When Does Measurement Invariance Matter?

Denny Borsboom, PhD

The question whether observed differences in psychometric test scores can be attributed to differences in the properties that such tests measure is relevant in many research domains; examples include the proper interpretation of differences in intelligence test scores across different generations of people,[1] gender differences in affectivity,[2] and crosscultural differences in personality.[3] This question also has generated some of the most conspicuous controversies in the social and life sciences, where the highest temperature in the many heated discussions around the topic has, without a doubt, been reached in the debate on IQ-score differences between ethnic groups in the United States.[4,5] Such debates are often unproductive because of a lack of unambiguous characterizations of concepts like "biased," "incomparable," and "culture-fair." Terms are easily coined, as is illustrated by Johnson's[6] count of no less than 55 types of measurement equivalence; however, it is often less easy to spell out their meaning in terms of their empirical consequences. However, without at least some degree of precision in one's conception of a term like "equivalence," it is difficult to have a scientifically productive debate, or even to agree on what aspects of empirical data are relevant for answering the questions involved.

It is for this reason that the establishment of concepts like measurement invariance and bias in an unambigous, formal framework with testable consequences[7–9] represents a theoretical development of great importance. Through this work, it has become clear that differences in raw scores (eg, IQ-scores) of different groups (eg, blacks and whites) cannot be used to infer group differences in theoretical attributes (eg, general intelligence) unless the test scores accord with a particular set of model invariance restrictions. Namely, the same attribute must relate to the same set of observations in the same way in each group. Statistically, this means that the mathematical function that relates latent variables to the observations must be the same in each of the groups involved in the comparison.[7,8] This idea has become known as the requirement of measurement invariance.

The theoretical definitions of measurement invariance and bias are very general, and apply to different models, such as item response theory (IRT) and factor models, in roughly the same way.[10,11] This does not hold for the empirical methods available for testing measurement invariance. In the past decades, psychometricians working on measurement invariance have produced many different statistical techniques to assess differential item functioning (DIF). These techniques usually employ different statistical assumptions, for instance, regarding the form of the relation between latent and observed variables and the shape of the population distribution on the latent variable, and employ different modeling strategies as well as selection criteria for flagging items as biased. For this reason, it is difficult to assess the consequences of choosing a particular technique; moreover, it is not always clear to what extent the choice of technique makes a difference with respect to the diagnosis of meaurement invariance and bias in applied situations.

For this reason, the articles on DIF collected here (by Crane et al;[12] Dorans and Kulick;[13] Jones;[14] Morales, Flowers, Gutierrez, Kleinman, and Teresi;[15] Edelen Orlando et al[16]) represent a useful project in the application of bias detection methods. Each set of authors analyzes the Mini-Mental State Examination (MMSE) for measurement invariance using the same data, albeit with different methods. Together, the articles provide a

From the Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Amsterdam, The Netherlands.
Reprints: Denny Borsboom, Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam. E-mail: d.borsboom@uva.nl.

ISSN: 0025-7079/06/4400-0176

sort of natural experiment on the robustness of biasing effects with respect to the choice of method. To what extent do these DIF detection methodologies converge in their conclusions regarding measurement invariance?

A CUP HALF FULL

Depending on one's perspective, the bias detection cup could be considered half full or half empty: The authors agree in their diagnosis on roughly half of the items (9 of 21). Items 1 (What is the year?), 4 (What day of the week is this?), 5 (What month of the year is this?), 12 (Serial 7s), and 21 (Copy design) are uniformly identified as being "pure" items, that is, as items not showing evidence of bias with respect to language in any direction. These items were used by Edelen Orlando et al[16] to establish a set of "anchor" items, which provided a matching criterion or benchmark for assessing the functioning of the other items. Thus, it does seem that their construction of an anchor set was successful, although it is somewhat surprising that Jones[14] diagnosed an additional anchor item (item 9) with bias.

