5-2006

# Predicting adverse impact and mean criterion performance in multistage selection

Wifried DE CORTE
*Ghent University*

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

Paul R. SACKETT
*University Minnesota - Twin Cities*

**DOI:** https://doi.org/10.1037/0021-9010.91.3.523

Published in Journal of Applied Psychology, Volume 91, Issue 3, May 2006, Pages 523-537
https://doi.org/10.1037/0021-9010.91.3.523

# Predicting Adverse Impact and Mean Criterion Performance in Multistage Selection

Wilfried De Corte and Filip Lievens
Ghent University

Paul R. Sackett
University of Minnesota, Twin Cities Campus

The authors present an analytical method to assess the average criterion performance of the selected candidates as well as the adverse impact and the cost of general multistage selection decisions. The method extends previous work on the analytical estimation of multistage selection outcomes to the case in which the applicant pool is a mixture of applicant populations that differ in their average performance on the selection predictors. Next, the method was used to conduct 3 studies of important issues practitioners and researchers have with multistage selection processes. Finally, the authors indicate how the method can be integrated into a broader analytical framework to design multistage selection decisions that achieve intended levels of selection cost, workforce quality, and workforce diversity.

*Keywords:* adverse impact, personnel selection, estimation procedure, quality selected workforce

The use of multiple predictors (tests, interviews, simulation exercises) in selection decisions is common. Whenever multiple predictors are used, decisions need to be made about how the predictor information is to be combined. Regression weights or job analytically determined importance weights are commonly used in settings where all predictors are administered to all applicants. Issues of adverse impact against members of protected groups may also come into play. For example, although criterion-related validity is maximized by using regression weights, it may be that alternative combinations of the same predictors can produce validity levels quite close to those produced by the optimal regression weights but with less adverse impact. Decision makers may, of course, vary in their willingness to accept a given reduction from optimal validity in return for a given gain in minority group selection.

In many cases, however, multiple predictors are used in a sequential, multiple-hurdle fashion. Although it is clearly optimal for purposes of predictive accuracy to administer all predictors to all candidates, cost and logistic concerns frequently make this infeasible. Thus, it is common to use multistage selection, in which a subset of predictors is used for initial screening, and additional predictors are used for subsequent selection decisions. Inexpensive predictors, predictors lending themselves to unproctored administration (e.g., application blanks), and predictors amenable to group administration are common candidates for use in initial screening decisions, whereas more expensive predictors requiring proctored and/or individualized administration are subsequently administered to a smaller, more manageable number of candidates.

Wilfried De Corte, Department of Data Analysis, Ghent University, Ghent, Belgium; Filip Lievens, Department of Personnel Management and Work and Organizational Psychology, Ghent University; Paul R. Sackett, Department of Psychology, University of Minnesota, Twin Cities Campus.

Correspondence concerning this article should be addressed to Wilfried De Corte, Department of Data Analysis, Ghent University, H. Dunantlaan 1, 9000, Ghent, Belgium. E-mail: wilfried.decorte@rug.ac.be

Given that the choice to use multistage selection is a pragmatic one, it is important to anticipate the effects on key selection outcomes of various choices made in multistage selection. Three outcomes of interest to many firms are the selection cost, the criterion performance of those selected (i.e., validity), and the minority hiring rates among those selected (i.e., adverse impact; Sackett, Schmitt, Ellingson, & Kabin, 2001). Given an applicant pool of a given size and composition, a fixed number of openings, and a decision to use a set of predictors, there are a variety of decisions to be made, each of which affects the selection cost, the mean performance of those selected, and the minority hiring rate. The first is determining which predictors to administer at an initial stage and which to administer at subsequent stages. The second is the cutoff score or selection ratio to apply to predictors used at an initial stage. At times, there may be a clear criterion-referenced basis for such a decision (as in the case of a job requiring a person to lift 35 lb). In many cases, though, validity evidence may indicate a linear test–criterion relationship, and there may be no single criterion level seen as a uniquely meaningful threshold. Thus, a decision is required about the proportion of the pool that will advance to subsequent stages in the selection procedure. The third is determining how final selection decisions should be made. Here the key decision is whether the predictors used in initial screening also play a part in the final-selection decision (i.e., if $A$ is administered at Stage 1 and $B$ at Stage 2, is the final selection done on the basis of $B$ only or on the basis of $A + B$)?

To resolve these decision problems, before the selection is actually performed, the practitioner must at least be able to determine the expected criterion performance and adverse impact that correspond to a particular set of choices with respect to the stage-specific selection rates and predictor weights. An analytical method to carry out this estimation for the general multistage selection scenario is not available, however (cf. Sackett & Roth, 1996). At present, only single-stage selections can be addressed analytically (cf. De Corte & Lievens, 2003; Schmitt, Rogers, Chan, Sheppard & Jennings, 1997), whereas Monte Carlo simulation

methods have thus far been used for selection scenarios that are limited to two stages (Sackett & Roth, 1996).

To overcome this limitation, as well as to avoid inherently variable estimates, which are typically the result of using simulation methods, the first aim of this article is to present a widely applicable, analytical method to assess the cost, the standardized average criterion performance, and the group-specific adverse impact ratios of intended multistage selections. We also make available a computer program that implements the method. The present method extends previous related work on the analytical computation of multistage selection outcomes (Cronbach & Gleser, 1965; De Corte, 1998) to the case in which the applicant pool is not homogeneous but rather is a mixture of several applicant groups (both majority and minority groups) that differ in terms of their average performance on the predictors. The method is also related to the commonly used simulation approach (e.g., Doverspike, Winter, Healy, & Barrett, 1996; Hattrup & Rock, 2002; Sackett & Roth, 1996) in that it is based on the same assumptions and that its application is contingent on identical information with respect to the predictors and the criterion dimensions. Yet, as compared with the simulation-based approach, the present analytical method has three distinct advantages. First, the results of the simulation-based approach vary over repeated applications on the same input data, whereas the analytical method always results in the same point estimate. Basically, the value obtained by the present method equals the average result that would be obtained over (infinitely) many repeated implementations of the simulation-based approach. Second, the computation of the analytical result is dramatically faster. As we explain later, this enables the integration of the method within a Monte Carlo procedure to handle situations in which the selection practitioner is uncertain about (some of) the values of the predictor and criterion characteristics of the intended selection decision. Finally, we argue in the final section of the article that the present method is required for the systematic design of multistage selection decisions that aim to achieve a given set of goals in terms of cost, workforce quality, and desired levels of workforce diversity.

The second aim of the article is to use the method to conduct three studies of important issues practitioners and researchers have with multistage selection processes. In the first study, we investigated the changes in group differences at each stage in the process. We estimated at each stage the progressive levels of predictor and predicted criterion mean group differences and adverse impact. These results extend results obtained by Roth, Bobko, Switzer, and Dean (2001) on two-stage selection. In the second study, we investigated the consequences of various decision options in multistage selection (cf. Sackett & Roth, 1996). We specifically investigated the trade-offs among cost, validity, and adverse impact as a function of the order of the predictors. In the third study, we applied the method to a Monte Carlo simulation to investigate the merits of alternative selection scenarios where there was uncertainty about specific features of the selection process.

The article is structured as follows. We start by presenting the basic features of the selection decision scenarios that the method intends to address and detail the objectives of the analytical method. Next, we provide a short, nontechnical description of the method and summarize the boundary conditions of its usage. We also briefly describe a computer program to implement the method. Then we discuss the applications introduced above. Finally, we reconsider the relevance of the method for the more optimal design of multistage selection decisions in which goals of selection cost, quality, and adverse impact are of importance.

## Method

### *Studied Selection Scenarios and Purpose of the Method*

The present method focuses primarily on the prototypical scenario that the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) labeled as the *fixed applicant pool.* This scenario occurs when a position is announced with an application deadline. Once that deadline is reached, the organization knows the size and makeup of the applicant pool, and we henceforth consider the case in which the candidates come from several applicant populations that differ in terms of their average scores on one or more available selection predictors and in terms of one or more relevant criterion dimensions. The standardized mean difference (i.e., the mean group difference) is used to express the extent to which the average predictor or criterion scores of the minority groups deviate from the corresponding averages of the majority group. The method focuses on multistage selection, in that at each selection stage, the remaining candidates are screened on a weighted composite of the predictors administered thus far and in that only the candidates who score sufficiently high on the composite predictor are retained for further scrutiny (or receive a job offer at the last stage). In what follows, a distinction is made between stage-specific retention rates, which indicate the proportion of the total number of applicants that are retained at the end of a given selection stage, and stage-specific selection rates, which refer to the proportion of remaining candidates selected at a given stage. Although retention and selection rates convey the same information in single-stage selections, it is important to distinguish between both rates in the context of multistage selection. Thus, in a three-stage selection in which 60% of the applicants are retained after the first stage, half of the remaining candidates are dropped at the second stage, and only one third of the remaining candidates receive a job offer, the stage-specific selection rates are .60, .50, and .33, whereas the corresponding stage-specific retention rates are .60, .30, and .10.

This characterization of the multistage selection scenario also includes situations in which the screening decisions are based on only the stage-specific predictors (in which case the earlier administered predictors receive a zero weight in the composite) as well as situations in which the retention decision in each stage is based on single-predictor information. In the latter case, the number of stages equals the number of predictors, and the predictor composites all reduce to the corresponding single predictor.

Given any particular set of choices of the selection practitioner with respect to (a) the total number of stages, (b) the stage-specific selection rates, and (c) the weights with which the predictors administered thus far will be used to determine the stage-specific composite predictor scores, the method determines for each stage the standardized average score of the retained applicants on a composite criterion that is a weighted combination, with user-specified weights, of the relevant criterion dimensions. Thus, stage-specific standardized average criterion scores are obtained for the overall sample as well as separately for each of the different applicant groups.

