

Selecting Predictor Subsets: Considering validity and adverse impact

Wilfried De Corte*, Paul Sackett** and Filip Lievens*

*Ghent University, Ghent, B-9000, Belgium. wilfried.decorte@ugent.be

**University of Minnesota, Twin Cities Campus, MN, USA

The paper proposes a procedure for designing Pareto-optimal selection systems considering validity, adverse impact and constraints on the number of predictors from a larger subset that can be included in an operational selection system. The procedure determines Pareto-optimal composites of a given maximum size thereby solving the dual task of identifying the predictors that will be included in the reduced set and determining the weights with which the retained predictors will be combined to the composite predictor. Compared with earlier proposals, the simultaneous consideration of both tasks makes it possible to combine several strategies for reducing adverse impact in a single procedure. In particular, the present approach allows integrating (a) investigating a large number of possible predictors (such as multitest battery of ability tests, or a collection of ability and nonability measures); (b) explicit predictor weighting within feasible test procedures of a given limited size.

1. Introduction

De Corte, Lievens, and Sackett (2007) first introduced the concept of Pareto-optimality to the personnel selection community with their treatment of the validity-adverse impact trade-off in creating composites of predictors. They noted that while there are a limitless number of possible sets of weights one might apply when combining predictors into a composite, only a small number of these are Pareto-optimal, meaning that one objective cannot be improved without harming the other. More specifically, a set of predictor weights is called Pareto-optimal when they result in a validity-adverse impact trade-off that can not be bettered because any other predictor weighting results in either (or both) lower validity or higher adverse impact. There is a set of Pareto-optimal weights for various attainable levels of validity, and, comparably, a set of Pareto-optimal weights for various attainable levels of adverse impact. Obtaining the set of Pareto-optimal solutions across the full spectrum from the validity-maximizing solution to the adverse impact-minimizing solution gives the researcher clear information about possible trade-offs between the two outcomes. In other words, it gives a clear picture of what could be gained regarding one outcome (e.g., adverse impact) if one were willing to accept a specified reduction in the other outcome (e.g., validity).

De Corte, Lievens, and Sackett (2006), De Corte *et al.* (2007) presented the setting in which one is considering a set of predictors for possible use. All predictors in the set are viewed as available for use, and there is no constraint on the number of predictors. In this paper we consider the setting in which feasibility constraints, related to for example total testing time or cost, limit the number of predictors that can be put into operational use, and in which one's task is to choose a smaller number of predictors from a larger initial set. In other words, we consider addressing both the predictor subset selection decision (i.e., which predictors should receive nonzero weights) and the predictor weighting decision (i.e., which nonzero weights should be assigned to the selected predictors) when designing optimal selection systems that must conform to given minimum feasibility requirements (cf. Kehoe, 2008).

Although predictor subset selection is quite common in the actual practice of personnel selection, methods, or algorithms for addressing the subset selection problem received little if any attention in the selection research literature during the past few decades. The obvious reason for this seems to be the paramount interest in maximizing validity when choosing between alternative predictor subsets. In that case the predictor subset selection decision is easily resolved by choosing the subset that shows the largest multiple correlation with

the criterion. However, the problem is less easily addressed when other concerns, such as adverse impact reduction, are also of interest, resulting in the validity-adverse impact dilemma that is now widely recognized as the most perplexing problem facing the practice of personnel selection today (Campion *et al.*, 2001). Moreover, because broadening the array of predictors under consideration is suggested as one of the most effective strategies for reducing adverse impact (e.g., Ployhart & Holtz, 2008), we see the investigation of larger numbers of initial predictors as a strategy that will be used with more frequency such that the problem of selecting predictor subsets for reasons of test feasibility is likely to occur with considerable frequency in the future.

To address the issue, we propose a method for identifying subset predictor composites that lead to a Pareto-optimal trade-off in terms of both validity and adverse impact reduction. Formally, while De Corte *et al.* (2007) provided a solution to the question of 'how does one identify the set of Pareto-optimal composites given N predictors?' the present method addresses the question of 'how does one identify the set of Pareto-optimal composites of at most n predictors from a larger set of N predictors while conforming to eventual other minimal feasibility requirements?' In presenting the new approach we will, for the sake of simplicity, focus on the situation in which only the attributes of selection quality and adverse impact are of concern, but the procedure can be extended in case of more than two valued attributes.

To situate the new procedure for deriving maximum size Pareto-optimal predictor composites we first describe earlier approaches to the selection of predictor subsets that aim to balance the concerns of adverse impact and selection quality. Then we present our new procedure and we use an example application to highlight the benefits of this procedure as compared with the results of the earlier approach. Finally, we summarize the results of three additional studies. These studies focus on (a) comparing the potential of predictor subsets of a different maximum size (b) the impact of varying predictor weight ranges and (c) the robustness of Pareto-optimal predictor subsets for variability in the predictor characteristics.

2. Earlier approaches

Although the pursuit of valued selection goals such as workforce diversity and quality received a lot of interest in the personnel selection literature, this research paid little attention to systematic procedures for selecting predictor subsets when balancing these goals within selection systems that have to meet certain minimum feasibility requirements in terms of overall cost, duration and so on. In fact, we know of only one study that addressed the issue. Johnson, Abrahams, and Held (2004)

focused on finding the three-predictor composite from the nine available ASVAB tests that 'would maximize validity and minimize adverse impact to the extent possible' (p. 4). To achieve this purpose, Johnson *et al.* propose using a single ad hoc metric to capture the combined benefits in terms of adverse impact and validity of different predictor subsets. The metric aims for evaluating the validity/effect size potential of composite formation in which it takes progressively larger increases in validity to offset increases in effect size and, hence, in adverse impact.

