The importance of measurement invariance in neurocognitive ability testing

Wicherts, J.M.

Published in:
The Clinical Neuropsychologist

Document version:
Publisher's PDF, also known as Version of record

DOI:
10.1080/13854046.2016.1205136

Publication date:
2016

Link to publication

Citation for published version (APA):
The importance of measurement invariance in neurocognitive ability testing

Jelte M. Wicherts

To cite this article: Jelte M. Wicherts (2016) The importance of measurement invariance in neurocognitive ability testing, The Clinical Neuropsychologist, 30:7, 1006-1016, DOI: 10.1080/13854046.2016.1205136

To link to this article: http://dx.doi.org/10.1080/13854046.2016.1205136

© 2016 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 30 Jun 2016.

Submit your article to this journal

Article views: 476

View related articles

View Crossmark data

Citing articles: 2 View citing articles
The importance of measurement invariance in neurocognitive ability testing

Jelte M. Wicherts

Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands

ABSTRACT

Objective: Neurocognitive test batteries such as recent editions of the Wechsler's Adult Intelligence Scale (WAIS-III/WAIS-IV) typically use nation-level population-based norms. The question is whether these batteries function in the same manner across different subgroups based on gender, age, educational background, socioeconomic status, ethnicity, mother tongue, or race. Here, the author argues that measurement invariance is a core issue in determining whether population-based norms are valid for different subgroups.

Method: The author introduces measurement invariance, argues why it is an important topic of study, discusses why invariance might fail in cognitive ability testing, and reviews a dozen studies of invariance of commonly used neurocognitive test batteries.

Results: In over half of the reviewed studies, IQ batteries were not found to be measurement invariant across groups based on ethnicity, gender, educational background, cohort, or age. Apart from age and cohort, test manuals do not take such lack of invariance into account in computing full-scale IQ scores or normed domain scores.

Conclusions: Measurement invariance is crucial for valid use of neurocognitive tests in clinical, educational, and professional practice. The appropriateness of population-based norms to particular subgroups should depend also on whether measurement invariance holds with respect to important subgroups.

Shuttleworth Edwards (in press) comments on several important challenges in using neurocognitive test batteries such as the third and fourth editions of the Wechsler’s Adult Intelligence Scale (WAIS; Wechsler, 1997, 2014) in a multicultural context. While focusing on South Africa, her arguments that population-based norms are insensitive for many relevant intra-national comparisons apply equally well to many other countries and testing contexts. For instance, in my home country, the Netherlands, first and later generation immigrants from Turkey or Morocco average lower scores on common IQ batteries (Evers, Te Nijenhuis, & Van der Flier, 2005; Te Nijenhuis, De Jong, Evers, & van der Flier, 2004), which raises the question of whether these IQ batteries are fair to these groups. More specifically, in neurocognitive ability testing, the question is whether we should apply the nation-level (i.e. ‘population-based’) norms or whether norms developed specifically for these groups would be more appropriate. Similar debates have continued in the U.S.A. with respect to

CONTACT Jelte M. Wicherts j.m.wicherts@uvt.nl

© 2016 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.
African-Americans and Hispanic Americans for decades. The question of the appropriateness of overarching, population-based norms for various social, cultural, or ethnic subgroups is expected to become increasingly relevant internationally because of increasing globalization and migration. For instance, if non-Western job applicants or refugees are tested in clinical or selection contexts, it is crucial that the tests are valid for these groups. The problems in these contexts have much in common.

Here, I take a generic perspective to the many issues raised by Prof. Shuttleworth-Edwards with respect to the WAIS-III and WAIS-IV by focusing on the core issue of measurement invariance. It is my goal to argue that the applicability of unitary norming (or population-based norms) in neurocognitive testing should depend on whether the test at hand shows measurement invariance with respect to relevant groupings within the population for which the norms were developed. To this end, I will (a) provide a gentle introduction to the psychometric notion of measurement invariance (for the most comprehensive discussion of invariance, see Millsap, 2009), (b) discuss why invariance might fail in cognitive ability tests and why it is an important topic of study, and (c) review some recent results of invariance testing of commonly used neurocognitive test batteries.

