Cambridge University Press
978-1-108-48522-7 — Assessing Measurement Invariance for Applied Research
Craig S. Wells
Excerpt
More Information

CHAPTER 1

Introduction

What Is Measurement Invariance?

The concept underlying measurement invariance is often introduced using a metaphoric example via physical measurements such as length or weight (Millsap, 2011). Suppose I developed an instrument to estimate the perimeter of any object. My instrument is invariant if it produces the same estimate of the object's perimeter, regardless of the object's shape. For example, if my instrument provides the same estimate of the perimeter for a circle and a rectangle that have the same true perimeter, then it is invariant. However, if for a circle and a rectangle of the same true perimeter my measure systematically overestimates the perimeters of rectangles, then my measure is not invariant across objects. The object's shape should be an irrelevant factor in that my instrument is expected to provide an accurate estimate of the perimeter, regardless of the object's shape. However, when we have a lack of measurement invariance, the estimated perimeter provided by my instrument is influenced not only by the true perimeter but also by the object's shape. When we lack measurement invariance, irrelevant factors systematically influence the estimates our instruments are designed to produce.

We can apply the concept of measurement invariance from physical variables to variables in the social sciences. To do so, let's suppose I have a constructed-response item, scored 0 to 10, that measures Grade 8 math proficiency. For the item to be invariant, the expected scores for students with the same math proficiency level should be equal, regardless of other variables such as country membership. However, if, for example, Korean students with the same math proficiency level as American students have higher expected scores than Americans, then the item lacks measurement invariance. In this case, an irrelevant factor (i.e., country membership) plays a role in estimating item performance beyond math proficiency. When using my non-invariant instrument to estimate the perimeter of an object, I need the estimate from my instrument, as well as the shape of the object, to provide an accurate estimate. The same is true for the non-invariant math item. To estimate accurately a student's math proficiency, I would need their response on the item and their country membership. For an invariant math item, however, I would only need their item response.

While the use of physical measurements can be useful for introducing the concept of measurement invariance, there are two important differences when extending the idea to constructs in the social sciences, such as math proficiency or depression. First, the variables we measure in the social sciences are latent and cannot be directly observed. Instead, we make inferences from our observations that are often based on responses to stimuli such as multiple-choice, Likert-type, or constructed-response items. As a result, we must deal with unreliability, which makes it more difficult to determine whether our measures (or items) are invariant. Second, in the physical world we can obtain a gold standard that provides very accurate measurements. The gold standard can be used to match object shapes based on their true perimeter, which then allows us to compare the estimates produced by my instrument between different object shapes of the same perimeter. Unfortunately, there are no gold standards in the social sciences for the latent variables we are measuring. Latent variables that are used to match students are flawed to a certain degree, which, again, makes it difficult to assess measurement invariance.

Measurement invariance in the social sciences essentially indicates that a measure (or its items) is behaving in the same manner for people from different groups. To assess measurement invariance, we compare the performance on the item or set of items between the groups while matching on the proficiency level of the latent variable. While the idea of the items behaving in the same way between groups is useful for conveying the essence of measurement invariance, it is too simple to provide an accurate technical definition to understand the statistical approaches for examining measurement invariance. To fully understand what I mean by an item being invariant across groups within a population, I will begin by first defining the functional relationship between the latent variable being measured, which I will denote as θ, and item performance, denoted Y. The general notation for a functional relationship can be expressed as $f(Y|\theta)$, which indicates that the response to the item or set of items is a function of the latent variable. For example, if an item is scored dichotomously ($Y = 0$ for an incorrect response, $Y = 1$ for a correct response), then $f(Y|\theta)$ refers to the probability of correctly answering the item given an examinee's level on the latent variable, and can be written as $P(Y=1|\theta)$. For an item in which measurement invariance is satisfied, the functional relationship is the same in both groups; that is,

$$P\left(Y=1\middle|\theta, G=g_1\right)=P\left(Y=1\middle|\theta, G=g_2\right). \tag{1.1}$$

G refers to group membership, with g_1 and g_2 representing two separate groups (e.g., Korea and America). Another way of expressing measurement invariance is that group membership does not provide any additional information about the item performance above and beyond θ (Millsap, 2011). In other words,

$$P(Y=1|\theta, G)=P(Y=1|\theta). \tag{1.2}$$

To illustrate the idea of measurement invariance graphically, Figure 1.1 provides an example of the functional relationships for two groups on a dichotomously scored item that is invariant. The horizontal axis represents the proficiency level on the latent variable – in this book, I will refer to the level on the latent variable as *proficiency*. The vertical axis provides the probability of a correct response. Because the item is invariant, the functional relationships for both groups are identical (i.e., the probability of a correct response given θ is identical in both groups). The proficiency distributions for each group are shown underneath the horizontal axis.

