

Running head: FAIR SELECTION

Measurement invariance versus selection invariance:  
Is fair selection possible?

Denny Borsboom  
Jan-Willem Romeijn  
Jelte M. Wicherts  
Department of Psychology  
University of Amsterdam

Send correspondence 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](mailto:d.borsboom@uva.nl)

**Abstract**

This paper 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 located at the higher end of the latent dimension, whereas specificity and negative predictive value will be higher in the group located 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 is shown to originate as a special case of Simpson's paradox, which arises because the observed score distribution is collapsed into an accept/reject dichotomy. Simulations show that 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.

**Measurement invariance versus selection  
invariance:  
Is fair selection possible?**

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, while he does not have the capacity to successfully complete his education, and Mary was diagnosed with an attention deficit hyperactive disorder, while she did 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 is possible to construct tests in such a fashion that they are as fair as possible, i.e., do not result in discrimination against particular groups.

Of course, the basic idea in selection is that people with higher ability levels should have higher chances of being selected for, say, college admission. Therefore, to discriminate between people, in the sense of distinguishing between their ability levels, is part and parcel of selection procedures. Consequently, if people of higher ability levels are more often encountered in a particular group, then more people from that group will be selected. Thus, if a test does not result in the selection of equal proportions of individuals in each group, that does not necessarily mean that something is wrong with the test. However, the proportion of selected individuals should be the same in subpopulations of people with equal ability, regardless of group membership: if you and I have the same ability to complete a college degree, then we should have the same probability of being admitted into college by a testing procedure, irrespective of any other characteristics on which we may differ, like sex, age, the color of our skin, or religious and political convictions.

In the literature on psychometric testing, this requirement is called *measurement invariance*. A test that fulfills the requirement of measurement invariance measures the same attribute in the same way in different subpopulations; hence, such testing procedures, although not infallible, can at least be considered to bring an element of fairness to the measurement procedure. Mathematically, the requirement of measurement invariance means that the function that relates psychological abilities to test scores should be invariant over groups (Mellenbergh, 1989; Meredith, 1993). In the past decades, several methods that allow for the detection of violation of this requirement have been developed (e.g., see Holland & Wainer, 1993; Millsap & Everson, 1993), and these methods are now commonly used in large testing programs. For instance, measurement invariance is routinely investigated by the Educational Testing Service, which is responsible for the widely used SAT.

It is widely believed, both in psychometric circles and beyond, that if measurement invariant tests are used, then selection processes are fair to different groups of people; in fact, as indicated above, this is one of the main motivations for their development. However, it is questionable whether this belief is correct. In a truly fair selection procedure, one would 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 efficiency (in terms of sensitivity and specificity) is invariant. Call this property *selection invariance*.

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 implies 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 that 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 efficiency across groups, and found that the more measurement invariance was violated, the greater were the differences in test efficiency. 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 efficiency would occur.

This intuition, however, is incorrect. The present paper shows that, for 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, while test efficiency 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 satisfied, the selection procedure must assign different odds of being selected to people with the same ability, but different group membership, in which case the test cannot be measurement invariant.

In sum, in many cases in which groups differ in measured ability, a test cannot exhibit both measurement invariance and selection invariance. The implications of this situation are serious, and will be discussed below. First, we discuss the relation between measurement invariance and selection invariance in greater detail. Because we focus on the situation in which measurement invariance holds, the present paper can be viewed as a companion paper to Millsap and Kwok (2004), who investigated essentially the same situation for the case in which measurement invariance is violated.

### Measurement invariance and selection invariance

The basic idea of psychometric selection is to identify suitable candidates on the basis of a test score. We will assume that the suitability of an individual depends on a single continuous latent variable  $\theta$ . Define a suitable candidate as an individual  $i$  with  $\theta_i \geq \theta_c$ , where  $\theta_c$  is the latent cutoff that separates individuals who one wants to reject from those one wants to select. Because the ability  $\theta$  is not directly observable, people are selected on the basis of a test score that is viewed as an observable indicator of  $\theta$ ; denote this test score  $X$ . We will assume that the latent variable is distributed according to the normal probability density function, denoted  $f(\cdot)$ , with mean  $\mu_g$  and standard deviation  $\sigma_g$  in group  $G = g$ , so that

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

Further, we assume that the test scores  $X$  are linearly related to the abilities  $\theta$  according to the regression line  $E(X|\theta) = a_g\theta$ , where  $a_g > 0$  is the regression coefficient for group  $g$  and  $\epsilon$  denotes the residual or error score:

$$X = a_g\theta + \epsilon. \quad (2)$$

Finally, we assume that the errors  $\epsilon$  are normally distributed with variance  $\sigma_\epsilon^2$  constant across levels of  $\theta$  (i.e., homoscedasticity).

These assumptions are consistent with the single factor model for continuous item responses (Mellenbergh, 1994); they will also approximate the common situation in which people are selected on the basis of a total test score composed of dichotomous items, which are considered to measure a single dominant continuous latent variable, provided that the number of items is reasonably large so that the relation between the total score and the latent variable is approximately linear.

Given these assumptions, the test score of an individual in group  $g$  with a certain true ability  $\theta$  is a random draw from the probability density function:

$$p(X|\theta) = f_{a_g\theta, \sigma_{\epsilon_g}}(X). \quad (3)$$

The selection procedure operates on the bivariate distribution of  $X$  and  $\theta$ , which for a single group is graphically represented in Figure 1. The objective of a selection procedure is to select those individuals whose ability  $\theta$  exceeds the threshold  $\theta = \theta_c$ ; in this paper we use the convention that  $\theta_c = 0$ . Individuals whose ability exceeds this threshold are called suitable, denoted  $S$ , others are unsuitable,  $\neg S$ . A standard selection procedure based on an ability test selects those individuals whose test score exceeds a certain threshold value  $X_c$ . Throughout this paper we adopt the convention that  $X_c = 0$ . Individuals whose test score exceeds this value and pass the test are accepted, denoted  $A$ , others are rejected,  $\neg A$ .

In any selection procedure, it may happen that an individual who is not suitable, having  $\theta < 0$ , passes the test by sheer luck,  $X \geq 0$ . In that case, the individual is accepted yet unsuitable:  $A \cap \neg S$ . Similarly, an individual who is suitable,  $\theta \geq 0$ , due to whatever circumstance, may not pass the test,  $X < 0$ . The individual is then suitable yet rejected:  $S \cap \neg A$ . The four quadrants in Figure 1 are thus occupied by four combinations of suitable and unsuitable accepted and rejected individuals (true positives, false positives, true negatives, and false negatives). The efficiency 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.

