude Test Battery. Section III: Development*. Washington, DC: Author.

Wee, S., Newman, D. A., & Joseph, D. L. (2014). More than g: selection quality and adverse

impact implications of considering second-stratum cognitive abilities. . *Journal of Applied Psychology*, *99*, 547-563. http://dx.doi.org/10.1037/a0035183

Footnotes

[1]In the paper and in the simulations we assume that the joint distribution of the predictors and the criterion in the total applicant population is a mixture of two multivariate distributions characterized by the same variance/covariance matrix but a different mean vector in both the minority and majority applicant population. It is well known that this implies a different variance/covariance matrix in the total population (e.g., Day, 1969) and that a mixture of normal distributions with a different mean is no longer normal. These consequences are of little if any practical consequence in the present context, however (cf. De Corte et al., 2011). Also, the assumption that the minority and the majority populations share the same variance/covariance matrix can easily be dropped.

[2]The T,F and S systems are computed by solving a constrained nonlinear program similar to the program used to calculate the PO systems in the decision-aid of De Corte et al. (2011). However, two nonlinear equality constraints are added to the program formulation to ensure that the T, F and S systems have the desired selection quality and diversity value.

[3]Wee et al. (2014) did not generate finite sized samples of predictor/criterion scores but sampled predictor/criterion correlation matrices and effect sizes, and inserted the thus obtained correlation/effect size data in the population formula of De Corte et al. (2006) to compute the selection system trade-off values in the validation condition. The obtained trade-offs therefor apply to validation settings involving the total applicant population. A similar remark also applies to the Song et al. (2017) study because this study used very large validation samples of 10,000 applicants as a proxy for the total applicant population.

[4]The procedure adopted in De Corte et al. (2011) anticipates on this limitation but leads to an unsatisfactory quantification of the achievement in the new setting of PO and other selection systems because it compares each system to a different set of dominated systems

and, hence, uses a different gauge for each of the systems.

[5]For general multi stage selections, the diversity trade-off can only be gauged by either the minority (majority) selection rate or the AIR. Also, although the issue on undefined validation diversity achievement values can again be resolved by adopting the convention to assign a value of one, this practice will distort the results when studying, for example, the effect of the validation applicant pool size on the validation diversity achievement of selection systems.

[6]In case of a multi stage selection system with a final selection ratio, as specified under C_v, that differs from the selection rate indicated under C_c we translated the intermediate retention rates of the system corresponding to C_c to matching retention rates under C_v such that the ratio between the overall selectivity rate under C_c and the overall selectivity rate under C_v is evenly distributed across the selection stages. As an example, consider the situation where under C_c the selection rates of a two stage selection system are equal to .30 after the first stage and .10 after the final stage, whereas the final selection rate equals .40 under C_v. In that case, the ratio between the overall selection rate under C_c and the overall selection rate under C_v is equal to ¼ and evenly distributing this ratio across the two stages implies being ½ less selective in both stages, implying selection rates of 0.60 and 0.40 in stage one and two for the translated system under C_v.

[7]Repeatedly solving the same optimization problem, starting from different initial values for the problem variables, typically resulted in solutions that either differed in the solution value for the objective function or (more often) violated the equality constraint imposed on the quality level of the selection system. However, note that the computation of the validation diversity achievement measure is perfectly reliable for validation conditions involving applicant populations because in that case both the objective function and the equality constraint of the optimization problems are analytic instead of nonanalytic functions

such that classic gradient based methods, that can routinely handle nonlinear equality constraints, can be used to solve the problems.

[8]To avoid problems with proportions equal to 0 or to 1, we adopted the suggestion of Smithson and Verkuilen (2006) to first transform the proportions p to $(p(n-1)+0.5)/n$, with n the sample size equal to the number of cases within the cells (i.e., n=500) and then apply the logit transformation. We also duplicated the analyses, applying beta regression (Cribari-Neto and Zeileis, 2010) to a number of simpler models, but found that both the present and the beta regression approach result in virtually the same effect size estimates for the effects in the models.

