Psychological Methods
2008, Vol. 13, No. 2, 75–98

1082-989X/08/$12.00 DOI: 10.1037/1082-989X.13.2.75

Measurement Invariance Versus Selection Invariance: Is Fair Selection Possible?

Denny Borsboom
University of Amsterdam

Jan-Willem Romeijn
Groningen University

Jelte M. Wicherts
University of Amsterdam

This article shows that measurement invariance (defined in terms of an invariant measurement model in different groups) is generally inconsistent with selection invariance (defined in terms of equal sensitivity and specificity across groups). In particular, when a unidimensional measurement instrument is used and group differences are present in the location but not in the variance of the latent distribution, sensitivity and positive predictive value will be higher in the group at the higher end of the latent dimension, whereas specificity and negative predictive value will be higher in the group at the lower end of the latent dimension. When latent variances are unequal, the differences in these quantities depend on the size of group differences in variances relative to the size of group differences in means. The effect originates as a special case of Simpson's paradox, which arises because the observed score distribution is collapsed into an accept–reject dichotomy. Simulations show the effect can be substantial in realistic situations. It is suggested that the effect may be partly responsible for overprediction in minority groups as typically found in empirical studies on differential academic performance. A methodological solution to the problem is suggested, and social policy implications are discussed.

Keywords: measurement invariance, selection invariance, fairness, testing, psychometrics

The extent to which our society is dominated and structured by psychometric selection processes can hardly be overestimated. College admission tests are used to determine who is admitted to university education, IQ tests to choose between job applicants, and diagnostic tests to determine what treatment one will receive upon experiencing psychological problems. Such selection procedures affect people's lives directly and profoundly. For many people, psychometric testing procedures may be among the most significant encounters with applied psychology that they will ever have.

Psychometric tests, however, are fallible instruments. Any decision based on a psychometric test is the result of a probabilistic inference, which implies that there always remains a chance that one has made the wrong decision: John was admitted into college, although he does not have the capacity to successfully complete his education, and Mary was diagnosed with attention-deficit hyperactivity disorder, although she does not actually suffer from this condition. Because such incorrect decisions may have an adverse impact on people's lives, it is important to gauge the probabilities with which such errors are made and, if possible, to control them in such a way that every tested person has a fair chance of a correct decision. Although it may be impossible to avoid incorrect decisions completely, it may nevertheless be possible to construct tests in such a fashion that they are unbiased, that is, that they do not result in discrimination against particular groups.

How does one establish that test scores are unbiased? The selection literature has long been dominated by conceptualizations of lack of bias that emphasize invariant predictive

Denny Borsboom and Jelte M. Wicherts, Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands; Jan-Willem Romeijn, Department of Philosophy, Groningen University, Groningen, The Netherlands.

Denny Borsboom's research was sponsored by NWO Innovational Research Grant 451-03-068. Jan-Willem Romeijn's research was sponsored by NWO Innovational Research Grant 275-20-013. Jelte M. Wicherts's research was sponsored by NWO Innovational Research Grant 451-07-016. We would like to thank Conor Dolan for his comments on an earlier draft of this article and Abe Hofman for his help in preparing the article.

Correspondence concerning this article should be addressed to Denny Borsboom, Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: d.borsboom@uva.nl

models; in fact, this property was used in the highly influential definition of test bias by Cleary (1968). This definition says that test scores are unbiased if the regression of some criterion (e.g., success on the job or educational achievement) on the test scores (e.g., IQ scores or SAT scores) is equal for different groups (e.g., for males and females or for different ethnic groups); bias occurs when there are differences in the regression function. This definition of bias is espoused in influential sources such as the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 79) and the *Principles for Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology, 2003, pp. 31–34).

However, contemporary psychometric work has delivered a very serious competitor to prediction invariance as the basis for definitions of test bias, namely, measurement invariance. Measurement invariance holds when not the relation between test and criterion but the relation between test scores and the latent attribute that these test scores measure is the same for different groups (Holland & Wainer, 1993; Mellenbergh, 1989; Meredith, 1993; Millsap & Everson, 1993; Millsap & Kwok, 2004). A test that fulfills the requirement of measurement invariance measures the same attribute in the same way in different subpopulations. Mathematically, the requirement of measurement invariance means that the function that relates psychological abilities to test scores is invariant over groups (Mellenbergh, 1989; Meredith, 1993). Therefore, measurement invariance ensures, conditional on the measured attribute, that individuals from different groups have the same test score distribution and hence the same probability of being admitted or rejected in a selection procedure.

It is interesting that the notions of predictive invariance and measurement invariance seem to have coexisted as a basis for defining test bias for quite a long time during the 1980s and 1990s, presumably because many people thought that they were conceptualizations of the same property (or nearly so). This misconception may have been fueled by two incorrect ideas about measurement invariance and prediction invariance: first, that criterion scores Y can be taken to be identical to the attribute θ measured by the test scores X; second, that if the regression of Y on X is the same for two groups, then the regression of X on Y is also the same. To one who holds these ideas, it will seem that prediction invariance and measurement invariance are really the same property: For if one believes that $f(Y \mid X) = f(Y \mid X,G)$ implies that $f(X \mid Y) = f(X \mid Y,G)$, and one also believes that $Y = \theta$, then one will think that $f(Y \mid X) = f(Y \mid X,G)$ (predictive invariance) implies $f(X \mid \theta) = f(X \mid \theta,G)$ (measurement invariance). Even though the difference between predictor–criterion and criterion–predictor regressions is, of course, a basic fact of statistics and has been recognized in the testing literature (e.g., Darlington, 1971), such a line of reasoning, or a very similar one, must be prevalent among psychologists; otherwise, it is hard to reconcile the fact that invariant test–criterion regressions continue to be adduced as evidence for measurement invariance (e.g., Aguinis & Smith, 2007; Evers, Te Nijenhuis, & Van der Flier, 2005; Gamliel & Cahan, 2007; Hunter & Schmidt, 2000; Rushton & Jensen, 2005).

This line of reasoning is incorrect, however. Specifically, Millsap (1997, 2007) has shown, in the context of the common factor model, that measurement invariance is inconsistent with prediction invariance; with the exclusion of some esoteric situations, satisfaction of measurement invariance implies a violation of prediction invariance and vice versa. Millsap's (1997) work therefore forces a choice between conceptualizations of bias in terms of prediction invariance or in terms of measurement invariance. One simply cannot hold on to the position that in general test scores have to be both measurement invariant and prediction invariant, for this position is logically inconsistent.

When faced with the choice between the properties of measurement invariance and prediction invariance in defining bias, we think that psychologists should favor measurement invariance. There are several reasons for this. First, as the work of Millsap (1997; see also Millsap, 2007) has shown, prediction invariance implies a violation of measurement invariance when groups differ on the attribute measured. Thus, when the regression of, say, job success on IQ scores is equal for two groups, but the groups are not equal in the distribution on the latent variable measured by the IQ scores (i.e., intelligence), then measurement invariance must be violated. This means that should one choose to equate lack of bias with prediction invariance, then one implicitly embraces the unreasonable position that, for instance, IQ tests on which people with the same level of intelligence have different expected IQ scores according to their ethnic background are unbiased. Thus, measurement invariance is consistent with the item response theory (IRT) approach of differential item functioning (DIF) detection, whereas predictive invariance is not. As an illustration, when groups differ in latent means but not in latent variances, one way to make a test unbiased in the predictive invariance sense is to add items that are uniformly biased (Mellenbergh, 1989) with respect to the lower scoring group (Millsap, 2007). This is clearly an undesirable consequence of a conceptualization of bias in terms of predictive relations.

Second, predictive invariance is itself a conceptually and temporally ambiguous concept. Predictive relations are relations between the test score and some criterion variable, which is normally temporally separated from the test score. This means that predictive invariance requires a choice of (a) which variable will function as the criterion and (b) at

what time point this variable will be measured. Both choices are obviously important determinants of the correlations one will find and hence will influence whether prediction invariance is satisfied, yet they are to a significant extent arbitrary. Definitions of bias in terms of measurement invariance do not suffer from such problems of ambiguity, because the relevant inferences concern the attribute measured by the test at the time of testing rather than some future state of affairs.

Third, further complications arise when test bias is defined on the basis of predictive invariance because that definition does not involve a causal measurement model. For instance, a common finding in differential prediction studies using cognitive ability tests is that criterion performance of lower scoring minority groups is overpredicted when the regression of the higher scoring majority group is used (Sackett & Wilk, 1994). Hence, minority groups perform worse on the criterion than is to be expected from their average test scores. This result is often proffered in support of the claim that these tests are biased against the majority group (e.g., Campbell, 1996; Hunter, Schmidt, & Rauschenberger, 1984; Sackett & Wilk, 1994). However, an entirely different picture might emerge if one were to reverse the regression by predicting the test scores from criterion performance (as an indicator of true ability) using the same data. Specifically, it is quite possible that predicted test scores of minority group members with a particular level of criterion performance might be lower than the predicted test scores of majority members with the same level of criterion performance (Birnbaum, 1979). But this implies the opposite result: The test is biased against the minority group. Such inconsistencies do not arise when test bias is defined in terms of a measurement model that dictates the causal direction of effects (e.g., test scores are the result of latent ability and not the other way around).

Thus, given that one has to choose between prediction invariance and measurement invariance when defining the concept of test bias, measurement invariance has better cards. Now the question arises as to whether one can connect the broader concept of fairness to the psychometric concept of test bias in an unambiguous manner, so that one could define *fair test use* as the use of tests that yield measurement-invariant test scores. Such a definition, if plausible, clearly would be very useful. However, in the present article we will argue that such a move presents serious problems and that the concept of fairness cannot be unambiguously attached to the concept of measurement invariance. We argue that even though the literature on measurement invariance must be considered a major stride forward in psychometric theorizing on test bias, it does not provide an unambiguous definition of fair test use. In fact, we doubt whether such a definition can be given at all.

The reason for this is that from a truly fair selection procedure, one should expect not only that people with the same ability have the same chance of being selected, regardless of their group membership, but also that the selection procedure works equally well for each group so that the test's accuracy is invariant; that is, its sensitivity, specificity, positive predictive value, and negative predictive value are equal for each group. Call this property *selection invariance*. Now, intuitively, one may expect that measurement invariance and selection invariance can coexist; several colleagues whom we probed for their intuition on this point in fact thought that the former implied the latter. This is because measurement invariance ensures that in every subpopulation of people with equal ability the proportion of incorrect decisions will be identical across groups, so in these subpopulations selection invariance holds. Hence, one may expect that selection invariance also holds for the proportion of incorrect decisions in the intact populations, because these can be conceptualized as the union of the subpopulations in question. Also, Millsap and Kwok (2004) evaluated the effect of violations of measurement invariance for differential test accuracy across groups and found that the more measurement invariance was violated, the greater were the differences in test accuracy. On the basis of these findings, one might be tempted to conclude that in the situation where measurement invariance does hold, no differences in test accuracy would occur.

This intuition, however, is incorrect. This article shows that for a very general class of situations in which latent differences between groups exist, the properties of measurement invariance and selection invariance are inconsistent. This somewhat counterintuitive result arises as a special case of Simpson's (1951) paradox. Measurement invariance is a property that applies to the expected test score for individuals of equal ability, whereas test accuracy is a property that applies to subpopulations that are aggregated over ability levels. In the process of aggregation, the independence of group membership and the probability of passing the test is destroyed, which in turn causes violations of selection invariance. On the other hand, if selection invariance is to be satisfied, the selection procedure must use different cut points for different groups, which is widely perceived as unfair (if not simply illegal). In sum, in many cases in which groups differ in the measured ability, a test cannot exhibit both measurement invariance and selection invariance unless different cut points are used, which compromises fairness. Because we focus on the situation in which measurement invariance holds, this article can be viewed as a companion article to Millsap and Kwok (2004), who investigated essentially the same situation for the case in which measurement invariance is violated.