Apart from these average criterion scores, the method also determines two sets of adverse impact (AI) ratios. The first set of AI ratios is stage-specific, whereas the second set of AI ratios is cumulative. The two sets of AI ratios are both considered because they convey different pieces of information on the adverse impact of the selection. The cumulative AI ratios report the adverse impact cumulated over all selection stages thus far completed, whereas the stage-specific ratios indicate the adverse impact for a given selection stage. As illustrated in the Study 1: Progressive Group Differences in Multistage Selection section, the combination of both indices allows for an improved analysis of adverse impact.

Finally, the method computes the predictor mean differences between subgroups after each stage. This improves and extends results obtained by Roth, Bobko, Switzer, and Dean (2001) on the effects of prior selection when using a first predictor on the subgroup differences with respect to a second, related predictor. Roth et al.'s (2001) results are improved because the present method provides an analytical computation of the effects of prior selection instead of an approximate, simulation-based estimation. The results are also extended because the method determines these effects not only in case of prior selection on a single predictor but also in case of prior sequential selection on several generally weighted predictor composites.

### *Boundary Conditions and Assumptions*

The application of the present method is contingent on a number of boundary conditions and assumptions. A first set of conditions relates to the features of the studied selection scenario described earlier and to the method of referral that is adopted in the different selection stages. In particular, the method aims at multistage selections from fixed applicant pools in which the candidates come from different populations. In addition, it is understood that at each stage, only (and all) candidates are retained who score at or above a certain cutoff score on the stage-specific predictor (or predictor composite). So, pass–fail decisions are made at each stage, resulting in a noncompensatory, multiple-hurdle selection process, as is often seen in practice.

Second, to derive the standardized average criterion scores of the selected applicants, the adverse impact ratios, and the predictor mean subgroup differences after previous screening, an assumption is made. In particular, it is assumed that the available predictors and the relevant criterion dimensions have a joint multivariate normal distribution with the same variance–covariance matrix but different means in the subgroup populations. This assumption generalizes the one that is currently used in the single-stage analytical method (e.g., De Corte & Lievens, 2003) and is identical to the assumption that underlies the simulation-based approximate calculations (e.g., Doverspike et al., 1996; Sackett & Roth, 1996).

Finally, the application of the method requires that certain selection input data are available. As discussed later in the *Limitations* section, these input data values must be chosen carefully because the results of the calculations depend on them. The data requirements, as well as the fact that the results are contingent on the values of the input data, are not unique to the present method, however. The use and the results of, for example, the simulation approach depend on the same data. More specifically, both the simulation and the present analytical method require data with respect to (a) the predictor–criterion correlations, (b) the mean subgroup differences for the predictors and the criteria, (c) the intercorrelations among the predictors and among the criteria, and (d) the proportional representation of the different applicant groups in the total applicant population. We see two possible bases for obtaining these data. First, estimates of the validities, intercorrelations, and mean subgroup differences of many popular predictor variables can be obtained from the numerous meta-analytic studies on this subject (e.g., Bobko et al., 1999; Hough, Oswald, & Ployhart, 2001; Ones & Anderson, 2002; Salgado, Anderson, Moscoso, Bertua, & De Fruyt, 2003; Schmidt & Hunter, 1998). Many of these studies also provide fairly accurate values for the subgroup mean differences and the intercorrelations of the most common aspects of job performance (e.g., Roth, Huffcutt, & Bobko, 2003). Second, the selection input parameter data might be retrieved from relevant values gathered during previous administrations of a selection system. Organizations have often been using a set of predictors for selection decisions for some years. Hence, local data on the mean subgroup differences, validities, and intercorrelations of these predictors are available.

### *Method and Implementation*

The appendix provides a detailed account of the method to derive the required results from the information on the intended selection scenario and the values of the selection input parameters. In general terms, the method proceeds in four steps. In the first step, the data on the predictor correlations are used together with the intended weighting scheme of the predictors to compute the correlation matrix of the stage-specific composite predictors at each stage. These computations are based on standard formulas to calculate the variance–covariance matrix of linear combinations of variables. Next, the method computes the cutoff values of the composite predictors that correspond to the intended retention rates at the end of each stage. To achieve this purpose, several nonlinear equations are solved, one for each selection stage. These cutoff values and the earlier computed variance–covariance matrix of the stage-specific (composite) predictors are subsequently used in the third step to determine at each stage and for each applicant population the proportion of retained applicants. All these proportions are obtained by evaluating the value of suitably truncated multinormal distributions with algorithms that generalize the thus far adopted approach to the computation of single-stage selection rates (e.g., Taylor & Russell, 1939) to the general case of multistage selection. The resulting proportions enable a straightforward computation of the stage-specific and the cumulative AI ratios of the intended selection with respect to each of the different applicant groups.

In the final step, the method calculates for each stage the average (composite) criterion score of the applicants that are retained at the end of the stage. These calculations are performed for each applicant group separately and for the total group of remaining applicants. As detailed in the appendix, these average criterion scores can be obtained by using a regression equation in which the average scores of the retained/selected applicants on the (thus far implemented) stage-specific predictors are combined according to the optimal regression weights for regressing the composite criterion on these predictors. Also, the required stage-specific average predictor scores are computed with formulas that evaluate the expectation of truncated multinormal distributions (cf. Muthén, 1990; Tallis, 1961) and by extending these formulas to the situation where the joint distribution of the predictors is a mixture of multinormals with the same variance–covariance matrix but different mean vectors, as is presently the case. In the special case of single-stage selection from a homogeneous applicant population, these formulas reduce to those used by Brogden (1949), Cronbach and Gleser (1965), Naylor and Shine (1965), and many others to compute the average criterion score of the selected applicants.

To implement the method, Wilfried De Corte wrote and compiled a computer program to an executable code that runs on a personal computer under the Windows 95/98, NT, XP, and 2000 operating systems. The computer program and a manual that describes the preparation of the input file and the actual usage of the program can freely be downloaded from the Internet (see De Corte, 2005). The documentation also contains an example application and provides further details on the output generated. Because of the increasing numerical complexity, the program is at present limited to the analysis of sequential selections with no more than 4 stages, 10 predictors, 5 criterion dimensions, and 5 different applicant populations, but these limitations should not pose a problem for most practical applications. As explained in the Study 3: Uncertainty in the Selection Parameter Data section, the program also provides the opportunity to embed the analytical computations within a Monte Carlo procedure to handle uncertainty in (some of) the selection parameter data.

## Study 1: Progressive Group Differences in Multistage Selection

We noted in the introductory section that the present method results in a comprehensive overview of both the intermediate and the final (expected) outcomes of an intended selection decision. Such an overview may interest the practitioner because it provides a detailed account of the expected applicant flow through the selection stages, thereby showing which stage, if any, is likely to

result in a disproportional retention/selection of certain applicant groups. In addition, the method helps to understand the causes of such eventual disproportional retention rates by calculating how the initial group differences on the predictors and the criterion evolve through the subsequent stages. To illustrate the potential of the method for addressing these issues and to familiarize the practitioner with the workings and the results of the method, we apply it to a representative example situation.

### *Selection Scenario and Parameter Data*

The application relates to a situation in which (a) the total applicant population was a mixture of White, Black, and Asian candidate populations (with mixture proportions of .70, .20, and .10, respectively), (b) four predictors were available to perform the selection (i.e., biodata [BI], a test of cognitive ability [CA], a measure of conscientiousness [CO], and a structured interview [SI]); and (c) the overall performance criterion was a weighted sum of two constituting dimensions, with weights of 3 and 1 for the task (job) performance and the contextual performance dimensions, respectively (cf. Borman, Penner, Allen, & Motowidlo, 2001; Motowidlo, Borman, & Schmit, 1997).

Table 1 displays the input parameter data used for the predictor and criterion mean subgroup differences, for the predictor validities and intercorrelations, as well as for the correlation between the two criterion dimensions. The reported data are based on the results of previous meta-analytic studies (cf. Bobko et al., 1999; Hough et al., 2001; Hunter & Hunter, 1984; Roth, Huffcutt, & Bobko, 2003; Salgado et al., 2003), and they correspond to uncorrected values (henceforth also referred to as *population estimates*) of the selection input parameters in the (unscreened) applicant populations. Uncorrected values were preferred over corrected estimates because the information was gathered from different sources that either did not always provide corrected values or used different corrections.

On the basis of these input data, the present method was applied to determine the effects of a three-stage selection scenario with retention rates of .60, .30, and .10 (and, hence, with selection rates of .60, .50, and .33) in the consecutive stages. In Stage 1, only the BI information was used to screen the candidates, whereas the retention decisions in Stages 2 and 3 were based on a regression-based composite of the CO and the CA predictors and the SI information, respectively. Thus, the Stage 1 predictor composite ($P_1$) equaled the BI predictor, the Stage 2 predictor ($P_2$) was a weighted composite with weights of .223 and .319 for CO and CA, and the Stage 3 composite predictor ($P_3$) corresponded to the SI.

### *Results*

The results of the application are summarized in Tables 2 and 3. Table 2 shows how the mean subgroup differences of the predictors and the composite predictors evolved through the sequential selection process.

On the basis of Table 2, we can infer that the Black–White and the Asian–White mean predictor and composite predictor differences changed substantially as a result of the previous screenings. For example, the initial mean Black–White difference on the CA predictor changed from a value of −1.000 in the unscreened group (cf. the prior-to-selection value) to the value of −0.520 (cf. the after Stage 3 value) in the finally selected part of this group. Thus, after the final stage, the mean difference on the CA predictor between the selected Black and the selected White applicants was no longer equal to −1.000 but had shrunk to a value of −0.520.

