1 Abstract

The methodological research question considered in this talk is the following: When we use a multigroup latent variable model to analyse data from multiple-item scales in a cross-national survey or other multigroup data, when and to what extent are comparative conclusions about the distributions of latent constructs across groups sensitive to the assumption that measurement equivalence holds for all of the items? We begin by briefly introducing the key terms in this statement.

In a cross-national social survey the same questions, translated into the local languages, are asked of respondents in several countries (for an overview of the field, see Smith 2010 and other chapters in Harkness et al. 2010). Analogous types of data arise also from comparative studies in educational and psychological assessment, such as cross-national programmes of educational testing where literacy and numeracy tests are administered to subjects in different countries to study levels of competence. A key purpose of such studies is to answer research questions about comparisons between different groups of subjects. While our work is primarily motivated by cross-national research where the groups of interest are populations of countries, the methodological questions and conclusions that we discuss apply more broadly (e.g. to ethnic groups within a country, or the same population but at different times). To emphasise this, we will typically use the term “group” rather than “country” in the discussions below.

Latent variable models operationalise the idea of measurement by multiple items. The best-known is the classical factor analysis model for continuous items. Here, however, we consider instead latent trait models for items which are treated as categorical variables. The motivation of this emphasis is that many survey items are of a categorical nature, as are typical test items in psychological and educational testing. In psychometric applications, latent trait models are more commonly known as item response theory (IRT) models.

One of the crucial issues in the design and analysis of multigroup studies is measurement equivalence of the items (also known, especially in psychometric applications, as absence of differential item functioning). This essentially means comparability: An item is equivalent across groups if it has the same meaning in all the groups in which it is used. Such comparability is in principle critical because if it fails, any observed differences between groups might reflect artifacts of measurement rather than true differences in the distributions of the constructs of interest.
In a multigroup latent variable model, non-equivalence of measurement is operationalised as an association between the group and an item, conditional on the latent variables. It then becomes possible to specify models with and without equivalence, by comparing them to assess the extent of non-equivalence, and, if necessary, to employ a model in which some items are non-equivalent. This is not yet done routinely in analyses of cross-national surveys, where measurement is most often assumed equivalent by default, but the use of non-equivalence models is also becoming increasingly common. Such models are, however, practically and conceptually problematic to the extent that we would still prefer to avoid them if possible. Lack of measurement equivalence in multigroup comparisons thus presents a dilemma for the data analyst, where neither ignoring non-equivalence nor allowing for it are fully appealing as general approaches.

A situation that would fortuitously resolve the dilemma would be one where comparative conclusions of main interest were actually not very sensitive to assumptions about the measurement, in particular to wrongly ignoring any non-equivalence in it. This will of course not be true in general, but an interesting question is whether it may be the case in circumstances that are commonly encountered in comparative research. The implications for practical data analysis would be very different if group comparisons were severely distorted even by small variation in measurement, than if they were only affected by the grossest violations of equivalence. This is the question considered in this article. In other words, if we act as if our measures function in the same way in all groups, but in reality they do not, when will this cause conclusions about substantively interesting comparisons between the groups to be seriously misleading?

We examine the question using numerical sensitivity analyses of a range of scenarios. Oberski (2013) proposes statistics which can be used to conveniently assess the sensitivity of parameter estimates of multigroup structural equation models fitted to observed data. Our analyses are implemented in a way which avoids the need for simulation and requires only one model fit for each scenario that we consider. This makes it possible to examine more cases and in a finer-grained manner than would be possible with a simulation study.

References

