Cluster bias: Testing measurement invariance in multilevel data

Jak, S.

Publication date
2013

INTRODUCTION

This thesis is about using structural equation modelling to detect and account for measurement bias in multilevel data. The basic concepts and their importance will be illustrated below, by using an example from educational research.

Suppose, a researcher is interested in the influence of students' motivation on their mathematical ability. After weeks of calling schools, she finds 200 teachers and 700 students willing to participate in her study. The students complete a motivation questionnaire with 10 items such as "I think learning math is good for me" and "I like math", scored on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). The children also make a mathematical ability test consisting of 60 items that can be answered correctly or incorrectly.

Before the researcher can test any hypothesis on the relation between motivation and mathematical ability, the researcher needs to know: are these measurements valid? Are differences in motivation and mathematical ability reflected by differences in the associated item scores (Borsboom, Mellenbergh & van Heerden, 2004)? A related issue is the question of measurement invariance: do the items measure the same attributes for different (groups of) respondents (Mellenbergh, 1989; Meredith, 1993; Oort, 1992, 1993)? If the mathematical ability test items indeed measure the same attribute in boys and girls, then boys and girls with equal mathematical ability should, on average, have identical observed scores. If this is not the case, we speak of measurement bias. For example, an item with a worded math problem may be easier to solve for girls, because girls are generally better in reading than boys (Wei et al., 2012). For that reason, with equal levels of mathematical ability, girls may have more correct answers on this item than boys will.

This thesis presents models and methods to investigate and account for measurement bias in multilevel data, such as data from children in school classes. One difficulty is that we do not have a direct measure of the (latent) variables of interest, such as mathematical ability or motivation, we have to work with observed item scores. The relationship between the observed item scores and motivation or mathematical ability can be represented by a measurement model, such as a linear factor model (Mellenbergh, 1994; Spearman, 1904, 1928). In this thesis we use factor models as measurement models, in which the variables that were intended to be measured are represented by continuous latent common factors, that capture all common variance in the observed scores. Each item is also affected by a unique factor that has a structural part (causing item specific variance), and a random part (measurement error) (Bollen, 1989).

The investigation of measurement bias should always be preceded by the establishment of a sensible measurement model. Chapter 1 serves as an introduction to the concept of measurement bias. Using two examples with data from a cognitive ability test, we show that measurement bias and multidimensionality are closely related. An item that shows measurement bias is multidimensional, as it taps into a dimension that was not intended to be measured. If this dimension is related to the variable with respect to which measurement bias is tested (often variables like gender, ethnicity, age), then the item is said to be biased with respect to this variable.

Another question that the researcher in the example above might ask is: Are motivation and mathematical ability measured identically in different classrooms? As she collected data from school children, who are clustered in classes, the data have a multilevel structure. Children's scores are affected by class level variables, such as teacher quality and classroom composition. Differences in these variables may lead to differences in the average scores of children in different classrooms, that are not accounted for by the common factor (mathematical ability or motivation in the example). In Chapter 2 we propose a test for cluster bias, which can be used to examine whether measurements are biased with respect to school class. This test can be used in more situations than with children in classes only. It can be used to investigate bias with respect to any clustering variable in multilevel data (e.g. data from people in countries, from patients in hospitals, or from children in families), hence the general term "test for cluster bias".

The motivation items from the example were scored on 7-point scales, which can be treated as continuous scores in a linear factor model (Dolan, 1994). The answers to the math questions were dichotomous (right/wrong), which needs to be taken into account in the measurement model. Chapter 3 extends the test for cluster bias to situations with dichotomous and ordinal item responses.

The cause of cluster bias in the mathematical ability test or in the motivation items is a class level variable, such as the mathematical ability of the teacher. If the researcher also measured the mathematical ability of teachers, she may test whether the differences between the classroom level math scores can be explained by mathematical ability of the teacher. This involves testing for measurement bias with respect to a class level variable. In the population, if there is no cluster bias, there is no bias with respect to any other class level variable. In Chapter 4 it is investigated whether the test for cluster bias indeed detects all bias that is caused by specific class level variables.

The researcher from the example has three types of so-called violators with respect to which the tests may be biased: student level variables (e.g. student's gender, student's ethnicity), the clustering variable (class) and class level variables (e.g. teacher quality, average student SES). In Chapter 5 we propose a 5-step approach to investigate bias with respect to these three types of violators.

This thesis will help researchers who analyse multilevel data to evaluate measurement bias in their research instruments in a systematic and valid way.