**Measurement invariance**

Measurement invariance starts with a simple thought experiment involving two persons who have the same underlying cognitive ability that a given test (or item) is presumed to measure. The persons are from two different groups and take the same test (or item). By studying invariance, we wish to determine whether this test (or item) is fair to both persons, or, in other words, whether the test (or item) functions equivalently (i.e. invariently) for both test takers. The groups can be defined in a number of ways, e.g. they can be based on a test taker’s gender, age, educational background, socioeconomic status, ethnicity, regional background, nationality, mother tongue, or race. Although one could group persons in multiethnic and socioeconomically heterogeneous countries such as South Africa in a number of ways, the measurement of ‘group’ is often not the main challenge (i.e. it can be operationalized on the basis of a self-report item). Yet the relevance of certain groupings depends on a host of reasons that Shuttleworth-Edwards discussed in her article within the South African context and that I will discuss in a wider sense below.

The ‘cognitive ability that the test or item is supposed to measure’ can also be defined in a number of ways, and this is where most of the contention in debates on testing fairness and validity resides. The trouble perhaps starts already with the fact that any given cognitive item is strongly dependent on a number of cognitive processes in a complex multivariate fashion, covering a great deal of cognitive ability factors in healthy populations (McGrew, 2009) and perhaps even more factors in unhealthy populations.

Modern psychometric theory offers tools to operationalize the targeted ability in a fairly unambiguous manner, that is, as the dimension underlying performance of a set of items or subtests. With their fondness of Greek, psychometricians often use theta to denote this latent dimension. Theta is often assumed to be normally distributed (which is sensible as it is affected by many biological, cultural, and social factors) and is seen as the underlying cause of the performance on the test or item. Measurement invariance requires that two persons with the same theta (i.e. targeted latent ability) should obtain the same expected
test or item score, regardless of the group they are in. If an item is scored dichotomously, measurement invariance with respect to our chosen group requires that the two equally able test takers from the two groups have same chance of answering the item correctly. Mellenbergh (1989) used this notion to define measurement invariance more generally in statistical terms. In his definition, an observed item score $X$ is measurement invariant with respect to group, if:

$$P(X|\theta, \text{group}) = P(X|\theta).$$  \hspace{1cm} (1)

The definition uses conditional distributions (indicated by $P(\cdot|\cdot)$) that describe the distribution of scores on $X$ after we have taken into account the scores on the latent cognitive ability within the groups. In this way, we allow for the possibility that groups differ in mean (or variance) of the targeted latent cognitive ability for many reasons besides measurement problems. Specifically, the definition states that the distribution of observed scores $X$, which is conditional on the latent cognitive ability ($\theta$), does not also depend on the grouping variable. In other words, knowing one’s $\theta$ renders group moot in predicting the performance on the item $X$. If the equation does not hold, invariance fails and group differences in $X$ cannot be solely attributed to group differences in the targeted cognitive ability ($\theta$). Such lack of invariance implies that $X$ is in part also a function of group besides $\theta$, which points at bias. The definition is general because it applies to all $\theta$ values and can also be used to study other types of distributions, including multivariate ability distributions that are common in cognitive ability testing.

This definition of measurement invariance underlies many different psychometric tests of invariance. In modern item response theory, measurement bias in individual items is denoted by differential item functioning (DIF; Holland & Wainer, 1993) and many well-established methods exist to study invariance of items (Millsap, 2009; Millsap & Everson, 1993). Similarly, testing of measurement invariance in the confirmatory factor analyses, where the term strict factorial invariance (Meredith, 1993) is common, has become increasingly popular in recent decades as a means of studying the fairness of cognitive ability batteries like the Wechsler scales. Although it is possible to study invariance both at the item level and at the subtest level in one analysis, most tests of measurement invariance in the literature focus either on items within a scale (using item response models) or on subtests in a larger factor model. The logic of these invariance tests is the same, but they differ in focus (item vs. subtest) and in the type of psychometric model (non-linear item response models vs. linear factor models).

