
Article:

Heesen, Remco ORCID: 0000-0003-3823-944X and Romeijn, Jan-Willem (2022)
Measurement invariance, selection invariance, and fair selection revisited. Psychological Methods, 28 (3). pp. 687-690. ISSN 1082-989X (In Press)

https://doi.org/10.1037/met0000491

Measurement Invariance, Selection Invariance, and Fair Selection Revisited

Remco Heesen*† Jan-Willem Romeijn†

Abstract

This note contains a corrective and a generalization of results by Borsboom et al. (2008), based on Heesen and Romeijn (2019). It highlights the relevance of insights from psychometrics beyond the context of psychological testing.

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

 The final article is available via its DOI: 10.1037/met0000491.

To contact the authors, please write to remco.heesen@uwa.edu.au or j.w.romeijn@rug.nl. RH's research was supported by the Dutch Research Council (NWO) under grant 016.Veni.195.141.

*Department of Philosophy, School of Humanities, University of Western Australia, Crawley, WA 6009, Australia.

†Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands.

Introduction

Borsboom et al. (2008) showed that we cannot combine two important fairness requirements for selection procedures. On the one hand we wish that selection procedures respect 'measurement invariance', meaning that they treat all individuals in the same way. Specifically, the probability for an individual to be selected may only depend on the latent ability that they are tested for; not on other characteristics of the individual. On the other hand they must be 'selection invariant', i.e., treat all groups within the population equally. In particular, if we partition the population into groups, we want the probabilities of misclassification to be the same in these groups. In short, the procedure may not discriminate individuals or groups on the basis of any other characteristic than the latent ability at issue.

Per Borsboom et al. (2008), fair selection is impossible if this is understood as a procedure that makes good on both requirements. This result is driven by the fact that groups will in general differ on the latent variable that is being selected for; the latent ability will correlate with other population characteristics. For example, the ability may be the command of a language, and this will give certain nationalities or ethnicities an edge. People from different groups thereby have different probabilities for being selected. The specific unfairness that ensues if we maintain measurement invariance is that the selection procedure does not work equally well for people from different groups: the procedure will incorrectly reject and accept members of these groups at different rates. Moreover, the arguably most impactful errors will be higher for more vulnerable groups.

This is not an isolated finding. We briefly review related results in the next section. We then provide a corrective and a generalization of the original results by Borsboom et al. (2008), based on Heesen and Romeijn (2019). This is timely as there is widespread concern over the transparency, adequacy and fairness of automated selection and classification procedures. Psychometrics is in an excellent position to advance this debate.

Similar Findings in Other Fields

Interestingly, the problem of implicit discrimination in Borsboom et al. (2008) was rediscovered, presumably independently, by Kleinberg et al. (2017) and Chouldechova (2017) in the context of the discussion on fairness in Artificial Intelligence (AI): a machine learner that fairly judges individuals can nevertheless discriminate at the group level. The result rightly received public attention (e.g., Angwin et al. 2016) and has inspired further discussion in the AI community (e.g., Barocas and Selbst 2016, Corbett-Davies et al. 2017, Barocas et al. 2019).

Since Kleinberg et al. (2017) and Chouldechova (2017) do not compare their results to those of Borsboom et al. (2008), we briefly do so here. At the heart of Kleinberg et al. (2017) and Chouldechova (2017)'s results is Bayes' theorem. In the terminology of Borsboom et al. (2008), which is explained in detail in the next section, they hold fixed the test sensitivity and specificity for two groups taking the same test (the likelihoods in Bayes' theorem), and then observe that if the base rates are different in the two groups, the positive and negative predictive values (the posteriors in Bayes' theorem) will be different as well. Thus, assuming that the base rates are different, if one wants equal predictive value across groups, one cannot have equal sensitivity and specificity, and vice versa. Because equal predictive values are identified as a necessary condition for fairness, requiring a test to be fair in this sense entails that it will not operate equally well for groups with different probabilities for the latent variable, i.e., different base rates, thereby making the test unfair in another sense. So Kleinberg et al. and Chouldechova are ultimately highlighting a form of base-rate neglect. In medicine and epidemiology, base-rate neglect has been widely reported and discussed (Casscells et al. 1978, Hoffrage et al. 2000).