The articles uniformly detect item bias in items 2 (What is the season?), 6 (What state are we in?), 7 (What city are we in?), and 17 (Repeating a phrase read by the interviewer). The first 3 of these items are easier for English-speaking patients, whereas the last is easier for Spanish-speaking patients. Finally, despite the presence of biasing effects at the item level, the total test score is virtually unbiased. Apparently, bias in both directions (ie, favoring English and Spanish speakers, respectively) averages out across items, leaving a net weight of 0 at the level of the total score. Thus, the MMSE provides a good illustration of bias cancellation. It is not clear how often bias cancellation occurs and, hence, it is interesting to see such a near-perfect instantiation of the phenomenon in a real research situation.

With respect to the remaining 12 items, the authors do not agree uniformly, likely because of variation in the employed criteria for DIF. These are indeed varied: Jones[14] uses improvement in model fit when biasing effects are added to the model in a stepwise fashion; Morales et al[15] use cut-off scores on NCDIF measures; Crane et al[12] use a Bonferroni correction in the detection of nonuniform DIF, but then switch to an effect size measure for the detection of uniform DIF; Dorans and Kulick[13] use ST-PDIF values; and Edelen Orlando et al[16] use conventional significance tests at alpha = 0.05 for the construction of an anchor test, but turn to a Benjamini-Hochberg correction when detecting DIF.

WILL THE TRUE CRITERION FOR DIF PLEASE STAND UP?

The fact that different criteria for DIF lead to different conclusions as to which items are biased is not suprising. However, it does raise a problem, because the choice of criterion will affect which items are flagged for DIF. So, which criterion should be used? Could there be such a thing as a "right" criterion?

It is doubtful whether there could be such a criterion, and the reason for this lies in a structural problem with the construction of concrete empirical criteria for the presence of DIF. This problem originates from the imposition of a categorical classification scheme (DIF/no-DIF) on what is essentially a matter of degree. DIF, usually viewed as the difference between parameters that govern the item response function in different groups, is defined on the continuum: It can theoretically assume every value of the real line. Measurement invariance, in contrast, occurs in only one situation, namely when the difference in question equals 0. All other values for the difference between parameters that govern the item response function, according to formal definitions, imply bias. This is because bias is defined simply as the negation of measurement invariance.[7,8] Thus, DIF is a matter of degree, but measurement invariance is not, because the former varies on the entire continuum, while the latter is defined as an ideal situation that corresponds to just one value of the difference between parameters, namely zero.[7,8] This type of hypothesis is called a point hypothesis. One can sense intuitively that the chances of encountering a true point hypothesis are slim, by considering that such a hypothesis says that 2 parameters differ by *exactly* 0.000 . . . , with an infinite number of zeroes behind the delimiter.

Statistical tests for detecting bias usually conceptualize the 'no-DIF' hypothesis as a null hypothesis, which is to be rejected if the *P* value associated with the data is below some chosen level of significance. However, like all point hypotheses, the "no-DIF" hypothesis usually will be false. It follows that with sufficiently large sample sizes, all items in all tests will probably be identified as biased with respect to all conceivable subpopulations. With more moderate sample sizes, the distinction between items with and without DIF is induced by a critical area of rejection, which is in turn a direct function of sample size and chosen significance level, both of which are arbitrary from a bias detection perspective.

If an effect size measure is used instead of a significance level, the question arises what amount of bias should be considered problematic. This matter is difficult to resolve, because even substantial levels of bias (in terms of effect sizes) at the item level need not interfere with research purposes. One example, nicely illustrated in the present articles, occurs when biasing effects cancel out. In that case, total test scores can sensibly be used for the comparison of populations, even though they are made up of items that are individually biased. Moreover, blind removal of the items with the most severe level of bias (for instance, the 4 items on which the present studies unanimously agree) may actually induce more, rather than less, bias at the test score level. This is because removing such items can disturb the equilibrium of biasing effects needed for cancellation to occur. In addition, blindly removing items may adversely affect content validity.

Evidently, whether bias is to be considered an important validity threat is not a straightforward function of *p*-values or effect sizes. The reason for this is that the importance of bias is partly a pragmatic issue: It depends on aspects of the research situation that are not statistical in nature, such as the purposes for which the test scores are being used.