The conception of a single metric to evaluate the adverse impact/selection quality potential of a predictor composite requires a number of difficult, ad hoc decisions, however. Thus, similar to the implementation of a multiattribute utility analysis approach (Aguinis & Harden, 2004; Roth & Bobko, 1997), it must be decided how the performance of a test composite with respect to the different concerns of adverse impact and validity should be scaled to 'utility points' and how these utilities should subsequently be weighed to an overall effectiveness value. Also, as shown in the next section, the single metric approach leads at best to only one Pareto-optimal trade-off composite from the entire set of such trade-offs and an altogether different procedure is needed to uncover the latter entire set. Thus while Johnson and colleagues work offered a useful initial foray into the validity-adverse impact trade-off, we believe that the approach we develop here is a useful next step in understanding and managing this trade-off.

3. Procedure for deriving pareto-optimal predictor subset composites

Our method requires certain data, but these data do not differ from the ones that are used in the single metric approach. Both start from estimates of the validity, the effect size and the intercorrelation of the predictors. These estimates are usually based on the results of an existing local validation study, but they can also reflect findings reported in the constantly growing number of meta-analytic studies on the validity and effect size of selection predictors (e.g., Bobko, Roth, & Potosky, 1999; Hough, Oswald & Ployhart, 2001; Potosky, Bobko, & Roth, 2005; Salgado, Anderson, Moscoso, Bertua, & De Fruyt, 2003; Schmidt & Hunter, 1998).

Finding or approximating the set of Pareto-optimal trade-offs is typically performed by means of multiobjective optimization methods (De Corte *et al.*, 2007). These methods can be divided in two categories: (a) classical methods, based on mathematical principles and (b) non-classical methods that follow some natural or physical process (Shukla & Deb, 2005). Classical methods are not very appropriate in the present context, however, because their implementation would require repeatedly

solving difficult mixed integer nonlinear programming problems in which the integer problem variables are zero-one variables that represent the nonpresence or presence of a predictor in the subset and the continuous problem variables correspond to the weights with which the predictors are combined into the composite predictor. We therefore prefer using a less computationally expensive nonclassical method. More specifically, we apply the multiobjective evolutionary scheme as described by Deb, Pratap, Agarwal, and Meyarivan (2001) because extensive evaluation studies indicate that this algorithm usually provides a good approximation of the set of Pareto-optimal trade-offs (Shukla & Deb, 2005).

For the present purposes, the implementation of the multiobjective evolutionary algorithm requires the computation of the validity and effect size of candidate Pareto-optimal composites. Except for the provision that the composites are formed from a maximum number, n , of the N available predictors, both quantities are calculated according to the usual formulas for the effect size and the validity of linear combinations of predictors. Thus, assuming standardized predictors for convenience, the validity of a candidate composite c , v_c , is determined as

$$v_c = \frac{\sum_i j_{ic} w_{ic} v_i}{n + \sum_i \sum_{k \neq i} j_{ic} j_{kc} w_{ic} w_{kc} r_{ik}}$$

where w_{ic} is the weight with which predictor i is used in the formation of the composite c ; v_i is the validity of the i th predictor; r_{ik} denotes the correlation between the predictors i and k ; and the zero-one indicator variable j_{ic} indicates whether the i th predictor is included (i.e., $j_{ic} = 1$) or not included ($j_{ic} = 0$) in the composite formation, with $\sum_i j_{ic} \leq n$ to assure that no more than n predictors are used in forming composite c . In turn, the effect size, d_c , of the composite c is evaluated as

$$d_c = \frac{\sum_i j_{ic} w_{ic} d_i}{n + \sum_i \sum_{k \neq i} j_{ic} j_{kc} w_{ic} w_{kc} r_{ik}}$$

where d_i corresponds to the effect size of predictor i .

At the end of the computations, the multiobjective evolutionary algorithm results in a representative sample of the Pareto-optimal validity/effect size trade-offs associated with predictor composites that can be formed using at most n of the N available predictors. Although the obtained sample of Pareto-optimal trade-offs between predictor composite validity and effect size provides only a point-wise approximation of the continuous Pareto-optimal curve, the approximation can be made as dense as required by generating a sufficient number of Pareto-optimal trade-offs and/or by applying curve fitting methods.

3.1. Further comments

Note that the present method can also be used when other than the present validity and effect size metrics are chosen to translate the goals of selection quality and work force diversity, and one may wonder whether these different choices will result in different solutions for the set of Pareto-optimal composites and associated trade-offs. The answer is that the obtained Pareto-optimal solutions will be the same provided that these other metrics induce either the same or the reverse order as the validity and the effect size measures on the total set of candidate composites. For example, the adverse impact ratio and the standardized mean difference (d) are monotonically related, and thus the choice of metric does not affect conclusions.

So, choosing different metrics than validity and effect size for translating the quality and diversity objective will only make a difference when at least one of the new metrics is not monotonically related to the corresponding original metric. This will be the case when using, for example, the expected job performance metric instead of the validity measure for the selection quality because these two metrics induce an identical order only in case that the applicants come from a single homogeneous population, whereas the induced orders differ when the total applicant group is a mixture of different applicant subgroups. However, even in the latter case, the induced orders are usually very similar. As a consequence, the Pareto-optimal predictor composites obtained by choosing the effect size and the validity metrics and the Pareto-optimal composites obtained by choosing the adverse impact ratio (or minority hiring rate) and the expected job performance measures are very similar as well.

As a final remark, it is noted that the above-described method is easily adapted when the selection feasibility requirements also include boundary conditions on, for example, total testing time or cost. In that case, one or more constraints that express the boundary conditions are added to the formulation of the multiobjective programming problem.