Table 1.
Predictor/Criterion Data for the Three Types of Selection Environment

Selection Environment 1 (cf. GATB, US Department of Labor, 1970)										
	Predictor	$d^{\#}$	1	2	3	4	5	6	7	8	9
1.	Intelligence	0.95									
2.	Verbal	0.87	0.74								
3.	Numerical	0.71	0.76	0.67							
4.	Spatial	0.74	0.64	0.46	0.54						
5.	Form Perc	0.54	0.64	0.47	0.58	0.59					
6.	Clerical	0.47	0.61	0.62	0.66	0.39	0.65				
7.	Motor	0.10	0.36	0.37	0.41	0.20	0.45	0.54			
8.	Finger	0.32	0.25	0.17	0.24	0.29	0.42	0.32	0.37		
9.	Manual	0.14	0.19	0.10	0.21	0.21	0.37	0.26	0.46	0.32	
	Criterion										
1.	Performance	0.38	0.29	0.27	0.35	0.26	0.31	0.34	0.30	0.24	0.25

Selection Environment 2 (cf. Johnson et al., 2004)											
	Predictor	$d^{\#}$	1	2	3	4	5	6	7	8	9
1.	General Science	1.008									
2.	Arithmetic Reasoning	0.725	.598								
3.	Verbal	0.684	.780	.629							
4.	Mathematics Knowledge	0.162[a]	.467	.694	.475						
5.	Mechan. Comprehension	0.992	.596	.620	.561	.413					
6.	Auto & Shop Info	1.213	.593	.432	.506	.090	.725				
7.	Electronics Information	0.797	.649	.516	.622	.335	.642	.757			
8.	Assembling Objects	0.502	.430	.532	.426	.456	.574	.348	.398		
9.	Coding Speed	0.178	.272	.373	.337	.415	.192	.029	.169	.294	
	Criterion										
1.	Performance	0.380	.522	.545	.561	.407	.545	.529	.525	.442	.341

[a] This *d* value seems low, but it is the value mentioned by Johnson et al.

Selection Environment 3 (cf. De Corte et al., 2011)							
	Predictor	$d^{\#}$	1	2	3	4	5
1.	Cognitive Ability	0.72					
2.	Structured Interview	0.32	.31				
3.	Conscientiousness	0.06	.03	.13			
4.	Biodata	0.57	.37	.17	.31		
5.	Integrity	0.04	.02	-.02	.34	.25	
	Criterion						
1.	Performance	0.38	.52	.48	.22	.32	.42

#: *d* corresponds to the standardized mean difference between the majority and the minority applicant populations.

Table 2.
Constraints Demarcating the Set of Feasible Selection Systems

Selection Environment 1
Only single stage selection systems are admissible. Any combination of the available predictors may be used as the selection composite provided that the weight of the predictors in the composite has a value between 0 and 1 (both included) and the composite is not less valid than the most valid predictor (i.e., has a validity not less than .35).
Selection Environment 2
A selection system is feasible if it corresponds to a two stage selection design in which a composite of the first four predictors is used in the first stage and a composite of the remaining five predictors is used in the second stage. The predictor weights in the composites must have a value between 0 and 1 (both included). The retention rate after the first stage must be between .3 and .6, .4 and .7, and .55 and .75 when the selection rate under C_0 equals .1, .2 and .4, respectively.
Selection Environment 3
A selection system is feasible if it corresponds to (a) a single stage design in which the final accept/reject decision is based on a weighed composite of the CA, CO, BI and IN predictors, (b) a two stage design where the candidates are first screened on the basis of a weighed composite of CA, CO and BI, and the remaining candidates are selected/ rejected using a weighed composite of the SI and IN predictors, or (c) a three stage design where the intermediate retention decisions involve top-down selection on a CA and IN composite and a CO and BI composite for the first and the second stage respectively, and the SI predictor is used in the final selection stage. The predictor weights in the composites must have a value between 0 and 1 (both included). For the two stage selection systems, the retention rate after the first stage must be in the range .3-.6, .35-.6, and .55-.75 when the selection rate under C_0 equals .1, .2 and .4, respectively. For the three stage selection systems, the retention rate after the first (the second) stage must be in the range .55-.70 (.25-40), .60-.75 (.35-.45) and .70-.80 (.50-.55) when the selection rate under C_0 equals .1, .2 and .4, respectively.