If we consider (approximately normally distributed) subtest scores, measurement invariance requires that the expected value on a subtest is equal for groups matched on $\theta$. So if we had two groups of respondents with the same mean underlying or latent ability level, these groups should obtain the same mean score on the subtest. Figure 1 depicts the core relation in the linear factor model. In the left panel in Figure 1, the two groups have the exact same latent ability distribution, and they obtain the same expected test score; hence, measurement invariance applies. Another way of looking at measurement invariance is to focus on the unobserved relation between the underlying (unobserved) latent cognitive ability ($\theta$) and the observable (often also called manifest) variable $X$ (i.e. the subtest score). In the linear confirmatory factor model, we use the term factor loading to denote the slope of the linear regression predicting $X$ from $\theta$. The linear regression model has a slope, an intercept, and a residual of which we only estimate its variance. Although the factor loading
is often considered in comparing the factorial comparability of groups in factor analysis, the intercepts are arguably even more important in testing invariance (Wicherts & Dolan, 2010). Equal factor loadings across groups are a necessary, but not a sufficient condition for measurement invariance.

In the right scenario in Figure 1, I present another example of the confirmatory factor model when the groups differ in mean latent (unobserved) ability. Here, we again have a situation of measurement invariance, because the actual group differences in the theta distributions (mean differences in cognitive ability) are correctly reflected in group differences in observed means on subtest X. This reminds us that measurement invariance can apply both to scenarios where there are group differences in latent cognitive ability distributions, but also to scenarios where such differences are absent; measurement invariance concerns the equality of the relation between the latent (unobserved) cognitive ability and the observed scores and does not concern the question of whether the groups are different at the latent level.

Figure 2 depicts two scenarios in which measurement invariance fails. Specifically, in the left scenario of Figure 2, the measurement intercepts differ between the two groups, although the factor loadings are identical across groups. The observed mean differences on X are inflated and larger than is to be expected from the group difference in the targeted cognitive ability. The implication of this so-called uniform measurement bias is that all scores of members of the lower scoring group are biased downwards. Going back to our two hypothetical test takers who were from the two different groups, but had the same theta value (i.e. latent ability), it is apparent that one of our test taker’s manifest score X is suspect and should not be interpreted in light of any common norm table that is based on the combined groups 1 and 2 (the other side of the coin is actually equally problematic because the person in the higher scoring group perhaps obtains scores on X that are too high due to some upward bias, leading to an overly positive assessment of this person’s cognitive ability). One can envision many different reasons for the different measurement properties of X in this scenario, and I discuss some explanations below.
The right-hand side of Figure 2 depicts a scenario in which both the measurement intercept and the regression slope (factor loading) are different between groups. Depending on the particular level of theta of our two hypothetical test takers, either the one or the other obtains scores on X that are too low. The expected value on X for persons who have theta value that is slightly above the mean depends on the group: In one group, high-scoring individuals will obtain scores that overestimate their latent theta ability, and low-scoring individuals will obtain scores that underestimate their latent theta ability; the opposite pattern will be present for the other group. Whatever variable(s) cause(s) this type of bias, it is clear that it does so differently across the two groups and across the different ability levels.

In sum, if a cognitive ability test fails the test of invariance, group differences are not a simple function of the targeted underlying cognitive abilities; rather, other factors affect test performance differently across these groups. Thus, a lack of test invariance negatively affects the quality of assessment and decisions made on the basis of the test scores (Millsap & Kwok, 2004). Moreover, failures of invariance at the subtest level highlight that the measurement model relating the targeted cognitive abilities to the observed subtest scores differs across groups. This has direct implications for the question of validity (Borsboom, Mellenbergh, & Heerden, 2004).