The mean subgroup difference values reported for the composite predictors can be interpreted in a similar way. Consider, for example, the mean subgroup difference values of $P_2$. In the initial, unscreened applicant pools, this composite showed mean subgroup differences from the White population of −0.871 and 0.118 for Black and Asian populations, respectively, whereas after the screening on the basis of $P_1$, in Stage 1, these values were −0.783 and 0.178, respectively. Therefore, depending on the nature of the previous screening, mean differences between retained groups on (composite) predictors may either decrease or increase.

Observe that the method took the above-reported changes of the predictor and predictor composite subgroup mean differences into account when computing the adverse impact and the average

Table 1
*Average Value and Standard Deviation of the Effect Sizes (i.e., Standardized Subgroup Mean Differences) and Intercorrelations of the Performance Predictors and the Performance Criteria*

| | Effect size $d$[a] | | | | Intercorrelation matrix | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Black | | Asian | | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | |
| Variable | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Predictor | | | | | | | | | | | | | | | | |
| 1. Biodata | −0.33 | 0.08 | −0.23 | 0.08 | — | | | | | | | | | | | |
| 2. Cognitive ability | −1.00 | 0.25 | 0.20 | 0.12 | .19 | .05 | — | | | | | | | | | |
| 3. Conscientiousness | −0.09 | 0.09 | −0.08 | 0.09 | .51 | .10 | .00 | .06 | — | | | | | | | |
| 4. Structured interview | −0.23 | 0.07 | −0.13 | 0.07 | .16 | .07 | .24 | .05 | .12 | .12 | — | | | | | |
| Criterion dimension | | | | | | | | | | | | | | | | |
| 5. Task performance | −0.45 | 0.07 | 0.10 | 0.15 | .28 | .07 | .30 | .08 | .18 | .04 | .30 | .04 | — | | | |
| 6. Contextual performance | 0.07 | 0.06 | 0.02 | 0.06 | .25 | .10 | .16 | .08 | .20 | .09 | .26 | .04 | .17 | .16 | — | |
| 7. Composite criterion | −0.39 | — | 0.10 | — | .33 | — | .32 | — | .22 | — | .35 | — | .95 | — | .45 | — |

*Note.* The composite criteria is a weighted sum of task performance (with a weight of 3) and contextual performance (with a weight of 1). Dashes under *SD* indicate that this type of value is not applicable.
[a] Standardized subgroup mean differences are relative to the White (majority) applicant population.

Table 2
*Predictor and Composite Predictor Standardized Subgroup Mean Differences*

| Group and predictor | Standardized subgroup mean difference | | | |
|---|---|---|---|---|
| | Prior to selection | After Stage 1 | After Stage 2 | After Stage 3 |
| **Predictor** | | | | |
| Black applicant group | | | | |
| Biodata | −0.330 | −0.134 | −0.033 | −0.044 |
| Cognitive ability | −1.000 | −0.963 | −0.489 | −0.520 |
| Conscientiousness | −0.090 | 0.010 | 0.275 | 0.252 |
| Structured interview | −0.230 | −0.199 | −0.063 | −0.021 |
| Asian applicant group | | | | |
| Biodata | −0.230 | −0.096 | −0.138 | −0.137 |
| Cognitive ability | 0.200 | 0.225 | 0.133 | 0.156 |
| Conscientiousness | −0.080 | −0.012 | −0.072 | −0.068 |
| Structured interview | −0.130 | −0.109 | −0.037 | −0.037 |
| **Composite predictor** | | | | |
| Black applicant group | | | | |
| $P_1$ | −0.330 | | | |
| $P_2$ | −0.871 | −0.783 | | |
| $P_3$ | −0.230 | | −0.063 | |
| Asian applicant group | | | | |
| $P_1$ | −0.230 | | | |
| $P_2$ | 0.118 | 0.178 | | |
| $P_3$ | −0.130 | | −0.137 | |

*Note.* $P_1$–$P_3$ are Stage 1–Stage 3 predictor composites.

criterion score values at the different selection stages. Therefore, the method not only improves and extends earlier approaches with regard to the determination of the effects of previous screening on the subgroup mean differences of subsequently administered predictors (cf. Roth et al., 2001), but also incorporates these effects into the calculation of the selection outcome values.

Table 3 presents information on two of the main outcomes of multistage selection decisions that are relevant to organizations. In particular, Table 3 presents the AI ratios (cumulative and stage specific) and the standardized average criterion score (stage-specific and total) of the candidates retained at each stage. Recall that in Stage 1, the top-scoring 60% of the candidates on $P_1$ (i.e., the BI predictor) were retained. This resulted in cumulative and stage-specific AI ratios of 0.797, 0.859, and 1.000 and in standardized average criterion scores of −0.058, 0.397, and 0.259 for the Black, the Asian, and the White applicant groups, respectively. At Stage 2, recall that in this illustration, the remaining candidates were screened on the basis of composite $P_2$ (i.e., a composite of CA and CO), and the mean difference on this composite between the Black and Asian applicants retained in Stage 1 and the similarly retained White applicants changed from −0.871 and 0.118 to −0.783 and 0.178, respectively (cf. Table 2). Selecting the top-scoring 50% of individuals at this stage produced stage-specific Black and Asian AI ratios of 0.432 and 1.137, respectively, whereas the corresponding cumulative AI ratios were 0.344 and 0.977. Additionally, the standardized average criterion scores of the Black, Asian, and White candidates that passed Stage 2 increased, as compared with the Stage 1 averages, to 0.356, 0.590, and 0.492, respectively, and the standardized average criterion score of all applicants retained changed from 0.218 at the end of Stage 1 to 0.492 at the end of Stage 2.

Finally, in the last stage, the remaining 30% (.60 × .50 = .30) of the candidates were screened using composite $P_3$ (equal to the SI predictor). Although this composite showed mean differences of −0.230 and −0.130 in the unscreened Black and Asian candidate populations, the corresponding mean differences at the beginning of Stage 3 were −0.063 and −0.137 (cf. Table 2). Furthermore, applying a selection rate of one in three in this stage to obtain the intended hiring rate of 0.10 led to stage-specific Black and Asian AI ratios of 0.930 and 0.854, respectively. The corresponding cumulative AI ratios were 0.320 and 0.834, respectively, whereas the standardized criterion averages of the applicants finally selected further rose to 0.656, 0.921, and 0.793 for the Black, the Asian, and the White candidate groups, respectively, resulting in an overall average criterion score of 0.795 for all selected applicants.

Apart from the results shown in Tables 2 and 3, the computer program completed the applicant flow analysis by computing, for each applicant group and for each stage, the stage-specific and the cumulative retention rates. The present method therefore provided a comprehensive account of all the major effects of general multistage selections. It must be remembered, though, that all the obtained results were conditional on the values used for the selection input parameters. Whereas this dependency is of no consequence when the method is used to determine the outcomes of a specific hypothetical selection scenario with deliberately chosen input values, it becomes a reason for concern when the method is applied to study the consequences of planned but not yet implemented selections. In that case, the relevant input values may be expected to vary from one application to the other, and we therefore address in Study 3 how to account for this variability in the selection input. However, we first explored how the present method may help to advance the understanding of multistage selection and, in particular, the understanding of the effects of different orders of predictor administration.

Table 3
*Adverse Impact Ratios and Average Criterion Scores of the Selected Applicants in a Three-Stage Selection*

| Stage | Black | Asian | White | Total |
|---|---|---|---|---|
| **Cumulative adverse impact ratios** | | | | |
| 1 | 0.797 | 0.859 | 1.000 | |
| 2 | 0.344 | 0.977 | 1.000 | |
| 3 | 0.320 | 0.834 | 1.000 | |
| **Stage-specific adverse impact ratios** | | | | |
| 1 | 0.797 | 0.859 | 1.000 | |
| 2 | 0.432 | 1.137 | 1.000 | |
| 3 | 0.930 | 0.854 | 1.000 | |
| **Standardized average criterion score selected applicants** | | | | |
| 1 | −0.058 | 0.397 | 0.259 | 0.218 |
| 2 | 0.356 | 0.590 | 0.492 | 0.492 |
| 3 | 0.656 | 0.921 | 0.793 | 0.795 |

## Study 2: Multistage Selection Design: Impact of Predictor Sequence

### *Purpose*

As noted in the introductory section, the present method can also be used to study issues related to multistage selection design, in the hope of deriving generally applicable design principles. To illustrate this usage, we implemented the method to address a key question that practitioners and researchers are likely to have about multistage selection design. More specifically, we used the method to study the effects of different ways of sequencing predictors. To keep the scope of the investigation feasible, albeit realistic, we focused on the situation in which two of the available predictors showed virtually identical validities but were substantially different in terms of (administration) cost and subgroup mean differences. Table 1 shows an example of this situation (see the data for when only the task performance criterion dimension is retained). In that case, both the CA and the SI predictors had a validity of .30 for predicting task performance, whereas the Black–White mean differences of the two predictors ranged from −1.00 (CA) to −0.23 (SI). Therefore, the second study considered only the task performance criterion and used these two predictors, together with the CO predictor, for exploring the effects of different administration orders. Similar to Study 1, the average quality of the applicants selected (as expressed in terms of the average criterion value of the employees selected) and adverse impact served as the two selection outcomes. In addition, the computer program was expanded with an option to calculate the cost of general multistage scenarios as a third selection outcome. A simple cost metric was used that accounted for only the costs associated with administering the predictors and, hence, did not consider, for example, development costs. In the scenarios that were studied here, there were two low-cost predictors (CA and CO), with an estimated administration cost of $35 per candidate, and one high-cost predictor (SI), with an estimated administration cost of $100 per candidate. Note that there is nothing sacrosanct about the values chosen for this illustration.