THE PRAGMATIC DIMENSION OF BIAS DETECTION

Shifting the question from "is test X biased?" to "does the amount of bias in test X matter?" changes the nature of inquiry. Although the first question can be viewed as an empirical one, the second cannot be so construed. Whether bias matters depends not just on the amount of bias, but also on the purposes of the researcher and on the source of the biasing effect. For instance, in medical diagnosis, where test scores may affect individuals' lives directly, bias can be expected to generally be more pertinent than in research where one is merely interested in establishing the direction of a correlation between 2 constructs in different populations. Different combinations of the purposes of the researcher and the source of the biasing effect therefore suggest different courses of action with respect to the items in question. To clarify this problem situation, the relative importance of DIF is discussed in terms of different purposes for which test scores can be used. Three possible uses of the test scores are considered: comparing group means, investigating the relations between variables within groups, and the selection of individuals.

Comparing Means Between Groups

If the goal of the researcher is to compare group means on the basis of observed test scores, bias can be a serious problem. Unless biasing effects cancel, mean group differences in observed scores need not reflect differences in the latent variables of interest because the observed scores are confounded. Here, the size of the biasing effects is crucial. Now, it is not possible to say anything sensible about what amount of bias should, in general, be deemed "too much," without relating it to other factors in the research situation. This is because the importance of the biasing effects can only be assessed by considering their size in relation to the size of targeted effects, where the term "targeted effects" indicates the effects one is interested in from a substantive viewpoint. Suppose that the test scores are biased, but the amount of bias is an order of magnitude smaller than the targeted effects. In that case, there is no real risk of confounding, however large the size of the biasing effect is in absolute terms. For if the targeted effect obtains, then it will swap biasing effects in any direction; if it does not obtain, then the biasing effect will not be large enough to lead the researcher to the erroneous conclusion that the targeted effect exists. Similarly, small biasing effects may be negligible in describing relationships qualitatively, ie, in terms of their direction. Although there is no clear rule about how small the biasing effect should be in relation to the targeted effect for it to be negligible, this can be reasonably investigated by studying the robustness of effects under various levels of measurement bias.

Thus, it is not possible to identify a given amount of bias as problematic in all situations and circumstances. For however large the biasing effect may be, if the theoretically predicted effect is an order of magnitude larger, bias will not be a problem. Of course, one can only properly assess whether this is the case if one has a substantive theory that is sufficiently detailed to make predictions on effect sizes. If the theory is not capable of making predictions more specific than 'the effect will be different from zero', then even the smallest amount of bias can lead the researcher to the wrong conclusion. Therefore, if theories are incapable of making predictions in terms of effect sizes, which is the rule rather than the exception in the social and life sciences, measurement invariance is always an issue. This does not mean that full measurement invariance is invariably necessary; for partial invariance, as discussed by Gregorich[17] and Meredith,[11] allows for latent mean comparisons to be made without full measurement invariance. Rather, it means that insight into the presence and strength of biasing effects is a prerequisite for sensible group comparisons to be made, and therefore a serious modeling exercise is called for. Thus, it is safe to say that, at least in the social and life sciences, research on measurement invariance (rather than measurement invariance itself) must be seen as a prerequisite for any inference from observed mean differences to latent mean differences to be made. It is evident that many researchers who engage in comparisons of groups do not currently follow this recommendation; often, measurement invariance is tacitly assumed rather than investigated. Hopefully, the currently available methodology for investigating measurement invariance will change this situation, so that checking for measurement invariance will become part of the standard analyses in studies on group differences.

Comparing Within-Group Relations

If one's research interest is not to compare means between groups, but rather to investigate how variables are related within different groups, then bias may be entirely irrelevant. For instance, suppose one is interested in the causal effect of alcohol intake on cognitive functioning, and that one examines this effect in separate groups, say, in Spanish and English speakers. Further suppose one uses test scores known to be biased against English speakers, with the size of the biasing effect being in the same order of magnitude as the expected effect of alcohol intake. In that situation, the presence of a biasing effect precludes a sensible comparison of the means of English and Spanish speakers. However, this does not imply that one cannot use the test scores to investigate whether alcohol intake influences cognitive functioning in each of the groups separately; clearly, this remains possible.