3.2. Implementation

To implement the present method for computing Pareto-optimal subset predictor composites a computer program, operating under the Windows operating system, was written. The program returns a summary description of the different Pareto-optimal composites and a tabular display of the corresponding Pareto-optimal trade-off values for the selection goals. The operational details for implementing each Pareto-optimal composite are provided as well. In particular, the output details for each composite the subset of used predictors and the weights with which these predictors are combined to the Pareto-optimal composite. In addition the program gen-

erates the necessary output for drawing the curve of Pareto-optimal goal trade-offs using the freely available graphical software in the R language (cf. <http://lib.stat.cmu.edu/R/CRAN>).

The computer program provides several options for tuning the computation of the Pareto-optimal composites to the particular requirements of the user. One set of control parameters offers the choice between different metrics for translating the selection quality and work force diversity goal. Thus, the user can choose between the metrics of composite validity and expected job performance for the selection quality goal whereas the composite effect size, the expected adverse impact ratio and the hiring rate in the minority applicant group can be chosen for translating the work force diversity objective. Still other parameters permit imposing bounds on the predictor weights and enforcing limits on the subset predictor costs or the subset administration times. Together these extensions insure that the present method can be widely applied. The executable code of the program and a manual that describes its usage are available from the first author. He may be e-mailed at: wilfried.decorte@ugent.be.

4. Illustration

To illustrate the potential of the proposed method, and the usage of the graphical summary results, we applied the procedure to study the Pareto-optimal trade-off curve in a situation where only a subset of the ASVAB subtests can be administered to applicants who come from two different populations. We choose the ASVAB example because this choice permits a direct comparison between the results obtained by the present procedure and those reported by Johnson *et al.* (2004) who implement the earlier discussed single metric approach. We note that the procedure developed here is applicable to any set of predictors. The current illustration focuses on a battery of cognitive tests, but the procedure will be useful in other settings as well, such as the examination of a broad range of predictor measures, including both cognitive and noncognitive predictors.

Table 1. Predictor effect sizes, validities, and intercorrelations

| Predictors | Effect size ^a | Validity | Intercorrelation matrix | | | | | | | | | | | | | | | |
|-----------------------------------|--------------------------|----------|-------------------------|------|------|------|------|------|------|-------|--|--|--|--|--|--|--|--|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | | | | | | | |
| 1. General science (GS) | 1.008 | .522 | | | | | | | | | | | | | | | | |
| 2. Arithmetic reasoning (AR) | 0.725 | .545 | .598 | | | | | | | | | | | | | | | |
| 3. Verbal (VE) | 0.684 | .561 | .780 | .629 | | | | | | | | | | | | | | |
| 4. Mathematics knowledge (MK) | 0.162 | .407 | .467 | .694 | .475 | | | | | | | | | | | | | |
| 5. Mechanical comprehension (MC) | 0.992 | .545 | .596 | .620 | .561 | .413 | | | | | | | | | | | | |
| 6. Auto and shop information (AS) | 1.213 | .529 | .593 | .432 | .506 | .090 | .725 | | | | | | | | | | | |
| 7. Electronics information (EI) | 0.797 | .525 | .649 | .516 | .622 | .335 | .642 | .757 | | | | | | | | | | |
| 8. Assembling objects (AO) | 0.602 | .442 | .430 | .532 | .426 | .456 | .574 | .348 | .398 | | | | | | | | | |
| 9. Coding speed (CS) | 0.178 | .341 | .272 | .373 | .337 | .415 | .192 | .029 | .169 | 0.294 | | | | | | | | |

Note. All Table 1 data are derived from Johnson *et al.* (2004). ^aEffect sizes are relative to the minority applicant population.

4.1. Data for the illustration

The application uses data from Johnson *et al.* (2004). The data include effect size (with respect to several minority populations), validity (as related to school performance within 32 Navy jobs), and intercorrelation of the following nine ASVAB subtests: (1) General Science (GS), (2) Arithmetic Reasoning (AR), (3) Verbal (VE), (4) Mathematics Knowledge (MK), (5) Mechanical Comprehension (MC), (6) Auto and Shop Information (AS), (7) Electronics Information (EI), (8) Assembling Objects (AO), and (9) Coding Speed (CS). For the present study only the intercorrelation matrix of the nine subtests computed from a large applicant sample, the effect size data comparing the Black and White applicant groups and the validities (corrected for range restriction and, therefore reflecting subtest validities in the applicant population) for one of the Navy jobs (i.e., the Aviation Boatswain's Mate-Equipment job; details on the job are available at <http://usmilitary.about.com/od/enlistedjob1/a/abe.htm>) will be used. The data are summarized in Table 1.

4.2. Obtaining maximum size pareto-optimal composites

We first applied our method to approximate the set of Pareto-optimal composites that can be formed using at most three of the nine ASVAB predictors. Only composites in which the selected predictors receive positive weights that vary between 1 and 10 are considered. Negative weights were not allowed because they lead to composites in which valued job related attributes are counted against the applicants; whereas weights smaller than one were excluded to fix the maximum ratio between predictor weights to 10, ensuring that the composites effectively relate to the selected predictors. The results of the analysis are reported in Table 2 and Figure 1. Table 2 details a representative selection of the obtained Pareto-optimal composites, indicating for each composite (a) the associated Pareto-optimal trade-off in terms of the composite validity, the composite effect size and the minority hiring rate, (b) the set of included