---

Insert Figure 1 about here

---

*The two group scenario*

The primary focus of this paper 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 paper.

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 groups differences in the distribution of  $\theta$ , 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  $\theta$ , 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  $\theta$ , the probability distribution or density of the observed scores does not depend on group membership. Thus, measurement invariance is defined by the restriction

$$p(X|\theta) = p(X|\theta \cap g). \quad (4)$$

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 (4) comes down to the equality of both  $a_g$  and  $\sigma_{\epsilon_g}$  across the levels of the grouping variable  $G$ . If these requirements are satisfied, then the selection situation for a two group scenario may be depicted as in Figure 2: the bivariate distributions of  $X$  and  $\theta$  lie on the same regression line.

---

Insert Figure 2 about here

---

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:

$$p(A|\theta) = p(A|\theta \cap g). \quad (5)$$

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 non-suitability, given rejection), sensitivity (the probability of acceptance, given suitability) and specificity (the probability of rejection, given non-suitability) be equal across groups. If these probabilities are equal across groups, then the test works equally well in each group; i.e., it has the same efficiency. This is what we call *selection invariance*. Formally, it is expressed by the following equalities:

$$\textit{Invariant Positive Predictive Value} \quad p(S|A \cap g) = p(S|A), \quad (6)$$

$$\textit{Invariant Negative Predictive Value} \quad p(\neg S|\neg A \cap g) = p(\neg S|\neg A), \quad (7)$$

$$\textit{Invariant Sensitivity} \quad p(A|S \cap g) = p(A|S), \quad (8)$$

$$\textit{Invariant Specificity} \quad p(\neg A|\neg S \cap g) = p(\neg A|\neg S). \quad (9)$$

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

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, i.e., 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:

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

Both the numerator and the denominator in this equation are integral expressions involving both the distribution of group  $g$  over the ability

parameter  $\theta$  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  $\mu_g$ , the mean of the distribution of group  $g$  over the ability: the higher the mean of the group's distribution function over  $\theta$ , the higher the probability of suitability given acceptance. For two groups  $H$  and  $L$  we therefore obtain:

$$\mu_H > \mu_L \quad \Rightarrow \quad p(S|A \cap H) > p(S|A \cap L). \quad (11)$$

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.<sup>1</sup>

We may wonder whether a similar result holds for test sensitivity, i.e., the probability of acceptance given suitability:

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

This is indeed so, as can be derived in roughly the same way. We have already established that  $p(S|A \cap g)$  gets larger with larger  $\mu_g$ . Therefore, to establish that  $p(A|S \cap g)$  gets larger with larger  $\mu_g$ , it suffices to show that the problem discussed above is symmetric with respect to interchanging the variables  $X$  and  $\theta$ .

There is an intuitive argument for the symmetry of the selection problem. First, the linear dependence of the test scores on ability entails a linear dependence of ability on test scores, with a regression line  $E(\theta|X) = \frac{X}{a}$  and residual errors distributed according to  $f_{\frac{X}{a}, \frac{\sigma_\epsilon}{a}}(\theta)$ . Second, given the linear dependence of test scores on ability, the normal distribution of the group over the ability parameter,  $f_{\mu_g, \sigma}(\theta)$ , translates into a normal distribution over the test scores,  $f_{m_g, s}(X)$  with mean  $m_g = a\mu_g$  and standard deviation  $s = f(\sigma_\epsilon, \sigma, a)$ , a function independent of  $\mu_g$ . This means that we can derive the same probabilistic inequalities after interchanging the variables  $X$  and  $\theta$ . A more detailed argument for both these claims is given in Appendix B.

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:

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

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:

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

and

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

This shows that, in a selection procedure that conforms to the present assumptions of a) multivariate 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, while 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.

#### *Generalization of the impossibility result*

In real world situations, groups may differ both in mean ability and in variance. However, the foregoing was based on the assumption that the variances of the latent variable distributions are equal across groups. We now relax this assumption. A direct consequence of doing this is that 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 investigate how the relation between measurement invariance and selection invariance behaves as a function of these differences.

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 obtains; i.e., 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,

$$\theta \rightarrow \alpha\theta, \quad (16)$$

$$a \rightarrow \frac{a}{\alpha}. \quad (17)$$

for  $\alpha \in (0, \infty)$ . Second, considering the transformation (17) while all else is kept fixed, both sensitivity and positive predictive value get larger for larger regression coefficients  $a$ . Both observations are proved in Appendix C.

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  $f_{a\theta,\sigma_\epsilon}(X)$  and denote the corresponding probability assignment over  $X$  and  $\theta$  by  $p_{a,\sigma_\epsilon}$ , or  $p_a$  for short. The first observation then comes down to the following statements:

$$\begin{aligned} p_a(S|A \cap g_{\mu,\sigma}) &= p_{\frac{a}{\alpha}}(S|A \cap g_{\alpha\mu,\alpha\sigma}), \\ p_a(A|S \cap g_{\mu,\sigma}) &= p_{\frac{a}{\alpha}}(A|S \cap g_{\alpha\mu,\alpha\sigma}). \end{aligned} \quad (18)$$

The second observation comes down to the following two statements:

$$\alpha < 1 \quad \Rightarrow \quad \begin{cases} p_{\frac{a}{\alpha}}(S|A \cap g_{\mu,\sigma}) > p_a(S|A \cap g_{\mu,\sigma}), \\ p_{\frac{a}{\alpha}}(A|S \cap g_{\mu,\sigma}) > p_a(A|S \cap g_{\mu,\sigma}). \end{cases} \quad (19)$$

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

Consider the selection problem for two groups,  $g_{\mu,\sigma}$  and  $g_{\mu',\sigma'}$  with  $\sigma \neq \sigma'$ , and a test characterized by a regression coefficient  $a$  and an error variance  $\sigma_\epsilon$ . We assume nothing on the order of  $\mu$  and  $\mu'$ , and we hold on to measurement invariance, so that  $a$  and  $\sigma_\epsilon$  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 (16) and (17) with  $\alpha = \frac{\sigma'}{\sigma}$  so that  $\alpha < 1$ . We choose to transform in the group with the larger standard deviation. For this group, equation (18) then gives us

$$\alpha = \frac{\sigma'}{\sigma} \quad \Rightarrow \quad \begin{cases} p_a(S|A \cap g_{\mu,\sigma}) = p_{\frac{a}{\alpha}}(S|A \cap g_{\alpha\mu,\alpha\sigma}), \\ p_a(A|S \cap g_{\mu,\sigma}) = p_{\frac{a}{\alpha}}(A|S \cap g_{\alpha\mu,\alpha\sigma}). \end{cases} \quad (20)$$