Hence, if the sole purpose is to determine the ordering of persons, made within each of the groups separately (eg, level of cognitive functioning), and to relate this ordering to another ordering, made within each of the groups separately (eg, level of alcohol intake), the presence of DIF need not constitute a validity threat. Whether it actually is a validity threat partly depends on the source of the biasing effects. In particular, if bias is caused by multidimensionality, then it is likely to be a problem not only for mean comparisons across groups but also in relating variables within a group. If bias originates from different sources, like translation or differential interpretation of items, it need not be a validity threat.

To illustrate this, first consider a situation where items do not show DIF because they are multidimensional, but because they effectively constitute different items in different populations (see Borsboomet al[18]). In the MMSE, this situa-

tion occurs with some items because they have been literally translated: Although literally translated items may measure the same attribute in each group, they can have very different psychometric properties. A good example is the item that asks the patient to repeat the phrase "no ifs, ands, or buts". The translation of this item is literally the same, but psychometrically the translated item is so different that it almost instantiates a different item altogether. Consequently, it is no wonder that bias is found: I do not speak Spanish at all, but I am sure I could repeat "no hay peros que valgan" with less trouble than the English original, even if I had no clue what the words mean.

In this case, the source of the biasing effect is not multidimensionality; the item in question may be a measure of the same latent variable in each group, and it may be unidimensional within each group (ie, depend only on the latent variable of interest). This may be so, even though group differences on observed means cannot be attributed to mean differences on the latent variable. These differences originate because the item, although literally equivalent, effectively functions as a different item in both groups. If one is merely interested in ordering people in each group separately, then such items are likely to enhance rather than lower the precision with which this is done. In that case, there is no need to discard such items.

It is perhaps useful to shortly elaborate on this observation, because some researchers are of the opinion that that bias is synonymous with multidimensionality. Obviously, this is not the point of view to which I subscribe. In my view, the model presented by Shealy and Stout[19] correctly captures the relation between bias and multidimensionality. In their view, multidimensionality is not synonymous with bias, but a possible explanation for it. Now, if the bias is caused by multidimensionality, then groups must differ in their population distribution on the additional dimension measured by the test; otherwise this dimension cannot differentially affect the item responses. This means that the additional dimension must be distinct from, but associated with, group membership. There is no reason to suppose such a situation to obtain in the example of the translated item (see Borsboom et al[18] for other examples).

However, if the source of the biasing effect is that item responses do depend on a second latent attribute, not targeted by the researcher, that the groups possess in uneven amounts (ie, DIF is due to multidimensionality[19]), then the situation is different. For instance, consider the item 'What state are we in?' and its Spanish equivalent. The biasing effect on this item does not seem to be induced by the fact that the item is translated. It is more likely to depend on other attributes that the groups possess in uneven degrees, such as level of education. Now, if this occurs, then the test scores depend on more than the latent variable of interest. In that case, the test scores are likely to violate unidimensionality in each of the groups separately as well as in a multigroup analysis. And that means that we are also inducing bias in the ordering of people *within* groups. Suppose that the test measures level of education in addition to cognitive functioning, in the sense that, when the level of cognitive functioning is controlled for, an effect of education remains. Since level of education varies within groups as well as between them, lower educated Spanish people will receive lower scores than higher educated Spanish people, even if they have the same level of cognitive functioning. This means that any within-group relation found to exist between the total scores on the MMSE and another variable may be confounded, because it may actually reflect a relation that involves level of education rather than cognitive functioning.

Thus, it is possible to use raw scores on items, that are biased in a between-group sense, to investigate relations between variables within groups, but the conditions under which this can be done safely are limited. In particular, it is feasible only if the biasing effect is not caused by factors that are likely to confound within-group orderings of persons as well as between-group mean differences. One way to investigate whether this is the case is to fit unidimensional models in each of the groups separately, in addition to doing a multigroup analysis. If an item turns out to be biased in the multigroup model, but undimensionality is satisfied in each of the groups separately, then the item need not be discarded if tests are only used for ordering people within each of the groups separately, and not for the comparison of group means.

However, it must be noted that the quantitative comparison of the magnitude of effects across groups will usually be problematic in the presence of all kinds of biasing effects. A well known and relatively simple problem of this kind occurs, for instance, when tests have differential reliability across groups. In this case, the effects of interventions on the observed scores will usually be different, because regression coefficients will be attenuated to a larger degree in the group where reliability is lower. This may thus lead to spurious interactions. Hence one must be very careful in quantitatively comparing the magnitude of effects across groups.