This article is structured as follows. First, we discuss the relation between measurement invariance and selection invariance in greater detail and provide proof of the incompatibility between measurement invariance and selection invariance. Second, we give a conceptual explanation of the

statistical mechanism that produces the inconsistency. Third, we present simulation studies to estimate the severity of the problem in realistic scenarios. We close the article by considering the theoretical, practical, and societal implications of this work.

Measurement Invariance and Selection Invariance

The basic idea of psychometric selection is to optimally select persons for differential treatment on the basis of their test scores. This may involve college admissions, where one aims to select the brightest candidates; personnel selection, where one aims to select the candidates who have the skills required for the job; or health care, where one has to decide whether a person is eligible for treatment. In all these contexts, invariance properties of the selection procedures across groups are crucially important, especially when groups vary in the properties assessed; these concerns may vary from sex differences in college admission test scores to ethnic differences in diagnostic test scores that relate to eligibility for medical treatment. We identify the persons that the selection procedure should select as *suitable* persons and persons that it should reject as *unsuitable*.

Throughout this article, we assume that individual differences in suitability can be represented as a single continuous latent variable θ. Define a suitable candidate as an individual i with $\theta_i \geq \theta_c$, where θ_c is the latent cutoff that separates individuals whom one wants to reject from those whom one wants to select. Because the variable θ is not directly observable, people are selected on the basis of a test score that is viewed as an observable indicator of θ; denote this test score X. We will assume that the latent variable is distributed according to the normal probability density function, denoted $f(.)$, with mean μ_g and standard deviation σ_g in group $G = g$, so that (1)

$$p(\theta) = f_{\mu_g,\sigma_g}(\theta). \tag{1}$$

Further, we assume that the test scores X are linearly related to the abilities θ according to the regression line $E(X \mid \theta = \tau_g + \lambda_g\theta)$, so that

$$X = \tau_g + \lambda_g\theta + \varepsilon_g, \tag{2}$$

where $\lambda_g > 0$ is the regression coefficient for group g, τ_g is the intercept, and ε_g denotes the residual or error score. Finally, we assume that the errors ε are normally distributed with variance σ_ε^2 constant across levels of θ (i.e., homoscedasticity).

These assumptions are consistent with the single-factor model for continuous item responses (Mellenbergh, 1994). However, when people are selected on the basis of a total test score composed of dichotomous items, these assumptions will never be exactly met, because the total test score is bounded. Nevertheless, provided that the number of items is reasonably large and that the item difficulties are well spread along the θ scale, the relation between the total score and the latent variable will approximate linearity. Similarly, we conjecture that a total test score constructed from Likert items will, in many situations, approximate the assumptions made, although this also depends on the distribution of item parameters over the θ scale and on the number of items used. Thus, because in these situations the present assumptions are approximated at best and because the proofs in this article do assume strict linearity, the exact extent to which they apply to any concrete alternative situations should be studied on a case-by-case basis; this may be done through analytical methods or through simulations.

Given the assumptions stated, the test score of an individual in group g with a certain true position θ is a random draw from the probability density function:

$$p(X|\theta) = f_{\tau_g+\lambda_g\theta,\sigma_{\varepsilon_g}}(X). \tag{3}$$

The selection procedure operates on the bivariate distribution of X and θ, which for a single group with intercept $\tau_g = 0$ may be represented graphically as in Figure 1. The objective of a selection procedure is to select those individuals whose ability θ exceeds the threshold $\theta = \theta_c$. Throughout this article we use the convention that $\theta_c = 0$. Note that this is not a restrictive choice because we are free to choose the zero point of the latent variable. Individuals whose ability exceeds the threshold $\theta = 0$ are called *suitable*, denoted S, others are *unsuitable*, denoted $\neg S$. A standard selection procedure based on an ability test selects those individuals whose test score exceeds a certain threshold value X_c. Individuals whose test score exceeds the value X_c and hence pass the test are called *accepted*, denoted A, others are *rejected*, denoted $\neg A$.

In any selection procedure, it may happen that an individual who is not suitable according to the attribute measured by the test, having $\theta < 0$, nevertheless passes the test,

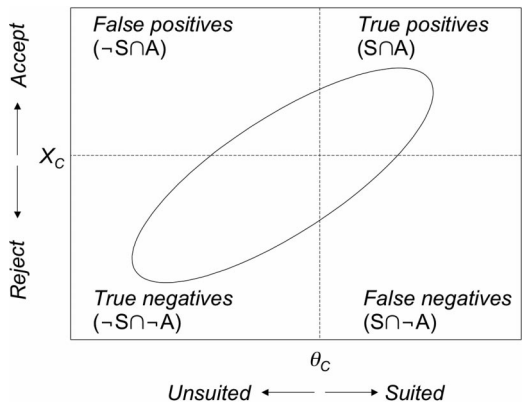


Figure 1. The selection problem. The figure shows the distribution of candidates on the selection variable X as a function of the latent ability θ.

$X \geq X_c$; that is, the person passes the test because of factors different from the attribute measured (e.g., luck, test-specific preparation, having an exceptionally good day, etc.). In that case, the individual is accepted yet unsuitable: $A \cap \neg S$. Similarly, an individual who is suitable, $\theta \geq 0$, may fail the test, $X < X_c$. The individual is then suitable yet rejected: $S \cap \neg A$. In the simplified situation of Figure 1, in which $\tau_g = 0$ and $X_c = 0$, the four quadrants are thus occupied by four combinations of suitable and unsuitable accepted and rejected individuals (true positives, false positives, true negatives, and false negatives). The accuracy of a selection procedure is determined by the distribution of individuals over these quadrants; obviously, one desires as many people as possible in the true positive and true negative quadrants.

The Two-Group Scenario

The primary focus of this article is on a situation that involves the selection of individuals from two groups that differ in the distribution of the latent variable. This is the same scenario as that considered by Millsap and Kwok (2004). In this section we consider the situation in which the variance of the latent distributions are equal but the means differ. The assumption of equal variances will be relaxed later in this article.

Denote the group with the higher mean as H and the group with the lower mean as L. In the context of fair selection, it is generally deemed important that, regardless of group differences in the distribution of θ, the relation between the test scores and the latent variable be invariant across groups. If this is not the case, then individuals from different groups, who have the same position on θ, have different expected test scores. This should be considered unfair, because it implies that, given a specific level of ability, group membership still influences the selection procedure. The requirement that the relation between observed scores and latent ability is invariant across groups is known as *measurement invariance*. Formally, measurement invariance is satisfied if and only if, conditional on θ, the probability distribution or density of the observed scores does not depend on group membership. Thus, measurement invariance is defined by the restriction

$$p(X \mid \theta) = p(X \mid \theta \cap g). \tag{4}$$

This requirement is identical to the conditions formulated by Mellenbergh (1989) and Meredith (1993). Further, given the fact that the probability function p has been assumed to be linear and the population distributions normal, the requirement of measurement invariance (Equation 4) comes down to the equality of τ_g, λ_g, and $\sigma_{\varepsilon g}$ across the levels of the grouping variable $G = \{L,H\}$.

Under the assumption of measurement invariance, we can introduce one further convention regarding test scores. By measurement invariance we have $\tau_L = \tau_H$. Throughout this article we will fix $\tau_L = \tau_H = X_c = 0$. Like the fixation of $\theta_c = 0$, this assumption does not lead to any loss of generality in our characterization of the selection procedure, because we are free to choose the scaling of the scores. So, if measurement invariance is satisfied and if this further convention is respected, Figure 1 is an adequate representation of the selection situation, and the situation for a two-group scenario may be depicted as in Figure 2. The bivariate distributions of X and θ lie on the same regression line.

Measurement invariance implies that the distribution of the test score, conditional on a given value of the latent variable, is invariant across groups. This, in turn, means that the probability of being accepted, determined by passing the test, does not vary among members of the different groups G conditional on their level of ability:

$$p(A \mid \theta) = p(A \mid \theta \cap g). \tag{5}$$

Equation 5 implies that for every level of ability, the false positive and false negative rates are exactly the same in both groups (see Appendix A for a proof). For this reason, it is tempting to think that measurement invariance entails equal sensitivity and specificity in each group. The next section shows that this is incorrect.

Measurement Invariance and Selection Invariance

It would seem that a fair selection procedure should be characterized by an identical number of selection errors across groups; that is, it should distribute its errors evenly. This requirement implies that the test's positive predictive value (the probability of suitability, given acceptance), negative predictive value (the probability of nonsuitability, given rejection), sensitivity (the probability of acceptance, given suitability), and specificity (the probability of rejection, given nonsuitability) be equal across groups. If these probabilities are equal across groups, then the test works equally well in each group; that is, it has the same accuracy.

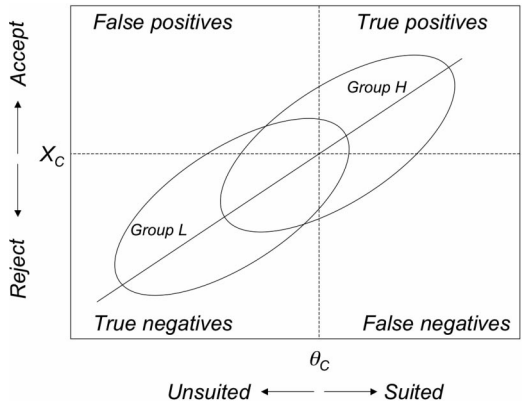


Figure 2. Selection in two groups H and L under measurement invariance.

This is what we call *selection invariance*. Formally, it is expressed by the following equalities:

Invariant positive predictive value:

$$p(S \mid A \cap g) = p(S \mid \mathrm{A}) \tag{6}$$

Invariant negative predictive value:

$$p(\neg S \mid \neg A \cap g) = p(\neg S \mid \neg A) \tag{7}$$

Invariant sensitivity: $p(A \mid S \cap g) = p(A \mid S)$ (8)

Invariant specificity: $p(\neg A \mid \neg S \cap g) = p(\neg A \mid \neg S)$ (9)

Analogous to the terminology of the theory of measurement invariance, we define *selection bias* as the violation of one or more of these equalities.

In the literature on fairness, similar quadrants have been defined in terms of the relation between test scores and a criterion (Cleary, 1968; Cole, 1973; Darlington, 1971; Hunter et al., 1984; Petersen & Novick, 1976; Thorndike, 1971). In this literature, Darlington (1971) and Petersen and Novick (1976) showed that several models defined in terms of invariance of (combinations of) the probabilities associated with the quadrants in Figure 2 were mutually inconsistent and inconsistent with the invariance of the regression of the criterion on the test score. Here, we show that an analogous inconsistency occurs when selection invariance and measurement invariance are both required to hold while groups differ in the distribution of θ. The main difference between the present analysis and the relevant 1970s literature is that we proceed from a latent variable perspective and define both measurement invariance (as per Mellenbergh, 1989; Meredith, 1993) and selection invariance from this perspective. Portions of our results can, however, be statistically translated into those of the 1970s literature under appropriate substitution of variables and functions relating them. For instance, if one substitutes the predictor for the latent variable, substitutes the criterion for the test score, assumes a common and homoscedastic regression function of the criterion on the test score across groups, and assumes equal variances in both test score and criterion, analogues to the statistical inequalities presented below can be found in the 1970s literature (e.g., under these substitutions Equation 11 specifies the violation of what Petersen and Novick called the *equal probability model*, given that what they call the *regression model* is true). In practical situations, however, the model we use is drastically different from the predictor–criterion model because (a) the predictor–criterion model does not include a measurement model and (b) the test score plays the role of independent variable in the 1970s literature, whereas in the latent variable approach it plays the role of dependent variable. Also, due to the complicated relation between measurement invariance and prediction invariance (Millsap, 1997), it is not straightforward to analyze the relation between the situation as viewed from a measurement perspective (the current analysis) and as viewed from a predictive perspective (the analysis as for instance presented in Darlington, 1971, and Petersen & Novick, 1976). We return to the relation between these approaches in the discussion.