### *Implementation*

To obtain a clear picture of the effects of different predictor sequencing, we focused on situations in which the total applicant group consisted of only Black (20%) and White (80%) candidates. For the same reason, all studied scenarios were characterized by an identical final retention rate of .20, and they all related to selections in which the criterion was limited to the task performance dimension. Finally, to gauge the overall selection cost, we examined a situation in which the total number of applicants was equal to 1,000. On the basis of these specifications and the Table 1 values for the predictors and the task performance criterion, we used the present method to analyze various scenarios in the hope of deriving a general rule for sequencing the predictors in multistage selection. As discussed in this study's *Results* section, the search was only partly successful because the applicability of the obtained rule depended on certain boundary conditions. We detail these conditions in this study's *Results* section, and we also explain how they affect the applicability of the derived rule.

### *Results*

We began the study by testing a basic set of six scenarios that related to two-stage selections, using only the CA and SI predictors. Table 4 presents these scenarios. Per scenario, the order in which the different predictors were used in the stages (cf. the details under the heading *Predictor sequence* in Table 4) and the stage-specific selection rates are given. Basically, these first six scenarios in Table 4 comprise three consecutive blocks of two scenarios. Within each of these blocks, the first (uneven numbered) scenario differs from the second (even numbered) scenario only in terms of the order with which the two high-validity predictors (CA and SI) were administered. This setup was chosen to enable a straightforward evaluation of the effects of a different sequencing of the two predictors. In addition, given that these effects were likely to vary depending on the rates of selectivity with which the CA and SI predictor were applied, we analyzed blocks in which the CA and SI predictors were used with the same level of selectivity (e.g., Scenarios 1 and 2).

Inspection of the results of the two-stage scenarios shows that the order in which equally valid predictors were administered had little if any effect on the average quality of the selected applicants, irrespective of the rate of selectivity with which the predictors were used. However, the results for the adverse impact criterion tell a different story. When predictors differed in subgroup mean differences, it seems that the predictor with the highest impact (here CA) was best administered first, provided that the level of selectivity with which the predictor was applied did not exceed that of the lower impact predictor (compare Scenarios 1–4 with Scenarios 5 and 6). This is a rather unexpected result, and at the end of this section, we show how a careful analysis of the boundary conditions under which it applies suggests an explanation of this finding.

Another interesting albeit much less surprising result from the two-stage scenarios is that giving more weight in the selection to the high-impact predictor (i.e., using this predictor with a lower selection rate) was associated with lower values for the adverse impact ratio. Finally, turning to the test cost measure, it can be verified that a judicious analysis of the expected return of alternative scenarios, as provided by the present method, may often be quite beneficial. In particular, this type of analysis may avoid the implementation of scenarios that are considerably more costly than others and yet offer virtually the same or an even worse return in terms of adverse impact and quality of the selected applicants (e.g., compare Scenarios 1 and 2 or Scenarios 3 and 6). Obviously, the latter observation relates to the present condition in which the high-impact predictor was substantially cheaper than the low-impact predictor, but this condition is likely to prevail in many selection applications because most high-validity, low-impact predictors (e.g., SIs and work sample tests) are more costly than readily available, roughly equally valid but higher impact CA predictors.

Next, the study moved to three-stage scenarios to further test the tentative rule of administering the highest impact predictor (here CA) first, provided that the level of selectivity with which the predictor is applied does not exceed that of the lower impact predictor. The difference between these three-stage scenarios and the two-stage scenarios is that a low-impact, low-validity predictor (CO) was used in the second stage. Inspection of Scenarios 7

Table 4
*Studied Multistage Selection Scenarios*

| Scenario | Selection rate | | | Predictor sequence | | | Average criterion score | Black AI ratio | Total test cost |
|---|---|---|---|---|---|---|---|---|---|
| | S1 | S2 | S3 | S1 | S2 | S3 | | | |
| 1 | .45[a] | .45 | — | CA | SI | — | 0.525 | 0.310 | 79,721 |
| 2 | .45[a] | .45 | — | SI | CA | — | 0.528 | 0.265 | 115,653 |
| 3 | .50 | .40 | — | CA | SI | — | 0.520 | 0.339 | 85,000 |
| 4 | .50 | .40 | — | SI | CA | — | 0.526 | 0.247 | 117,500 |
| 5 | .33[a] | .67[a] | — | CA | SI | — | 0.527 | 0.251 | 68,333 |
| 6 | .33[a] | .67[a] | — | SI | CA | — | 0.521 | 0.332 | 111,667 |
| 7 | .50 | .80 | .50 | CA | CO | SI | 0.532 | 0.333 | 92,500 |
| 8 | .50 | .80 | .50 | SI | CO | CA | 0.535 | 0.290 | 131,500 |
| 9 | .70 | .41[a] | .70 | CA | CO | SI | 0.468 | 0.456 | 88,071 |
| 10 | .70 | .41[a] | .70 | SI | CO | CA | 0.463 | 0.424 | 134,500 |
| 11 | .67[a] | .60 | .50 | CA | CO | SI | 0.506 | 0.431 | 98,333 |
| 12 | .67[a] | .60 | .50 | SI | CO | CA | 0.514 | 0.301 | 137,333 |
| 13 | .50 | .60 | .67[a] | CA | CO | SI | 0.520 | 0.330 | 82,500 |
| 14 | .50 | .60 | .67[a] | SI | CO | CA | 0.511 | 0.381 | 128,000 |
| 15 | .80 | .50 | .50 | CO | CA | SI | 0.533 | 0.332 | 103,000 |
| 16 | .80 | .50 | .50 | CO | SI | CA | 0.536 | 0.289 | 129,000 |
| 17 | .41[a] | .70 | .70 | CO | CA | SI | 0.469 | 0.453 | 77,857 |
| 18 | .41[a] | .70 | .70 | CO | SI | CA | 0.472 | 0.419 | 85,816 |
| 19 | .60 | .67[a] | .50 | CO | CA | SI | 0.507 | 0.429 | 96,000 |
| 20 | .60 | .67[a] | .50 | CO | SI | CA | 0.519 | 0.299 | 109,000 |
| 21 | .60 | .50 | .67[a] | CO | CA | SI | 0.520 | 0.328 | 86,000 |
| 22 | .60 | .50 | .67[a] | CO | SI | CA | 0.516 | 0.378 | 105,500 |

*Note.* S1–S3 = Stages 1–3; AI = adverse impact; CA = cognitive ability; SI = structured interview; CO = conscientiousness. Dashes indicate that this type of value is not applicable.
[a] Proportion was rounded.

through 14 in Table 4 reveals that the tentative rule seems to generalize to such three-stage selection strategies. Again, a comparison of the scenarios within each of the four blocks (i.e., comparing Scenario 7 with 8 and 9 with 10, etc.) shows that the order of CA and SI predictor administration had almost no effect on the average criterion score of the selected employees. However, in line with the two-stage scenarios, the same comparisons revealed that the adverse impact ratio was higher when the high-impact predictor was used first, as long as its associated level of selectivity was less than or equal to the selectivity level with which the low-impact predictor was implemented. Hence, it is again true that costly scenarios may often be outperformed in terms of quality and adverse impact by other, cheaper scenarios. Finally, the second group of scenarios showed that assigning more weight to the low-impact predictors (i.e., applying the predictors with a low selection rate, such that they tended to dominate the overall selection) increased the proportion of minority hires, but the selected employees had a lower average criterion score, especially in cases where the low-validity, low-impact predictor (the CO predictor in this example) was predominant (cf. Scenario 9).

The generality of the rule was further scrutinized by examining three-stage scenarios wherein the CO predictor was added in the first stage (cf. the Scenarios 15 and 22). Results of this third group of scenarios largely duplicated those of the second group of scenarios, except with respect to the selection cost measure. Because these scenarios differed from the corresponding scenarios in the second group only in terms of the staging of the CO predictor (e.g., Scenario 15 applied the same selection rates for the three predictors as Scenario 7 did, but Scenario 15 implemented the CO predictor in the first stage instead of the second; Scenarios 16 and 8 corresponded in the same way, etc.), this finding suggests that the staging order of a low-validity, low-impact predictor is relevant only as far as the total cost of the selection procedure is concerned.

Finally, we examined whether the rule still held (a) for other than the present .80 versus .20 majority–minority composition of the total applicant pool, (b) when the final retention rate of the selection differed from the present value of .20, and (c) when the BI predictor was substituted for the CO measure. Generally, these additional analyses (detailed results are available from Wilfried De Corte) confirmed the tentative principle that, in terms of adverse impact, it is better to sequence the high-impact, valid predictor before the low-impact, equally (or almost equally) valid predictor (instead of the reverse order), provided that the high-impact predictor is not applied more selectively than the low-impact predictor.

However, additional analyses also identified conditions wherein the rule about the sequencing of equally valid predictors was no longer generally applicable. In particular, we found that our tentative rule depends on the level of correlation between the two valid predictors (i.e., the CA and SI predictors in the example). When the predictors do not correlate, the advantage of using first the high-impact predictor vanishes at equal levels of selectivity, and the advantage is substantially reduced in situations where the high-impact predictor is applied with a lower selectivity level than that used for the low-impact predictor. The boundary condition on the correlation between the equally valid predictors also suggests an explanation for the finding about the preferred sequencing of these predictors. When the two predictors have a positive correlation, the mean group difference on the predictor that is applied after the first predictor selection is smaller in the retained groups

than the corresponding differences in the initial groups (cf. the Study 1: Progressive Group Differences in Multistage Selection section and the results presented in Roth et al., 2001). However, this reduction in mean group difference is more substantial when the mean group difference on the first predictor is large compared with the corresponding reduction when the first predictor shows only a small mean group difference. To illustrate this, consider Scenarios 1 and 2 in Table 4. In Scenario 1, the application of the high-impact CA predictor in the first stage reduces the Black–White mean difference on the SI predictor in the retained groups from the initial value of −0.230 to a value of −0.057, whereas the reverse application order of the predictors results in the much smaller reduction from −1.000 to −0.963 for the corresponding difference on the CA predictor. So, applying a high-impact predictor before a low-impact predictor lowers the mean difference on the latter predictor more considerably compared with the corresponding reduction when the predictors are applied in the reverse order. As a consequence, the latter order is expected to result in a lower value of the cumulative AI ratio than that obtained when the high-impact predictor is administered first.