Selection

The discussion has so far concentrated on research purposes that often serve a primarily scientific interest, or, in Meredith's[11] terms, are in the area of "basic" research. However, test scores on the MMSE are not just used for scientific purposes; they are also used to make decisions about individuals.

It has since long been acknowledged that, when tests are to be used in the selection of individuals rather than for scientific research on population characteristics (eg, means and correlations), they should conform to higher psychometric standards because of the danger of bias. These standards were originally set in terms of equal reliabilities and validity coefficients in different groups.[20] Since the development of the theory on measurement invariance, however, the stakes have been raised considerably. Meredith[8] defines fairness (in terms of equal conditional distributions of the predictors, given the true criterion score, across groups) and equity (as equal conditional distributions of the criterion, given the true predictor scores, across groups). These definitions seem reasonable. Meredith[9,11] has shown that, unless measurement invariance holds, fairness and equity cannot exist in principle. Thus, when the purpose of test use is the selection of individuals, measurement invariance is a necessary condition for fair selection procedures.

There seems to be little room for negotiation on the importance of measurement invariance in a selection context. Conditions that may lead a researcher to conclude that violations of measurement invariance do not pose a serious validity threat, given the research purposes, are not usually applicable to the situation where one is selecting individuals. A good example is bias cancellation. Biasing effects may cancel, but if they do, they cancel at the level of (conditional) population distributions, rather than at the level of the individual person. So, while it may be true that, if bias cancellation occurs, conclusions regarding group differences in means can remain valid even though some items are biased, this will not satisfy the individual patient who has been misdiagnosed because he or she happened to belong to a particular ethnic group.

The conditions to be met for fairness and equity to exist, as discussed by Meredith,[11] are in fact somewhat stricter than measurement invariance alone because measurement invariance is a necessary but not a sufficient condition for fairness and equity. Meredith's[9,11] work thus indicates that very careful psychometric analyses are crucial for developing fair and equitable selection procedures. Naturally, the point remains that measurement invariance will usually be violated to *some* degree—however small—so that every selection procedure will be also unfair to *some* degree (see also Millsap and Kwok[21]). Further, the requirements discussed by Meredith[11] are so strict that one might expect truly fair and equitable selection procedures to be rare, if they exist at all. However, this does not alter the point that, when the purpose of test use is selection, the *minimization* of bias must be considered a primary goal in developing and evaluating selection procedures. Millsap and Kwok[21] give an accessible treatment of the question to what extent violations of measurement invariance lead to bias in selection, that can be used by applied researchers to evaluate the severity of such violations for selection biases.

Of course, given that some degree of bias is likely to be present in any real selection situation, the question arises how high a degree of bias is still acceptable. This question is, in my view, very difficult to answer in general terms. Fortunately, however, such an answer may not always be necessary. Because selection often is unavoidable, abandoning one selection instrument commonly implies using another, so that the proper strategy is unambiguous: One should use the instrument that shows the least amount of bias. Needless to say, one cannot make an informed choice in this regard if one has no information on the degree to which different instruments are biased, which once more underscores the fact that investigating measurement invariance should be routinely done in every situation where fair selection is at stake.

DISCUSSION

Although the present collection of articles on DIF detection[12–16] shows considerable agreement on the size and direction of biasing effects in the MMSE, some disagreement is also evident. This disagreement is likely to be due to the use of different empirical criteria for the diagnosis of DIF. Instead of searching for the 'right' criterion, the present commentary has attempted to go beyond the question whether items have DIF, by evaluating when and how DIF matters in different situations. When comparing means, biasing effects may be negligible if they are small, and in this case violations of measurement invariance need not be a serious validity threat. However, it is not the absolute size of the biasing effects that matters, but their size relative to targeted effects; hence, in the absence of a theory that says how large the targeted effects are supposed to be, tests for measurement invariance are necessary for evaluating differences in group means. When relations between variables are studied within different groups separately, bias may be ignorable if the source of the biasing effect does not confound within-group orderings. This, for instance, may occur when the biasing effect is due to a source that varies between groups, but not within groups. Such a situation may arise quite naturally if items are translated literally. However, if the biasing factors have effects both within and between groups, relations found within groups may be as confounded as mean differences between groups, so that bias again becomes a validity threat. Finally, if tests are used for selecting individuals, then the minimization of bias should always be a primary goal in developing selection procedures.

