Practical Significance of Item Response Theory Model Misfit

Crisan, Daniela

DOI:
10.33612/diss.128084616

Chapter 1

Introduction

1.1. Context

In assessment in education, psychology, and health research, tests and questionnaires play an important role. Item response theory (IRT; Embretson & Reise, 2000; Meijer & Tendeiro, 2018; Sijtsma & Molenaar, 2002) models are used to construct new tests and questionnaires, and to evaluate the psychometric quality of existing ones. Due to the increasing availability of easy-to-use software, not only test development companies, but also researchers use IRT for the development and evaluation of their instruments.

IRT consists of a class of models through which proficiency or trait (denoted θ) levels, as well as characteristics of individual items, can be estimated and the quality of measurement can be assessed. Unidimensional IRT models, which are the focus of this thesis, assume that θ is unidimensional, that the items are locally independent for a fixed θ level, and that the probability of answering an item correctly or endorsing an item is an increasing function of θ. Models differ in terms of the functional relation between the data and stochastic outcomes (e.g., the probability of endorsing an item). Arguments in favor of the plausibility of IRT assumptions can be based on substantive considerations, but are often based on empirical evidence. In fact, before results can be drawn from a fitted IRT model, researchers should report fit measures to show the model they use describe the data fairly well, so that, for example, estimated θ levels can be trusted.

Albeit checking model assumptions is an important step in IRT applications, these models and their underlying assumptions represent ideals about data that are almost never met in practice. As Funder (1997, pp. 32-33) discussed in a more general context, "There are only two kinds of data: Terrible data that are ambiguous, potentially misleading, incomplete, and imprecise. The second kind is No Data." Because IRT models never fit the data perfectly, researchers who apply these models are left in uncertainty with respect to which model to choose or whether or not to remove misfitting items. These types of decisions are often not straightforward. For example, more flexible models can be used, which may show better fit to the data. However, more flexible models also have more parameters, which require more complex estimation methods and can be less stable across replications (see Molenaar, 1997a). Furthermore, the literature on existing measures shows that often researchers are less inclined to choose more complex IRT models, while struggling with

poorly fitting items that are already part of an operational instrument. Removing the poorly fitting items can be problematic and sometimes not even possible. Examples can be found in educational measurement where students are administered tests and where removing items may affect total scores on exams. Another example is clinical measurement, where standard instruments cannot be altered because of score comparability.

Given the practical restrictions in model choice and item removal, the question is "Does it matter?" Does it matter, from a practical point of view, which model we choose? Or whether we keep items in the test that seem to violate some model assumptions? In this thesis I try to answer these types of questions.

1.2. Topic of the thesis

Researchers and practitioners need evidence about the stability of the main conclusions of empirical research in which IRT models are used (Molenaar, 1997a). For example:

a) Are the main conclusions derived from the use of an instrument (e.g., a test or a questionnaire) similar with or without bad items?
b) Is there a difference in the main conclusions derived from an instrument with or without misfitting item-score patterns?
c) If there are differences, how large and how consequential are they?

In this thesis I investigated these types of questions using both simulated and empirical data. The overarching topic of this thesis is the *practical* significance of misfit, that is, "the extent to which the decisions made from test scores are robust against the misfit of the IRT models" (Sinharay & Haberman, 2014, p. 23).

First, I investigate, through both simulated and empirical data, whether the main conclusions in research hold under different IRT models and especially under different violations of IRT model assumptions. Although these questions are not always easy to answer because decisions are often based not only on test scores, but also on other sources of information, such as interviews, this should not withhold researchers from investigating them (Sinharay & Haberman, 2014). Second, I investigate a tool that can be used as an effect size measure of misfit and which provides researchers with information as to what extent model choice and violations of model assumptions are consequential. This effect size measure can help researchers to decide whether or not to

remove items from a test and it can contribute to the first question I try to answer in this thesis.

As several authors have emphasized (e.g., Sinharay & Haberman, 2014; Steinberg & Thissen, 2006), practical significance of misfit cannot be decided based on a misfit statistic and can only be answered within a research context. Therefore, in this thesis I considered outcomes that are specific in several assessment contexts: (1) educational assessment, (2) clinical assessment, (3) personnel selection, and (4) personality assessment.

1.3. Outline of the thesis

Following this introduction chapter, in Chapter 2 I investigate, based on simulated data, the practical consequences of violations of the unidimensionality assumption on selection decisions in the context of educational assessment. Specifically, I am interested in the effects of model violations on the rank-ordering of examinees, on the overlap between sets of selected examinees based on different scoring methods, and on the predictive validity of the test. In Chapter 3, I examine the consequences of ignoring violations of assumptions that underlie the use of classical sum-scores to score individuals on an often-used clinical measurement instrument. In particular, I focus on the assessment of attention problems in children and whether psychometrically more refined models could improve the prediction of relevant outcomes in adulthood.

In Chapter 4, I shift my attention to Mokken scale analysis (MSA; e.g., Sijtsma & van der Ark, 2017; Wind, 2017). MSA is a popular framework within IRT to evaluate the psychometric quality of questionnaires and their individual items, in various contexts such as personality, clinical psychology and health-related measurement, education, or human resources and marketing. Although many empirical papers report the extent to which sets of items (i.e., existing scales or tests) form the so-called Mokken scales, much less attention has been devoted to the effects of violations of commonly used rules-of-thumb in MSA on practical decisions. I investigate the practical consequences of retaining or removing items with psychometric properties that do not comply with these rules-of-thumb in MSA.

Chapters 2 through 4 are aimed at investigating the effects of model violations on outcome variables. In Chapter 5, I investigate the usefulness of a summary measure, the *Crit* index, as an effect size measure for the severity of violations of some basic assumptions underlying commonly-used IRT models.

The *Crit* index was proposed two decades ago as a measure intended to help researchers and practitioners to quantify violations of some scaling assumptions. Based on predefined rules-of-thumb, one decides the degree to which items violate such assumptions. Although the *Crit* index is currently implemented in several software programs, it is unclear how sensitive and how specific the measure and its rules-of-thumb are to detecting misfit of various types. In this chapter, I conduct a simulation study in order to address this concern. Finally, in Chapter 6 I provide an overarching discussion of the results from the previous chapters, and provide some practical guidelines for researchers and practitioners in the field of psychometric testing.

The chapters in this thesis are written as separate research papers. As a result, there is some overlap in the content of the chapters.