This is superficially similar to the results in Borsboom et al. (2008) and to the generalization presented here. But there are notable differences. Borsboom et al. (2008) make the stronger assumption of measurement invari-

ance, which requires that for any fixed value of the underlying continuous latent characteristic the test operates equally well, independently of group membership. We get a correspondingly stronger result: assuming different distributions over the latent characteristic for the two groups, all four of the test error rates (sensitivity, specificity, and positive and negative predictive value) differ across groups. Once again, requiring the test to be fair in one sense, this time requiring measurement invariance, the test fails to operate equally well for groups that differ on the latent characteristic. However, in the result from Borsboom et al. (2008), the requirement of fairness is more specific, and the differences in test quality between the two groups are more dramatic.

In labor market economics, there has long been awareness that selection methods for job allocation can be discriminatory (cf. Fang and Moro 2011). The focus of these discussions is on the impact of factoring in other characteristics explicitly: if skills are latent, employers will seek proxies and end up selecting on the basis of demographics that are known to correlate with skills. However, we are not aware of results that match those by Borsboom et al. (2008), which pertain to the implicit discriminatory nature of selection.

In philosophy, the above insights have been imported in several debates. For example, Heesen and Romeijn (2019) apply the present results to scientific peer review, viewed as a selection procedure, and suggest that they may lead to a conservative bias. Stewart and Nielsen (2020) and Stewart (2020) take Kleinberg et al. (2017) as their starting point for, respectively, discussions of testimonial injustice and assessment in general. We will not review these discussions here. Rather, by drawing attention to these connections, we hope to stimulate further work applying insights from psychometrics in other fields.

To facilitate a more easy uptake of the results from Borsboom et al. (2008) in these various fields, in what follows we present the original results in corrected form, ironing out several inaccuracies. Next we present a more

general and hence more widely applicable version of the result due to Heesen and Romeijn (2019). The result has a more succinct proof than the original, which can be found in their appendix.

The Setup

We briefly rehearse the formal setup of Borsboom et al. (2008). Assume a population of individuals who differentially possess some latent characteristic θ. We distinguish two groups in the population (H and L). In each of these groups, the distribution of the latent characteristic is Gaussian (or 'normal'), but the mean and variance of this distribution may differ between the groups. We write μ_g for the mean and σ_g for the standard deviation of θ in group g (where either $g = H$ or $g = L$).

An individual is considered *suitable* if her individual value of the latent characteristic exceeds a threshold value θ_c. This yields a binary division of the population into suitable individuals ($\theta \geq \theta_c$, marked S) and unsuitable individuals ($\theta < \theta_c$, marked $\neg S$).

We would like to select suitable individuals, but we do not observe the value of the latent characteristic directly. We instead rely on a test. An individual's test score X is assumed to be linearly related to the latent characteristic, subject to some random error. More precisely, for an individual in group g,

$$X = \tau_g + \lambda_g \theta + \varepsilon_g, \tag{1}$$

where $\lambda_g > 0$ is the regression coefficient, τ_g the intercept, and ε_g the error term. Errors are assumed to be Gaussian with mean zero and standard deviation $\sigma_{\varepsilon,g}$.

We select individuals based on a threshold X_c on the test scores. An individual is *accepted* (event A) if $X \geq X_c$ and rejected (event $\neg A$) if $X < X_c$.

The requirement of *measurement invariance* states that, conditional on the true value of the latent characteristic θ, the probability distribution of

the test scores should be independent of group membership (formally: $X \mid \theta \sim X \mid \theta \cap g$ for any g). In the present context, this amounts to the requirement that $\tau_H = \tau_L$, $\lambda_H = \lambda_L$, and $\sigma_{\varepsilon,H} = \sigma_{\varepsilon,L}$. Borsboom et al. (2008, p. 79) further assume $\theta_c = X_c = \tau_H = \tau_L = 0$, supposedly without loss of generality.