To see that measurement invariance and selection invariance are mutually inconsistent properties in case the locations of the latent distributions differ, first consider the positive predictive value, that is, the probability of suitability given acceptance, as represented in Equation 6. As a function of the continuous distributions assumed for the test scores and the ability, the relevant expression for the probability in question is as follows:

$$p(S|A \cap g) = \frac{p(S \cap A|g)}{p(A|g)} = \frac{p(\theta \geq 0 \text{ and } X \geq 0|g)}{p(X \geq 0|g)}. \tag{10}$$

Both the numerator and the denominator in this equation are integral expressions involving both the distribution of group g over the ability parameter θ and the distributions over test scores given an ability. These integrals cannot be solved analytically. However, it can be shown that their quotient is a monotonically increasing function of μ_g, the mean of the distribution of group g over the ability: The higher the mean of the group's distribution function over θ, the higher the probability of suitability given acceptance. For two groups H and L we therefore obtain this:

$$\mu_H > \mu_L \Rightarrow p(S \mid A \cap H) > p(S \mid A \cap L). \tag{11}$$

The details of the proof are in Appendix A. This shows that if measurement invariance holds, then the positive predictive value necessarily varies over the groups considered: In the group with higher mean ability, the positive predictive value will be higher than in the group with lower mean ability.

We may wonder whether a similar result holds for test sensitivity, that is, the probability of acceptance given suitability:

$$p(A|S \cap g) = \frac{p(S \cap A|g)}{p(S|g)} = \frac{p(\theta \geq 0 \text{ and } X \geq 0|g)}{p(\theta \geq 0|g)}. \tag{12}$$

This is indeed so, as can be derived in roughly the same way. We have already established that $p(S \mid A \cap g)$ gets larger with larger μ_g. Therefore, to establish that $p(A \mid S \cap g)$ gets larger with larger μ_g, it suffices to show that the problem discussed above is symmetric with respect to interchanging the variables X and θ. Appendix B establishes that this is indeed the case. This means that the inequality derived in Appendix A (i.e., Equation 11) is invariant under substitution of θ for X and of X for θ.

Thus, with respect to the probability of acceptance given suitability, the same relation holds as was established for the probability of suitability given acceptance. Namely,

$$\mu_H > \mu_L \Rightarrow p(A \mid S \cap H) > p(A \mid S \cap L). \quad (13)$$

The sensitivity of the testing procedure will be better for the group with higher mean ability.

The inverse results hold for negative predictive value and specificity because they are the mirror images of positive predictive value and sensitivity. Hence, we further have

$$\mu_H > \mu_L \Rightarrow p(\neg S \mid \neg A \cap H) < p(\neg S \mid \neg A \cap L), \quad (14)$$

and

$$\mu_H > \mu_L \Rightarrow p(\neg A \mid \neg S \cap H) < p(\neg A \mid \neg S \cap L). \quad (15)$$

The specificity and negative predictive value of the testing procedure will be better for the group with lower mean ability.

The following explanation of the results, as well as the proofs that underlie them, may be instructive. Examine Figure 2 and consider the probability mass located in the true positives quadrant. This is roughly the surface of the ellipse in that quadrant. Compare this surface to the entire surface above the line $X = X_c$. The ratio of these two quantities is the positive predictive value. Appendix A proves that this ratio is always higher for group H, as can be expected from inspecting the figure. Now rotate Figure 2 counterclockwise 90° (this is the symmetry argument of Appendix B), and compare the ratio of probability mass located in the true positives quadrant to that above the line $\theta = \theta_c$. This ratio is test sensitivity, and clearly it is higher for group H as well.

Why is this true in general? Imagine that the joint distribution moves up along the regression line (i.e., in the direction of higher θ and X). One can sense that sensitivity and positive predictive value will both get higher; they are monotonically increasing in θ. Therefore, because group L and group H differ only in their location along the regression line, moving group L up that line until it coincides with group H will result in higher sensitivity as well as higher positive predictive value. Hence these quantities are always higher for the group with higher mean ability. If one turns the figure counterclockwise another 90°, one can do the same for negative predictive value, and turning it a last 90° for specificity. In this orientation, moving H in the direction of L can be seen to lead to a higher specificity and negative predictive value. Hence these quantities are always higher for the group with lower mean ability. For the current distributional assumptions, Appendixes A and B show that these results are generally true; for further background to the proofs see also Kotz, Balakrishnan, and Johnson (2000).

This shows that, in a selection procedure that conforms to the present assumptions of (a) bivariate normality, (b) equal variances of the ability distributions across groups, and (c) measurement invariance, we get the following result: In the group with higher mean ability, the procedure will have a higher sensitivity and positive predictive value, whereas in the group with lower mean ability, it will have a higher specificity and negative predictive value. Under the assumptions stated, it is impossible to simultaneously satisfy measurement invariance and selection invariance. The test does better in accepting members from the group with higher mean ability, and it does better in rejecting members from the group with lower mean ability.

Generalization of the Impossibility Result

In real-world situations, groups may differ both in mean ability and in variance. However, the foregoing conclusion was based on the assumption that the variances of the latent variable distributions are equal across groups. When this assumption is relaxed, the impossibility result may no longer hold in all situations; whether simultaneous satisfaction of measurement and selection invariance is possible depends on how the means and variances differ over groups. In this section, we describe how the relation between measurement invariance and selection invariance behaves as a function of these differences.

The situation where variances differ is rather more complicated than the case of equal variances, and we therefore discuss the full analysis of this case in Appendixes C and D. Appendix C gives an explanation of the situation, whereas Appendix D contains the proof on which that explanation is based. Here we focus on the impossibility results themselves. For two groups with μ,σ and μ',σ', respectively, we assume without loss of generality that $\sigma' < \sigma$. The results then fall into two categories. First,

$$\frac{\sigma'}{\sigma}\mu \geq \mu' \Rightarrow \begin{cases} p(S|A \cap g_{\mu,\sigma}) > p(S|A \cap g_{\mu',\sigma'}), \\ p(A|S \cap g_{\mu,\sigma}) > p(A|S \cap g_{\mu',\sigma'}). \end{cases} \quad (16)$$

This result is proved in detail in the appendixes. Second,

$$\frac{\sigma'}{\sigma}\mu \leq \mu' \Rightarrow \begin{cases} p(\neg S|\neg A \cap g_{\mu,\sigma}) < p(\neg S|\neg A \cap g_{\mu',\sigma'}), \\ p(\neg A|\neg S \cap g_{\mu,\sigma}) < p(\neg A|\neg S \cap g_{\mu',\sigma'}). \end{cases} \quad (17)$$

As explained in Appendix C, these latter inequalities are basically the same as the former. We employ the same equations in deriving them but we transform θ into $-\theta$.

These results widen the scope of the inequalities derived in the previous section; if we choose $\sigma = \sigma'$, we obtain the original inequalities. Equation 16 says that if in the group with the larger variance we transform the latent scale via $\theta' = \sigma'\theta/\sigma$ and the resulting mean of θ' lies above the untransformed mean of the other group, then sensitivity and positive predic-

tive value will be larger in the group with the larger variance.[1] Equation 17 implies that if this transformed mean lies below the untransformed mean of the other group, then specificity and negative predictive value will be larger in the group with the smaller variance. Thus we can either prove the inequality of sensitivity and positive predictive value or the inequality of specificity and negative predictive value.

This means that measurement invariance and selection invariance are inconsistent in all situations, because at least one of the pairs of sensitivity and specificity or positive and negative predictive value will not be invariant. Note that this does not mean that for any combination of values of $\frac{\sigma'}{\sigma}$, μ, and μ' only one of the two invariances will be violated. Our conditions are sufficient for proving the inequalities at issue but not necessary for the inequalities themselves, so the failure of a condition does not entail the failure of the inequalities. If for example $\sigma'/\sigma \ll 1$ while $\mu\sigma'/\sigma < \mu'$, the inequalities for sensitivity and positive predictive value still hold. More generally, we conjecture that both the invariance of sensitivity and specificity and the invariance of positive and negative predictive value are violated almost everywhere in the space $\frac{\sigma'}{\sigma} \times \mu \times \mu'$, except for on two planes where only one of the two invariances holds. These planes intersect along the line where $\frac{\sigma'}{\sigma} = 1$ and $\mu = \mu'$, that is, the situation where the latent distributions of the groups are identical. Unfortunately we do not have a proof of this conjecture. However, for present purposes it is sufficient that we have proved that selection invariance, in the general sense, will be violated if measurement invariance is satisfied.

Explaining the Paradox

We have shown that under a wide range of conditions, the satisfaction of measurement invariance entails the violation of selection invariance. The paradoxical nature of this result can be clearly seen when one compares Equation 5 to Equation 13: Although it is true, for any two subpopulations in H and L with equal θ, that the probability of passing the test given their level of ability is identical (per measurement invariance); for the intact populations H and L, which are the unions of these subpopulations, the probability of passing the test, given suitability, is different (selection bias). This section aims to offer a conceptual explanation of the mechanism that produces these inconsistencies.

First, consider test sensitivity. The discrepancy in test sensitivity can be understood as the result of a continuous version of Simpson's paradox (Simpson, 1951; Wainer & Brown, 2004; Yule, 1903). Simpson's paradox describes how effects that are observed in subpopulations can differ dramatically from effects observed in the aggregate of the subpopulations. Consider the case in which we follow 50 women and 50 men in their search for a job in either the fire or the police department. Say that the police department is looking to hire 40 people, whereas the fire department is looking to hire only 15, and further imagine that at the fire department 10 women and 40 men apply, whereas at the police department 40 women and 10 men apply. Now assume that for police and firefighters alike, women and men have an equal success rates, or probability of getting a job, namely .8 for the police and .3 for the firefighters. So the probability of getting a job is independent of gender. Nevertheless, if we do the math we find that $(.8 \times 40) + (.3 \times 10) = 35$ women will find a job, whereas only $(.8 \times 10) + (.3 \times 40) = 20$ men will. It looks as if on the aggregate level of government services, gender and job opportunity are not independent and that men are being treated unfairly. In the subpopulations of the police and the firefighters, however, there is perfect independence, and hence no unfairness at all. The apparent dependence results from the fact that gender is not independent of department: Many more men decide to apply at the fire department whereas the firefighters offer fewer jobs.

The present case is structurally similar to the above case. By measurement invariance, we have independence of the probability of incorrect decisions and group membership in all subpopulations with equal θ: For each value of θ the probabilities of false rejection and false acceptance are the same for the different subpopulations L and H. So the values of θ in the selection setting are analogous to the separate departments of the above example. But we find that aggregating over different values of θ destroys the independence of false rejection and group membership, in the same way as aggregating over different departments destroyed the independence of job opportunity and gender in the above example. To see how this works in detail, note that most incorrect selection decisions would occur around the cutoff score $X = 0$. The Gaussian on the left side of Figure 3 shows the probability of a false rejection for a point of θ that is associated with an expected test score slightly above the cutoff. As can be seen, this chance is considerable, although it will always be less than .50 for values of $\theta > 0$. The point is that with increasing

[1] To see the exact reach of this result, consider the conditions $\sigma' < \sigma$ and $\frac{\sigma'}{\sigma}\mu > \mu'$. The condition $\sigma' < \sigma$ can be fulfilled without loss of generality. For the condition $\frac{\sigma'}{\sigma}\mu > \mu'$, it is useful to distinguish between cases in which $\mu > \mu'$ and cases in which $\mu < \mu'$. In the case $\mu > \mu'$, the condition is met unless both μ and μ' lie far enough above the threshold $\theta = 0$ so that $\frac{\sigma'}{\sigma} > \frac{\mu'}{\mu}$. In the case $\mu < \mu'$, however, the condition is met only if μ' and μ lie far enough below the threshold $\theta = 0$ so that $\frac{\sigma'}{\sigma} < \frac{\mu'}{\mu}$.