**What causes failures of measurement invariance?**

When invariance fails, we know that the latent cognitive ability that the (sub)test (or item) is supposed to measure cannot explain all observed group differences on that (sub)test (or item) given the particular psychometric model used to test invariance. In a general sense, this means that somewhere the psychometric model shows misfit. A finding that an item in a scale fails to show invariance could perhaps be attributed to some group difference that is very specific to that item. For instance, an arithmetic item that uses a story about a credit card debt might not work for test takers with little knowledge of credit cards, leading to bias when comparing groups that differ in knowledge of credit cards, such as those differing in

---

**Figure 2.** Linear factor model in cases where measurement invariance does not hold. Notes: The scenario on the left shows a case of uniform bias disadvantaging the low-scoring group. The scenario on the right depicts a scenario of non-uniform bias and the resulting bias for a test taker from the low-scoring group whose value of theta is just above the mean of theta in the high-scoring group.
socioeconomic status. In the IRT literature, DIF is often interpreted in terms of item content, but there may be many other causes of bias. For valid test use and for (later) test development, it is quite informative to know why certain items or (sub)tests fail tests of invariance. Factors causing bias in an entire subtest can be diverse and relate to practical problems in how the test is administered, cultural issues related to item or test content, educational differences, language issues, but also cognitive or emotional factors. For instance, test anxiety due to negative stereotypes might cause bias on mathematical tests (Wicherts, Dolan, & Hessen, 2005), and test-taking strategies can greatly alter how one takes a fluid reasoning task such as the Raven’s Progressive Matrices tests (Fox & Mitchum, 2012). If such strategy use differs across groups, we would expect invariance to fail.

Wicherts and Dolan (2010) discuss numerous examples of potential reasons for intercept differences in confirmatory factor analyses of IQ batteries, and these include issues such as test-wiseness (familiarity with testing), test-taking strategies (e.g. tendencies to guess), familiarity with the test or item content, and abilities that are tapped by a certain subtest and that are distinct from the targeted latent ability. For instance, in a study of the invariance of a widely used Dutch children’s IQ test (the RAKIT) with respect to ethnic minorities, Wicherts and Dolan (2010) found a verbal meaning subtest to show lower averages among ethnic minorities, supposedly because of an additional ethnic difference on knowledge of Dutch words among ethnic minorities that went beyond the group difference on the latent ability factor on which this subtest loaded. The effect of the bias on this subtest on the computation of overall IQ was estimated to be around 7 IQ points, highlighting not only the severity of the bias, but also the need to use specific norms for this group of test takers. Moreover, Wicherts and Dolan (2010) found uniform bias on one of the indicators of a memory factor disadvantaging ethnic minorities. They attributed the bias on this subtest, called ‘Learning Names,’ to the fact that it involved names of fairy tales that might have been less familiar for ethnic minority children, thereby heightening the difficulty level of the test considerably.

Even fluid reasoning tests once considered to be culture-free, such as the Raven’s Progressive Matrices, are not immune to biasing factors. Notably, Wicherts, Dolan, Carlson, and van der Maas (2010) found that sub-Saharan African test takers often obtain lower average scores than Westerners on the various version of the Raven’s test. And although reliabilities of the scales were often fairly high, the convergent validity appeared to be lower in sub-Saharan African samples as opposed to studies in the Western world. Moreover, ten studies of the factorial nature of the Raven’s test among Africans showed that the test often measured multiple underlying factors instead of the one fluid reasoning factor, for which it is commonly used. Until more research suggests otherwise (but see Taylor, 2008), this lack of invariance of the various versions of the Raven’s test indicates that Western-based norms are not appropriate for use among sub-Saharan Africans.

In the context of the Wechsler’s scales, Shuttleworth-Edwards discusses several reasons for deviation of IQ (sub)test scores from the test taker’s ‘true intellectual capacity’. From the perspective of measurement invariance, the latter term relates to the latent cognitive ability or cognitive abilities that the IQ subtests are supposed to measure. The prescribed domain scores of the Wechsler scales (Verbal Comprehension (VC), Working Memory, Processing Speed, and Perceptual Organization) that are measured by different subtests offer a way to more explicitly model and test this notion in the confirmatory factor model. In fact, several recent studies have used this framework to test invariance with respect to groups of neurocognitive batteries like the Wechsler’s scales.
Table 1 lists a number of studies of invariance of IQ batteries in common use. The list is not intended to be exhaustive or fully representative of all studies of measurement invariance in neurocognitive IQ batteries (such a review would require much more journal space); rather, it is meant to highlight that failures of measurement invariance based on confirmatory factor analysis are quite commonly reported even for IQ batteries that are widely used in clinical, professional, and educational practice. This is worrisome, because it shows that in these tests, some grouping (that is potentially relevant for clinical practice) led to differences in observed IQ (sub)test scores that cannot be solely due to the latent cognitive abilities that these tests are supposed to measure. Perhaps these instances of failures of measurement invariance should lead to the use of within-group norms, corrections of bias at the subtest level, or later revisions of the subtests to restore appropriate levels of measurement invariance.