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 (19) we may then derive the conditional inequalities

$$\alpha < 1 \quad \Rightarrow \quad \begin{cases} p_{\frac{a}{\alpha}}(S|A \cap g_{\alpha\mu,\sigma'}) > p_a(S|A \cap g_{\alpha\mu,\sigma'}), \\ p_{\frac{a}{\alpha}}(A|S \cap g_{\alpha\mu,\sigma'}) > p_a(A|S \cap g_{\alpha\mu,\sigma'}). \end{cases} \quad (21)$$

Recall that  $\sigma' = \alpha\sigma$ . Therefore the expressions on the right side of Equation (20) are identical to the expressions on the left side of Equation (21).

Because of the employed transformations, the expressions on the right side of Equation (21) have variance  $\sigma'$  and regression parameter  $a$ , 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:

$$\alpha\mu > \mu' \quad \Rightarrow \quad \begin{cases} p_a(S|A \cap g_{\alpha\mu,\sigma'}) > p_a(S|A \cap g_{\mu',\sigma'}), \\ p_a(A|S \cap g_{\alpha\mu,\sigma'}) > p_a(A|S \cap g_{\mu',\sigma'}). \end{cases} \quad (22)$$

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 (20) and (21) is a decrease of sensitivity and positive predictive value, as can be seen from the inequalities in (21). Thus, if the inequalities in Equation (22) 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,  $p_a(A|S \cap g_{\mu,\sigma}) = p_{\frac{a}{\alpha}}(A|S \cap g_{\alpha\mu,\sigma'}) > p_a(A|S \cap g_{\alpha\mu,\sigma'}) \geq p_a(A|S \cap g_{\mu',\sigma'})$ . Hence it follows that  $p_a(A|S \cap g_{\mu,\sigma}) > p_a(A|S \cap g_{\mu',\sigma'})$ , 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 (20) to (22) and collect the conditions. We then obtain

$$\sigma' < \sigma \quad \text{and} \quad \frac{\sigma'}{\sigma}\mu \geq \mu' \quad \Rightarrow \quad \begin{cases} p_a(S|A \cap g_{\mu,\sigma}) > p_a(S|A \cap g_{\mu',\sigma'}), \\ p_a(A|S \cap g_{\mu,\sigma}) > p_a(A|S \cap g_{\mu',\sigma'}). \end{cases} \quad (23)$$

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

$$\sigma' < \sigma \quad \text{and} \quad \frac{\sigma'}{\sigma}\mu \leq \mu' \quad \Rightarrow \quad \begin{cases} p_a(\neg S|\neg A \cap g_{\mu,\sigma}) < p_a(\neg S|\neg A \cap g_{\mu',\sigma'}), \\ p_a(\neg A|\neg S \cap g_{\mu,\sigma}) < p_a(\neg A|\neg S \cap g_{\mu',\sigma'}). \end{cases} \quad (24)$$

For this we employ the Equations (20) to (22), but we transform  $\theta \rightarrow -\theta$ .

These results widen the scope of the inequalities derived in the previous section; if we choose  $\sigma = \sigma'$ , we obtain the original inequalities. Equation (23) says that, if the transformation of the latent variable leads to a scenario in which the transformed mean lies above the mean of the other group, then sensitivity and positive predictive value will be larger in the group with the larger variance.<sup>2</sup> Equation (24) implies that, if the transformed mean lies below the 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  $\frac{\sigma'}{\sigma}$ ,  $\mu$  and  $\mu'$ , 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  $\alpha \ll 1$  while  $\alpha\mu < \mu'$ , 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 enough 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 that, for any two subpopulations in  $H$  and  $L$  with equal  $\theta$ , the probability of passing the test given their level of ability is identical (*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. In the present case, we have independence of acceptance and group membership in all subpopulations with equal  $\theta$ , but aggregating over different values of  $\theta$  destroys this independence. This happens as follows. 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  $\theta$  which 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 ability, the chance of such an error becomes increasingly small. This can be seen from the Gaussian on the right.

---

Insert Figure 3 about here

---

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 towards the left of the figure than the group of suitable members from  $H$ . 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.<sup>3</sup> 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  $H$ .

To understand why groups differ in positive predictive value, i.e., in the probability of suitability given acceptance, it is instructive to focus on the regression of  $\theta$  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  $\theta$  is the same for members of different groups, the expected value of  $\theta$  given a particular test score  $X$  is lower for those in the low-scoring 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 Kelley's 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  $\theta$  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’); this, in turn, is a direct consequence of the situation that leads to Simpson’s paradox as discussed above.

### 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 one standard deviation difference in mean latent ability between groups and equal group size. We varied the selection ratio as the top 5%, top 10%, top 25% and top 50% (these percentages apply to the combined population). For each of these scenarios we examine 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); we varied this parameter by adding different amounts of white noise to the test scores, corresponding to different values of  $\sigma_\epsilon$  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.

---

Insert Figure 4 about here

---

---

Insert Figure 5 about here

---

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%; and 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, while the converse is true for the probability of false negatives. In the right panel of Figure 4, it can further be seen that the differences in test sensitivity are relatively stable across different levels of reliability, while 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  $\theta$ , the difference remains constant over all reliability ranges (after an initial divergence from the point of perfect reliability). On the other hand, the effect of differential regression towards 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 paper 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 paper 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 while they should not have been. There are reasons to suspect that such a situation does, in fact, obtain 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 drop-out rates among selected members from minority groups. Such increased drop-out rates have indeed been observed. In differential prediction studies, criterion performance (e.g., study success as reflected in freshmen GPA) 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 drop-out 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 U.S., 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). Since 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; and 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 the present paper. Thus, 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*

The present paper 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 to individuals. 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 fair to groups. This condition applies to aggregate populations and therefore is best guaranteed by selection invariance. The present paper elaborates on the conflicting implications of these requirements. While we think that this elucidates 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 o.k. 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.’