Table 3.
Study Design Factors

	Selection Situation Factors
1:	Selection Environment, with 3 levels (see Tables 1 and 2 for details)
2:	Selection Rate under C_c, with 3 levels: .10, .20 and .40
3:	Proportional Representation of the Majority Group under C_c, with 2 levels: .50 and .80
	Factors Differentiating between the Calibration Conditions C_c and the Validation Conditions C_v
4:	Normality vs non-normality of the Joint Distribution under C_v of Predictors and Criterion (i.e., the nature of the joint predictors/criterion score distribution in the parent majority and minority validation population from which applicant pools are sampled), with 3 levels: 1. multivariate normal, 2. moderate non normal with marginal distributions having skew 0.75 and kurtosis 4, and 3. strong non normal with marginal distributions having skew 2.00 and kurtosis 9.00
5:	Difference of the Mean and Correlation Structure of the Joint Population Predictor/Criterion Score Distribution under C_c versus C_v with 3 levels: identical, moderate difference (i.e., absolute difference between corresponding mean elements of .10 and .05 absolute difference between corresponding correlation elements), and strong difference (i.e., absolute difference between corresponding mean elements of .20 and .10 absolute difference between corresponding correlation elements),
6:	Selection Rate under C_v, with 3 levels: 1. identical to the C_c selection rate, 2. 50 percent higher than the C_c selection rate, and 3. 50 percent smaller than the C_c selection rate.
7:	Proportional Representation of the Majority Group under C_v, with 2 levels: 1. Identical, and 2. Different (e.g., the Proportional Representation of the Majority Group under C_v equals 0.5 when the Proportional Representation of the Majority Group under $C_c$equals 0.80)
8:	Size Applicant Pool under C_v, with 4 levels: 80, 250, 800 and 2500
	Selection System Factors
9:	Selection System Type with 6 levels: 1. PO with 100 percent estimated performance value (i.e., 100 percent performance value under C_c) (P systems), 2. Non PO with 0 percent estimated performance value (W systems), 3. Non PO with 25 percent estimated performance value (T systems), 4. Non PO with 50 percent estimated performance value (F systems), 5. Non PO with 75 percent estimated performance value (S systems), and 6. Unit weighing system (U systems)
10:	Diversity Trade-off Value Selection System (under C_c) with 4 levels: level 1 lowest, level 4 highest

Table 4.
Validation Quality Achievement of PO Systems: Global and According to the Relative Diversity of the Selection System (DIV, with levels 1 to 4 coding for increasing degrees of diversity), Selection Environment (SEN, with levels 1 to 3 coding for the environments 1 to 3; see Tables 1 and 2), Size of the Applicant Pool (SIZ, with level 1: 80 applicants; level 2: 250 applicants; level 3: 800 applicants; and level 4: 2500 applicants).

	Validation Quality Achievement			
Overall	.661			
	Level			
	1	2	3	4
DIV	.746	.715	.646	.538
SEN	.619	.612	.753	
SIZ	.612	.625	.671	.736

Table 5.

Average Validation Quality Achievement of PO and non PO Systems (0 Percent Calibration Quality Achievement, 0 CA, until 75 Percent Calibration Quality Achievement , 75 CA): Overall, by Environment, by Size of the Applicant Pool and by Relative Diversity of the Selection System.

		Selection System Type					
		0 CA	25 CA	50 CA	75 CA	PO	Unit
Overall		0.287	0.383	0.472	0.567	0.661	0.502
Environment	1	0.436	0.483	0.529	0.575	0.619	0.516
	2	0.282	0.366	0.452	0.533	0.612	0.457
	3	0.144	0.300	0.434	0.593	0.753	0.535
Size Pool	1 (80)	0.381	0.439	0.494	0.553	0.612	0.519
	2 (250)	0.307	0.388	0.463	0.544	0.625	0.494
	3 (800)	0.251	0.358	0.457	0.565	0.671	0.492
	4 (2500)	0.211	0.346	0.473	0.606	0.736	0.505
Diversity	1 (low)	0.345	0.445	0.540	0.648	0.746	0.515
	2	0.299	0.406	0.506	0.614	0.715	0.536
	3	0.258	0.355	0.450	0.543	0.646	0.559
	4 (high)	0.247	0.325	0.392	0.464	0.538	0.399