Several of the studies of invariance in Table 1 concerned the Wechsler scales and groupings that have clear relevance to the use of these scales within different countries. In two studies (one in Spain and the other in the Netherlands), the WAIS-III was found to be partially measurement invariant across sex (Dolan et al., 2006; van der Sluis et al., 2006). These studies showed sex differences on subtests that could not be explained by sex differences on the relevant domain scores in the WAIS-III. Notably, Arithmetic showed larger sex differences in favor of males than one would expect on the basis of any potential sex differences on the latent ability Working Memory, and Information showed a similar male advantage that did not align with the negligible sex difference on the VC factor. If one's goal is to measure VC, females are disadvantaged because of this uniform bias on the Information subtest. In a recent study, the authors did conclude that the WISC-IV showed invariance with respect to gender (Chen et al., 2015), highlighting that one should not overgeneralize (non) invariance but rather study it for all relevant versions and contexts.

Other studies have concerned age differences on the Wechsler scales and found invariance to be generally tenable (Niileksela et al., 2013). This is interesting in light of the current discussion, because test developers typically present IQ norms by age group. If the Wechsler scales indeed show invariance with respect to age groups, this implies that its subtests measure the same cognitive abilities in the same manner over age groups. With age-based norms, this might mean that two differently aged persons with the same level of cognitive functioning would obtain different IQ scores, raising the somewhat unorthodox question whether it always is sensible to actually use age-based norms.
Much of the discussion about group-based norming concerns ethnic group comparisons. Several studies have tested invariance with respect to ethnic groups of common IQ batteries. Here, the picture is quite mixed, with some U.S. studies supporting measurement invariance at the scale level (Dolan, 2000), and others, including some in South Africa and the Netherlands, showing failures of measurement invariance (Dolan et al., 2004; Wicherts & Dolan, 2010). At the very least, these results highlight that one cannot assume invariance with respect to ethnic groups.

Another interesting grouping variable relates to educational levels. Despite the strong relations between educational level and IQ test performance, it is not common practice to use IQ norms for specific educational levels (at least for adults). But if these cognitive test batteries are not invariant across educational levels, subtests might show bias with respect to certain educational levels. One recent study found the Italian WAIS-R to be invariant across educational levels (Tommasi et al., 2015), whereas a study involving the Spanish WAIS-IV (Abad et al., in press) showed several subtests to display non-uniform bias with respect to certain educational levels. This highlighted that these subtests measured cognitive ability differently across test takers from different educational backgrounds. Relevant questions are here as follows: Would it not be more appropriate to use separate norms for different educational levels if the tests function in a different manner across the groups? Or should we correct for the bias somehow and use the same norms across educational levels?

In clinical practice, it is important that the IQ tests also display measurement invariance with respect to clinical groups in which they are commonly used. Several studies showed that invariance does not always hold for these types of groupings in recent editions of Wechsler scales in the U.S. (Chen & Zhu, 2012; Reynolds et al., 2013). For instance, Reynolds et al. (2013) found that several of the subtests of the WAIS-IV did not show measurement invariance when comparing normal samples to a sample with intellectual disabilities.