## Study 3: Uncertainty in the Selection Parameter Data

### *Extension of the Method*

Up to this point, we have proceeded as if values of key input parameters (e.g., predictor–criterion correlations, interpredictor correlations) were known with certainty. However, there are multiple sources of uncertainty for many of the input parameters. One is true (i.e., nonartifactual) variability in mean subgroup difference and correlation measures when using meta-analytic estimates as input to our procedure. Consider, for example, the scenario in which the mean predictor–criterion correlation is .30, with a residual standard deviation of .05 after removing variability due to sampling error and other artifacts. Thus, there are as-yet-unidentified substantive or methodological features that result in variability from situation to situation. A given user cannot count on obtaining the mean value of .30 in his or her organization. Thus, when examining the effects of using that predictor as part of a sequential selection system, it is reasonable to ask whether a given pattern of findings (e.g., one sequencing of predictors produces higher mean performance among those selected than another sequencing) holds throughout the range of possible predictor–criterion correlations that the user might plausibly obtain.

The previous example focuses on nonartifactual variation. A user might also be interested in the effects of artifactual variation, such as that due to sampling error. Even if population validity were known with certainty, obtained values would vary from setting to setting because of sampling error. Thus, a user may be interested in knowing whether the advantage of one sequencing of predictors over another holds across the range of predictor–criterion correlation values that might result from sampling error.

To handle the inevitable variability in the input parameter data, we propose to integrate the present analytical method within a Monte Carlo simulation procedure in which the calculations are repeated many times. In each such repetition, a value for each data input parameter is randomly drawn from the distribution that represents the expected variability of that parameter, and the combination of the thus obtained input parameter values is used to calculate the selection outcome values. The frequency distribution of the values obtained over the entire set of replications, for example, for the average quality of the selected workforce, can then be regarded as an approximation of the distribution function of this selection outcome (cf., e.g., Rich & Boudreau, 1987). The frequency distribution and, in particular, selected percentile values of this distribution can therefore be used to determine, for example, a 90% probability interval for the selection outcome indices. We chose not to label these intervals *confidence intervals* because they do not relate to a parameter of a statistical model (cf. Stuart, Ord, & Arnold, 1999).

To actually apply the Monte Carlo procedure, the appropriate distributions of the data input parameters must be specified. Similar to previous applications of the procedure to assess the utility of selection decisions under uncertainty (e.g., Rich & Boudreau, 1987), the program that implements the extended method provides two options. The first option is based on rectangular sampling distributions for the input variables, requiring the specification of lower and upper bound values of these input parameters, whereas the second option assumes normal distributions, in which case the average value and the standard deviation of the input parameter must be provided. The type of distribution and even more so the values of the parameters of the distributions should be chosen carefully, however, because the results of the Monte Carlo extension are adequate only in so far as the distributions are representative for the variability that is actually of interest. Thus, in applications in which artifactual sources of variability (e.g., sampling error) as well as true (i.e., nonartifactual) variability are of interest, one may prefer normal distributions and equate the expected value and the standard deviation of the distribution to the average and the uncorrected standard deviation of the input parameter, respectively, as reported in previous summary studies. The probability intervals mentioned earlier can then be perceived as estimates of the corresponding total variability intervals of the selection outcomes. In other applications, certain possible sources of variability may not apply, such that less liberal values for the standard deviation parameter are indicated, leading to a probability interval for the selection outcomes that must be interpreted accordingly.

The Monte Carlo approach is not only useful for addressing the issue of variability; it also permits studying whether the relative standing of alternative selection scenarios is consistent over the set of likely values for the selection input variables. Although a selection scenario may on average (i.e., over the entire set of Monte Carlo replications) lead to, for example, a higher selection quality than a second scenario, the same conclusion need not apply at the level of each individual replication. Suppose, for example, that Scenario A leads on average to a selected workforce that scores 0.1 standard units higher than the workforce obtained under Scenario B, but the latter scenario outperforms Scenario A in 45% of the replications. In that case, the relative standing of the two scenarios is not really consistent over the set of likely value combinations for the selection input parameters, and it would therefore be incorrect to conclude that Scenario A is consistently better than Scenario B. If, on the other hand, a great majority of the replications show a better selection quality for Scenario A, the practitioner may be quite confident that this scenario will also outperform Scenario B when applied to his or her intended selection decision.

As illustrated in this example, the issue of consistency in the relative standing of alternative scenarios is of crucial importance when one intends to apply the present method within a decision-making context. In particular, the use of the method as a tool to decide between alternative selection scenarios critically depends on whether the relative position of these alternative scenarios is more or less stable over the expected variability in the selection parameter data. The next example therefore shows how the Monte Carlo extension of the method can be used to investigate this consistency issue. The example also clarifies the computation of the probability intervals of the relevant selection outcomes that we introduced earlier.

### *Illustration*

For reasons of convenience, the example focuses on Scenarios 7 through 14 (cf. Table 4) that were analyzed in the previous section as part of the study on the impact of different predictor sequencing. As reported previously, these eight scenarios all relate to a situation in which the total applicant pool is composed of 80% White (majority) and 20% Black (minority) candidates. Also, all scenarios share the same final retention rate of .20, which is attained after a three-stage selection process. The purpose is to test whether the relative position of these eight scenarios is maintained over a large set of possible value combinations for the selection input parameters. To achieve this purpose and to assure a rather severe test of the consistency issue, we applied the extended Monte Carlo version of the method, using values for the distribution parameters of the input variables that correspond to the condition of high uncertainty for the parameter values, including both sampling error and variability due to systematic effect sources. More specifically, at each of a total of 10,000 replication samples, the value of each input parameter was randomly drawn from a normal distribution with expectation equal to the value of the parameter, as shown in Table 1. The standard deviations of the input parameter distributions (cf. the values reported between brackets in Table 1) are based on results presented by Bobko et al. (1999); Hattrup, Rock, and Scalia (1997); Hough et al. (2001); Hunter and Hunter (1984); McManus and Kelly (1999); Murphy and Shiarella (1997); Roth et al. (2003); and Salgado et al. (2003). These standard deviations reflect the uncorrected variability of the corresponding input parameter. Also, some scale parameter values and, in particular, those related to the standard deviation of the mean Asian–White subgroup differences on Predictors 1, 3, and 4 and the contextual performance criterion, were chosen to be equal to the corresponding numbers for the Black subgroup because no other appropriate estimates could be found in the literature.

The results of these analyses are reported in Tables 5 and 6. Table 5 displays, for each scenario, the mean (as computed over the 10,000 Monte Carlo samples) of the final-stage, cumulative AI ratio and the standardized average criterion score of all finally selected applicants.

As expected, the latter mean values are virtually identical to the corresponding standardized average scores as computed on the basis of the mean value of the data input parameters (cf. the average criterion score reported in Table 4 for the scenarios). Alternatively, the Black subgroup's average AI ratios in Table 5 are all somewhat larger than the corresponding Table 4 values. This is because equal offsets from the subgroup mean difference values result in unequal offsets for the AI ratio value. Thus, a change of, for example, the CA mean subgroup difference from $-1.0$ to $-0.9$ results, all else being equal, in a higher increase of the AI ratio as compared with the decrease associated with a change in the mean subgroup difference from $-1.0$ to $-1.1$.

Table 5 also details the 90% probability intervals for these outcomes as based on the Monte Carlo replications. Because these replications mimic the variability of the selection input variables over both sampling error and other sources of variability, these probability intervals can be regarded as total variability intervals for the corresponding selection outcomes of quality and adverse impact. By and large these intervals show that the outcomes vary substantially over the studied set of possible value combinations for the selection input parameters. For most of the analyzed selection scenarios, there is also a considerable overlap between the corresponding probability intervals, suggesting that the relative

Table 5
*Mean Value (Over 10,000 Monte Carlo Samples) of the Standardized Average Criterion Score and the Adverse Impact (AI) Ratio, as Well as the Corresponding 90% Probability Intervals for the Eight Studied Selection Scenarios*

| | Selection rate | | | Predictor sequence | | | Average criterion score | Black AI ratio | 90% probability interval | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | S1 | S2 | S3 | S1 | S2 | S3 | | | Average score | Black AI ratio |
| 7 | .50 | .80 | .50 | CA | CO | SI | 0.532 | 0.340 | 0.432–0.631 | 0.195–0.510 |
| 8 | .50 | .80 | .50 | SI | CO | CA | 0.534 | 0.300 | 0.422–0.644 | 0.154–0.478 |
| 9 | .70 | .41[a] | .70 | CA | CO | SI | 0.468 | 0.460 | 0.383–0.554 | 0.309–0.624 |
| 10 | .70 | .41[a] | .70 | SI | CO | CA | 0.463 | 0.429 | 0.368–0.556 | 0.269–0.604 |
| 11 | .67[a] | .60 | .50 | CA | CO | SI | 0.507 | 0.434 | 0.422–0.592 | 0.291–0.590 |
| 12 | .67[a] | .60 | .50 | SI | CO | CA | 0.513 | 0.311 | 0.398–0.623 | 0.160–0.495 |
| 13 | .50 | .60 | .67[a] | CA | CO | SI | 0.519 | 0.338 | 0.415–0.619 | 0.188–0.516 |
| 14 | .50 | .60 | .67[a] | SI | CO | CA | 0.510 | 0.386 | 0.417–0.603 | 0.236–0.555 |

*Note.* S1–S3 = Stages 1–3; CA = cognitive ability; CO = conscientiousness; SI = structured interview.
[a] Proportion was rounded.

