Measurement 46 (2013) 3228–3237

Contents lists available at SciVerse ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application

Guido Makransky [a,*], Cees A.W. Glas [b,1]

[a] University of Southern Denmark, Denmark
[b] University of Twente, Netherlands

ARTICLE INFO

Article history:
Received 27 March 2013
Received in revised form 29 May 2013
Accepted 13 June 2013
Available online 25 June 2013

Keywords:
Psychometrics
Computerized adaptive testing (CAT)
Item response theory (IRT)
Measurement invariance (MI)
Differential item functioning (DIF)

ABSTRACT

Many important decisions are made based on the results of tests administered under different conditions in the fields of educational and psychological testing. Inaccurate inferences are often made if the property of measurement invariance (MI) is not assessed across these conditions. The importance of MI is even greater when test respondents are compared based on their responses to different items, such as the case in computerized adaptive testing (CAT), because the existence of items that exhibit differential item functioning (DIF) can produce bias within a group as well as between groups. This article demonstrates a straightforward psychometric method for conducting a test of measurement invariance (MI) and illustrates a method for modeling DIF by assigning group-specific item parameters in the framework of IRT. The article exemplifies two applications of the method for a CAT used in a high stakes international organizational assessment context. These examples pertain to context effects due to the test administration method (computer based linear test vs. CAT), and the context effects due to language in a CAT.

1. Introduction

High stakes testing is widely used to make important decisions in the contexts of educational and psychological measurement. Instruments designed to measure constructs such as knowledge, abilities, attitudes, personality traits, and educational attainment are often administered under different conditions. Examples of these different conditions include the use of different languages [3], assessment across different time points [13], or across different mediums of test administration [30]. The degree to which measurements conducted under these different conditions yield measures of the same attribute is known as measurement invariance (MI; [18]). Vandenberg and Lance [33] highlight the fact that inaccurate inferences are often made if MI is not assessed across these conditions. Nevertheless, MI is rarely tested in operational settings [33].

This is particularly problematic in international assessment contexts when comparisons are made across a large number of languages versions of a test. Factors such as globalization and the increasing mobility of the world's workforce, the emergence of complex multicultural societies such as the European Union, and the continuous increase in the number of countries participating in international comparative assessments are examples of trends that have enhanced the importance of ensuring linguistic and cultural equivalence across various language versions of assessment instruments. Evidence also suggests that the need for multi-language versions of achievement, aptitude, and personality tests, and surveys will continue to grow (e.g., [6,7,14–16,24]). Inaccurate inferences that can have large implications for individuals, organizations, and society can be made when MI exists across testing conditions. Therefore, methods are needed to easily assess and model language differences in tests

* Corresponding author. Tel.: +45 51432444.
E-mail addresses: gmakransky@health.sdu.dk (G. Makransky), c.a.w.glas@utwente.nl (C.A.W. Glas).
[1] Tel.: +31 53 489 3565.

0263-2241/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.measurement.2013.06.020

that are used across cultures, in order to provide an easy and flexible means of modeling these differences.

MI can be checked by assessing if individuals at the same trait level but from different subgroups have unequal probabilities of endorsing or responding correctly to an item. This is known as assessment of differential item functioning (DIF). There is a distinction made between uniform and non-uniform DIF. Uniform DIF occurs when one group consistently has a greater probability of endorsing or responding correctly to an item across all the levels of the latent trait. In contrast non-uniform DIF does not occur equally at all points on the latent trait, but may only be evident at high or low levels of the construct.

Although the assessment of DIF is necessary in many assessment contexts, the importance of DIF is even greater when test respondents are compared based on their responses to different items, such as the case in computerized adaptive testing (CAT, [38]). This is the case because respondents are administered different items; therefore, the existence of items that exhibit DIF can produce bias within a group as well as between groups in a CAT. The additional effect within groups occurs because not all respondents are administered the DIF items. So some are disadvantaged and others are not. Furthermore, fewer items are typically administered in a CAT. An item that exhibits DIF can consequently have a large effect on the test result. An item that exhibits DIF can also have major repercussions in a CAT because the sequence of items administered to the examinees depends in part on their responses to that item.