Where does all this leave us with regard to the implications for everyday research? It appears that measurement invariance is rarely explicitly investigated. Many researchers simply assume their measures to be invariant across groups, without checking this assumption. In combination with the fact that most instances of applied research focus on the rejection of point hypotheses with regard to observed scores (ie, "population means of test score X are equal" or "the regression of test score X on Y is the same across groups"), this is potentially problematic, because in this case even subtle violations of measurement invariance may lead to spurious conclusions. However, it is unclear how often such erroneous conclusions are drawn because measurement invariance is not routinely investigated. Obviously, this is undesirable. Fortunately, in view of the presently available methodological techniques, this situation need not persist. Hence, investigating measurement invariance should now become a routine part of research into the structure of group differences. Hopefully, the present volume will contribute to the swift incorporation of bias detection techniques into the standard methodological toolbox of the scientific researcher.

REFERENCES

1. Wicherts JM, Dolan CV, Hessen DJ, et al. Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*. 2004;32:509–537.
2. Smith LJ, Reise SP. Gender differences on negative affectivity: an IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction Scale. *J Pers Soc Psychol*. 1998;75:1350–1362.
3. Huang DC, Church TA, Katigbak MS. Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *J Cross Cult Psychol*. 1997;28:192–218.
4. Herrnstein RJ, Murray CA. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press; 1994.
5. Neisser U, Boodoo G, Bouchard TJ, et al. Intelligence: knowns and unknowns. *Am Psychol*. 1991;51:77–101.

6. Johnson TP. Methods and frameworks for crosscultural measurement. *Med Care*. 2006;44(Suppl 3):S17–S20.
7. Lord FM. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum; 1980.
8. Mellenbergh GJ. Item bias and item response theory. *Int J Ed Res*. 1989;13:127–143.
9. Meredith W. Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*. 1993;58:525–543.
10. Mellenbergh GJ. Generalized linear item response theory. *Psychol Bull*. 1994;115:300–307.
11. Meredith W, Teresi JA. An essay on measurement and factorial invariance. *Med Care*. 2006;44(Suppl 3):S69–S77.
12. Crane PK, Gibbons LE, Jolley L, et al. Differential item functioning analysis with ordinal logistic regression techniques DIFdetect and difwithpar. *Med Care*. 2006;44(Suppl 3):S115–S123.
13. Dorans NJ, Kulick E. Differential item functioning on the Mini-Mental State Examination: an application of the Mantel-Haenszel and standardization procedures. *Med Care*. 2006;44(Suppl 3):S107–S114.
14. Jones RN. Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: detecting differential item functioning using MIMIC modeling. *Med Care*. 2006;44(Suppl 3):S124–S133.
15. Morales LS, Flowers C, Gutierrez P, et al. Item and scale differential functioning of the Mini-Mental State Exam assessed using the differential item and test functioning (DFIT) framework. *Med Care*. 2006; 44(Suppl 3):S143–S151.
16. Edelen Orlando M, Thissen D, Teresi JA, et al. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the Mini-Mental State Examination. *Med Care*. 2006;44(Suppl 3):S134–S142.
17. Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Med Care*. 2006; 44(Suppl 3):S78–S94.
18. Borsboom D, Mellenbergh GJ, Van Heerden J. Different kinds of DIF: a distinction between absolute and relative forms of measurement invariance and bias. *App Psychol Meas*. 2002;26:433–450.
19. Shealy R, Stout W. An item response theory model for test bias. In: Holland P, Wainer H, eds. *Differential Item Functioning*. Hillsdale, NJ: Erlbaum; 1993:197–239.
20. Cronbach LJ. *Essentials of Psychological Testing*. New York: Harper & Row; 1990.
21. Millsap RE, Kwok O. Evaluating the impact of partial factorial invariance on selection in two populations. *Psychol Methods*. 2004;9:93–115.