Table 6
*Comparison of the Studied Selection Scenarios Over the Monte Carlo Replications: Difference in Average Outcome and Percentage With Which the Row Scenario Dominates the Column Scenario for Selection Quality (Upper Triangle) and Black Adverse Impact Ratio (Lower Triangle)*

| | Scenario | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Scenario | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 7 | | −0.002 | 0.064 | 0.070 | 0.026 | 0.019 | 0.013 | 0.022 |
| | | 43.6 | 97.7 | 98.1 | 89.0 | 80.1 | 79.1 | 88.8 |
| 8 | −0.041 | | 0.066 | 0.072 | 0.028 | 0.021 | 0.015 | 0.024 |
| | 0.2 | | 96.4 | 97.7 | 81.2 | 92.0 | 86.7 | 84.5 |
| 9 | 0.120 | 0.160 | | 0.005 | −0.039 | −0.045 | −0.051 | −0.042 |
| | 99.9 | 100.0 | | 65.8 | 1.5 | 7.8 | 1.7 | 0.5 |
| 10 | 0.088 | 0.129 | −0.031 | | −0.044 | −0.050 | −0.057 | −0.048 |
| | 99.4 | 100.0 | 3.9 | | 3.4 | 2.9 | 0.6 | 0.2 |
| 11 | 0.094 | 0.134 | −0.026 | 0.006 | | −0.006 | −0.013 | −0.004 |
| | 100.0 | 100.0 | 7.3 | 60.5 | | 42.3 | 20.3 | 37.6 |
| 12 | −0.030 | 0.011 | −0.149 | −0.118 | −0.124 | | −0.006 | 0.003 |
| | 7.0 | 84.7 | 0.0 | 0.0 | 0.1 | | 30.2 | 55.2 |
| 13 | −0.002 | 0.038 | −0.122 | −0.090 | −0.096 | 0.028 | | 0.009 |
| | 40.8 | 99.8 | 0.0 | 0.0 | 0.0 | 98.8 | | 73.8 |
| 14 | 0.046 | 0.087 | −0.074 | −0.042 | −0.048 | 0.076 | 0.048 | |
| | 99.1 | 100.0 | 0.0 | 0.7 | 0.5 | 99.9 | 100.0 | |

*Note*. The uppermost number within each cell corresponds to the difference in average outcome (upper triangle) or to the difference in Black adverse impact ratio (lower triangle) between the row and the column scenario.

standing of the scenarios may often differ from one set of input parameter values to the other.

To test for this presumed inconsistency, the percentage of Monte Carlo samples in which the first scenario dominated the second scenario was tabulated for each pair of selection scenarios and for each selection outcome. Table 6 summarizes the results of this tabulation. Table 6 also mentions, per pair of scenarios, the difference between the outcome value obtained for the first (row) and the second (column) scenarios.

The basic finding that emerges from Table 6 is that scenarios that differ at least noticeably for a given outcome are fairly consistently different for each of the examined combinations of values for the selection input parameters. For example, with respect to the selection quality outcome (upper triangle of Table 6), Scenario 7 results in a workforce quality that is only 0.026 standard units higher than the quality associated with Scenario 11. Yet, despite this very small difference, Scenario 7 dominates Scenario 11 in 89% of the individual Monte Carlo replications. Even the order within pairs of scenarios that differ between 0.01 and 0.02 standard units is very well preserved because the scenario with the largest outcome of the pair on average outperforms the lower outcome scenario on nearly 80% or more of the replications. Regarding comparisons between more different scenarios, the data are even more conclusive, indicating that scenarios that show a practically relevant difference are indeed consistently different. Note that it is not surprising that there is no straightforward linear relationship between the difference in average outcome and the corresponding consistency percentage because the latter percentage also depends on whether the scenarios differ in terms of one or several design aspects (e.g., the stage-specific selection rates, the order of the predictors).

The results on the consistency with which the scenarios differ in terms of the adverse impact outcome (lower triangle of Table 6) further corroborate the above derived trend. All this suggests that the relative standing of alternative selection scenarios in terms of both selection quality and adverse impact is consistent over variability in the selection input values. Although the absolute value of the outcomes associated with a scenario are quite dependent on the input parameter values, the relative positions of the scenarios in relation to one another is much more robust. As a consequence, we propose that the present method to assess selection outcomes is also practically relevant when the input parameter values are fairly uncertain.

## General Discussion

### *Methodological Contribution*

In terms of methodological contributions, the article presented a method and a related computer program that enable selection researchers and practitioners to explore the consequences of an intended multistage selection decision in terms of the selection cost, the selection quality, and the level of adverse impact when the applicants come from populations that have a different average score on the selection predictors. Although the method is based on assumptions identical to those of the existing simulation approach and its application is contingent on the same information (as the simulation approach) with respect to the nature of the intended selection and the characteristics of the predictors and the criteria, it has some important added benefits. To start with, it produces a point estimate of the major selection outcomes. Repeated applications on the same input data obtained by a simulation-based approach can at best approximate this point estimate.

Next, because the analytical method is computationally very fast, the method is easily integrated within a Monte Carlo approach to handle uncertainty, and variability in particular, in the values of the predictor–criterion parameter values. This integration results in two further benefits. First, it permits the derivation of appropriate probability intervals for the selection outcome variables, thereby representing the variability in the results that one may expect to obtain from one application of the intended selection to the other. Second, the integration makes it possible to study whether the relative positions of selection scenarios, in terms of the associated outcomes, are consistent over a range of possible value combinations for the selection input parameters.

The method is also generally applicable and comprehensive because it can deal with more than two applicant groups and more than two selection stages. In addition, the method can be extended to deal with the eventuality that the applicant populations not only have different predictor and criterion averages, but also show different predictor and criterion variances. Finally, with some minor modifications and provided that suitable additional data are available, the principles of the method can also be used to assess other measures of selection quality, such as the success ratio and the utility of the intended selection.

### *Substantive Contribution*

In terms of substantive contributions, the article explored a key issue in multistage selection decisions about the sequencing of equally valid predictors. To favor acceptable levels of adverse impact, it may seem reasonable to reserve the high-impact predictor for the last stage. The present program was used to examine the viability of this intuition across a fairly comprehensive set of scenarios. Results showed that the opposite seems to be true in most cases. In fact, an important new substantive finding was that if predictors have roughly equal validity but substantially different subgroup mean differences, it is better to sequence the high-impact, valid predictor before the low-impact predictor (instead of the reverse order). Moreover, the study revealed that it is better to use the high-impact predictor before the low-impact predictor, not only when the high-impact predictor is used with a substantially lower level of selectivity than that applied to the low-impact predictor, but also when there are equal levels of selectivity. We also discovered that this rule generalizes over various proportions in the majority–minority composition of the total applicant pool, over different rates of final selectivity, and over settings in which a third predictor is used either before or in between the two predictors. Although this rule is neither intuitively nor logically evident, it is not totally unexpected because Sackett and Roth (1996, p. 564) also observed that use of the high-impact predictor at the initial stage need not necessarily lead to higher minority hiring rates. Yet, the present findings provide a more precise indication about the conditions under which this will be the case.

On the basis of this information, one might be tempted to conclude that it is indeed feasible for multistage selections to derive generally applicable design principles that go beyond the sort of guidelines that follow from sheer logic. Examples of such logic-based rules are the following: (a) Assigning more weight to valid predictors as compared with less valid predictors (i.e., using the valid predictors more selectively than the less valid predictors) leads to increased average criterion scores for the selected applicants (cf. the results presented earlier as well as those of Sackett & Roth, 1996), and (b) using stage-specific predictors that combine the information of all the already administered predictors leads to a similar increase as compared with the use of only the stage-specific predictor information (cf. Sackett & Roth, 1996). This conclusion is premature, however. In particular, we recommend that researchers use extreme caution when endeavoring to search for general rules of thumb in multistage selection. In fact, in other analyses, we discovered that there are boundary conditions to the "general" rule. The rule about the sequencing of equally valid predictors is of only limited generalizability because it requires further amendment depending, among other things, on the level of intercorrelation between the predictors. When the predictors do not correlate, the advantage of using the high-impact predictor first vanishes at equal levels of selectivity, and the advantage is substantially reduced in situations in which the high-impact predictor is applied with a lower selectivity level than that used for the low-impact predictor.

The fact that the present rule on the sequencing of predictors is of only limited generalizability is not an exception. We analyzed a great number of multistage scenarios, searching for guidelines that generalize over a broad class of settings. Yet, time and again we found (as already proposed by Sackett & Roth, 1996) that "there are no simple rules that can be offered about which approach to hurdle based selection is preferred" (p. 569). Informative design principles are typically contingent on both the generic and the specific characteristics of the selection situation at hand. Although this is a rather unfortunate conclusion, it underscores at the same time the importance of the present method and computer program because they provide the means to allow researchers and practitioners to study the merits of alternative designs, given the particularities of the situation at hand.

As a last substantive contribution, the results of the Monte Carlo extension of the method indicate that the rank order of alternative selection scenarios, according to their quality or adverse impact, is fairly robust for variability in the selection input parameter values. This finding adds to the practical value of the method because it indicates that its application does not critically depend on the availability of accurate sample estimates for the selection input parameters. Reasonably approximate values, which are offered by meta-analyses, seem to be sufficient. Of course, future research with the same as well as different sets of predictors is needed to confirm this.