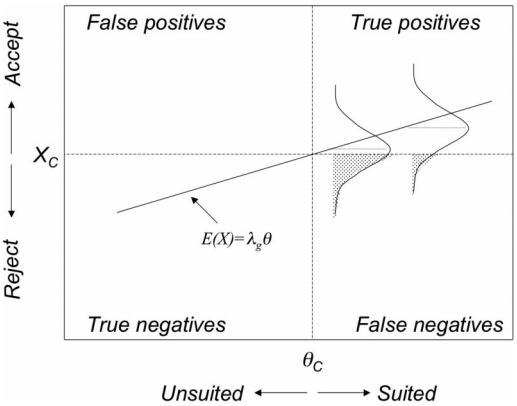


Figure 3. Error distributions for points close to and far from the cutoff score. The probability of selection errors decreases as the distance from the cutoff increases.

ability, the chance of such an error becomes increasingly small. This can be seen from the Gaussian on the right.

Now, the reason for the existence of group differences in sensitivity lies in the fact that in the population of suitable candidates (i.e., candidates with $\theta > 0$), there are proportionally more suitable members from population *L* than there are members from population *H* in the direct neighborhood of the cutoff score: The group of suitable members from *L* is distributed more toward the left of the figure than the group of suitable members from *H*.[2] A different way of viewing this is by noting that among suitable candidates, there are proportionally more members from *H* than members from *L* among those far above the cutoff.[3] As a result, incorrect rejections will occur more frequently in the subpopulation of suitable members from *L*, which lowers sensitivity in that population. The same argument, but now considered for the subpopulation of unsuitable candidates (i.e., candidates with $\theta \leq 0$) explains why specificity will be higher among members of *L*: There will be proportionally fewer members of *L* than members of *H* in the neighborhood of the cutoff score. Since the probability of the relevant false decision (i.e., incorrect acceptance) is highest for those who score in the neighborhood of $X = 0$, and there are proportionally more such candidates among unsuitable members of *H*, specificity will be lower for population *H*.

It is useful to give a simple numerical example to illustrate this process. We will consider just two points on the θ scale. Suppose that 80% of population *L* is located at the point $\theta = 1$ and 20% is located at $\theta = 2$ and that for population *H* this pattern is reversed: 20% is located at $\theta = 1$ and 80% at $\theta = 2$. Assume, as before, that the cutoff for suitability is at $\theta = 0$, so that all of the candidates are suitable (the only errors that we can make are false negatives). Suppose the common regression line that connects θ to the test score *X* is $E(X) = \theta$, with homoscedastic errors being distributed normally with unit variance. That this regression line is the same for both groups means that the test scores are measurement invariant with respect to the group variable. Then the conditional distribution of *X* given that $\theta = 1$ is $N(1,1)$, equal for candidates in *L* and in *H*; the conditional distribution of *X* given that $\theta = 2$ is $N(2,1)$, also equal for candidates in *L* and in *H*.

Now we select only candidates with scores $X > 0$. The associated probabilities of a false negative, as evaluated on the relevant normal distributions, are approximately $P(X \leq 0 \mid \theta = 1) = .16$ and $P(X \leq 0 \mid \theta = 2) = .02$, respectively; again these figures are equal for candidates in *L* and *H*. However, the probability of a false negative occurring in group *L* is not the same as in group *H*, due to the different distributions of θ in the groups. In group *L* the probability of a false negative equals $(.8 \times .16) + (.2 \times .02) = .13$, whereas in group *H* the probability of a false negative equals $(.2 \times .16) + (.8 \times .02) = .05$. Exactly as in Simpson's paradox, aggregating across the levels of θ creates a dependence between the group variable and the probability of false negatives. Given the assumptions of normality, linearity, and homoscedasticity employed in this article, integrating over infinitely many such subpopulations, instead of two, yields the impossibility results of the previous section.

To understand why groups differ in positive predictive value, that is, in the probability of suitability given acceptance, it is instructive to focus on the regression of θ on *X* (see also Appendix B, which formalizes this point). The following counterintuitive fact now emerges: Whereas the regression of test scores on latent trait scores is identical across groups (i.e., measurement invariance), the regression of latent trait scores on test scores is not. The regression of trait scores on test scores is not group-invariant because (due to mean group differences in ability) the intercept in the regression of the low-scoring group will be lower. Thus, although the expected value of *X* given θ is the same for members of different groups, the expected value of θ given a particular test score *X* is lower for those in the low-scoring

[2] Note that this is true in general only for the case where the variances in the groups are equal; if they are unequal, the mechanism is the same but the direction of the effect depends on the relation between means and variances, as derived in Appendixes C and D. For clarity of exposition, we restrict ourselves to the case of equal variances here.

[3] Note that the word *proportionally* is very important here. Clearly, the conclusion does not apply to absolute numbers of candidates but only to their number relative to the entire number of candidates within a group. If the sizes of the groups are different, there may be more candidates from the larger group in every region of the selection space. Similarly, even with equal group size, the absolute number of false decisions may go either way, depending on the position of the group's mean μ_g relative to the cutoff point. The inequalities are only concerned with the proportions of groups relative to being suitable, accepted, and so forth.

group. This counterintuitive effect was dubbed Kelley's paradox by Wainer (2000, 2005), because the effect follows from Kelley's (1927) classic formula for estimating ability scores from test scores. In contrast to the group differences in sensitivity and specificity discussed above, this effect also occurs without dichotomization of latent and observed variables in suited–unsuited and accepted–rejected categories.

In Kelley's (1927) formula, trait score estimation is a function of the test score and the mean ability score of the particular group, which is weighted by the test's unreliability. The effect is due to variance in the test scores that is unrelated to the latent trait (i.e., residual variance) and does not occur for the special case that θ and X are perfectly correlated. However, in case of a less-than-perfect relation between test scores and the latent trait scores, group differences in mean latent ability are necessarily underestimated by the mean differences in test score. This effect penalizes members of population L when ability is regressed on test scores. The key concept here is differential regression to the mean: The effect of regression to the mean will be stronger for population L, because in that group, selection has occurred proportionally more often on high scores that are not due to high ability (i.e., measurement errors).

Seriousness of Violations of Selection Invariance

We have shown that selection invariance and measurement invariance are inconsistent for almost all situations and have explained the mechanism that produces these inconsistencies. We now address the question how large these effects are for realistic selection scenarios.

The seriousness of violations of selection invariance, given measurement invariance, depends on the percentage of variance in test scores unrelated to the trait (i.e., unreliability), the value of the selection ratio, and the size of group differences in mean latent ability. To examine the seriousness of violations of selection invariance, we simulated data for a two-group scenario with a 1 *SD* difference in mean latent ability between groups and equal group size. We varied the selection ratio as the top 5% and top 25% (these percentages apply to the combined population). For each of these scenarios we examined the between-group differences in sensitivity, positive predictive value, specificity, and negative predictive value as a function of reliability. Reliability was defined as the ratio of true score variance to total variance (Lord & Novick, 1968) in the combined population; we varied this parameter by adding different amounts of white noise to the test scores, corresponding to different values of σ_ε in Equation 3. Correct acceptance was defined as acceptance of a candidate, given that the candidate occupies a position in the top s percent of the combined latent distribution, where s equals the selection ratio for the relevant scenario; the other relevant probabilities were computed analogously. The results of the simulations are graphically depicted in Figures 4 and 5.

The simulations show that, for the chosen parameter settings, the amount of selection bias implied by measurement invariance can be substantial. Differences in sensitivity are consistently between 5% and 10%; when less reliable tests are used for selection, differences in positive predictive value may be as large as 20%. The differences for specificity and negative predictive value are less pronounced and become serious only for less extreme selection ratios. As is to be expected from the theoretical work above, for all levels of reliability below 1.0, the probability of false positives is lower for the group with higher mean ability, whereas the converse is true for the probability of false negatives. In Figure 4, it can further be seen that the differences in test sensitivity are relatively stable across different levels of reliability, whereas the differences in positive predictive value increase rapidly as reliability drops. This differential effect of reliability originates as follows. Because the difference between groups in test sensitivity depends only on the distribution of θ, the difference remains constant over all reliability ranges (after an initial diver-

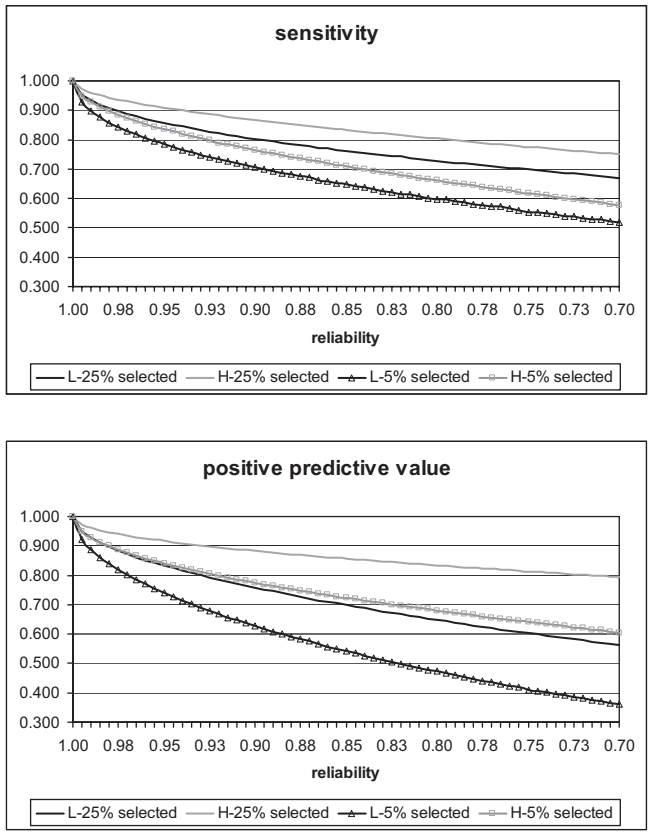
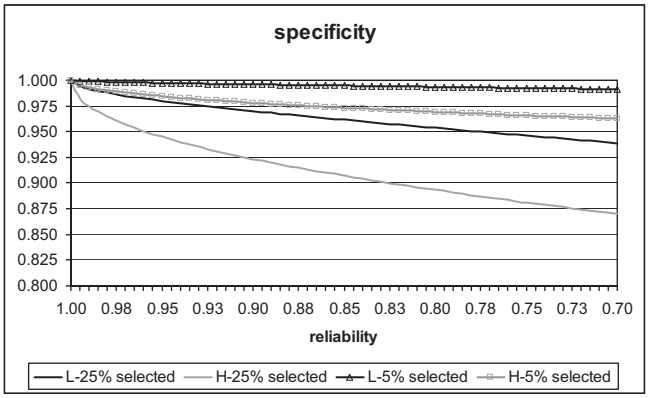


Figure 4. Sensitivity and positive predictive value as a function of reliability. The two dark lines represent group H (for selection ratios of 5% and 25%), and the two light gray lines represent group L (for the same selection ratios).



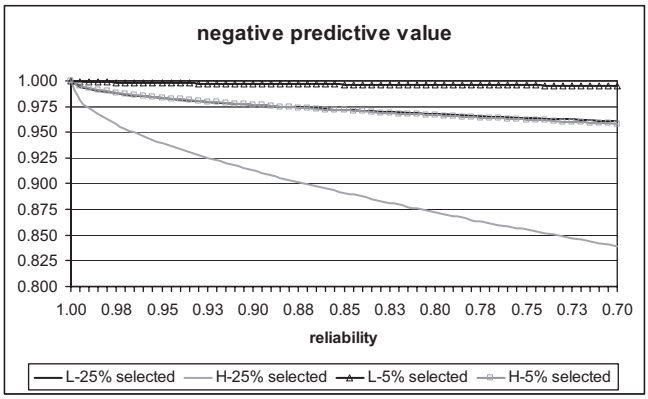


Figure 5. Specificity and negative predictive value as a function of reliability. The two dark lines represent group *H* (for selection ratios of 5% and 25%), and the two light gray lines represent group *L* (for the same selection ratios). Note that to improve readability, the *y*-axis has been scaled differently from that in Figure 4.

gence from the point of perfect reliability). On the other hand, the effect of differential regression toward the mean, which causes the difference in positive predictive value, increases in severity when reliability decreases and when selection becomes more extreme. This explains why the difference in positive predictive value increases for lower reliability as well as for more extreme selection ratios.