Figure 1.
Quality/Diversity Trade-offs Achieved by Various Selection Systems for Situation S_0 under Calibration Conditions C_c (Panel A) and Validation Conditions C_v (Panel B). Under C_c the systems with trade-offs P1 to P4 are Pareto optimal (100 percent calibration quality achievement); the systems with trade-offs W1 to W4 are the worst possible (0 percent calibration quality achievement); and the systems with trade-offs S1 to S4, F1 to F4, and T1 to T4 have 75, 50 and 25 percent calibration quality achievement, respectively. The systems with trade-offs U1 to U4 correspond to unit weighed selection systems. In panel B, the same symbols identify the trade-offs achieved by the systems under conditions C_v, whereas the symbols B1 to B4 and W1 and W4 show the best possible and the worst possible corresponding trade-offs that can be achieved under C_v.

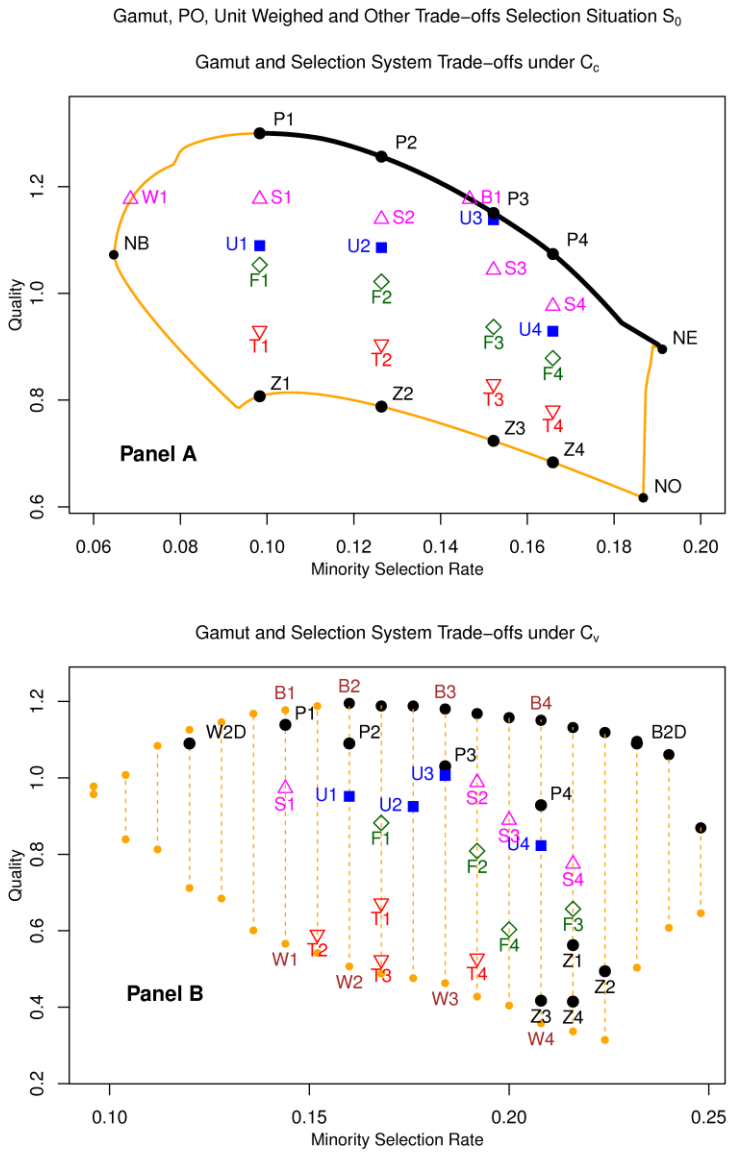

Figure 2.
Average Diversity/Quality Shrinkage in a Validation Condition Involving the Applicant Population