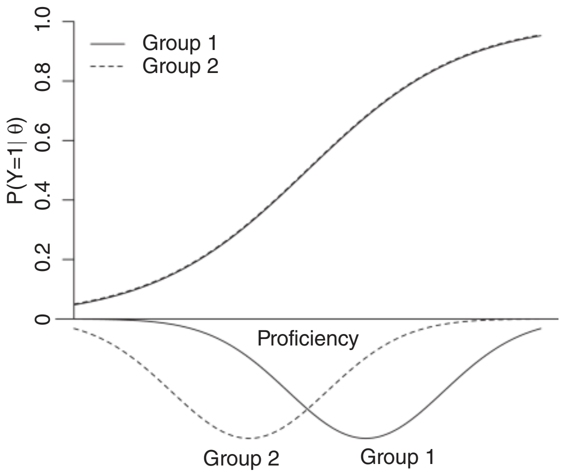


Figure 1.1 An example of functional relationships for two groups on a dichotomously scored item that is invariant.

Cambridge University Press
978-1-108-48522-7 — Assessing Measurement Invariance for Applied Research
Craig S. Wells
Excerpt
More Information

We can see that Group 1 has a higher proficiency than Group 2. The difference in proficiency distributions between the groups highlights the idea that matching on proficiency is an important aspect of the definition and assessment of measurement invariance. If we do not control for differences on θ between the groups, then differences in item performance may be due to true differences on the latent variable, not necessarily a lack of measurement invariance. The difference between latent variable distributions is referred to as *impact*. For example, since Group 1 has a higher mean θ distribution than Group 2, then Group 1 would have, on average, performed better on the item than Group 2, even if the functional relationship was identical, as shown in Figure 1.1. As a result, the proportion of examinees in Group 1 who answered the item correctly would have been higher compared to Group 2. However, once we control for differences in proficiency by conditioning on θ, item performance is identical. The fact that we want to control for differences in the latent variable before we compare item performance highlights the idea that we are not willing to assume the groups have the same θ distributions when assessing measurement invariance.

Figure 1.2 illustrates an item that lacks measurement invariance. In this case, the probability of a correct response conditioned on θ is higher for

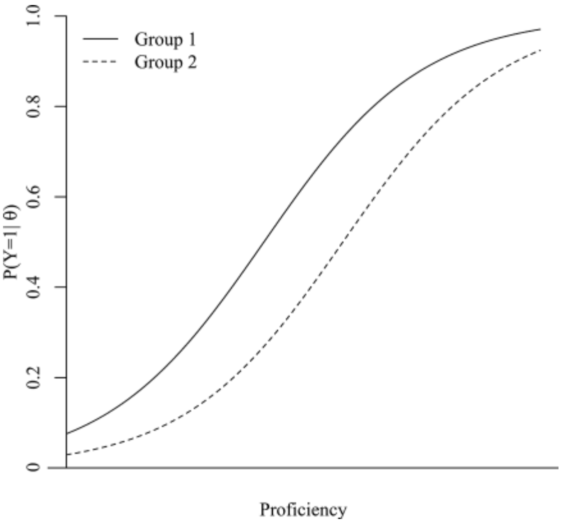


Figure 1.2 An example of functional relationships for two groups on a dichotomously scored item that lacks invariance.