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 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 paper 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; i.e., 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  $\theta$  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 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 since the persons located there do not play a significant role in producing differences in test efficiency. 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; and since 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  $\sigma_\epsilon \rightarrow \alpha\sigma_\epsilon$  is equivalent to transformation (17). 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 since they are defined on different sets of items.

Local increases in reliability would be effective in diminishing the problem and would not share the drawbacks of the first two options. Because, with respect to selection invariance, the most important region of  $\theta$  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  $\theta$  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. This point is defined by the selection ratio the selecting institution entertains, together with the mean and variance of the ability distribution in the applicant population, and hence will differ across institutions.

---

Insert Figure 6 about here

---

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. This 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  $\theta$ , 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 the present paper 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 paper may be addressed in future research.

First, we have assumed multivariate normality throughout this paper, 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 multivariate normal. The influence of nonlinearity and non-normality on selection properties thus requires attention.

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 paper. This could happen if group differences are reversed on different variables, so that the group which 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.

### References

- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.
- Holland, P. W. & Wainer, H. (1993). *Differential Item Functioning* (pp. 197-239). Hillsdale, New Jersey: Erlbaum.
- Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: Macmillan.
- 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., and Werts, C. E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement, 8*, 1-4.
- 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., & 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, 9*, 93-115.
- Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.
- Reilly, R. R. (1973). A note on minority group bias studies. *Psychological Bulletin, 80*, 130-133.
- 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.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613-629.
- Wainer, H. (2000). Kelley's paradox. *Chance, 13*, 47-48.
- Wainer, H. (2005). *Graphic discovery*. Princeton: 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.
- Willingham, W. W., Pollack, J. M., and 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

$$\begin{aligned}
 p(A|\theta \cap g) &= p(X \geq 0|\theta \cap g) \\
 &= \int_0^\infty f_{a_g\theta, \sigma_{\epsilon_g}}(X) dX \\
 &= \int_{-\infty}^{a_g\theta} f_{0, \sigma_{\epsilon_g}}(X) dX = N_{0, \sigma_{\epsilon_g}}(a_g\theta), \tag{25}
 \end{aligned}$$

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  $\theta$ ,  $p(A|\theta \cap g)$  is the same for each value of  $G$  thus comes down to the requirement that  $N_{0, \sigma_{\epsilon_g}}(a_g\theta)$  is the same for each value of  $G$ , which is true by assumption, since  $a_g$  and  $\sigma_{\epsilon_g}$  are constant.

Using expression (25), we can write out the numerator and the denominator in equation (10) as

$$p(S \cap A|g) = \int_0^\infty f_{\mu_g, \sigma}(\theta) N_{0, \sigma_\epsilon}(a\theta) d\theta, \tag{26}$$

$$p(A|g) = \int_{-\infty}^\infty f_{\mu_g, \sigma}(\theta) N_{0, \sigma_\epsilon}(a\theta) d\theta, \tag{27}$$

the sole difference between numerator and denominator being the integration boundaries. Note that, because of the assumption of measurement invariance, the parameters  $a$  and  $\sigma_\epsilon$  do not carry an index  $g$ . Only  $\mu_g$  varies over the groups.

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

$$\frac{d}{d\mu_g} [p(S|A \cap g)] > 0. \tag{28}$$

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  $V$ . In other words, the procedure is then not selection invariant due to an inequality of positive predictive value.

To derive equation (28), first recall the differentiation rule for quotients,

$$\frac{d}{dz} \left[ \frac{A}{B} \right] = \frac{\frac{dA}{dz} B - A \frac{dB}{dz}}{\left( \frac{dB}{dz} \right)^2}. \tag{29}$$

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

$$p(A|g) \frac{d}{d\mu_g} [p(S \cap A|g)] - p(S \cap A|g) \frac{d}{d\mu_g} [p(A|g)] > 0 \quad (30)$$

to derive the inequality (28). Rewriting this we find

$$\frac{dp(S \cap A|g)/d\mu_g}{p(S \cap A|g)} > \frac{dp(A|g)/d\mu_g}{p(A|g)}. \quad (31)$$

The remainder of the derivation focuses on this inequality.

To arrive at inequality (28), note first that we may apply the differentiation operation within the integral expressions. From the functional form of the normal density,

$$f_{\mu_g, \sigma}(\theta) \propto \exp -\frac{(\theta - \mu_g)^2}{2\sigma^2}, \quad (32)$$

we can derive that

$$\frac{df_{\mu_g, \sigma}(\theta)}{d\mu_g} = \left( \frac{\theta - \mu_g}{\sigma} \right) f_{\mu_g, \sigma}(\theta). \quad (33)$$

Applying this to the terms in the inequality (31), eliminating the factor  $1/\sigma$  on either side, we obtain

$$\frac{\int_0^\infty (\theta - \mu_g) fN d\theta}{\int_0^\infty fN d\theta} > \frac{\int_{-\infty}^\infty (\theta - \mu_g) fN d\theta}{\int_{-\infty}^\infty fN d\theta}, \quad (34)$$

where we have written  $fN$  to abbreviate  $f_{\mu_g, \sigma}(\theta)N_{0, \sigma_\epsilon}(a\theta)$ . The integration variable  $\theta$  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 (28) to hold, it is sufficient that the following inequality holds:

$$\frac{\int_{\mu_g}^\infty \theta' fN d\theta'}{\int_{\mu_g}^\infty fN d\theta'} > \frac{\int_{-\infty}^\infty \theta' fN d\theta'}{\int_{\mu_g}^\infty fN d\theta'}. \quad (35)$$

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

$$E_{[\mu_g, \infty)}[\theta'] > E_{(-\infty, \infty)}[\theta']. \quad (36)$$

But this inequality is evident: the expectation value of  $\theta'$  over the interval  $[\mu_g, \infty)$  is always strictly larger than the expectation value of  $\theta'$  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 problem with respect to $X$ and $\theta$

The selection problem is essentially characterized by a normal distribution of a group over the ability parameter  $\theta$  and a linear dependence linking the ability  $\theta$  to normal distributions over a test score  $X$ . We will show that the very 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  $\theta$ .