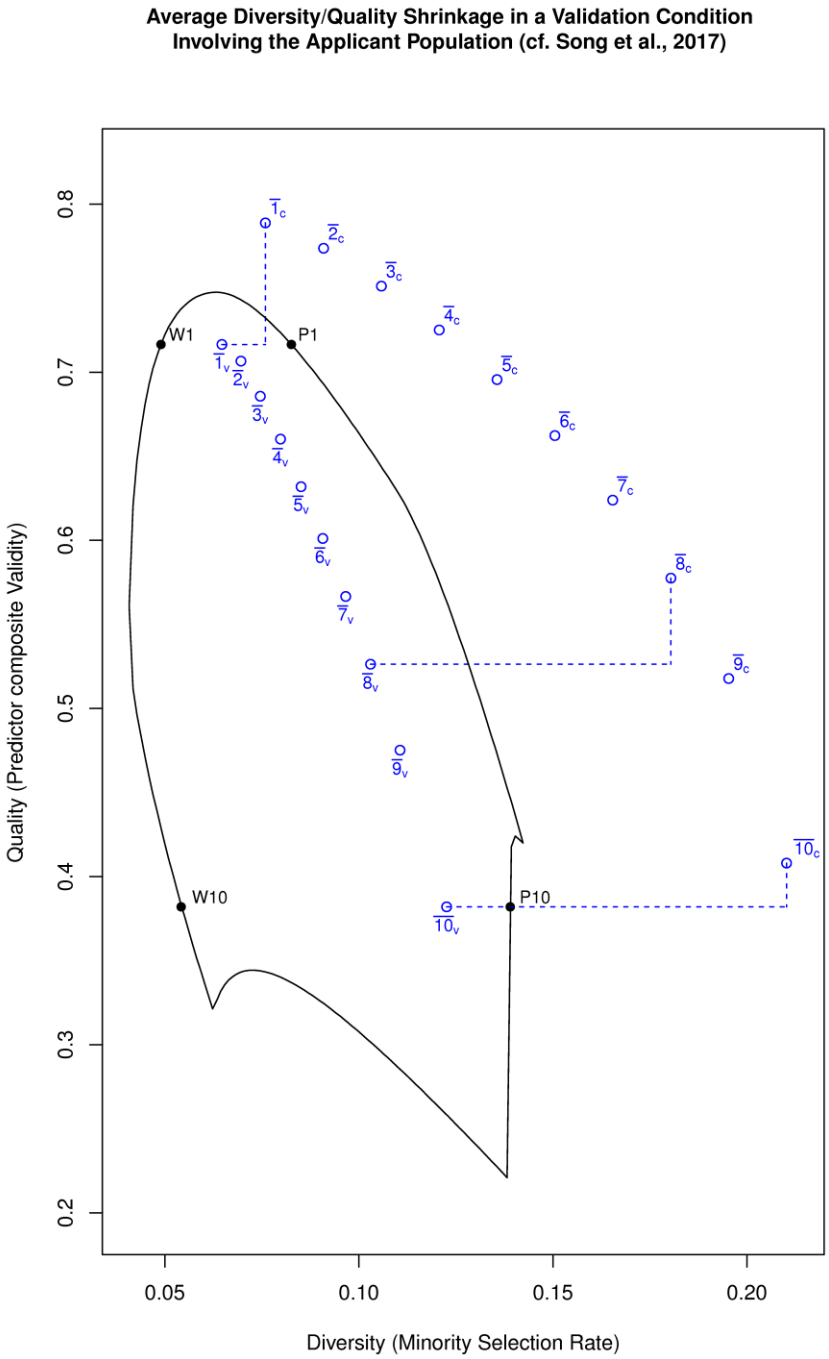

Figure 3.
Sampling Variability Validation Quality Achievement of PO Selection Systems. The square dots on the horizontal axis indicate the .10 quantile of the density, whereas the circle dots show the .90 quantile.