Group 1, indicating that the item is relatively easier for Group 1. In other words, Group 1 examinees of the same proficiency level as examinees from Group 2 have a higher probability of answering the item correctly. When measurement invariance does not hold, as shown in Figure 1.2, then the functional relationships for Groups 1 and 2 are not the same (i.e., $f(Y|\theta, G = g_1) \neq f(Y|\theta, G = g_2)$). Therefore, to explain item performance we need proficiency and group membership. In the case of non-invariance, the item is functioning differentially between the groups; in other words, the item is exhibiting differential item functioning (DIF). In this book, I will refer to a lack of measurement invariance as DIF. In fact, many of the statistical techniques used to assess measurement invariance are traditionally referred to as DIF methods.

The concept of measurement invariance can be applied to polytomously scored items; that is, items that have more than two score points (e.g., partial-credit or Likert-type items). For a polytomous item, Y could refer to the probability of responding to a category, or it could refer to the expected score on the item. For example, Figure 1.3 illustrates an invariant (top plot) and non-invariant (bottom plot) functional relationship for a polytomous item with five score categories. The vertical axis ranges from 0 to 4 and represents the expected scores on the polytomous item conditioned on θ (i.e., $E(Y = y|\theta)$). The expected item scores conditioned on proficiency are identical when invariance is satisfied but differ when the property of invariance is not satisfied.

Measurement invariance can also be extended to compare performance on a subset of items from a test (e.g., items that represent a content domain). In this case, the functional relationship looks a lot like a polytomous item in that we are comparing the expected score conditioned on the latent variable. When a scale based on a subset of items from a test lacks invariance, we often refer to it as *differential bundle functioning* (DBF). A special case of DBF is when we examine the performance of all items on a test. In this case, we are examining the invariance at the test score level. When the invariance is violated at the test score level, we refer to it as *differential test functioning* (DTF).

Why Should We Assess Measurement Invariance?

There are two basic reasons for why we should care about whether a test and its items are invariant across groups in a population. The first reason pertains to test validity in that the presence of DIF can impede test score interpretations and uses of the test. The *Standards for Educational and*

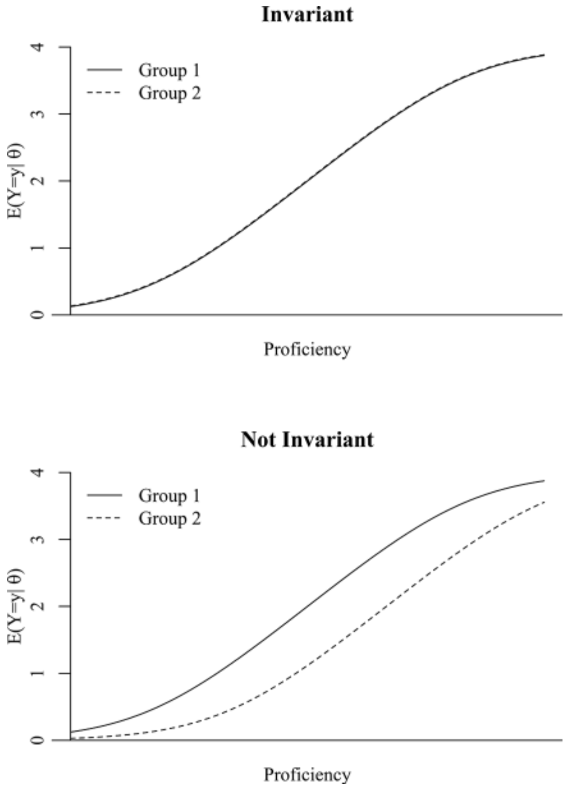


Figure 1.3 An illustration of an invariant and non-invariant polytomous item.

Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) describe five sources of validity: content, response processes, internal structure, relation to other variables, and consequences. Providing evidence to support measurement invariance is one of the aspects of internal structure. The presence of DIF is an indication that there may be a construct-irrelevant factor (or factors) influencing item performance. The consequences of having items that lack invariance in a test can be severe in some cases. For example, a lack of invariance at the item level can manifest to the test score level, leading to unfair comparisons of examinees from different groups. If the DIF is large enough, examinees may be placed into the wrong performance category (imagine how disheartening it would be, after working diligently to successfully build skills in, say, math, to be placed into a performance category below your expectation because of something other than math proficiency). A lack of invariance is not just a validity concern for large-scale assessments but for any test in which a decision is being made using a test score, such as remediation plans for struggling students in schools, determining whether an intervention is effective for a student, assigning grades or performance descriptors to students' report cards, etc. The examples I have discussed so far have pertained to educational tests. However, the importance of measurement invariance also applies to non-cognitive tests such as psychological inventories, attitudinal measures, and observational measures. In fact, it is important to establish measurement invariance for any measure prior to making any group comparisons using its results. Essentially, any time we are planning on using a score from an instrument, we should collect evidence of measurement invariance so that we can be confident that no construct-irrelevant factor is playing a meaningful role in our interpretations and uses.