When compared with the results for sensitivity, differences in specificity depend on the extremity of the selection ratio in the opposite way: The difference in the correct rejection rates becomes larger as the selection ratio becomes less extreme. This is because as the selection ratio becomes less extreme, the cutoff score becomes lower. As a result, the cutoff score moves into the lower tail of the distribution of population with higher mean ability; hence there are proportionally more unsuitable members of that population directly below the latent cutoff score, and incorrect acceptance occurs proportionally more often. In accordance, differences in negative predictive value become larger as the selection ratio becomes less extreme. Like the positive predictive value, the negative predictive value is sensitive to reliability as well; hence, the effects on negative predictive value become larger as tests become less reliable.

Discussion

The theoretical work reported in this article has shown that if latent differences between groups exist and measurement-invariant tests are used, selection procedures will produce different rates of incorrect decisions in these groups. Hence, with the exclusion of cases in which latent population differences are absent, measurement invariance and selection invariance are mutually inconsistent. Further, the reported simulation results suggest that the use of measurement-invariant tests for selection purposes leads to considerable violations of selection invariance. The results of this article therefore do not concern a mere statistical curiosity but imply the presence of a potentially serious problem in selection situations. We now turn to some empirical, methodological, and social policy implications of this problem.

Empirical Implications: Overprediction

The empirical situation implied by the present work is that in a group with higher mean ability, there will be more people who should have been accepted but were not, and in a group with lower mean ability, there will be more people who were accepted but should not have been. There are reasons to suspect that such a situation does, in fact, occur in our society. In educational selection situations, minority groups are often observed to achieve lower scores on the tests that are used for selection, which, if measurement invariance holds, may be due to differences in the location of the ability distributions between these groups. On purely statistical grounds, we would then expect higher dropout rates among selected members from minority groups. Such increased dropout rates have indeed been observed. In differential prediction studies, criterion performance (e.g., study success as reflected in freshmen grade point average) is often regressed on ability test scores. The typical result in these studies is that the regression line of low-scoring minority groups is lower than the regression line of majority groups, indicating that the former group has lower criterion scores than would be expected from their test scores (Neisser et al., 1996; Sackett & Wilk, 1994; Willingham, Pollack, & Lewis, 2002). This so-called *overprediction effect* could be accounted for by differential regression towards the mean, as discussed above (see also Linn & Werts, 1971; Millsap, 1997, 1998; Reilly, 1973).

Of course, there may be many other causes of increased dropout rates among minority groups (e.g., stereotype threat; Steele, 1997). For one thing, the assumption of a latent difference, which drives the statistical mechanism discussed here, may be false in real selection situations; overprediction may, in such cases, be caused by different factors. However, it is

important that in the interpretation of research findings the effects of this purely statistical mechanism are also recognized. One reason for this is that regardless of other explanations of the overprediction phenomenon, if there exist latent differences between groups, we can be certain that differential regression to the mean actually occurs.

The size of this effect depends for a large part on the reliability of the scores that are yielded by the measurement instrument used for selection. Now, while inspecting the simulations reported here, one may feel that with respect to, say, college admissions in the United States, the problem cannot be very large; for the reliability of (subtest) scores on selection instruments like the SAT is considerable in the general population (typically over .80, often over .90). However, it should be noted that reliability is a population-dependent concept (Mellenbergh, 1996) that is sensitive to restriction of range as it occurs in subpopulations with comparable abilities (as an illustration, in the limiting case of a subpopulation of people with equal ability reliability equals zero by definition; e.g., see Borsboom, 2005).

Because the SAT admittance policies for U.S. universities are known to the public, a significant degree of self-selection occurs prior to the formal selection process as carried out by universities (see Linn, 1983, for some illustrative empirical examples of the effects of self-selection). This suggests that the reliability of the SAT score within populations that apply to the same university may be considerably lower than the reliability of the SAT scores in the general population; of course it is the reliability in the self-selected applicant populations, not the reliability in the general population, that drives the mechanisms discussed in this article. In other words, in the general population reliability may be, say, .95 (which renders the amount of selection bias, given measurement invariance, negligible, as illustrated in the simulations), whereas the reliability in a self-selected population of people who apply for admission at a given educational institute may be lower (such that the amount of selection bias, given measurement invariance, may be quite high). Note also that in gauging these effects it will be important to examine reliability estimates within each group separately, because differential reliabilities may play a role in producing selection bias as well. Our current simulations do not address this issue because the combination of (strict) measurement invariance and equal latent variances implies equal reliabilities across groups.

In conclusion, it would be unwise to dismiss the mechanism discussed above purely on the basis of reliability estimates that apply to the general population. In any practical situation, evaluation of whether the problem actually occurs requires the inspection of reliability estimates as they apply to the subpopulation that will actually be subjected to the selection procedure in question. These figures are likely to differ substantially over different universities and campuses, depending on how serious the restriction of range induced by self-selection is.

Social Policy Implications: Fairness

This article is largely based on the juxtaposition of two intuitions that most people have with respect to fairness in selection. On the one hand, we want selection procedures to be fair in the sense that any two individuals with the same ability have an equal probability of passing the test. The conditions that best protect such fairness are provided through the requirement of measurement invariance, because measurement invariance operates conditional on levels of ability. On the other hand, we want selection procedures to be equally accurate for different groups. This condition applies to aggregate populations and therefore is best guaranteed by selection invariance. In one interpretation, the conditions of measurement invariance and selection invariance could be paired to conceptions of fairness that apply at the level of the individual and at the level of the group, respectively. This article elaborates on the conflicting implications of these requirements, in that sense it shows that unless one is prepared to give up either invariance at the individual or at the group level, the notion of fair test use cannot be consistently connected to psychometric invariance properties at both levels simultaneously.

Although we think that the results presented in this article elucidate the workings of selection procedures with respect to different selection scenarios, it also presents us with a significant problem. How should we deal with this situation? What are the implications for social policy?

One problem that immediately presents itself is the following. A direct implication that follows from the difference in positive predictive value across groups is that in the estimation of trait scores, prior information on group membership increases the precision of the estimates. This raises the question of whether one should use this information in selection. The argument for doing this is that it results in a more accurate selection procedure, and on this basis some have indeed argued for incorporating information on group membership in selection procedures (e.g., Miller, 1994). The argument against such practice is that it implies the use of different cutoff scores in different groups, which is commonly viewed as unfair and discriminatory. Howard Wainer (personal communication, October 28, 2005) articulated the case against using prior information on group membership in selection forcefully by drawing attention to the distinction between the use of tests as measurement instruments and the use of tests as contests:

> "I think a key issue in this is the use of the group membership information. If it is for diagnosis (e.g., educational diagnosis—what is the best course of study for a particular child; medical diagnosis—what disease do you have) using prior information is fine—and maybe even imperative. This is the so-called testing

> for the purpose of measurement. But for selection (testing as contest) it is not okay. If you and I get the same score on a selection test it is not proper to take me over you because my mother might have had more education. Measurements must be as accurate as possible. Contests must be as fair as possible. So using priors is fine for measurement, but not for contests. [This is] subtle perhaps, but critically important."

The distinction between tests as measurements and tests as contests is, in our view, useful, although different opinions on this matter are certainly possible. Also, we are of the opinion that the ethical implications of using prior information on group membership in selection outweigh the technical benefits of increased precision, and hence we tend to agree with Wainer on ideological grounds. We should, however, be very careful in evaluating the effects of any given policy in selection, even when ethically it is prima facie beyond contention. Even when we agree that tests as contests should be as fair as possible to individuals and accordingly hold on to the requirement of measurement invariance, the implication is that the sensitivity and positive predictive value will differ over populations. This may have negative consequences in itself.

For instance, consider again the scenario as it may occur in selection situations that involve minority groups. As explained in the previous paragraph, there are empirical reasons to consider the possibility that accepted members from some minority groups may include a larger number of false positives, which may be partly responsible for the observed increased dropout rates among members of such groups. The presence of more false positives among minorities may create a perceived empirical basis for prejudice.

Suppose that people notice that among minority groups that (for whatever reason) have a lower position on the dimensions that determine academic performance, there are more candidates who were incorrectly accepted. Such a perception may promote and sustain negative attitudes towards minorities. This, in turn, may affect the performance of minority groups adversely so that the improvement of the social status of these groups is hampered. Because such a result may itself have a negative effect on the determinants of academic performance of new generations of minorities, a vicious circle of reciprocal negative effects looms. Even when our intentions are good, our actions may have adverse consequences.

Thus, even when we choose to treat tests as contests and focus on measurement invariance as the preferred conceptualization of fairness in selection, the violation of selection invariance that is entailed by this decision may have negative consequences in itself. Our ideological inclinations simply will not make such problems go away. Therefore it would seem best to have a methodological procedure that minimizes the violations of selection invariance, conditional on measurement invariance. The next section evaluates a number of possible methodological solutions that could be considered.

Methodological Implications: Can the Problem Be Solved?

The effects on selection invariance that are discussed in this article can, in general, be countered by using more reliable test scores. In particular, we can design tests in which the variance of the errors around the regression line is made smaller. Three methodological procedures can be followed to achieve this. First, one could try to improve test reliability across the board, that is, for all populations involved in selection. Second, one could try to improve test reliability selectively, for instance, in the group with lower mean ability. Third, one could try to improve test reliability locally, by improving reliability in the region of the latent ability θ that is most influential in producing selection bias.

It is well known that reliability increases with test length. The first option could therefore be realized by creating longer tests. This option, however, is often not viable. Test constructors are already doing all they can to make test scores as reliable as possible, and moreover they are faced with practical and financial limitations to test length. In addition, increasing reliability across the board is inefficient; especially at the extremes of the latent dimension, it is not necessary, because the persons located there do not play a significant role in producing differences in test accuracy. Moreover, as indicated above, it is the reliability of test scores within applicant populations, not the general population, that drives the violation of selection invariance; because applicant subpopulations will typically be self-selected and hence more homogenous in ability level, adding items that increase reliability in the general population may have negligible effects on the reliability of test scores in actual applicant populations.

Selective increases in reliability in the group with lower mean ability could also be envisioned to decrease the problem in question, at least insofar as the differences in sensitivity and positive predictive value are concerned. An advantage of this solution is that it allows for an analytic evaluation: For sensitivity and positive predictive value the transformation $\sigma_\varepsilon \rightarrow \alpha\sigma_\varepsilon$ is equivalent to Transformation C2 (see Appendix C), where α represents the amount by which the variance of the conditional distribution of test scores, given a position on the latent variable, is changed by using a longer or shorter test. However, implementing this solution would involve the use of longer tests for the group with lower mean ability. Such a procedure has two obvious drawbacks. First, although it will decrease differences in sensitivity and positive predictive value, it will simultaneously increase differences in specificity and negative predictive value. Second, the procedure would involve differential treatment of the members of each population, which may be perceived as discriminatory; also, the cutoff scores would be different because they are defined on different sets of items.

Local increases in reliability would be effective in dimin-

ishing the problem and would not share the drawbacks of the first two options. Because with respect to selection invariance the most important region of θ is the region around the cutoff, selection procedures could be tailored to be more reliable in that region. Classical approaches to reliability are not suitable for achieving this goal because they are population dependent, but item response theory (IRT) approaches are. IRT models evaluate measurement precision as a continuous function of the latent dimension known as the test information function (see Mellenbergh, 1996, for a discussion of the difference between conceptualizations of measurement precision in classical and modern test theory). The form of the test information function can be influenced by administering different items.

In the present situation, measurement precision could be targeted at the value of θ that is estimated to be the cutoff score (i.e., the value that produces the right selection ratio). This could be done by administering extra items with difficulty parameters close to that point on the latent scale. Such extra items could be administered, for instance, when the available test scores indicate that the tested person is located in the relevant region of the latent dimension. It is perhaps useful to note that such a procedure is not the same as that followed in widely used adaptive testing programs; these programs increase measurement precision at the location of a person's estimated latent ability, whereas what would be needed here is a procedure that increases measurement precision at a fixed point of the latent scale. Rather, the procedure is an implementation of sequential mastery testing as studied in the IRT literature (e.g., Eggen & Straetmans, 2000; Lewis & Sheehan, 1990; Spray & Reckase, 1996; Vos, 1999; Weiss & Kingsbury, 1984). Although in the IRT literature the focus is on gaining efficiency in deciding whether a person is a master or nonmaster (because fewer items have to administered to those far above the cutoff) rather than on producing selective increases in measurement precision, the statistical machinery in place could be used for either purpose.