Finally, there is one type of grouping that test developers have taken very seriously, and it concerns time. It has been well known that IQ norms become obsolete fairly quickly by a phenomenon called the Flynn effect (Flynn, 1984, 1987). In a set of studies, we found that the comparisons of different cohorts (e.g. young adults taking the WAIS either in the 1960s or in the 2000s) were associated with failures of measurement invariance (Wicherts et al., 2004). This result showed that after as little as a few decades, the many sociocultural changes within a given country can lead commonly used IQ batteries to change measurement properties. Again, in this context, test developers feel the need to develop novel norms, but many other group comparisons in Table 1 show similar failures of invariance across groups that might even show larger differences in mean IQ than those seen in studies of the Flynn effect. This raises the question of how invariance testing can inform decisions on which norms to use for a given group.

The relevance of measurement invariance for norming

It has become common practice to conduct factor analyses and other psychometric analyses (like item response models) during the development and validation phases of neurocognitive test batteries. For instance, the most recent edition of the test manual of the South African WAIS-IV (Wechsler, 2014) reports the fit of the proposed four-factor model involving the four domain scores (VC, Perceptual Organization, Processing Speed, and Working Memory) in the overall standardization sample (but no subgroup analyses). This model aligns with the
norms presented in the manual for these four domains. In answering the question of the appropriateness of these norms for certain subgroups within the standardization sample, it is warranted to ask whether this factor model also fits for these subgroups (e.g. ethnic groups, or groups with substandard education) and whether the measurement parameters in this model (i.e. factor loadings, measurement intercepts, and residual variances) are invariant across these different subgroups. Such tests for invariance are needed to ascertain whether the substantive meaning of the subtest scores in terms of domain scores is valid throughout the standardization sample or for any other group in which the battery is used.

In my view, a failure of measurement invariance with respect to some of the subgroups in the standardization sample should lead one to ask, in line with Prof. Shuttleworth-Edward's thesis based on other arguments, whether it is appropriate to use these overall (population-based) norms for such subgroups. A failure of invariance means that the mean group differences cannot be interpreted solely in terms of the targeted latent cognitive abilities. If measurement invariance fails, a person’s domain score should be interpreted in a way that corrects for the measurement bias that is evident in invariance analyses with the subgroup to which this person belongs. A drastic way to do this is to develop subgroup norms. However, there are other ways to deal with the bias; if we understand why invariance of a certain subtest fails, later revisions might improve it. Also, in some cases in which only particular subtests fail the invariance test, it might be warranted to use a partial invariance model (Byrne, Shavelson, & Muthén, 1989) and either discount that subtest score from computing a domain score, or to correct the bias before using it in the computation of the domain score (Wicherts & Dolan, 2010).

It has been long recognized that the psychological test scores need to be valid and reliable, and this is reflected in many professional standards and even in legislation in several countries. An integral part of validity (a test measures what it is supposed to measure) of a cognitive test is that there should not be factors other than the targeted latent ability that differentially affect certain distinguishable groups of test takers. When tests are used for diagnoses and selection purposes, the issue of measurement invariance is crucial (Borsboom, 2006).

I wholeheartedly agree with Prof. Shuttleworth-Edwards that we should be cautious in using norms for groups if we have not yet studied rigorously whether all subtests are appropriate for these groups. I also agree with her that in a heterogeneous country such as South Africa, there are many reasons to suspect that there are issues in measurement of cognitive abilities that might lower validity for some groups, particularly (historically) disadvantaged ones. I would like to add to her thesis that measurement invariance tests provide an excellent way to empirically test whether subtests (and items) in neurocognitive batteries function in the same manner across different subgroups and that such evidence can and should inform us on the appropriateness of certain norms. Measurement invariance testing is certainly not the only way forward in heightening validity and test fairness of neurocognitive ability tests, but the methods to do so are well established and currently not used to their full potential.

Disclosure statement

No potential conflict of interest was reported by the author.
**Funding**

This work was supported by the Netherlands Organization for Scientific Research (NWO) [VIDI grant 452-11-004].

**References**


Tommasi, M., Pezzuti, L., Colom, R., Abad, F. J., Saggino, A., & Orsini, A. (2015). Increased educational level is related with higher IQ scores but lower g-variance: Evidence from the standardization of the WAIS-R for Italy. Intelligence, 50, 68–74. doi:http://dx.doi.org/10.1016/j.intell.2015.02.005