Table 2. Tabular overview selected Pareto-optimal trade-offs from a maximum size composite of at most three predictors

| Trade-off | Trade-off | | Minority hiring rate ^a | Number minority hired ^b | Predictor weight | | | Predictor identity | | |
|-----------|-----------|-------------|-----------------------------------|------------------------------------|------------------|------|-----|--------------------|---|---|
| | Validity | Effect size | | | | | | | | |
| 1 | .41 | 0.16 | .256 | 10 | / | / | 1.0 | / | / | 4 |
| 2 | .42 | 0.17 | .254 | 10 | / | 10.0 | 1.0 | / | 4 | 9 |
| 3 | .43 | 0.18 | .252 | 10 | / | 7.9 | 1.7 | / | 4 | 9 |
| 4 | .45 | 0.20 | .247 | 10 | / | 6.0 | 3.9 | / | 4 | 9 |
| 5 | .47 | 0.24 | .236 | 9 | 1.0 | 10.0 | 5.0 | 3 | 4 | 9 |
| 6 | .48 | 0.26 | .231 | 9 | 1.5 | 9.7 | 5.1 | 3 | 4 | 9 |
| 7 | .49 | 0.28 | .226 | 9 | 2.0 | 9.6 | 5.4 | 3 | 4 | 9 |
| 8 | .50 | 0.30 | .221 | 9 | 2.6 | 9.7 | 5.3 | 3 | 4 | 9 |
| 9 | .52 | 0.32 | .216 | 9 | 3.3 | 9.7 | 5.1 | 3 | 4 | 9 |
| 10 | .53 | 0.35 | .209 | 8 | 3.0 | 7.0 | 4.0 | 3 | 4 | 9 |
| 11 | .54 | 0.37 | .204 | 8 | 3.7 | 6.9 | 4.3 | 3 | 4 | 9 |
| 12 | .55 | 0.40 | .197 | 8 | 6.2 | 9.5 | 5.9 | 3 | 4 | 9 |
| 13 | .56 | 0.42 | .192 | 8 | 5.8 | 7.6 | 4.5 | 3 | 4 | 9 |
| 14 | .57 | 0.45 | .186 | 7 | 7.9 | 8.4 | 5.8 | 3 | 4 | 9 |
| 15 | .58 | 0.48 | .179 | 7 | 9.7 | 8.5 | 5.7 | 3 | 4 | 9 |
| 16 | .58 | 0.51 | .172 | 7 | 7.8 | 5.5 | 3.9 | 3 | 4 | 9 |
| 17 | .59 | 0.55 | .164 | 7 | 9.7 | 5.1 | 3.8 | 3 | 4 | 9 |
| 18 | .60 | 0.58 | .158 | 6 | 8.1 | 9.7 | 4.3 | 3 | 4 | 7 |
| 19 | .60 | 0.62 | .149 | 6 | 8.7 | 9.9 | 5.8 | 3 | 4 | 7 |
| 20 | .61 | 0.65 | .144 | 6 | 7.9 | 7.3 | 5.3 | 3 | 4 | 7 |
| 21 | .62 | 0.68 | .138 | 6 | 8.8 | 10.0 | 3.7 | 3 | 4 | 6 |
| 22 | .63 | 0.72 | .130 | 5 | 8.0 | 9.1 | 4.0 | 3 | 4 | 6 |
| 23 | .63 | 0.75 | .125 | 5 | 7.9 | 9.1 | 4.8 | 3 | 4 | 6 |
| 24 | .64 | 0.79 | .118 | 5 | 8.1 | 9.2 | 5.5 | 3 | 4 | 6 |
| 25 | .65 | 0.83 | .111 | 4 | 7.9 | 8.6 | 6.1 | 3 | 4 | 6 |
| 26 | .65 | 0.86 | .106 | 4 | 8.0 | 8.7 | 7.2 | 3 | 4 | 6 |
| 27 | .66 | 0.91 | .098 | 4 | 6.8 | 8.2 | 4.4 | 4 | 6 | 9 |
| 28 | .67 | 0.95 | .092 | 4 | 5.2 | 7.3 | 3.7 | 4 | 6 | 9 |
| 29 | .67 | 1.01 | .083 | 3 | 4.1 | 6.8 | 3.1 | 4 | 6 | 9 |
| 30 | .67 | 1.03 | .080 | 3 | 5.2 | 9.4 | 4.0 | 4 | 6 | 9 |

Note. ^aMinority Hiring Rate for a .30 selection rate from a .80 majority and a .20 minority group. ^bBased on a total applicant sample of 200 candidates. The solidus (/) means not applicable.

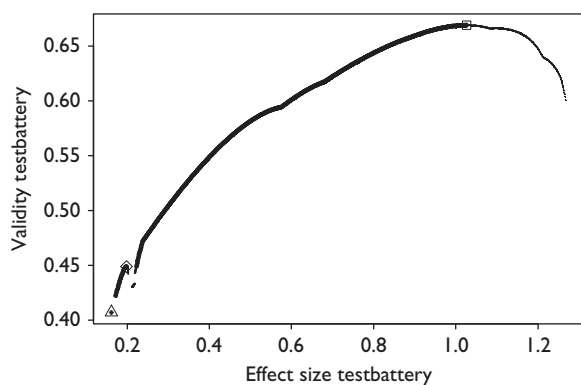


Figure 1. Pareto-optimal validity/effect size trade-off curve for maximum size composites (test batteries) from at most three ASVAB predictors. Continuous predictor weight range between 1 and 10.

predictors, and (c) the weights used to combine the predictors to the composite. Figure 1 displays the entire Pareto-optimal validity/effect size trade-off curve for the example application. The curve consists of a single point (cf. the lower left triangle point) and two line segments (both drawn as thick solid lines). The lower left point

corresponds to the only Pareto-optimal validity/effect size trade-off that can be achieved when using a single predictor in the composite. The short line segment represents the Pareto-optimal trade-offs obtainable when using two predictors in the composite formation and the upper right line segment summarizes the Pareto-optimal trade-offs for three predictor composites.

The gaps in the Pareto-optimal curve are instructive for detailing the difference between Pareto-optimal and other trade-offs. The gaps indicate that no Pareto-optimal trade-off is possible for certain effect size (or validity) values, although other (non-Pareto-optimal) trade-offs may be possible for these effect size (validity) values. Thus, the gaps show that there is no Pareto-optimal trade-off for an effect size value of, for example, .21 in the present application. However, this does not mean that it is not possible to achieve a (nonoptimal) trade-off with a value of .21 for the effect size. In fact, many such trade-offs, all showing an effect size value of .21, are possible in the present example, but the highest possible validity among all these trade-offs is only .44, which is smaller than the .45 validity value of the Pareto-optimal trade-off obtained for an effect size value of .20

(cf. the diamond shaped point on the Pareto curve in Figure 1).