It is useful to note that the cutoff score used here need not have any substantive meaning, in the sense that it separates masters from nonmasters. Rather the goal would be to reach a simple pass–fail decision. The cut point on the θ scale would be defined by the selection ratio that the selecting institution entertains, together with the mean and variance of the ability distribution in the applicant population, and hence would generally differ across institutions. Standard IRT algorithms for sequential mastery testing could then be used to achieve local increases in measurement precision at the point of the cutoff score for θ. Such a procedure may serve to approximate selection invariance to a greater degree while maintaining measurement invariance and therefore could be used to steer a middle course between protecting fairness at the level of the individual and at the level of the group.

A selective improvement of reliability along these lines would result in a joint distribution of test and ability scores that is "squeezed" in the region around the cutoff. The joint distribution would therefore look like the one represented in Figure 6. For obvious reasons, we have come to refer to this as the dog bone method. The method would not counter selection bias entirely but would decrease the difference in incorrect decision across groups, while it retains the property of measurement invariance: The conditional distribution of the test score, given θ, would still be identical across groups. Also, it would not require differential treatment of members of different populations; members of both groups are administered extra items if they score in the relevant region. Finally, because different items could be administered by institutions that employ different selection ratios and deal with applicant populations centered at different positions of the latent scale, the procedure can be efficiently tailored to particular selection situations.

Future Work

Although this article has outlined some of the most important aspects of the relation between measurement invariance and selection invariance, several questions that are beyond the scope of this article may be addressed in future research.

First, we have assumed bivariate normality throughout this article, but in most selection procedures this assumption will be approximately true at best. One reason for this is that test scores are usually bounded; hence the relation between the expected test score and the latent variable will be nonlinear, and the joint distribution of ability and test scores will not be bivariate normal. The influence of nonlinearity and nonnormality on selection properties thus requires attention. For one, it looks like the linearity of the relation between X and θ is not essential. Rather the proof hinges on

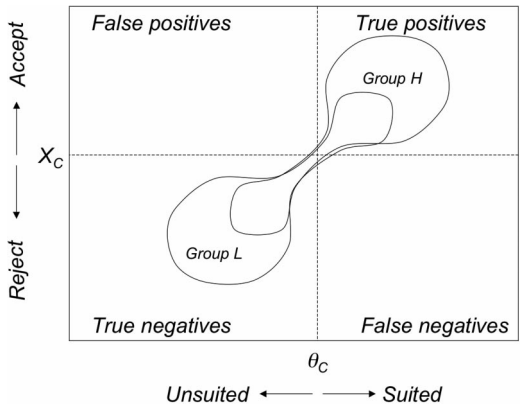


Figure 6. Minimization of selection bias by means of selective increases in reliability. Adaptive testing may be used to create an increase in reliability in the neighborhood of the cutoff score, which diminishes the differences in test accuracy across groups.

this relation being monotonic. An analogous proof may perhaps be constructed on weaker assumptions.

Second, we have restricted ourselves to a scenario where selection is based on a single test score. In reality, however, selection usually takes place on more than one variable. Hence the problem becomes multidimensional. It could be the case that incorporating sets of variables in the selection process decreases the violations of selection invariance discussed in this article. This could happen if group differences are reversed on different variables, so that the group that is located at the low end of one variable is located at the high end of another. In cases where group differences are in the same direction on different variables, however, the problems could be expected to be aggravated; the same holds for situations where some of the additional measures used are less reliable. Investigating the mechanisms of multidimensional selection may thus be fruitful, although we expect that analytic approaches will quickly become mathematically forbidding. In such cases, simulation studies may be used to shed some light on this issue.

Third, the viability and effect of incorporating local improvements in measurement precision, which was suggested as a possible methodological solution to the problem of selection bias, can be studied from an IRT perspective. One important question is how many additional items one would need to establish a reasonable approximation to selection invariance. The answer to this question will probably differ according to the selection situation considered and hence may be best addressed in particular selection contexts with known psychometric qualities. Another important point is that using many items with an item difficulty located at the cutoff point would serve to strengthen the nonlinearity of the relation between the test score and the latent variable; hence the effects of this on selection invariance should be investigated as well. Finally, various results from the IRT literature on sequential mastery testing may be utilized to further study this method and judge its viability.

Fourth, the relation between the current conceptualization of selection invariance and previous conceptualizations of that property (e.g., Petersen & Novick, 1976) should be studied in greater detail. In much literature the criterion score functions to define whether a person is a true positive or a false positive. In our conceptualization these quadrants are defined in terms of the latent dimension measured through the test. Now suppose, for instance, that in addition to the test score, we also have a criterion measure that depends on θ or on a related latent variable. Assume that we have measurement invariance and group differences in the distribution of θ, so the selection invariance in our terms is violated. Does this entail that selection invariance, as defined through the criterion scores, is violated as well? Conversely, suppose that measurement invariance is violated in such a way that selection invariance is satisfied. Does this entail that selection invariance, as defined through the criterion scores, will be satisfied as well? Under reasonable assumptions, we think that these questions are amenable to systematic investigation. We conjecture, in fact, that selection invariance as defined through the criterion will generally fail to align with selection invariance as defined in terms of the latent variable, and some preliminary simulations suggest that this is indeed the case. One would also expect this from the fact that prediction invariance does not align with measurement invariance (Millsap, 1997); normally, if measurement invariance is satisfied, prediction invariance is not, and one may therefore expect that situation at the predictive level will be very different from that at the measurement level. For instance, if measurement invariance is violated, then selection invariance in our definition may or may not be violated; predictor–criterion regressions may or may not be invariant across groups, and selection invariance as defined through the criterion scores may or may not be violated. If measurement invariance is satisfied, then selection invariance as we define it will normally be violated, and prediction invariance will be too, but selection invariance as defined through the criterion scores may or may not be violated (see Petersen & Novick, 1976). The implication is that for one who takes the measurement model seriously in selection situations, results that apply to the predictor–criterion relation must generally be deemed uninformative with respect to the situation at the measurement level. Clearly, this has important consequences for test practices. Thus, the relation between predictive models and measurement models is one that invites and deserves further study.

References

Aguinis, H., & Smith, M. E. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology, 60,* 165–199.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Psychological Association.

Birnbaum, M. H. (1979). Procedures for the detection and correction of salary inequities. In T. R. Pezzullo & B. E. Brittingham (Eds.), *Salary equity* (pp. 121–144). Lexington, MA: Lexington Books.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge, United Kingdom: Cambridge University Press.

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behavior, 49,* 122–158.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cole, N. S. (1973). Bias in selection. *Journal of Educational Measurement, 10,* 237–255.

Darlington, R. B. (1971). Another look at "culture fairness." *Journal of Educational Measurement, 8,* 71–82.

Eggen, T. J. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713–734.

Evers, A., Te Nijenhuis, J., & Van der Flier, H. (2005). Ethnic bias and fairness in personnel selection: Evidence and consequences. In A. Evers, N. Anderson, & O. F. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 306–328). Oxford, United Kingdom: Blackwell.

Gamliel, E., & Cahan, S. (2007). Mind the gap: Between-group differences and fair test use. *International Journal of Selection and Assessment, 15,* 273–282.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Erlbaum.

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests. *Psychology, Public Policy, and Law, 6,* 151–158.

Hunter, J. E., Schmidt, F. L., & Rauschenberger, J. (1984). Methodological and statistical issues in the study of bias in mental testing. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 41–99). New York: Plenum Press.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: Macmillan.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous multivariate distributions* (2nd ed.). New York: Wiley.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367–386.

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement, 20,* 1–15.

Linn, R. L., & Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8,* 1–4.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Oxford, United Kingdom: Addison-Wesley.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127–143.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 19,* 223–236.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1,* 293–299.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525–543.

Miller, E. M. (1994). The relevance of group membership for personnel selection: A demonstration using Bayes' theorem. *Journal of Social, Political, and Economic Studies, 19,* 323–359.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248–260.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33,* 403–424.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72,* 461–473.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing bias. *Applied Psychological Measurement, 17,* 297–334.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 99,* 93–115.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13,* 3–29.

Reilly, R. R. (1973). A note on minority group bias studies. *Psychological Bulletin, 80,* 130–133.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11,* 235–294.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in pre-employment testing. *American Psychologist, 49,* 929–954.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, 13,* 238–241.

Society for Industrial and Organizational Psychology. (2003). *Principles for validation and use of personnel selection procedures.* Retrieved from http://siop.org/_principles/principles.pdf

Spray, J. A., & Reckase, M. D. (1996). Comparison of SRPT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21,* 405–414.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

Thorndike, R. L. (1971). Concepts of culture fairness. *Journal of Educational Measurement, 8,* 63–70.

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics, 24,* 271–292.

Wainer, H. (2000). Kelley's paradox. *Chance, 13,* 47–48.

Wainer, H. (2005). *Graphic discovery.* Princeton, NJ: Princeton University Press.

Wainer, H., & Brown, L. M. (2004). Two statistical paradoxes in the interpretation of group differences: Illustrated with medical school admission and licensing data. *The American Statistician, 58,* 117–123.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement, 39,* 1–37.

Yule, G. H. (1903). Notes on the theory of association of attributes in statistics. *Biometrika, 2,* 121–134.

Appendix A

Failure of Selection Invariance

We first prove Equation 5. Because of Equation 3 we have

$$p(A|\theta \cap g) = p(X \geq 0|\theta \cap g) = \int_0^{\infty} f_{\lambda_g\theta,\sigma_{\varepsilon_g}}(X)dX = \int_{-\infty}^{\lambda_g\theta} f_{0,\sigma_{\varepsilon_g}}(X)dX = N_{0,\sigma_{\varepsilon_g}}(\lambda_g\theta), \quad \text{(A1)}$$

where $f_{m,s}$ and $N_{m,s}$ refer to the Gaussian density and the normal distribution with mean m and standard deviation s, respectively. The requirement that over all values of θ, $p(A \mid \theta \cap g)$ is the same for each value of G thus comes down to the requirement that $N_{0,\sigma_{\varepsilon_g}}(\lambda_g\theta)$ is the same for each value of G, which is true by assumption, since λ_g and σ_{ε_g} are constant.

Using Expression A1, we can write out the numerator and the denominator in Equation 10 as

$$p(S \cap A|g) = \int_0^{\infty} f_{\mu_g,\sigma}(\theta)\, N_{0,\sigma_\varepsilon}(\lambda\theta)d\theta, \quad \text{(A2)}$$

$$p(A|g) = \int_{-\infty}^{\infty} f_{\mu_g,\sigma}(\theta)\, N_{0,\sigma_\varepsilon}(\lambda\theta)d\theta, \quad \text{(A3)}$$

the sole difference between numerator and denominator being the integration boundaries. Note that because of the assumption of measurement invariance, the parameters λ and σ_ε do not carry an index g. Only μ_g varies over the groups.

Unfortunately there is no analytic solution for the quotient of the above integrals. Nevertheless we may study $p(S \mid A \cap g)$ as a function of the parameter μ_g. Specifically, we can differentiate this function with respect to μ_g and prove that

$$\frac{d}{d\mu_g}[p(S|A \cap g)] > 0. \quad \text{(A4)}$$

This means that of two groups L and H, if $\mu_L < \mu_H$, the probability of suitability given acceptance is strictly larger for H than for L. In other words, the procedure is then not selection invariant due to an inequality of positive predictive value.