Most of the current research on DIF has focused on developing sophisticated statistical methods for detecting or "flagging" items that function differently across groups [37]. Most literature assumes that items that produce DIF between groups should be identified and eliminated from the test. This approach is also evident in the International Test Commission (ITC) guidelines. According to the ITC test adaptation guideline D.9: "Test developers/publishers should provide statistical evidence about the equivalence of items in all intended populations". In his interpretation of the guideline, Hambleton [15] states that when performance is not equivalent, a sound reason must be available or the item should be deleted from the test.

Eliminating items that exhibit DIF can have two disadvantages. The first is that the items could in fact measure important components of the construct across conditions, but do so in a different way. The elimination of these items can leave gaps in the measurement of the construct that can make it difficult to maintain the validity of the test. The second disadvantage is that it can become very difficult to obtain a large number of DIF free items when tests exist across a large number of conditions (e.g., languages). Remaining DIF items can still have consequences for the conclusions that are made based on test results even when efforts are made to eliminate the worst DIF items, simply because the comparison of results across such a large sample of diverse conditions can make it difficult to obtain a meaningful DIF free scale. Kreiner [21] illustrates this in reading data from the 2006 Program for International Student Attainment (PISA) survey. Kreiner's results provide evidence of DIF that affect the ranking of countries even though considerable effort is made to document the elimination of items that exhibit DIF across countries for this scale.

An alternative approach to the elimination of items could be to investigate if the items that exhibit DIF actually measure the same construct across conditions even if they do so in a different way. In Item Response Theory (IRT), such differences can be modeled by group-specific item parameters (e.g., [31,36]). This approach is only valid if it can be explicitly shown that the responses to the items given in the two groups pertain to the same latent variable. In other words, the construct that is being measured must remain the same in both groups. This can be shown by investigating if the same IRT model holds for the entire set of response data [9]. The reasoning behind this approach is that items can have slightly different true parameters across conditions. These differences can be modeled when there is statistical evidence to support the hypothesis that the items measure the same construct across the conditions.

The main objective of this article is to present a straightforward method that can be used for investigating MI and for modeling DIF with group-specific item parameters. We demonstrate the method by illustrating two examples: The investigation of MI arising from context effects based on the test administration method (computer based linear test vs. CAT), and context effects due to language in a CAT. The possibility of modeling DIF with group-specific item parameters is specifically relevant for CAT's because of the risk of within as well as between-group bias. Also, large item banks in CAT's mean that the number of items that may exhibit DIF is typically larger. Therefore, we focus on the application of the method for a CAT in this article; however, the method can be used broadly across fixed and adaptive tests.

The remainder of the article follows the following format. First, we will describe a method used to investigate MI and introduce the possibility of modeling DIF with group-specific item parameters (virtual items), as an alternative to eliminating items that exhibit DIF. Next, we will illustrate two applications of the methodology that are typically necessary when developing a CAT in an international organizational context. These include context effects due to the test administration method (computer based linear test vs. CAT), and context effects due to language in a CAT. Test administration context effects are typically relevant when a CAT is developed from an existing linear test. Language context effects are common in international testing where tests are frequently adapted across cultures. Finally, we discuss practical issues, and look ahead at possible future applications.

1.1. Modeling DIF with group-specific item parameters

Research on the topic of MI is typically conducted in the framework of confirmatory factor analysis (CFA) or IRT (e.g., [26,27,32]). IRT approaches investigate MI by assessing item response functions, whereas the CFA approaches focus on item loadings. Therefore, IRT approaches provide more detail when the equivalence of a single scale or specific scale items is of interest [25]. Therefore, the present