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

$$\begin{aligned}
 p_{\text{test}}(X, \theta) &= f_{a\theta, \sigma_\epsilon}(X) \\
 &= f_{0, \sigma_\epsilon}(X - a\theta) \\
 &= f_{0, \frac{\sigma_\epsilon}{a}}\left(\frac{X}{a} - \theta\right) \\
 &= f_{\frac{X}{a}, \frac{\sigma_\epsilon}{a}}(\theta).
 \end{aligned} \tag{37}$$

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  $\theta$  and by looking at normal distributions over  $\theta$  given  $X$ . The test may just as well be characterized by a linear dependence linking scores  $X$  to normal distributions over ability  $\theta$  as by the converse dependence.

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

$$p(\theta, X) = f_{\mu_G, \sigma}(\theta) f_{a\theta, \sigma_\epsilon}(X). \tag{38}$$

To establish the symmetry of the selection problem with respect to  $X$  and  $\theta$ , we must show that under the assumption of a certain test  $p_{\text{test}}$ , 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,  $s_L = s_H$ .

This requires us to compute the functional form of the marginal distribution for  $X$ . To this aim, write

$$\begin{aligned}
p(X) &= \int_{-\infty}^{\infty} p(X, \theta) d\theta \\
&= \int_{-\infty}^{\infty} f_{\mu_g, \sigma}(\theta) f_{a\theta, \sigma_\epsilon}(X) d\theta \\
&= \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} \left( \frac{\theta - \mu_g}{\sigma} \right)^2 - \frac{1}{2} \left( \frac{X - a\theta}{\sigma_\epsilon} \right)^2 \right] d\theta. \tag{39}
\end{aligned}$$

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

$$\begin{aligned}
p(X) &\sim \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} \theta'^2 - \frac{1}{2} \left( \frac{X - a(\sigma\theta' + \mu_g)}{\sigma_\epsilon} \right)^2 \right] d\theta' \\
&= \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} \theta'^2 - \frac{1}{2} \left( \frac{X - a\mu_g - a\sigma\theta'}{\sigma_\epsilon} \right)^2 \right] d\theta'. \tag{40}
\end{aligned}$$

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

$$\begin{aligned}
p(X) &\sim \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} \theta'^2 - \frac{1}{2} \left( X' - \frac{a\sigma}{\sigma_\epsilon} \theta' \right)^2 \right] d\theta' \\
&= \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} \theta'^2 - \frac{1}{2} (X' - \gamma\theta')^2 \right] d\theta' \tag{41}
\end{aligned}$$

Writing out this expression and collecting the terms we find

$$\begin{aligned}
p(X) &\sim \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} (1 + \gamma^2) \theta'^2 + \gamma\theta' X' - \frac{1}{2} X'^2 \right] d\theta' \\
&= \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} \left( (1 + \gamma^2)^{1/2} \theta' - \frac{\gamma}{(1 + \gamma^2)^{1/2}} X' \right)^2 \right. \\
&\quad \left. - \frac{\gamma^2}{1 + \gamma^2} X'^2 - \frac{1}{2} X'^2 \right] d\theta' \\
&= \int_{-\infty}^{\infty} \exp \left[ -\frac{1}{2} \left( (1 + \gamma^2)^{1/2} \theta' - \frac{\gamma}{(1 + \gamma^2)^{1/2}} X' \right)^2 \right] d\theta' \\
&\quad \times \exp \left[ -\left( \frac{1}{2} + \frac{\gamma^2}{1 + \gamma^2} \right) X'^2 \right]. \tag{42}
\end{aligned}$$

We can now substitute once again,  $\theta'' = (1 + \gamma^2)^{1/2} \theta' - \frac{\gamma}{(1 + \gamma^2)^{1/2}} X'$  and integrate out  $\theta''$ . Note that  $X'$  may here be treated as constant.

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

$$\begin{aligned}
p(X) &\sim \exp \left[ \left( \frac{1}{2} + \frac{\gamma^2}{1 + \gamma^2} \right) X'^2 \right] \\
&= \exp \left[ \left( \frac{X - a\mu}{\sigma_\epsilon} \right)^2 \left( \frac{1}{2} + \frac{\gamma^2}{1 + \gamma^2} \right) \right]. \tag{43}
\end{aligned}$$

The distribution over test scores comes out normally distributed with a mean  $m_g = a\mu_g$  and the rather elaborate standard deviation  $s = \sigma_\epsilon \left( \frac{1}{2} + \frac{\gamma^2}{1+\gamma^2} \right)^{-1/2}$  with  $\gamma = \frac{a\sigma}{\sigma_\epsilon}$ . 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  $\sigma$ ,  $\sigma_\epsilon$  and  $a$  are all identical for both groups, the standard deviation  $s$  is the same for both groups too.

### Appendix C

#### Scale transformations

We first establish the observation expressed in equations (??) and (??). After the combined transformations (16) and (17), the probability over  $X$  and  $\theta$  can be written as

$$\begin{aligned}
p(X, \theta) &= f_{\alpha\mu, \alpha\sigma}(\alpha\theta) f_{\frac{a}{\alpha}, \alpha\theta, \sigma_\epsilon}(X) \\
&\sim \exp\left[\frac{1}{2}\left(\frac{\alpha\theta - \alpha\mu}{\alpha\sigma}\right)^2\right] f_{a\theta, \sigma_\epsilon}(X) \\
&\sim \exp\left[\frac{1}{2}\left(\frac{\theta - \mu}{\sigma}\right)^2\right] f_{a\theta, \sigma_\epsilon}(X) \\
&\sim f_{\mu, \sigma}(\theta) f_{a\theta, \sigma_\epsilon}(X).
\end{aligned} \tag{44}$$

The scale transformation thus leaves the probability assignment over  $X$  and  $\theta$  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 (19). We first deal with the probability of acceptance given suitability,  $p(A|S \cap g)$ . Note that for a group  $G_{\mu, \sigma}$

$$\begin{aligned}
p_a(S|g_{\mu, \sigma}) &= \int_0^\infty f_{\mu, \sigma}(\theta) \left[ \int_{-\infty}^\infty f_{a\theta, \sigma_\epsilon}(X) dX \right] d\theta \\
&= \int_0^\infty f_{\mu, \sigma}(\theta) d\theta \\
&= 1 - N_{\mu, \sigma}(0),
\end{aligned} \tag{45}$$

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

$$\begin{aligned}
p_a(A \cap S|g_{\mu, \sigma}) &= \int_0^\infty f_{\mu, \sigma}(\theta) \left[ \int_0^\infty f_{a\theta, \sigma_\epsilon}(X) dX \right] d\theta \\
&= \int_0^\infty f_{\mu, \sigma}(\theta) N_{0, \sigma_\epsilon}(a\theta) d\theta.
\end{aligned} \tag{46}$$