To derive Equation A4, first consider the differentiation rule for quotients,

$$\frac{d}{dz}\left[\frac{A}{B}\right] = \frac{\frac{dA}{dz}B - A\frac{dB}{dz}}{B^2}. \quad \text{(A5)}$$

Because the denominator of this expression is always positive, we are only required to show that

$$p(A|g)\frac{d}{d\mu_g}[p(S \cap A|g)] - p(S \cap A|g)\frac{d}{d\mu_g}[p(A|g)] > 0 \quad \text{(A6)}$$

(*Appendixes continue*)

to derive Inequality A4. Rewriting this we find

$$\frac{dp(S \cap A|g)/d\mu_g}{p(S \cap A|g)} > \frac{dp(A|g)/d\mu_g}{p(A|g)}. \tag{A7}$$

The remainder of the derivation focuses on this inequality.

To arrive at Inequality A7, note first that we may apply the differentiation operation within the integral expressions. From the functional form of the normal density,

$$f_{\mu_g,\sigma}(\theta) \propto \exp - \frac{(\theta - \mu_g)^2}{2\sigma^2}, \tag{A8}$$

we can derive that

$$\frac{df_{\mu_g,\sigma}(\theta)}{d\mu_g} = \left(\frac{\theta - \mu_g}{\sigma^2}\right) f_{\mu_g,\sigma}(\theta). \tag{A9}$$

Applying this to the terms in Inequality A7, eliminating the factor $1/\sigma^2$ on either side, we obtain

$$\frac{\int_0^\infty (\theta - \mu_g)\, fNd\theta}{\int_0^\infty fNd\theta} > \frac{\int_\infty^\infty (\theta - \mu_g)\, fNd\theta}{\int_{-\infty}^\infty fNd\theta}, \tag{A10}$$

where we have written fN to abbreviate $f_{\mu_g,\sigma}(\theta) N_{0,\sigma_\varepsilon}(\lambda\theta)$. The integration variable θ can now be substituted with $\theta' = \theta - \mu_g$. On the left side of the inequality, this entails a shift in the integration boundary.

It thus turns out that for Inequality A4 to hold, it is sufficient that the following inequality holds:

$$\frac{\int_{\mu_g}^\infty \theta' fNd\theta'}{\int_{\mu_g}^\infty fNd\theta'} > \frac{\int_{-\infty}^\infty \theta' fNd\theta'}{\int_{-\infty}^\infty fNd\theta'}. \tag{A11}$$

Interestingly, this is the same as an inequality of expectation values for θ' under the distribution fN:

$$E_{[\mu_g,\infty)}[\theta'] > E_{(-\infty,\infty)}[\theta']. \tag{A12}$$

But this inequality is evident: The expectation value of θ' over the interval $[\mu_g,\infty)$ is always strictly larger than the expectation value of θ' over the interval $(-\infty,\infty)$, if only there is some probability mass in the domain $(-\infty,\mu_g)$. And this is in fact the case for the function fN.

Appendix B

Symmetry of the Problem with Respect to X and θ

The symmetry argument establishes that the inequalities of Appendix A are invariant under interchanging θ and X. The proofs in Appendix A are constructed for a selection problem characterized by a normal distribution of a group over the ability parameter θ and a linear dependence linking the ability θ to normal distributions over a test score X. What we first have to show is that the same problem is characterized by a normal distribution of a group over the scores X and a linear dependence linking scores X to normal distributions over ability θ. Then we ask whether this reparameterized function has the same properties as the original function, in the sense of having more probability mass in the true positives quadrant relative to the combined true positives and false negatives quadrants, that is, whether the case of sensitivity is susceptible to the same proof as given in Appendix A. This will be the case if the order of the latent means, μ_g, is preserved in the observed means and if the variances of the observed scores remain identical.

To establish this, consider the test as a linear dependence linking ability to normal distributions over a test score. It is captured completely by a single function p_{test} over θ and X, referring to the event of a person with ability θ receiving a test score X. Note that p_{test} differs from the full probability assignment p by the fact that we have not fixed a marginal probability over θ. Now we may reformulate the function as follows:

$$p_{\text{test}}(X, \theta) = f_{\lambda\theta,\sigma_\varepsilon}(X) = f_{0,\sigma_\varepsilon}(X - \lambda\theta) = f_{0,(\sigma_\varepsilon/\lambda)}\left(\frac{X}{\lambda} - \theta\right) = f_{X/\lambda,\sigma_\varepsilon/\lambda}(\theta). \tag{B1}$$

The point of this reformulation is that we can describe the very same function defining the test by looking at normal distributions over X given θ and by looking at normal distributions over θ given X. The test may just as well be characterized by a linear dependence linking scores X to normal distributions over ability θ as by the converse dependence.

Now assume the normal distribution of a group $G = g$ over abilities θ, $f_{\mu_g,\sigma}(\theta)$, as the marginal probability over θ. The full probability of the event of a member of group g with ability θ receiving a test score X can now be written as the product of this marginal and the above function:

$$p(\theta, X) = f_{\mu_g,\sigma}(\theta) f_{\lambda\theta,\sigma_\varepsilon}(X). \tag{B2}$$

To establish the symmetry of the selection problem with respect to interchanging X and θ, we must now show that under the assumption of a certain test p_{test}, the marginal distribution with respect to X is again the normal, and that a difference of means in ability, $\mu_L < \mu_H$, translates to a difference of means in test scores, $m_L < m_H$, while the standard deviations are again equal, $s_L = s_H$. If this is indeed the case, then we can use the results of Appendix A to establish that differences in sensitivity are in the same direction as differences in positive predictive value.

This requires us to compute the functional form of the marginal distribution for X. To this aim, we will write out the marginal probability $p(X)$ in such a way that we can integrate out the variable θ by means of a number of suitably chosen substitutions. The selection problem determines that

$$p(X) = \int_{-\infty}^{\infty} p(X, \theta) d\theta = \int_{-\infty}^{\infty} f_{\mu_g,\sigma}(\theta) f_{\lambda\theta,\sigma_\varepsilon}(X) d\theta = \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{\theta - \mu_g}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{X - \lambda\theta}{\sigma_\varepsilon}\right)^2\right] d\theta. \tag{B3}$$

(*Appendixes continue*)

We may substitute $\theta' = \frac{\theta - \mu_g}{\sigma}$ and write

$$p(X) = \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\theta'^2 - \frac{1}{2}\left(\frac{X - \lambda(\sigma\theta' + \mu_g)}{\sigma_\varepsilon}\right)^2\right] d\theta'$$

$$= \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\theta'^2 - \frac{1}{2}\left(\frac{X - \lambda\mu_g - \lambda\sigma\theta'}{\sigma_\varepsilon}\right)^2\right] d\theta'. \quad \text{(B4)}$$

Now we can further substitute $X' = \frac{X - \lambda\mu_g}{\sigma_\varepsilon}$ and $\gamma = \frac{\lambda\sigma}{\sigma_\varepsilon}$, so that

$$p(X) = \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\theta'^2 - \frac{1}{2}\left(X' - \frac{\lambda\sigma}{\sigma_\varepsilon}\theta'\right)^2\right] d\theta' = \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\theta'^2 - \frac{1}{2}(X' - \gamma\theta')^2\right] d\theta'. \quad \text{(B5)}$$

Writing out this expression and collecting the terms we find

$$p(X) = \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}(1 + \gamma^2)\theta'^2 + \gamma\theta' X' - \frac{1}{2}X'^2\right] d\theta'$$

$$= \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left((1 + \gamma^2)^{1/2}\theta' - \frac{\gamma}{(1 + \gamma^2)^{1/2}}X'\right)^2 - \frac{\gamma^2}{1 + \gamma^2}X'^2 - \frac{1}{2}X'^2\right] d\theta'$$

$$= \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left((1 + \gamma^2)^{1/2}\theta' - \frac{\gamma}{(1 + \gamma^2)^{1/2}}X'\right)^2\right] d\theta' \times \exp\left[-\left(\frac{1}{2} + \frac{\gamma^2}{1 + \gamma^2}\right)X'^2\right]. \quad \text{(B6)}$$

We can now substitute, once again, $\theta'' = (1 + \gamma^2)^{1/2}\theta' - \frac{\gamma}{(1 + \gamma^2)^{1/2}}X'$ and integrate out θ''. Note that X' may here be treated as constant. The integration runs over a different θ'' at each X', but these variables θ'' only differ in a simple translation, and this translation is nullified because the integration runs from $-\infty$ to ∞.

The functional form of the marginal distribution has now been computed. We end up with

$$p(X) = \exp\left[\left(\frac{1}{2} + \frac{\gamma^2}{1 + \gamma^2}\right)X'^2\right] = \exp\left[\left(\frac{X - \lambda\mu}{\sigma_\varepsilon}\right)^2\left(\frac{1}{2} + \frac{\gamma^2}{1 + \gamma^2}\right)\right]. \quad \text{(B7)}$$

The distribution over test scores comes out normally distributed with a mean $m_\varepsilon = \lambda\mu_g$ and the rather elaborate standard deviation

$$s = \sigma_\varepsilon\left(\frac{1}{2} + \frac{\gamma^2}{1 + \gamma^2}\right)^{-1/2}, \quad \text{(B8)}$$

with $\gamma = \frac{\lambda\sigma}{\sigma_\varepsilon}$. The order in the means of the two groups L and H is preserved: If for the means over the abilities we have $\mu_L < \mu_H$, then also we have $m_L < m_H$ for the means over the test scores. Moreover, since σ, σ_ε, and λ are all identical for both groups, the standard deviation s is the same for both groups too. Hence the probabilistic inequalities derived in Appendix A are invariant under interchanging X and θ, and Equation 13 follows.

Appendix C

Generalized Inequality

The strategy we use to deal with the situation in which groups differ in both means and variances is the following. We apply a transformation to the latent variable distribution in one group in order to remove the difference in variances. This causes the variance of the latent variable distribution to be the same across groups but alters the mean and regression coefficient in the transformed group. We then scale the regression coefficient back to its original value. This gives a counterfactual situation, in which both the variances and the regression parameter are the same in both groups. We then use the results derived in the previous section to evaluate whether, in that counterfactual situation, a difference in, say, sensitivity exists.

If so, we consider the direction of the effect of the employed transformations on sensitivity to evaluate whether we can generalize the conclusion from the transformed case to the untransformed case. For instance, if we know that the set of transformations made sensitivity smaller than it originally was, while in the resulting situation sensitivity was still larger in the transformed group when compared to the untransformed group, then sensitivity must also have been greater in that group before the transformation started. Finally, we derive a set of conditions under which this situation is obtained, that is, the set of conditions that describes the class of situations in which measurement invariance and selection invariance are provably inconsistent.

We first concentrate on sensitivity and positive predictive value, as expressed in Equations 10 and 12. We make the following two observations concerning these probabilities. First, they are invariant under the combined scale transformations of the ability parameter and the regression coefficient,

$$\theta \rightarrow \alpha\theta, \tag{C1}$$

$$\lambda \rightarrow \frac{\lambda}{\alpha}. \tag{C2}$$

for $\alpha \in (0,\infty)$. Second, considering Transformation C2 while all else is kept fixed, both sensitivity and positive predictive value get larger for larger regression coefficients λ. Both observations are proved in Appendix D.

We can write down the observations more formally as follows. Consider a group g with a density $f_{\mu,\sigma}(\theta)$, and denote this group with $g_{\mu,\sigma}$. Further, consider a test characterized by $f_{\lambda\theta,\sigma_\varepsilon}(X)$ and denote the corresponding probability assignment over X and θ by $p_{\lambda,\sigma_\varepsilon}$, or p_λ for short. The first observation then comes down to the following statements:

$$p_\lambda(S|A \cap g_{\mu,\sigma}) = p_{\lambda/\alpha}(S|A \cap g_{\alpha\mu,\alpha\sigma}), \tag{C3}$$

$$p_\lambda(A|S \cap g_{\mu,\sigma}) = p_{\lambda/\alpha}(A|S \cap g_{\alpha\mu,\alpha\sigma}).$$

The second observation comes down to the following two statements:

$$\alpha < 1 \Rightarrow \begin{cases} p_{\lambda/\alpha}(S|A \cap g_{\mu,\sigma}) > p_\lambda(S|A \cap g_{\mu,\sigma}), \\ p_{\lambda/\alpha}(A|S \cap g_{\mu,\sigma}) > p_\lambda(A|S \cap g_{\mu,\sigma}). \end{cases} \tag{C4}$$

Both these pairs of observations are used in the proof concerning two groups with differing means and variances.