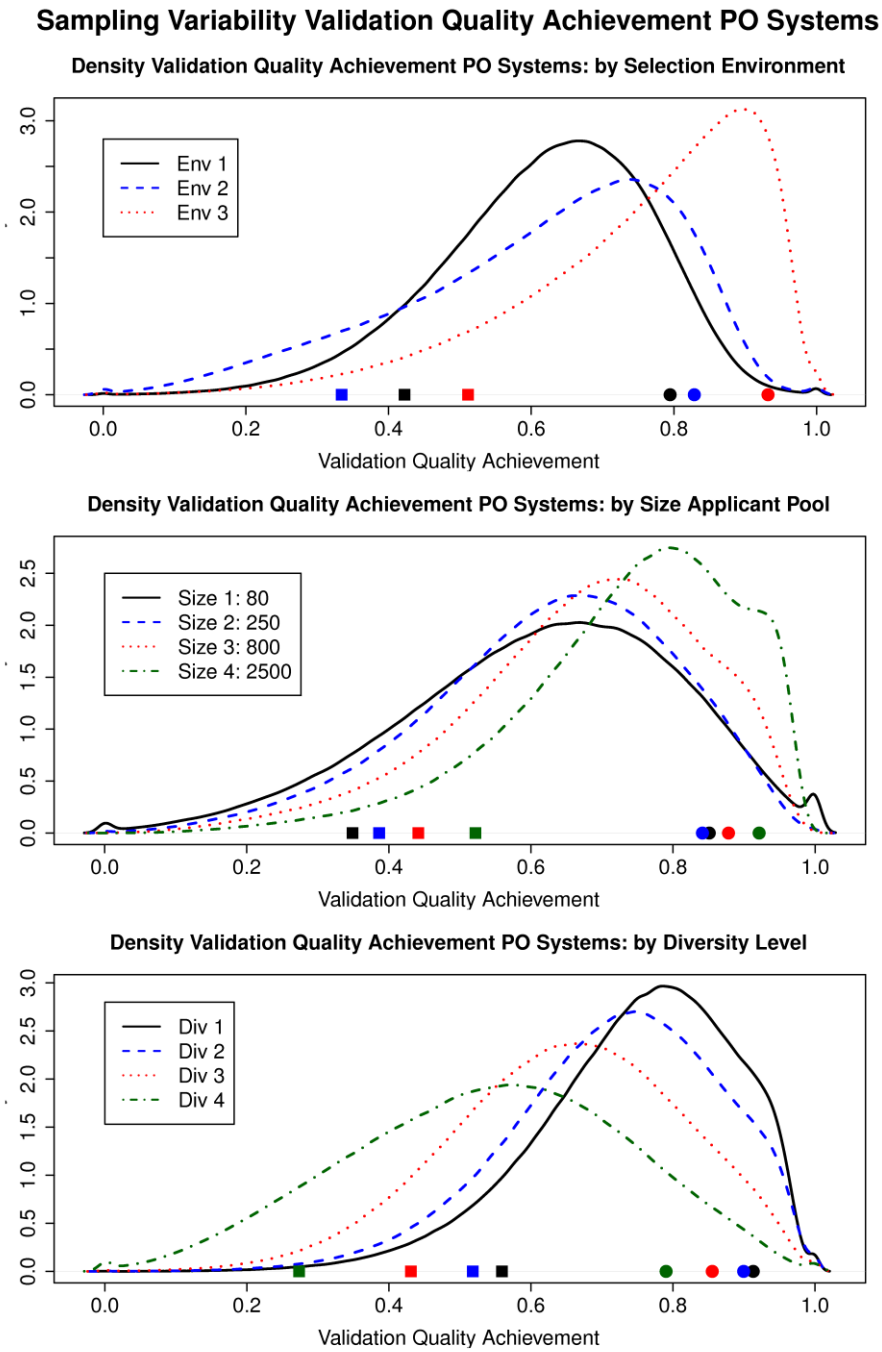

Figure 4.
Sampling Variability Validation Quality Achievement of PO and non PO Systems (P SYS: PO System; Z SYS: zero percent calibration quality achievement system; T SYS: 25 percent calibration quality achievement system; F SYS: 50 percent calibration quality achievement system; S SYS: 75 percent calibration quality achievement system; U SYS: unit weight system