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

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

$$\begin{aligned}
p_{\frac{a}{\alpha}}(A \cap S|g_{\mu,\sigma}) &= \int_0^{\infty} f_{\mu,\sigma}(\theta) N_{0,\sigma_\epsilon} \left( \frac{a}{\alpha} \theta \right) d\theta \\
&= \int_0^{\infty} f_{\mu,\sigma}(\theta) \left[ N_{0,\sigma_\epsilon}(a\theta) + \left( N_{0,\sigma_\epsilon} \left( \frac{a}{\alpha} \theta \right) - N_{0,\sigma_\epsilon}(a\theta) \right) \right] d\theta \\
&= p_a(A \cap S|g_{\mu,\sigma}) \\
&\quad + \int_0^{\infty} f_{\mu,\sigma}(\theta) \left[ N_{0,\sigma_\epsilon} \left( \frac{a}{\alpha} \theta \right) - N_{0,\sigma_\epsilon}(a\theta) \right] d\theta. \quad (47)
\end{aligned}$$

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

$$\alpha < 1 \quad \Rightarrow \quad p_{\frac{a}{\alpha}}(A \cap S|g_{\mu,\sigma}) > p_a(A \cap S|g_{\mu,\sigma}). \quad (48)$$

Because the denominator in the quotient of equation (12), as expressed in equation (45), does not depend on  $a$ , the above comes down to (19) for the probability of acceptance given suitability.

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

$$\frac{p_{\frac{a}{\alpha}}(A \cap S|g_{\mu,\sigma})}{p_{\frac{a}{\alpha}}(A|g_{\mu,\sigma})} > \frac{p_a(A \cap S|g_{\mu,\sigma})}{p_a(A|g_{\mu,\sigma})} \quad (49)$$

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

$$p_{\frac{a}{\alpha}}(A \cap S|g_{\mu,\sigma}) p_a(A \cap \neg S|g_{\mu,\sigma}) > p_a(A \cap S|g_{\mu,\sigma}) p_{\frac{a}{\alpha}}(A \cap \neg S|g_{\mu,\sigma}). \quad (50)$$

But because of equation (48), this holds if we have

$$p_a(A \cap \neg S|g_{\mu,\sigma}) \geq p_{\frac{a}{\alpha}}(A \cap \neg S|g_{\mu,\sigma}). \quad (51)$$

Writing integral expressions for these probabilities in much the same way as in (47), we find

$$\begin{aligned}
p_a(A \cap \neg S|g_{\mu,\sigma}) &= \int_{-\infty}^0 f_{\mu,\sigma}(\theta) N_{0,\sigma_\epsilon} \left( \frac{a}{\alpha} \theta \right) d\theta \\
&= \int_{-\infty}^0 f_{\mu,\sigma}(\theta) \left[ N_{0,\sigma_\epsilon} \left( \frac{a}{\alpha} \theta \right) + \left( N_{0,\sigma_\epsilon}(a\theta) - N_{0,\sigma_\epsilon} \left( \frac{a}{\alpha} \theta \right) \right) \right] d\theta \\
&= p_a(A \cap \neg S|g_{\mu,\sigma}) \\
&\quad + \int_{-\infty}^0 f_{\mu,\sigma}(\theta) \left[ N_{0,\sigma_\epsilon}(a\theta) - N_{0,\sigma_\epsilon} \left( \frac{a}{\alpha} \theta \right) \right] d\theta. \quad (52)
\end{aligned}$$

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

## Notes

<sup>1</sup>An intuitive way for the reader to see this is to try and draw a picture like Figure 2, where the bivariate distributions (ellipses) lie on the same regression line (measurement invariance) and the probability of suitability given acceptance (roughly, the ratio of the surface of the part of the ellipse, that lies in the top right quadrant, to the surface of the part, that lies above the line  $X = X_c$ ) is also the same (selection invariance); the reader will find that this ratio always comes out greater for group  $H$ .

<sup>2</sup>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  $\mu$  and  $\mu'$  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  $\mu'$  and  $\mu$  lie far enough below the threshold  $\theta = 0$  so that  $\frac{\sigma'}{\sigma} < \frac{\mu'}{\mu}$ .

<sup>3</sup>Note that the word ‘proportionally’ is very important here; 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 size of the groups is different, there may be more candidates from the larger group in every region of the selection space.

### Figure Captions

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

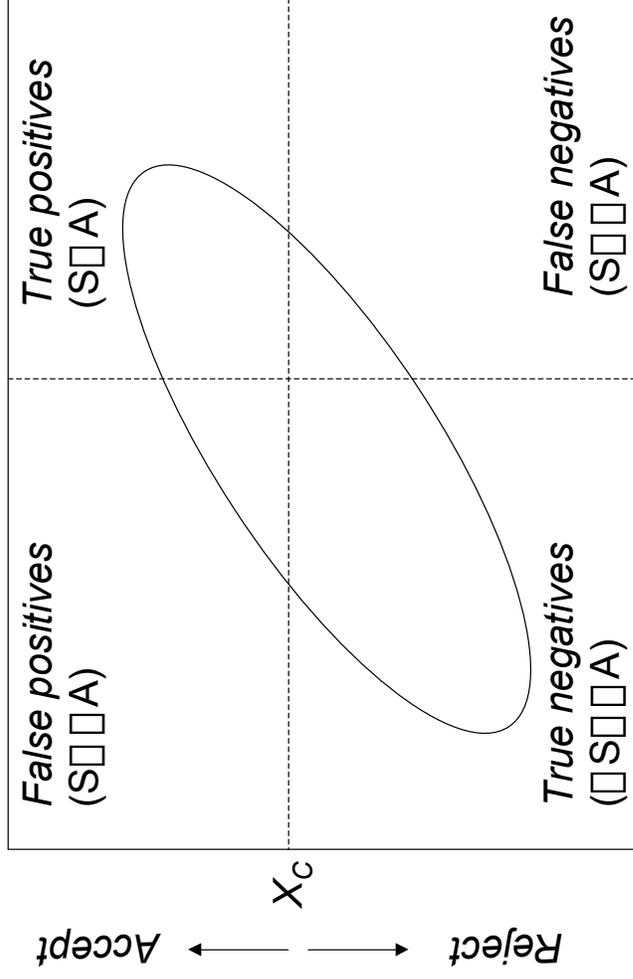
*Figure 2.* Selection in two groups  $H$  and  $L$  under measurement invariance.