Selection invariance requires instead that the error rates of the selection process are the same across groups. The relevant quantities here are the *positive predictive value* $p(S \mid A)$, the *negative predictive value* $p(\neg S \mid \neg A)$, the *sensitivity* $p(A \mid S)$, and the *specificity* $p(\neg A \mid \neg S)$.

Results

The main result of Borsboom et al. (2008) is that, in general, measurement invariance and selection invariance cannot be achieved simultaneously. More specifically, they claim to show two things.

First, if measurement invariance obtains and the two groups differ (only) in their means, i.e., $\mu_H > \mu_L$ and $\sigma_H = \sigma_L$, then selection invariance fails in that the test will have greater positive predictive value and sensitivity for group H:

$$p(S \mid A \cap H) > p(S \mid A \cap L) \quad \text{and} \quad p(A \mid S \cap H) > p(A \mid S \cap L). \tag{2}$$

As a corollary, group L will experience greater negative predictive value and specificity.

Second, if measurement invariance obtains and the two groups differ in both mean and variance, selection invariance fails as well. More specifically, if $\mu/\sigma \geq \mu'/\sigma'$ and $\sigma > \sigma'$ then positive predictive value and sensitivity will be greater for the group with mean μ and standard deviation σ. On the other hand, if $\mu'/\sigma' \geq \mu/\sigma$ and $\sigma > \sigma'$ then negative predictive value and specificity will be greater for the group with mean μ' and standard deviation σ'.

There are a couple of issues with the second result. First, the inequalities for negative predictive value and specificity are backwards. Contrary to the claim in the previous paragraph, if $\mu'/\sigma' \geq \mu/\sigma$ and $\sigma > \sigma'$ then negative predictive value and specificity will be greater for the group with mean μ and standard deviation σ. The numbered equation (17) in the original paper should actually read:

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

The second issue is that the paper encourages the slightly misleading suggestion that there is something special about the ratio between the mean and the standard deviation. But this turns out to be a consequence of the not completely innocent assumption that $\theta_c = 0$. If we repeat the proofs without that assumption (we do not provide this here, but the claim is a special case of the results discussed in the next section), we find that the direction of the inequalities depends on whether

$$\frac{\mu - \theta_c}{\sigma} \geq \frac{\mu' - \theta_c}{\sigma'} \quad \text{or} \quad \frac{\mu - \theta_c}{\sigma} \leq \frac{\mu' - \theta_c}{\sigma'}. \tag{3}$$

The two issues just identified are the only ones that affect the results of Borsboom et al. (2008). That said, there are some minor errors in the proofs of that paper that we wish to highlight while we are at it.

First, there is a typo in equations (A11) and (A12): all three occurrences of μ_g should in fact read $-\mu_g$.

Second, Borsboom et al. (2008, appendix B) aims to identify the marginal distribution of X (within a group g) and finds that X is normally distributed with mean $\lambda_g \mu_g$ and standard deviation

$$s = \sigma_{\varepsilon,g} \left(\frac{1}{2} + \frac{\lambda_g^2 \sigma_g^2}{\sigma_{\varepsilon,g}^2 + \lambda_g^2 \sigma_g^2} \right)^{-1/2}. \tag{B8}$$

This is the result of a small mistake earlier in the proof: in equation (B6) they write $\exp[-(\frac{1}{2}+\frac{\gamma^2}{1+\gamma^2})X'^2]$ which should have been $\exp[-(\frac{1}{2}-\frac{1}{2}\frac{\gamma^2}{1+\gamma^2})X'^2]$. The correct standard deviation (independently verified using moment-generating functions) is $\sqrt{\sigma^2_{\varepsilon,g}+\lambda^2_g\sigma^2_g}$.

Third, there are some typos in appendix D. In equation (D1) there is a minus sign missing inside both sets of square brackets. The second line of p. 98 refers to equation (C6) but should refer to equation (C4). And the line between equations (D7) and (D8) should refer to equation (D5) rather than equation (D1).

Improved Results

It turns out that the incompatibility between measurement invariance and selection invariance holds under more general assumptions than the ones made by Borsboom et al. (2008). We drop all structural assumptions about the test and instead assume just that it accepts or rejects individuals and is responsive to the latent characteristic. Measurement invariance then amounts to the requirement that $p(A \mid \theta) = p(A \mid \theta \cap g)$ for any g. Responsiveness to the latent characteristic is captured in the assumption that $p(A \mid \theta)$ is a strictly increasing function of θ.