### *Limitations*

The present method is not without limitations. First, identical to the simulation approach, the application of the method and the accuracy of the obtained results are contingent on a number of boundary conditions. We summarized these conditions in the *Boundary Conditions and Assumptions* section, and we recall here the assumption about the joint distribution of the predictor–criterion scores in the different applicant populations. When this assumed distribution is a poor approximation of the predictor–criterion distribution in the studied applicant samples, the results are systematically inaccurate.

Second, the effects of eventual refusal to accept a job offer by some of the selected candidates or applicant withdrawal along the process remain unaccounted for. However, the incorporation of the

effects of applicant withdrawal/refusal requires fairly detailed information about the underlying process of refusal and its eventual relation with performance on the predictor variables (Murphy, 1986; Ryan, Sacco, McFarland, & Kriska, 2000; Schmit & Ryan, 1997). This information is usually not available, or it may be situation specific and therefore less suitable to model within a generally applicable procedure, whether the method is analytical or based on simulation.

Third, the earlier conclusion of robustness for the uncertainty in the input parameters should not be mistaken as an excuse to pay less attention to the values used for the correlations and mean subgroup differences of the predictors and the criterion dimensions. Instead, the practitioner who intends to use these methods to perform what-if analyses of alternative selection scenarios is encouraged to ensure that the input data on the proportional composition of the total applicant group and the predictor–criterion correlations and mean subgroup differences are as accurate as possible. Accurate estimates constitute the best guarantee that practitioners can avoid poor or even misleading results. Given that some inaccuracy of the predictor–criterion data might be expected, we strongly recommend using the Monte Carlo extension of the analytical method. Accordingly, users receive information about the potential variability of the results over different actual applications of an intended scenario.

### *Implications for Future Research*

Now that this study presented an analytical method to assess the adverse impact and the average criterion performance of selected applicants for general multistage selection decisions, the next step for future research should be to use this method to achieve another important objective: the design of multistage selection decisions that aim to achieve a given set of goals in terms of workforce quality and desired levels of workforce diversity. As observed throughout this article, it seems that this design issue cannot be fully addressed by means of general rules of thumb because the available evidence shows that these rules are either limited in scope and guidance or overly complex to match the particularities of the selection at hand. Instead of using the present method to continue the search for such general rules of thumb, we believe that this method can better be integrated in a much more direct approach to the design of multistage selections, namely the approach of multicriteria optimization.

In the present context, the method of multicriteria optimization could assist the design problem by providing both a tabular and a pictorial overview of all the feasible selection scenarios that result in an optimal trade-off between the typically conflicting goals of low-cost, low-adverse impact, and high quality of the selected applicants. In the language of multicriteria optimization, these optimal trade-off scenarios are also called *Pareto optimal* in the sense that any other feasible scenario that differs from them results in either a decrease of the selection quality or an increase of the cost or the level of adverse impact, and the entire set of optimal scenarios (and, in particular, the set of associated trade-offs) is usually referred to as the *Pareto surface* (Keeney & Raiffa, 1993). Given a tabular or, preferably, a pictorial representation of this Pareto surface, the practitioner could then decide on the preferred optimal trade-off and the associated scenario.

Despite the potential of the multicriteria optimization approach and its popularity in other domains (e.g., engineering, economics, and management science), it has not yet been proposed to assist in the design of selection scenarios. One plausible reason for this omission is that the multicriteria optimization approach requires that the effects of alternative design choices on the relevant objectives be computed analytically instead of only approximately, as was previously possible with the simulation-based method. The present method has removed this barrier so that the final step toward the implementation of the multicriteria optimization approach to the design of more optimal multistage selection scenarios comes within reach.

## References


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52,* 561–589.

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Assessment and Selection, 9,* 52–69.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology, 2,* 171–183.

Cronbach, L., & Gleser, G. (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

De Corte, W. (1998). Estimating and maximizing the utility of sequential selection decisions with a probationary period. *British Journal of Mathematical and Statistical Psychology, 51,* 101–121.

De Corte, W. (2005). *CAIMSGUZ program* [Computer software and manual]. Retrieved March 2, 2006, from http://users.ugent.be/~wdecorte/software.html

De Corte, W., & Lievens, F. (2003). A practical procedure to estimate the quality and the adverse impact of single-stage selection decisions. *International Journal of Assessment and Selection, 11,* 89–97.

Doverspike, D., Winter, J., Healy, M., & Barrett, G. (1996). Simulation as a method of illustrating the impact of differential weights on personnel selection outcomes. *Human Performance, 9,* 259–273.

Hattrup, K., & Rock, J. (2002). A comparison of predictor-based and criterion-based methods for weighing predictors to reduce adverse impact. *Applied H. R. M. Research, 7,* 22–38.

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

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

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives.* Cambridge, England: Cambridge University Press.

McManus, M. A., & Kelly, M. L. (1999). Personality measures and biodata: Evidence regarding their incremental predictive value in the life insurance industry. *Personnel Psychology, 52,* 137–148.

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10,* 71–83.

Murphy, K. R. (1986). When your top choice turns you down. Effect of rejected offers on the utility of selection tests. *Psychological Bulletin, 99,* 133–138.

Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate framework for studying test validity. *Personnel Psychology, 50,* 823–854.

Muthén, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology, 43,* 131–143.

Naylor, J. C., & Shine, L. C. (1965). A table for determining the increase in mean criterion score obtained by using a selection device. *Journal of Industrial Psychology, 3,* 33–42.

Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology, 75,* 255–276.

Rich, J. S., & Boudreau, J. (1987). The effects of variability and risk in selection and utility analysis: An empirical comparison. *Personnel Psychology, 40,* 55–84.

Roth, P. L., Bobko, P., Switzer, F. S., III, & Dean, M. A. (2001). Prior selection causes biased estimates of standardized ethnic group differences: Simulation and analysis. *Personnel Psychology, 54,* 591–617.

Roth, P. L., Huffcutt, A. L., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88,* 694–706.

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, S. D. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology, 85,* 163–179.

Sackett, P. R., & Roth, L. (1996). Multistage selection strategies: A Monte Carlo investigation of effects on performance and minority hiring. *Personnel Psychology, 49,* 549–572.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist, 56,* 302–318.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56,* 573–605.

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

Schmit, M. J., & Ryan, A. M. (1997). Applicant withdrawal: The role of test-taking attitudes and racial differences. *Personnel Psychology, 50,* 855–876.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82,* 719–730.

Stuart, A., Ord, K., & Arnold, S. (1999). *Kendall's advanced theory of statistics: Vol. 2A: Classical inference and the linear model.* London: Arnold.

Tallis, G. M. (1961). The moment generating function of the truncated multinormal distribution. *Journal of the Royal Statistical Society, Series B, 23,* 223–229.

Taylor, H. C., & Russell, J. F. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

# Appendix

## Calculation of the Adverse Impact (AI) Ratio and the Standardized Average Criterion Score in Multistage Selection

In this appendix, we detail the computation in general multistage selection of the cumulative and stage-specific AI ratios, $\mathbf{A}^{(r)}$ and $\mathbf{A}^{(h)}$, as well as the stage-specific standardized average criterion scores, $\overline{\mathbf{C}}^{*(s)} = (\overline{C}^{*(s)}_1, \ldots, \overline{C}^{*(s)}_S)'$ (with $S$ the total number of stages), of the applicants retained thus far. We also briefly indicate how the mean subgroup differences, after previous screenings, of the predictors and the composite predictors can be determined. Throughout, the symbol $j$, with $j = 1, \ldots, J$, is used to identify the different applicant populations, and the majority (reference) population arbitrarily corresponds to the last (i.e., the $J$th) applicant population. Also, the vector $\mathbf{q} = (q_1, \ldots, q_J)'$ summarizes the proportional representation of the different applicant groups in the total applicant population.

To obtain the required quantities, we invoke the assumption introduced earlier that the predictors $\mathbf{X} = (X_1, \ldots, X_r, \ldots, X_R)'$ and the criterion dimensions $\mathbf{Y} = (Y_1, \ldots, Y_t, \ldots, Y_T)'$ have a joint multivariate normal distribution in the different applicant populations. The assumption implies that the composite predictors $\mathbf{P} = (P_1, \ldots, P_S)'$ and the composite criterion $C$ also follow a multivariate normal distribution with the same variance/covariance matrix but with different mean vectors in these populations. Also, when we assume, without loss of generality, that $(\mathbf{X}', \mathbf{Y}')'$ have a multivariate standard normal distribution in the majority applicant group, and when we rescale the stage-specific predictors and the composite criterion within each population to the corresponding unit variance predictors, $\mathbf{P}^* = (P^*_1, \ldots, P^*_s, \ldots P^*_S)'$, and unit variance composite criterion, $C^*$, we find that the joint distribution of $\mathbf{P}^*$ and $C^*$ is standard $S + 1$-variate normal with correlation matrix $\mathbf{R}_{\mathbf{P}^*C^*}$ in the majority group: $(\mathbf{P}^*, C^*) \sim \mathrm{N}_{S+1}(\mathbf{0}, \mathbf{R}_{\mathbf{P}^*C^*})$. Alternatively, the joint distribution of $\mathbf{P}^*$ and $C^*$ in the $j$th minority applicant group can be written as $(\mathbf{P}^*, C^*) \sim \mathrm{N}_{S+1}[(\mathbf{d}'_{\mathbf{P}^*j}, d_{C^*j})', \mathbf{R}_{\mathbf{P}^*C^*}]$.