In other instances, the gaps may indicate that no trade-offs (neither Pareto-optimal nor other) are possible for certain effect size or validity values. Thus, in the present example, no trade-off is possible for an effect size value of, for example, .165. To clarify the instances where portions of the gaps indicate nonexistence of a Pareto-optimal trade-off versus nonexistence of any kind of trade-off, we added thin line segments to Figure 1. The trade-offs on these thin line segments are not Pareto-optimal but show the maximum achievable validity for the corresponding effect size value. Because Pareto-optimal trade-offs also show maximum validity at their corresponding effect size value, the concatenation of the thin and the thick line segments represents the set of maximum achievable validity/effect size trade-offs. The figure shows that many of these maximum achievable trade-offs are Pareto-optimal (cf. the trade-offs on the thick line segments), but others (i.e., the trade-offs on the thin line segments) are not. Thus, the thin line segments show that the trade-off with values of, for example, .21 and .44 for effect size and validity is a maximum achievable, but not a Pareto-optimal trade-off because it is dominated by the .20, .45 Pareto-optimal trade-off. The upper right thin line segment further indicates that none of the maximum validity trade-offs associated with the highest effect size values are Pareto-optimal. All maximum validity trade-offs corresponding to effect size values in the range between 1.04 and 1.27 are dominated by the Pareto-optimal trade-off with values of .67 and 1.03 for the validity and effect size, respectively (cf. the square point in Figure 1). Finally, note that the gaps that remain after considering both the thick and the thin line segments refer to effect size (validity) values for which no trade-off is feasible.

In general, the gaps in the maximum achievable or the Pareto-optimal curve are caused by the restrictions imposed on the weighting of the predictors when forming the composite predictor. In the present example only predictor weights that vary between 1 and 10 are permissible. If the predictor weights would be allowed to vary between 0 and 10, the gaps would disappear but the Pareto-optimal trade-off curve and the maximum possible trade-off curve would not necessarily coincide.

Leaving behind the distinction between maximum possible and Pareto-optimal trade-off curves, the key message of Table 2 and Figure 1 is that using a maximum size composite of at most three predictors can lead to substantially different Pareto-optimal validity/effect size trade-off values. Consider, for example, using only predictor 4. The resulting, single predictor composite shows a Pareto-optimal validity/effect size trade-off value of 0.41 and 0.16 (cf. trade-off number one in Table 1 and the lower left triangle point in Figure 1). At the other extreme the battery, including predictors 4, 6, and 9

with weights 5.2, 9.4, and 4.0, offers a Pareto-optimal validity/effect size trade-off value of 0.67 and 1.03 (cf. trade-off number 30 in Table 1 and the trade-off point marked with a square on Figure 1). All other Pareto-optimal composites show a more balanced trade-off between composite validity and effect size.

As witnessed by Table 2 and Figure 1, the obvious strength of the present method is that it enables the uncovering of all Pareto-optimal predictor composites that can be formed using a maximum number of predictors. When addressing the issue of composite formation only these composites should be considered because all other composites are dominated by at least one member of the Pareto-optimal set. To fully appreciate this unique contribution, the set of Pareto-optimal trade-offs can be compared with the entire set of all achievable trade-offs as is done in Figure 2. In this figure, the upper thick line segments represent the Pareto-optimal trade-off curve, the concatenation of the upper thin and thick line segments corresponds to the maximum possible trade-off curve, and the lower thin line segments represent the minimum achievable trade-off curve (i.e., the curve that indicates the minimum possible validity achievable for a given effect size value). So, the area enclosed between the upper and lower line segments represents all validity/effect size trade-offs that can be obtained using any weighted combination (with weights between 1 and 10) of at most three predictors from the 9 ASVAB test collection. We also added two dashed lines on the figure to indicate four trade-offs; two of these show the same effect size (end points of the vertical dashed line); whereas the other two have the same validity (end points of the horizontal dashed line). From the location of the first pair of trade-offs (both with effect size equal to 0.70) with respect to the validity axis, it can be seen that the



Figure 2. Validity/effect size trade-offs achievable for maximum size composites of three ASVAB predictors. The upper, thick part of the curve summarizes the set of Pareto-optimal validity/adverse impact ratio trade-offs. Continuous predictor weight range between 1 and 10.

Pareto-optimal composite has a validity of .62, whereas the corresponding dominated composite has a validity of only .48. Similarly, looking at the second pair (both with validity equal to .57) it is shown that the Pareto-optimal composite has an effect size of only .47 as compared with an effect size value of 1.26 for the corresponding dominated composite.

The important difference in validity and effect size, observed within the two selected pairs of trade-offs convincingly demonstrates the importance of choosing (a) the appropriate subset of predictors and (b) the optimal weighting of the selected predictors when forming predictor composites. Although the importance of these choices is well-known (cf. Ployhart & Holtz, 2008), we emphasize that only the present method offers the possibility to fully appreciate the consequences in terms of validity and effect size of using other than Pareto-optimal predictor composites.

4.3. Choosing different metrics for the selection goals

In the subsection 'Further comments' we noted that different, but monotonically related, metrics for the same selection goal will lead to the same set of Pareto-optimal subset predictor composites. Figure 2 illustrates this feature by also considering the minority hiring rate (cf. the upper horizontal axis in the figure) as another measure for the work force diversity selection goal besides the thus far used effect size metric. In the application, the minority hiring rate reflects the proportion of selected minority applicants for a top-down selection with a 30% selection rate from an applicant group that comprises 80% majority, and 20% minority applicants. Because the minority hiring rate is monotonically related to the composite effect size, using either metric for the work force diversity goal results in the same solution for the set of Pareto-optimal subset predictor composites. However, when the set is computed using the validity and the minority hiring rate metrics, the Pareto-optimal selection goal trade-offs can now be expressed in terms of these measures instead of in terms of validity and composite effect size.