(*Appendixes continue*)

Consider the selection problem for two groups, $g_{\mu,\sigma}$ and $g_{\mu',\sigma'}$ with $\sigma \neq \sigma'$, and a test characterized by regression coefficient λ and an error variance σ_ε^2. We assume nothing on the order of μ and μ', and we hold on to measurement invariance, so that σ and σ_ε are the same for both groups. We further assume $\sigma > \sigma'$ without loss of generality. We may then employ the above observations to derive more general inequalities for the probabilities involved in selection invariance.

The first observation allows us to transform the standard deviation in one group so that it equals the standard deviation in the other group. For this purpose we use the combined Transformations C1 and C2 with $\alpha = \dfrac{\sigma'}{\sigma}$ so that $\alpha < 1$. We choose to transform in the group with the larger standard deviation. For this group, Equation C3 then gives us

$$\alpha = \frac{\sigma'}{\sigma} \Rightarrow \begin{cases} p_\lambda(S|A \cap g_{\mu,\sigma}) = p_{\lambda/\alpha}(S|A \cap g_{\alpha\mu,\alpha\sigma}), \\ p_\lambda(A|S \cap g_{\mu,\sigma}) = p_{\lambda/\alpha}(A|S \cap g_{\alpha\mu,\alpha\sigma}). \end{cases} \tag{C5}$$

We now compare these probabilities to a situation where the standard deviations and means have been transformed as above, but the regression parameter equals its original value. From Equation C4 we may then derive the conditional inequalities

$$\alpha < 1 \Rightarrow \begin{cases} p_{\lambda/\alpha}(S|A \cap g_{\alpha\mu,\sigma'}) > p_\lambda(S|A \cap g_{\alpha\mu,\sigma'}), \\ p_{\lambda/\alpha}(A|S \cap g_{\alpha\mu,\sigma'}) > p_\lambda(A|S \cap g_{\alpha\mu,\sigma'}). \end{cases} \tag{C6}$$

Recall that $\sigma' = \alpha\sigma$. Therefore the expressions on the right side of Equation C5 are identical to the expressions on the left side of Equation C6.

Because of the employed transformations, the expressions on the right side of Equation C6 have variance σ' and regression parameter λ, allowing us to compare the expressions to those for the untransformed group $g_{\mu',\sigma'}$. Specifically, we can apply the results of the previous section concerning acceptance and suitability as follows:

$$\alpha\mu > \mu' \Rightarrow \begin{cases} p_\lambda(S|A \cap g_{\alpha\mu,\sigma'}) > p_\lambda(S|A \cap g_{\mu',\sigma'}), \\ p_\lambda(A|S \cap g_{\alpha\mu,\sigma'}) > p_\lambda(A|S \cap g_{\mu',\sigma'}). \end{cases} \tag{C7}$$

These are just Conditions 11 and 13, used here to compare the transformed situation in one group with the untransformed situation in the other group. Clearly, if $\alpha\mu = \mu'$, the two terms in the above are equal.

The joint effect of Transformations C5 and C6 is a decrease of sensitivity and positive predictive value, as can be seen from the inequalities in C6. Thus, if the inequalities in Equation C7 hold, which will be the case under the condition $\alpha\mu > \mu'$, the inequalities must hold for the original, untransformed case as well. With regard to sensitivity, for example, we have established that, under the conditions stated, $p_\lambda(A \mid S \cap g_{\mu,\sigma}) = p_{\lambda/\alpha}(A|S \cap g_{\alpha\mu,\sigma'}) > p_\lambda(A \mid S \cap g_{\alpha\mu,\sigma}) \geq p_\lambda(A \mid S \cap g_{\mu',\sigma'})$. Hence it follows that $p_\lambda(A \mid S \cap g_{\mu,\sigma}) > p_\lambda(A \mid S \cap g_{\mu',\sigma'})$, which is the inconsistency between measurement invariance and selection invariance with regard to sensitivity.

To establish the reach of these results, we may now concatenate Equations C5 to C7 and collect the conditions. We then obtain

$$\sigma' < \sigma \text{ and } \frac{\sigma'}{\sigma}\mu \geq \mu' \Rightarrow \begin{cases} p_\lambda(S|A \cap g_{\mu,\sigma}) > p_\lambda(S|A \cap g_{\mu',\sigma'}), \\ p_\lambda(A|S \cap g_{\mu,\sigma}) > p_\lambda(A|S \cap g_{\mu',\sigma'}). \end{cases} \tag{C8}$$

Furthermore, by inverting the above reasoning we can also prove that

$$\sigma' < \sigma \text{ and } \frac{\sigma'}{\sigma}\mu \leq \mu' \Rightarrow \begin{cases} p_\lambda(\neg S|\neg A \cap g_{\mu,\sigma}) < p_\lambda(\neg S|\neg A \cap g_{\mu',\sigma'}), \\ p_\lambda(\neg A|\neg S \cap g_{\mu,\sigma}) < p_\lambda(\neg A|\neg S \cap g_{\mu',\sigma'}). \end{cases} \tag{C9}$$

For this we employ Equations C5 to C7, but we transform $\theta \rightarrow -\theta$.

Appendix D

Scale Transformations

We first establish the observation expressed in Equations C1 and C2. After the combined transformations C1 and C2, the probability over X and θ can be written as

$$p(X, \theta) = f_{\alpha\mu,\alpha\sigma}(\alpha\theta) f_{(\lambda/\alpha)\alpha\theta,\sigma_\varepsilon}(X) = \exp\left[\frac{1}{2}\left(\frac{\alpha\theta - \alpha\mu}{\alpha\sigma}\right)^2\right] f_{\lambda\theta,\sigma_\varepsilon}(X) = \exp\left[\frac{1}{2}\left(\frac{\theta - \mu}{\sigma}\right)^2\right] f_{\lambda\theta,\sigma_\varepsilon}(X)$$
$$= f_{\mu,\sigma}(\theta) f_{\lambda\theta,\sigma_\varepsilon}(X). \quad \text{(D1)}$$

The scale transformation thus leaves the probability assignment over X and θ invariant. Integrals such as the probabilities of suitability given acceptance and acceptance given suitability are therefore left invariant under the transformations as well.

We now establish the observation concerning stand-alone scale transformations of the regression parameter, as expressed in Equation C4. We first deal with the probability of acceptance given suitability, $p(A \mid S \cap g)$. Note that for a group $g_{\mu,\sigma}$

$$p_\lambda(S|g_{\mu,\sigma}) = \int_0^\infty f_{\mu,\sigma}(\theta)\left[\int_{-\infty}^\infty f_{\lambda\theta,\sigma_\varepsilon}(X)dX\right]d\theta = \int_0^\infty f_{\mu,\sigma}(\theta)d\theta = 1 - N_{\mu,\sigma}(0), \quad \text{(D2)}$$

where $N_{\mu,\sigma}$ is again the normal distribution. This integral does not depend on λ. Recall also that

$$p_\lambda(A \cap S|g_{\mu,\sigma}) = \int_0^\infty f_{\mu,\sigma}(\theta)\left[\int_0^\infty f_{\lambda\theta,\sigma_\varepsilon}(X)dX\right]d\theta = \int_0^\infty f_{\mu,\sigma}(\theta) N_{0,\sigma_\varepsilon}(\lambda\theta)d\theta. \quad \text{(D3)}$$

The probability of acceptance given suitability, $p(A \mid S \cap g_{\mu,\sigma})$, is the quotient of these two expressions, as given in Equation 12.

Consider the probability of passing given suitability for the transformed regression parameter $\frac{\lambda}{\alpha}$. We can write

$$p_{\lambda/\alpha}(A \cap S|g_{\mu,\sigma}) = \int_0^\infty f_{\mu,\sigma}(\theta) N_{0,\sigma_\varepsilon}\left(\frac{\lambda}{\alpha}\theta\right)d\theta = \int_0^\infty f_{\mu,\sigma}(\theta)\left[N_{0,\sigma_\varepsilon}(\lambda\theta) + \left(N_{0,\sigma_\varepsilon}\left(\frac{\lambda}{\alpha}\theta\right)\right.\right.$$
$$\left.\left. - N_{0,\sigma_\varepsilon}(\lambda\theta)\right)\right]d\theta = p_\lambda(A \cap S|g_{\mu,\sigma}) + \int_0^\infty f_{\mu,\sigma}(\theta)\left[N_{0,\sigma_\varepsilon}\left(\frac{\lambda}{\alpha}\theta\right) - N_{0,\sigma_\varepsilon}(\lambda\theta)\right]d\theta. \quad \text{(D4)}$$

Because the normal distribution N_{0,σ_ε} is a monotonically increasing function, the latter integral is strictly positive if $\frac{\lambda}{\alpha}\theta > \lambda\theta$. And because in the above integral the domain is $\theta \geq 0$, this condition is equivalent to $\alpha < 1$. Therefore,

$$\alpha < 1 \Rightarrow p_{\lambda/\alpha}(A \cap S|g_{\mu,\sigma}) > p_\lambda(A \cap S|g_{\mu,\sigma}). \quad \text{(D5)}$$

(*Appendixes continue*)

Because the denominator in the quotient of Equation 12, as expressed in Equation D2, does not depend on λ, the above comes down to Equation C6 for the probability of acceptance given suitability.

It is now easy to deal with the probability of suitability given acceptance. We want to obtain

$$\frac{p_{\lambda/\alpha}(A \cap S|g_{\mu,\sigma})}{p_{\lambda/\alpha}(A|g_{\mu,\sigma})} > \frac{p_{\lambda}(A \cap S|g_{\mu,\sigma})}{p_{\lambda}(A|g_{\mu,\sigma})}. \tag{D6}$$

With some algebra, using the fact that $p(A \mid g) = p(A \cap S \mid g) + p(A \cap \neg S \mid g)$, this is equivalent to

$$p_{\lambda/\alpha}(A \cap S|g_{\mu,\sigma})\, p_{\lambda}(A \cap \neg S|g_{\mu,\sigma}) > p_{\lambda}(A \cap S|g_{\mu,\sigma})\, p_{\lambda/\alpha}(A \cap \neg S|g_{\mu,\sigma}). \tag{D7}$$

But because of Equation D1, this holds if we have

$$p_{\lambda}(A \cap \neg S|g_{\mu,\sigma}) \geq p_{\lambda/\alpha}(A \cap \neg S|g_{\mu,\sigma}). \tag{D8}$$

Writing integral expressions for these probabilities in much the same way as in Equation D4, we find

$$p_{\lambda}(A \cap \neg S|g_{\mu,\sigma}) = \int_{-\infty}^{0} f_{\mu,\sigma}(\theta)\, N_{0,\sigma_\varepsilon}\left(\frac{\lambda}{\alpha}\,\theta\right) d\theta = \int_{-\infty}^{0} f_{\mu,\sigma}(\theta)\left[N_{0,\sigma_\varepsilon}\left(\frac{\lambda}{\alpha}\,\theta\right) + \left(N_{0,\sigma_\varepsilon}(\lambda\theta) \right.\right.$$

$$\left.\left. - N_{0,\sigma_\varepsilon}\left(\frac{\lambda}{\alpha}\,\theta\right)\right)\right] d\theta = p_{\lambda/\alpha}(A \cap \neg S|g_{\mu,\sigma}) + \int_{-\infty}^{0} f_{\mu,\sigma}(\theta)\left[N_{0,\sigma_\varepsilon}(\lambda\theta) - N_{0,\sigma_\varepsilon}\left(\frac{\lambda}{\alpha}\,\theta\right)\right] d\theta. \tag{D9}$$

Because in this integral the domain is $\theta < 0$, the inequality is again satisfied if $\alpha < 1$.

Received February 28, 2007
Revision received March 28, 2008
Accepted March 31, 2008 ■