We also drop the assumption that the latent characteristic follows a Gaussian distribution. We instead assume that there is a (shared) log-concave density function f such that, for each group g, the density function f_g is given by

$$f_g(\theta) = \frac{1}{\sigma_g} f\left(\frac{\theta - \mu_g}{\sigma_g}\right). \tag{4}$$

The family of log-concave density functions is a non-parametric family that includes, e.g., the uniform and exponential distributions (Saumard and Wellner 2014). Since the Gaussian density function is log-concave, this assumption is a strict generalization of the one made by Borsboom et al. (2008).

The density function f may exist for all real numbers (e.g., the Gaussian), on a half-line (e.g., the exponential), or a finite interval (e.g., the uniform). To avoid edge cases, we assume throughout this section that θ_c is chosen such that $0 < p(S \mid g) < 1$ for at least one group g.

The first result then generalizes as follows (Heesen and Romeijn 2019, theorem 3). Assuming measurement invariance, if $\mu_H > \mu_L$ and $\sigma_H = \sigma_L$, then

$$p(S \mid A \cap H) > p(S \mid A \cap L) \quad \text{and} \quad p(A \mid S \cap H) \geq p(A \mid S \cap L). \tag{5}$$

The latter inequality is strict unless the right tail of f is exponential. Under the same conditions we also have

$$p(\neg S \mid \neg A \cap L) > p(\neg S \mid \neg A \cap H) \quad \text{and} \quad p(\neg A \mid \neg S \cap L) \geq p(\neg A \mid \neg S \cap H). \tag{6}$$

The second result also generalizes once the factors mentioned in the previous sector are taken into account (Heesen and Romeijn 2019, theorem 5). If measurement invariance is satisfied, then

$$\frac{\mu_H - \theta_c}{\sigma_H} \geq \frac{\mu_L - \theta_c}{\sigma_L} \quad \& \quad \sigma_H > \sigma_L \Rightarrow \begin{cases} p(S \mid A \cap H) > p(S \mid A \cap L), \\ p(A \mid S \cap H) > p(A \mid S \cap L). \end{cases} \tag{7}$$

And conversely,

$$\frac{\mu_H - \theta_c}{\sigma_H} \geq \frac{\mu_L - \theta_c}{\sigma_L} \quad \& \quad \sigma_L > \sigma_H \Rightarrow \begin{cases} p(\neg S \mid \neg A \cap L) > p(\neg S \mid \neg A \cap H), \\ p(\neg A \mid \neg S \cap L) > p(\neg A \mid \neg S \cap H). \end{cases} \tag{8}$$

Thus the results from Borsboom et al. (2008) are ultimately seen to hold in a significantly more general mathematical setting.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *Propublica*. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, accessed January 31, 2022.

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

Barocas, S. and Selbst, A. D. (2016). Big Data's disparate impact. *California Law Review*, 104(3):671–732.

Borsboom, D., Romeijn, J.-W., and Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, 13(2):75–98.

Casscells, W., Schoenberger, A., and Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory results. *The New England Journal of Medicine*, 299:999–1001.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.

Fang, H. and Moro, A. (2011). Theories of statistical discrimination and affirmative action: A survey. In Benhabib, J., Bisin, A., and Jackson, M. O., editors, *Handbook of Social Economics*, volume 1, chapter 5, pages 133–200. Elsevier.

Heesen, R. and Romeijn, J.-W. (2019). Epistemic diversity and editor decisions: A statistical Matthew effect. *Philosophers' Imprint*, 19(39):1–20.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. *Science*, 290(5500):2261–2262.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Papadimitriou, C. H., editor, *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, pages 43:1–43:23.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8:45–114.

Stewart, R. T. (2020). Identity and the limits of fair assessment. Manuscript.

Stewart, R. T. and Nielsen, M. (2020). On the possibility of testimonial justice. *Australasian Journal of Philosophy*, 98(4):732–746.