To illustrate this feature, consider the diamond shaped trade-off on the Pareto-optimal trade-off curve in Figure 2. This is trade-off number 13 in Table 2, obtained by using a Pareto-optimal composite from predictors 3, 4, and 9, with weights equal to 5.8, 7.6, and 4.5, respectively. In both the table and Figure 2 it is shown that the composite shows trade-off values of .56 for validity, .42 for composite effect size, and .192 for minority hiring rate. Referring to the upper horizontal axis of Figure 2 (the axis labelled minority hiring rate) it is further noticed that the Pareto-optimal validity-minority hiring rate trade-offs vary between .41 versus .256 and .67 versus .080, respectively. So, if one were willing to accept a 39%

loss in maximum possible validity (i.e., $.67 - .41 = .26$ and $.26 / .67 \times 100 = 39$), a gain of 220% in minority hiring rate (i.e., $(.256 - .080) / .080 \times 100$) can be obtained and the resulting validity-minority hiring rate trade-off of .41 and .256 is still Pareto-optimal. The present application therefore illustrates that it is possible to have Pareto-optimal composites that, compared with other Pareto-optimal composites, show a substantial increase in minority representation by accepting a much smaller decrease in validity. Other Pareto-optimal composites, that show less extreme but nevertheless still impressive increases in minority representation at a considerably lower validity decrease, are possible as well. Thus, one of these intermediate composites (i.e., composite number 18 in Table 2) indicates that, compared with the maximum validity Pareto-optimal trade-off, the minority hiring rate can almost be doubled (from .080 to .158) by accepting a 10% drop in the maximum possible validity (i.e., from .67 to .60).

Because the above reported relative increases in minority hiring rate could lead to misleading claims about the true impact of choosing different Pareto-optimal predictor composites, especially when the base rate minority hiring proportion (i.e., the minority hiring rate for the maximum validity composite) is low, column five of Table 2 further details the different Pareto-optimal composites in terms of the selected number of minority applicants in case that the total applicant group comprises a fairly typical number of 200 candidates. The additional information shows that only three minority applicants will be hired when using the maximum validity Pareto-optimal composite (cf. trade-off number 30 in Table 2), whereas this number is more than tripled to a total of 10 minority hires (out of 40 minority applicants) in case of the minimum effect size composite (cf. trade-off number 1 in Table 2). Also, more balanced Pareto-optimal composites (e.g., composite and trade-off number 18 in Table 2) may still double the number of minority hires from three to six. So, even using absolute indices instead of relative indices and focusing on typical total applicant group sizes, the expected variation in minority hiring over the different Pareto-optimal composites remains considerable. Choosing a more balanced composite such as, for example, composite number 18 in Table 2 instead of the maximum validity composite will have a real effect on minority representation and this effect will accrue with each repetition of a similar selection.

4.4. Relation with the single metric approach

The results of our application also clarify the relationship between the present procedure and the approach in which the composite formation is decided using a single metric. The study of Johnson *et al.* (2004) offers a typical example of the latter strategy in the context of fixed size composite formation. In this study, the single metric

approach, detailed in the section 'Earlier Approaches,' is applied to all possible composites obtained by equal weighting of three of the nine ASVAB predictors and the composite with the highest combined validity/transformed effect size sum is chosen. The resulting, preferred composite is the equally weighted combination of the VE, MK, and CS subtests (cf. the predictors 3, 4, and 9 in Table 1). The combined sum value of the composite equals 1.45 and the composite is characterized by a .560, .438 trade-off value for the validity and the effect size measure, respectively.

To assess the merits of the composite proposed by Johnson *et al.* (2004), we plotted the validity/effect size trade-off value of the composite (cf. the triangle point in Figure 2) on the graph of the Pareto-optimal trade-off curve. The plot indicates that the proposed composite, although performing very well, is not Pareto-optimal. As an example, it is bettered by a composite in which the VE, MK, and CS predictors receive weights of 6.2, 7.7, and 4.8, respectively, as this composite shows the same .560 validity level but a slightly smaller .430 effect size value.

It is no coincidence that our method shows trade-offs that dominate the trade-off proposed by Johnson *et al.* (2004) because these authors consider only three predictor composites in which the predictors receive equal weights. However, even in case that the predictors weights would be allowed to vary between 1 and 10, their single metric would still result in only one element of the Pareto-optimal curve derived by our method. Using a single metric to express the combined validity/effect size merit of a predictor composite will always result in such a single point, provided that the metric is a weighted sum of appropriate order preserving transforms of the initial validity and effect size measures. Before deciding in favor of this particular element it seems good practice to first compare the merits of the single metric proposal to the other achievable trade-offs. We therefore recommend using the present method as a routine step in the development of a single metric proposal.

5. Additional analyses

We conducted three ancillary studies (detailed results are available from the first author). The first additional study focused on comparing the potential of differently sized predictor subsets (i.e., maximum size composites from 2, 3, 4, and all 9 ASVAB predictors). Obviously, higher maximum size composites had a better validity/effect size potential than lower maximum size composites. However, from a cost perspective, it was interesting that the difference in validity/effect size potential between consecutive maximum size composites diminished for higher maximum size composites: Pareto-optimal composites from more than four predictors resulted in

validity/effect size trade-offs that were only marginally better than the trade-offs associated with composites from at most four predictors.

Second, we analyzed the impact of different predictor weight conditions (equal weights, integer weights, continuously varying weights) on the validity/effect size potential of predictor composite formation. When the ratio between the highest admissible weight and the lowest admissible weight equalled five or more, restricting the weights to integer values hardly affected the validity/effect size potential of predictor composite formation. However, Pareto-optimal trade-offs with low associated composite effect size and validity values were increasingly excluded when the predictor weight range was more restricted.