*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.

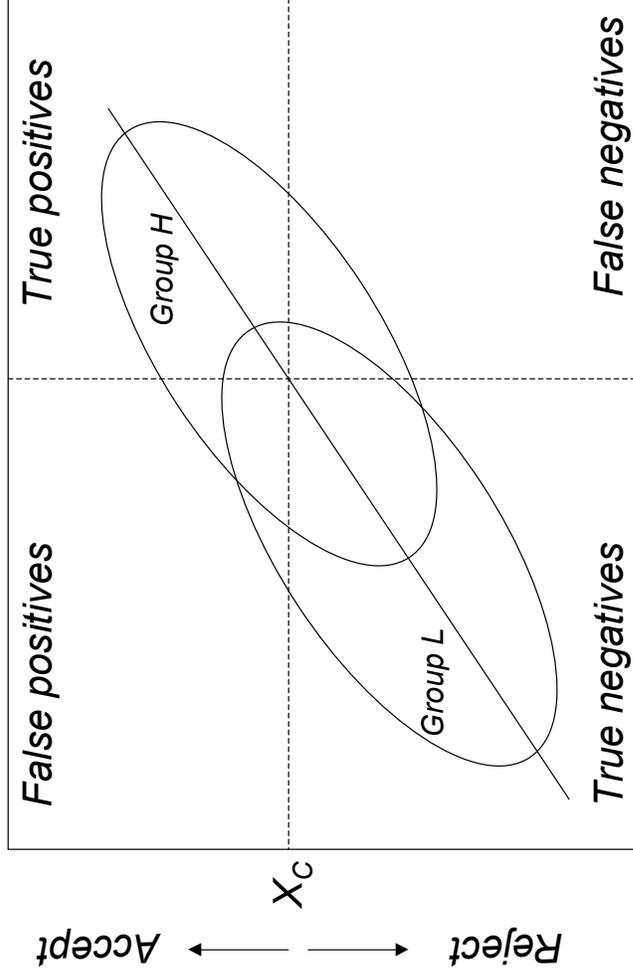
*Figure 4.* Differences in positive predictive value and sensitivity between groups. For different values of the selection ratio (top to bottom), the left panel shows positive predictive value and the right panel sensitivity for a low ( $L$ ) and high ( $H$ ) scoring group, as a function of reliability.

*Figure 5.* Differences in negative predictive value and specificity between groups. For different values of the selection ratio (top to bottom), the left panel shows negative predictive value and the right panel specificity for a low ( $L$ ) and high ( $H$ ) scoring group, as a function of reliability.

*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 efficiency across groups.



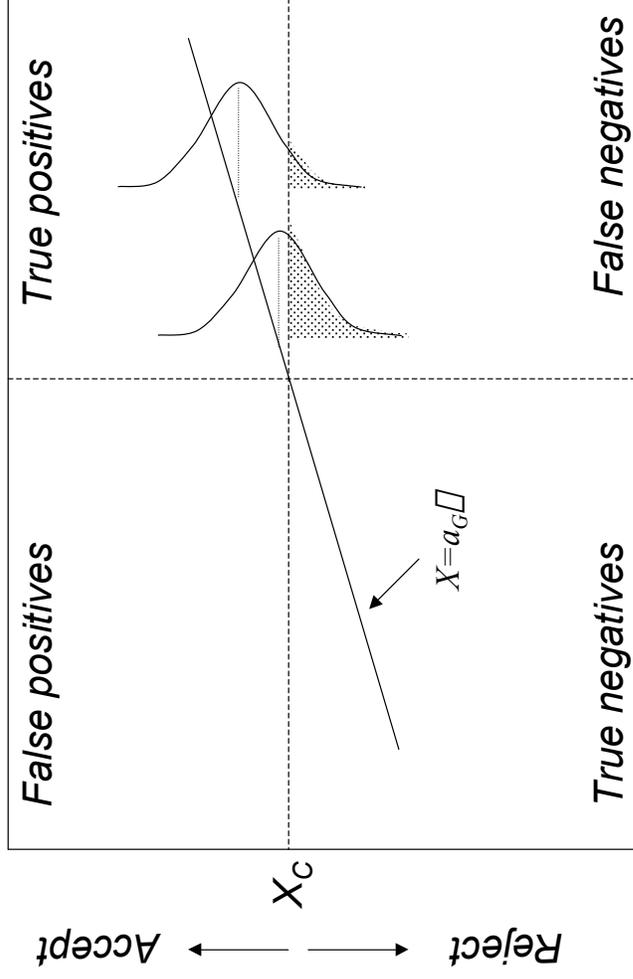
Unsuited ← ————— → Suited  
 Δ<sub>c</sub>



Reject  $\rightarrow$   
 Accept  $\leftarrow$

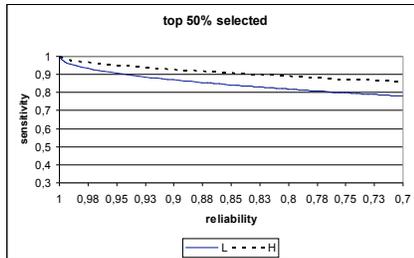
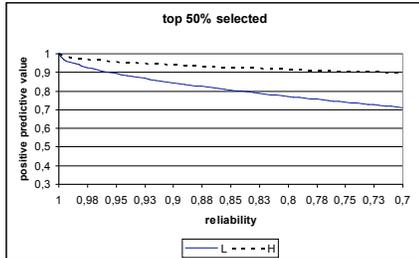
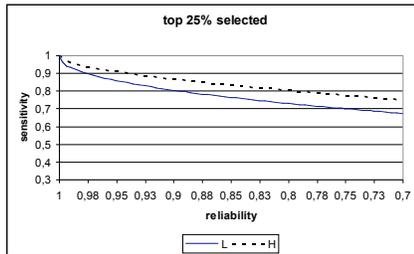
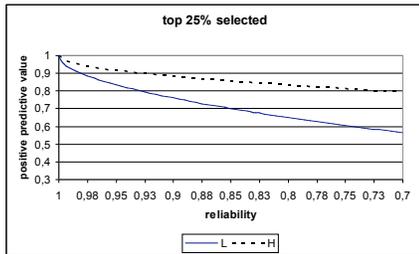
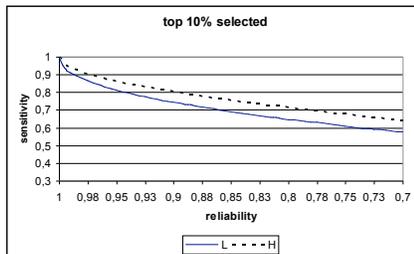
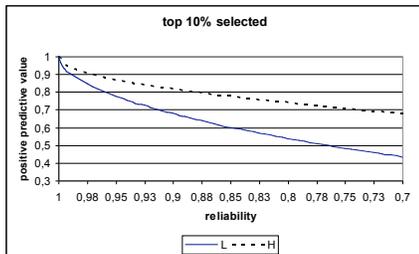
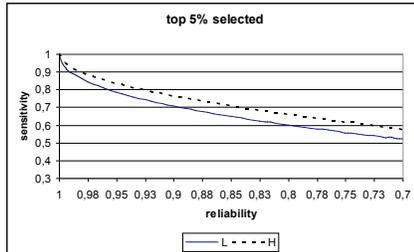
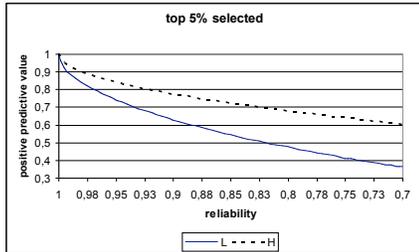
$\Delta c$