Using the vector $\mathbf{b}_s$ to summarize the weights with which the predictors administered thus far are combined to the stage-specific predictor composite at stage $s$, $P_s$, we determine the general elements $R_{P_s^*P_{s'}^*}$ and $R_{P_s^*C^*}$ of $\mathbf{R}_{\mathbf{P}^*C^*}$ as

$$R_{P_s^*P_{s'}^*} = \frac{(\mathbf{b}'_s, \mathbf{0}')\mathbf{R}_{\mathbf{X}}(\mathbf{b}'_{s'}, \mathbf{0}')'}{\sqrt{(\mathbf{b}'_s, \mathbf{0}')\mathbf{R}_{\mathbf{X}}(\mathbf{b}'_s, \mathbf{0}')'}\sqrt{(\mathbf{b}'_{s'}, \mathbf{0}')\mathbf{R}_{\mathbf{X}}(\mathbf{b}'_{s'}, \mathbf{0}')'}}$$

and

$$R_{P_s^*C^*} = \frac{(\mathbf{b}'_S, \mathbf{0}')\mathbf{R}_{\mathbf{XY}}(\mathbf{0}', \mathbf{w}')'}{\sqrt{(\mathbf{b}'_s, \mathbf{0}')\mathbf{R}_{\mathbf{X}}(\mathbf{b}'_s, \mathbf{0}')'}\sqrt{\mathbf{w}'\mathbf{R}_{\mathbf{Y}}\mathbf{w}}},$$

respectively, where $\mathbf{0}$ is a zero vector of the appropriate order, $\mathbf{w}$ represents the vector of preassigned weights to the $T$ criterion dimensions, $\mathbf{R}_{\mathbf{X}}$ and $\mathbf{R}_{\mathbf{Y}}$ are the correlation matrices of the predictors and the criterion dimensions, and $\mathbf{R}_{\mathbf{XY}}$ is the joint correlation matrix of $\mathbf{X}$ and $\mathbf{Y}$. Alternatively, the general element $d_{P_s^*j}$ of $\mathbf{d}_{\mathbf{P}^*j}$ equals $(\mathbf{b}'_s, \mathbf{0}')\mathbf{d}_{\mathbf{X}j}/\sqrt{(\mathbf{b}'_s, \mathbf{0}')\mathbf{R}_{\mathbf{X}}(\mathbf{b}'_s, \mathbf{0}')'}$, whereas $d_{C^*j} = \mathbf{w}'\mathbf{d}_{\mathbf{Y}j}/\sqrt{\mathbf{w}'\mathbf{R}_{\mathbf{Y}}\mathbf{w}}$, where the vector $\mathbf{d}_{\mathbf{X}j} = (d_{X_1j}, \ldots, d_{X_rj}, \ldots, d_{X_Rj})'$ indicates the mean subgroup differences, relative to the reference (majority) group, of the $R$ selection predictors for applicant group $j$; whereas the vector $\mathbf{d}_{\mathbf{Y}j} = (d_{Y_1j}, \ldots, d_{Y_tj}, \ldots, d_{Y_Tj})'$ refers to the mean subgroup differences for the same group (and again relative to the majority group) of the $T$ criterion dimensions.

Given the above results, the retention rate for the $j$th applicant group at the end of the $s$th stage, $r_{sj}$, can be equated to the value of a suitably truncated standard $s$-variate normal distribution. More specifically, using the notation $\Phi(c_1, \ldots, c_s; \mathbf{R})$ to refer to the upper tail probability of the $s$-variate standard normal distribution with correlation matrix $\mathbf{R}$ evaluated at the cutoff values $c_1, \ldots, c_s$, the value of $r_{sj}$ can be determined as

$$r_{sj} = \Phi(p^*_{c1} - d_{P_{1j}^*}, \ldots, p^*_{cs} - d_{P_{sj}^*}; \mathbf{R}^{(s)}_{\mathbf{P}^*}),$$

where $\mathbf{R}^{(s)}_{\mathbf{P}^*}$ denotes the correlation matrix of the first $s$ composite predictors, and the composite predictor cutoffs, $\mathbf{p}^*_c = (p^*_{c1}, \ldots, p^*_{cs})'$, have values such that the intended, stage-specific retention rate for the total applicant group, $r_s$, is attained. Because the latter retention rate is equal to $\Sigma_j q_j r_{sj}$, it follows that the entire set of composite predictor cutoffs $\mathbf{p}^*_c$ can be sequentially computed as the solution values of the following system of $S$ equations:

$$\begin{aligned} \sum_j q_j \Phi(p^*_{c1} - d_{P_{1j}^*}; \mathbf{R}^{(1)}_{\mathbf{P}^*}) &= r_1 \\ \vdots &= \vdots \\ \sum_j q_j \Phi(p^*_{c1} - d_{P_{1j}^*}, \ldots, p^*_{cS} - d_{P_{Sj}^*}; \mathbf{R}^{(S)}_{\mathbf{P}^*}) &= r_S . \end{aligned}$$

Once the composite predictor cutoff values have been determined, the computation of, for example, the cumulative adverse impact ratios $\mathbf{A}^{(r)} = [(a^{(r)}_{sj})]$ is straightforward. As the cumulative AI ratio of the intended selection for the $j$th applicant population at the end of stage $s$, $a^{(r)}_{sj}$, is equal to the ratio between the stage $s$ retention rate for this group and the corresponding rate for the last (i.e., the majority) group, its value can be calculated as

$$a^{(r)}_{sj} = \frac{r_{sj}}{r_{sJ}} = \frac{\Phi(p^*_{c1} - d_{P_{1j}^*}, \ldots, p^*_{cs} - d_{P_{sj}^*}; \mathbf{R}^{(s)}_{\mathbf{P}^*})}{\Phi(p^*_{c1}, \ldots, p^*_{cs}; \mathbf{R}^{(s)}_{\mathbf{P}^*})},$$

because, by earlier convention, $d_{\mathbf{P}_J^*} = (d_{P_{1J}^*}, \ldots, d_{P_{SJ}^*})' = \mathbf{0}$.

To determine the value of $\overline{C}^{*(s)}_s$, we invoke the equations of Muthén (1990) and Tallis (1961) on the moments of the truncated multinormal distribution. From these equations, and adopting the simplified notation $p^*_{c1j}, \ldots, p^*_{csj}$ for the cutoffs $p^*_{c1} - d_{P_1^*j}, \ldots, p^*_{cS} - d_{P_{Sj}^*}$, it follows that the average score after stage $s$ of the retained applicants from the $j$th population, $\overline{C}^*_{sj}$, can be computed as

$$\overline{C}^*_{sj} = d_{C_j^*} + \frac{\sum_l^s R_{P_l^*C^*}\phi(p^*_{clj})\Phi(f^{(1)}_{lj}, \ldots, f^{(l-1)}_{lj}, f^{(l+1)}_{lj}, \ldots, f^{(s)}_{lj}; \mathbf{R}^{[l,s]}_{\mathbf{P}^*})}{\Phi(p^*_{c1J}, \ldots, p^*_{csj}; \mathbf{R}^{(s)}_{\mathbf{P}^*})},$$

where $\phi(\cdot)$ represents the standard normal density, $\mathbf{R}^{[l,s]}_{\mathbf{P}^*}$ denotes the matrix of partial correlations of $P^*_1, \ldots, P^*_{l-1}, P^*_{l+1}, \ldots, P^*_s$ controlling for $P^*_l$, and, for $1 \le k \le s$ and $k \ne l$,

$$f^{(k)}_{lj} = \frac{p^*_{ckj} - R_{P_k^*P_l^*}p^*_{clj}}{\sqrt{1 - R^2_{P_k^*P_l^*}}},$$

with $R_{P_k^*P_l^*}$ the correlation between the composite predictors $P^*_k$ and $P^*_l$. It can be shown that the thus obtained averages correspond to

$$\overline{C}^*_{sj} = d_{C_j^*} + \sum_{l=1}^{s} \beta_{ls}\overline{P}^*_{slj}$$

where $\beta_{ls}$ is the regression weight of predictor $P^*_l$ when regressing $C^*$ on the predictors $P^*_1, \ldots, P^*_s$, and $\overline{P}^*_{slj}$ is the average score of the group $j$ retained applicants on predictor $P^*_l$ after the first $s$ stages. The averages $\overline{C}^*_{sj}$ (with $j = 1, \ldots J$) can then be combined with the group-specific retention rates, $r_{s1}, \ldots, r_{sJ}$, and the proportional representation values, $q_1, \ldots q_J$, to derive the average criterion score of the stage $s$ selected applicants, $\overline{C}^*_s$.

$$\overline{C}_s^* = \sum_j q_j r_{sj} \overline{C}_{sj}^* / r_s.$$

Finally, the standardized average criterion score of the stage $s$ selected applicants, $\overline{C}_s^{*(s)}$, can be computed as

$$\overline{C}_s^{*(s)} = \frac{\overline{C}_s^* - \overline{C}^*}{\sigma_{C^*}},$$

where $\overline{C}^* = \Sigma_j q_j d_{Cj^*}$ and $\sigma_{C^*} = \sqrt{1 + \Sigma_j q_j (d_{Cj^*} - \overline{C}^*)^2}$ correspond to the average and the standard deviation of the composite criterion $C^*$ in the total applicant population.

With some minor modifications, part of the above procedure can also be applied to calculate the mean subgroup differences, after previous screenings, of the predictors and the predictor composites. Thus, the mean difference after stage $s$ on, for example, predictor $X_r$ between the retained part of minority group $j$ and the retained part of the majority group $J$ is obtained as $X_{rsj} - X_{rsJ}$, where, for example,

$$\overline{X}_{rsj} = d_{X_r j} + \frac{\sum_l^s R_{P_l^* X_r} \phi(p_{clj}^*) \Phi(f_{lj}^{(1)}, \ldots, f_{lj}^{(l-1)}, f_{lj}^{(l+1)}, \ldots, f_{lj}^{(s)}; \mathbf{R}_{\mathbf{P}^*}^{[l,s]})}{\Phi(p_{c1j}^*, \ldots, p_{csj}^*; \mathbf{R}_{\mathbf{P}^*}^{(s)})}.$$