In a third set of analyses, we examined the sensitivity of the Pareto-optimal predictor subsets for variability in predictor parameter values (i.e., validities, effect sizes, and intercorrelations) via Monte Carlo simulation methods. There was substantial support that predictor combinations resulting in Pareto-optimal trade-offs between validity and effect size were robust for reasonably different predictor data. The simulation results also revealed that the difference in sensitivity to parameter variability between Pareto-optimal trade-offs (which correspond to differentially weighted composites) and unit weighted trade-offs (corresponding to unit weighted predictor composites) was small, with the Pareto-optimal trade-offs being somewhat less sensitive than the unit weighted trade-offs.

6. General discussion

6.1. Methodical contribution

The most effective strategies for addressing the validity-adverse impact dilemma, such as measuring the full range of relevant cognitive and noncognitive KSAOs or using alternative predictor measurement methods, all result in considering a larger number of predictors for inclusion when designing planned selections. However, the implementation of a substantial number of predictors often conflicts with feasibility concerns about total testing time and costs as well as with other logistical concerns that favor reducing the number of administered predictors. As a consequence decisions about (a) which predictors to include in the test battery and (b) the weighting of the chosen predictors to the predictor composite are becoming increasingly important. To help deciding these issues, we presented a method to uncover the set of Pareto-optimal trade-offs between validity and effect size that can be achieved when forming composites from maximum size subsets of a larger number of available predictors. The method thereby provides a summary of best possible practices for choosing and combining available predictors to composites that optimally balance the

validity and the diversity concerns. By including limits on the number of chosen predictors and allowing eventual additional constraints on total test cost and/or time, the method also explicitly aims for selection systems that are feasible to implement.

The method offers a significant methodical contribution to the selection literature because no other systematic procedure for achieving the same purposes is presently available. Although the presentation focused on the validity and the effect size as measures for the selection quality and the selection diversity objectives, we showed that other metrics are equally possible. The method and its implementation by means of a multiobjective evolutionary algorithm can also be adapted in case that only fixed size (instead of maximum size) composites are acceptable and/or that the applicant group contains candidates from more than one minority group.

In contrast to the ad hoc, single metric approach, the procedure does not lead to a single preferred predictor subset and a corresponding single predictor composite but to a summary of Pareto-optimal predictor composites associated with possibly different predictor subsets. Although this may be perceived as a drawback, the following observations suggest a more qualified position. First, the procedure eliminates all predictor composites of a given maximum size that are not Pareto-optimal, thereby substantially reducing the set of acceptable solutions. Second, the results of the procedure can be used to assess the merits of any particular single metric proposal. Without our procedure there is no inventory of the optimal trade-offs that can be achieved and, hence, no yardstick to evaluate an ad hoc based proposal. Finally, the fact that the method results in an overview of all Pareto-optimal solutions instead of a single best solution emphasizes the important role left for the decision maker when addressing the issues of predictor subset selection and predictor weighing. Rather than eliminating the contribution of the decision maker, the method provides through the enumeration of the Pareto-optimal trade-offs between quality and diversity detailed information so that a truly informed decision becomes possible. The method is therefore best characterized as a decision aid and not as a procedure that takes the decision making away from the selection practitioner.

By embedding our method within a Monte Carlo simulation procedure, it could be verified that Pareto-optimal trade-offs and the corresponding differentially weighted predictor composites show acceptable transportability to predictor systems with fairly different predictor parameter values. Such transportability is essential because the determination of Pareto-optimal trade-offs necessarily starts from predictor parameter values that will typically deviate from the unknown values of the future, intended application. In fact, without this transportability, the computation of Pareto-optimal trade-offs would not have much real practical value.

6.2. Substantive contribution

The example application and the results depicted in Figure 2 emphasize the substantial gains in terms of both validity and minority hiring rate that can be achieved when using Pareto-optimal instead of other less well chosen predictor composites. Although unit weighted composites are often quite useful (cf., Bobko, Roth, & Buster, 2007), we found that these composites are no match for Pareto-optimal composites when predictor information is available and the number of predictors in the composite is small to modest. Even in terms of transportability (i.e., sensitivity to variability/uncertainty in the predictor parameter values), Pareto-optimal composites and the associated trade-offs perform at least as well as unit-weighted or regression based composites.

So, the key message of the paper is that selection practitioners should consider using our method when addressing predictor subset and predictor weighting decisions instead of making trial based choices. To facilitate this usage, we make available a computer implementation of the method. Using heuristic rules of thumb for addressing the same decisions offers no real alternative to the method. In fact, the decision aid provided by our method renders the pursuit of such heuristic rules much less pressing. Compared with our method these heuristics can offer only vague guidance as to the choice and the weighting of the predictors. Also, using these heuristics in any particular setting requires the same predictor information as our method because otherwise it is not possible to decide which heuristics are applicable and which are not. So, given the present method we see little reason for the pursuit of rules of thumb about predictor weighting and predictor subset selection.

The example application showed a substantial range in validity-minority hiring rate trade-offs for the different Pareto-optimal predictor composites. In general, the quality and the range of these trade-offs will depend on the validity, intercorrelation and effect size characteristics of the predictors in the initial total predictor set and one could consider searching for patterns in these characteristics that affect the quality of the achievable Pareto-optimal trade-offs. However, we believe that this effort is of less practical value because it will typically not be possible to implement the eventual findings from such a search. Most selection applications start from a given total set of available predictors with a particular validity, intercorrelation, and effect size pattern. Even where constructing new predictors is possible, there is no guaranty that these new predictors will result in the preferred pattern because it is very difficult to construct predictors with preconceived validity, effect size, and intercorrelation values. As a consequence, we believe it to be more fruitful to explore the possibilities of given initial test batteries than to search for principles about

optimal validity, effect size, and intercorrelation patterns in hypothetical test batteries. It seems better to focus on what can be achieved in practice with an available test battery than on what could be realized with more ideal but probably unattainable batteries, especially when the nature of these ideal batteries is already known when both selection quality and work force diversity are of importance. For in that case, the ideal test battery consists of a single test with zero effect size and maximum validity.