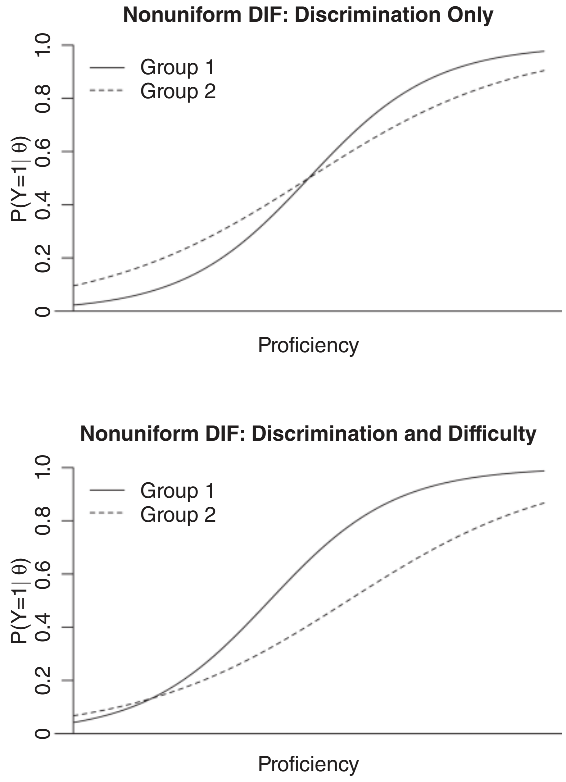
In addition to the direct effect that DIF can have on test score interpretation and use, it can also indirectly influence validity through its deleterious effect on measurement processes. For example, DIF can disrupt a scale score via its negative effect on score equating or scaling when using item response theory (IRT). A common goal in many testing programs is to establish a stable scale over time with the goal of measuring improvement (e.g., the proportion of proficient students within Grade 8 math increases over consecutive years) and growth (each student demonstrates improved proficiency over grades). Tests contain items that are common between testing occasions (e.g., administration years) that are used to link scales so that the scales have the same meaning. If some of the common items contain DIF, then the equating or scaling can be corrupted, which results in an unstable scale. This type of DIF, where the groups are defined by testing occasion, is referred to as *item parameter drift* in that some of the items become easier or harder over time after controlling for proficiency differences (Goldstein, 1983). A consequence of item parameter drift is that inferences drawn from test scores may be inaccurate (e.g., examinees may be placed into the wrong performance categories).

A second purpose for assessing measurement invariance is when we have substantive research questions pertaining to how populations may differ on a latent variable. For example, suppose we want to compare geographic regions on a math test. In addition to examining mean differences between regions, assessing measurement could provide useful information about how the items are functioning across the regions. We could find that certain domains of items are relatively harder for a particular region, suggesting that perhaps that group did not have the same opportunity to learn the content. Assessing measurement invariance in this context could also be useful for psychological latent variables. For instance, we may be interested in comparing gender groups on a measure of aggressiveness. Items that are flagged as DIF may provide insight into differences between the groups.

Forms of DIF

It is helpful to have nomenclature to classify the types of non-invariance. There are two basic forms of DIF: uniform and nonuniform (Mellenbergh, 1982). Uniform DIF occurs when the functional relationship differs between the groups consistently or uniformly across the proficiency scale. The plot shown in Figure 1.2 provides an example of uniform DIF. In this case, the probability of a correct response for Group 1 is higher compared to Group 2 throughout the proficiency scale. At the item level, the difference in the functional relationships for uniform DIF is defined only by the item difficulty, whereas the item discrimination is the same in both groups. As I will describe in Chapter 3, where I address IRT, the item discrimination is related to the slope of the functional relationship curve. In uniform DIF, the curve shifts to the right or left for one of the groups, while the slope remains the same.