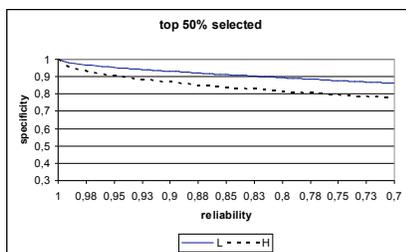
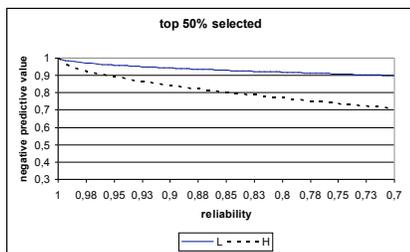
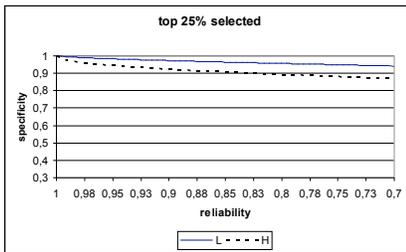
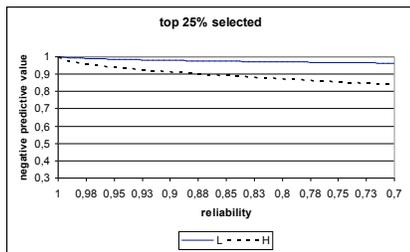
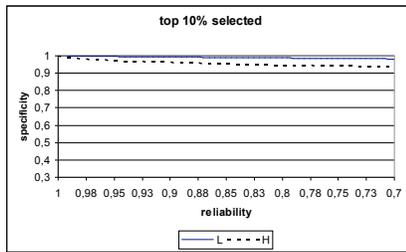
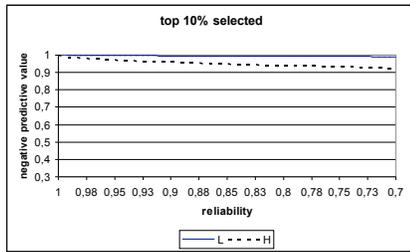
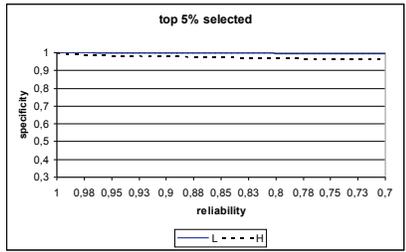
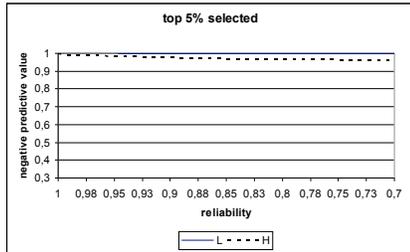
Unsuited  $\leftarrow$   $\rightarrow$  Suited

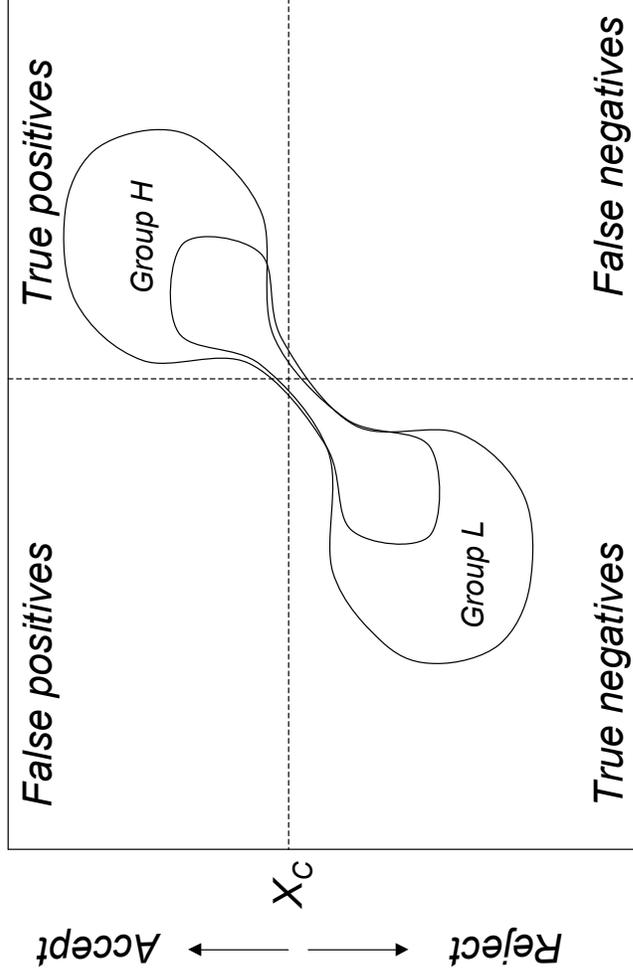


$\Delta c$

Unsuited ← → Suited







$\Delta_c$

Unsuited ← ————— → Suited