6.3. Limitations and future research suggestions

As mentioned in the section 'Procedure for Deriving Pareto-optimal Predictor Subset Composites,' our method requires certain data on the effect sizes, intercorrelations, and validities of the available predictors. However, these data requirements are not specific to our method but are shared by all previous efforts for gauging the consequences of selection design decisions on valued selection goals (e.g., De Corte *et al.*, 2007; Doverspike *et al.*, 1996; Finch *et al.*, 2009). Even the unsystematic, trial and error comparison of alternative predictor subsets depends on this type of predictor information. Although thus far published meta-analytic studies provide useful results, more detailed information on the validities, effect sizes and correlations of selection predictors, especially related to unscreened applicant populations, is still very much in need. We therefore repeat the plea for continuing meta-analytic research on predictor and criterion characteristics and for supplementing these efforts with local validity studies.

The present method provides systematic guidance to the selection practitioner in choosing and weighting selection predictors when both the goals of selection quality and work force diversity are valued. Although both decisions are important when designing a future planned selection, other decisions are often required as well. These other decisions about, for example, the nature of the selection rule, the sequencing of the predictors in case of a multistage, noncompensatory selection rule and the retention rates preferably used in the intermediate selection stages remain out of scope, however. So, whereas the present method offers an integration of the decision aid proposed by De Corte *et al.* (2007) for obtaining optimally weighted predictor composites with a systematic procedure for choosing optimal predictor subsets within given minimum feasibility requirements, still further work is needed to arrive at a decision aid that covers the entire selection design process.

Other future research avenues include comparing the Pareto-optimal selection outcomes achieved by our method with the corresponding outcomes obtained under, for example, different targeted recruiting strategies (cf. Newman & Lyon, 2009). Integrating recruitment

and job refusal information within the present decision aid offers a further challenge.

6.4. General conclusion

To address the validity-adverse impact quandary the selection research literature advises using an increasing number of predictors, assessing the broad spectrum of relevant KSAOs. At the same time, the organizational context often imposes feasibility constraints related to cost and time that favor using only a limited number of predictors. Because of this conflict predictor subset selection and predictor weighting decisions are becoming increasingly important when designing planned selections. Yet, no systematic procedure is presently available to guide these decisions. The paper therefore proposes a method that provides a direct and effective answer to the predictor subset and the predictor weighting problems. The availability of a computer implementation makes the method generally and easily applicable as well. We urge selection practitioners and researchers to use the method when studying selections where both selection quality and work force diversity are valued; either alone, or in combination with other approaches to address the validity-adverse impact dilemma.

References

- Aguinis, H. and Harden, E. (2004). Will banding benefit my organization? An application of multi-attribute utility theory. In Aguinis, H. (Ed.), *Test score banding in human resource selection* (pp. 193–216). Westport, CT: Praeger.
- Bobko, P., Roth, P.L. and Buster, M.A. (2007). The usefulness of unit weights in creating composite scores – A literature review, application to content validity, and meta-analysis. *Organizational Research Methods*, 11, 689–709.
- Bobko, P., Roth, P.L. and Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, 52, 561–589.
- Campion, M.A., Outtz, J.L., Zedeck, S., Schmidt, F.L., Kehoe, J.F. and Murphy, K.R., et al. (2001). The controversy over score banding in personnel selection: Answers to 10 key questions. *Personnel Psychology*, 54, 149–185.
- Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. (2001). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6, 182–197.
- De Corte, W., Lievens, F. and Sackett, P.R. (2006). Predicting adverse impact and multistage mean criterion performance in selection. *Journal of Applied Psychology*, 91, 523–537.
- De Corte, W., Lievens, F. and Sackett, P.R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, 92, 1380–1393.
- Doverspike, D., Winter, J.L., Healy, M.C. and Barrett, G.V. (1996). Simulations as a method of illustrating the impact of

- differential weights on personnel selection outcomes. *Human Performance*, 9, 259–273.
- Finch, D.M., Edwards, B.D. and Wallace, J.C. (2009). Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology*, 94, 318–340.
- Hough, L.M., Oswald, F.L. and Ployhart, R.E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, 9, 152–194.
- Johnson, J.W., Abrahams, N. and Held, J.D. (2004). A procedure for selecting predictors considering validity and adverse impact. Poster presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago.
- Kehoe, J.F. (2008). Commentary on pareto-optimality as a rationale for adverse impact reduction: What would organizations do? *International Journal of Assessment and Selection*, 16, 195–200.
- Newman, D.A. and Lyon, J.S. (2009). Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability and diversity. *Journal of Applied Psychology*, 94, 298–317.
- Ployhart, R.E. and Holtz, B.C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, 61, 153–172.
- Potosky, D., Bobko, P. and Roth, P.L. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of Assessment and Selection*, 13, 304–315.
- Roth, P.L. and Bobko, P. (1997). A research agenda for multi-attribute utility analysis in human resource management. *Human Resource Management Review*, 7, 341–368.
- Salgado, J.F., Anderson, N., Moscoso, S., Bertua, C. and De Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities as predictors of work behaviors: A European meta-analysis. *Personnel Psychology*, 56, 573–605.
- Schmidt, F.L. and Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.
- Shukla, P.K. and Deb, K. (2005). Comparing classical generating methods with an evolutionary multi-objective optimization method. *Lecture Notes in Computer Science*, 3410, 311–325.