Nonuniform DIF occurs when the lack of invariance is due to the discrimination between the groups, regardless of whether the difficulty differs between the groups. Whereas for uniform DIF the item can only be harder or easier for one of the groups, nonuniform DIF can take on many forms. Figure 1.4 provides two examples of nonuniform DIF. In the top plot, the DIF is defined only by the difference in discrimination between the two groups; in this case, the curve is flatter for Group 2, indicating a less discriminating item compared to Group 1. The difference between Groups 1 and 2 in answering the item correctly depends on the θ value; for lower θ values, the item is relatively easier for Group 2, whereas for higher θ values, the item is relatively harder compared to Group 1. The bottom plot in Figure 1.4 provides another example of nonuniform DIF, but in this case the item differs with respect to discrimination and difficulty such that the item is less discriminating and more difficult in Group 2. When testing for DIF, our goal is often not only to detect DIF but also to describe the nature or form of the DIF.


Nonuniform DIF: Discrimination Only

Nonuniform DIF: Discrimination and Difficulty

Figure 1.4 Two examples of nonuniform DIF.

Another important factor to consider when describing DIF is whether the set of DIF items is consistently harder or easier for one of the groups. When the DIF is consistent across items (e.g., the DIF items are all harder in one of the groups), then it is referred to as *unidirectional DIF*. If, on the other hand, some of the DIF items are easier in one group, while some of the other DIF items are harder, then that is referred to as *bidirectional DIF*. The reason it is helpful to make this distinction is that the effect of unidirectional DIF can often pose a more serious risk to psychometric procedures such as equating and making test score comparisons. In addition, unidirectional DIF can also make it more difficult to detect DIF items in that the DIF has a larger impact on the latent variable used to match examinees (see discussion on purification procedures for further details and how to mitigate the effect of unidirectional DIF).

Classification of DIF Detection Methods

The statistical techniques used to assess measurement invariance that we will explore in this book can be classified under three general approaches. Each of the approaches differs with respect to how the latent variable used to match examinees is measured. The first class of DIF detection methods, referred to as *observed-score* methods, uses the raw score as a proxy for θ. The raw scores are used to match examinees when comparing item performance. For example, measurement invariance is assessed by comparing item performance (e.g., the proportion correct for a dichotomously scored item) for examinees from different groups with the same raw score. Observed-score methods have the advantage of providing effect sizes to classify a detected item as nontrivial DIF. The observed-score methods addressed in this book include the Mantel–Haenszel procedure (Holland, 1985; Holland & Thayer, 1988), the standardization DIF method (Dorans & Kulick, 1986; Dorans & Holland, 1993), logistic regression (Swaminathan & Rogers, 1990), and the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993a, 1993b). I will describe the observed-score methods in Chapter 2.

The second class of methods uses a nonlinear latent variable model to define θ and subsequently the functional relationship. These methods rely on IRT models. The plots shown in Figures 1.1–1.4 are examples of item response functions provided by IRT models. One of the advantages of using IRT to examine DIF is that the models provide a convenient evaluation of DIF that is consistent with the definition of DIF. The IRT methods addressed in this book include *b*-plot, Lord's chi-square (Lord, 1977, 1980), the likelihood-ratio test (Thissen, Steinberg, & Wainer, 1993), Raju's area measure (Raju, 1988, 1990), and differential functioning of items and tests (DFIT; Raju, van der Linden, & Fleer, 1995). I will describe the basic ideas of IRT in Chapter 3 and the IRT-based DIF methods in Chapter 4.

The third class of methods uses a linear latent variable model via *confirmatory factor analysis* (CFA). Although there is a strong relationship between CFA and IRT, and the methods used to examine DIF are similar in some respects, they are distinct in important ways. For example, CFA and IRT evaluate the fit of the respective latent variable model using very different approaches and statistics. One of the advantages of using CFA to assess measurement invariance is that it provides a comprehensive evaluation of the data structure and can easily accommodate complicated multidimensional models. The methods we will examine in this book include