Journal of School Psychology 60 (2017) 41–63

Contents lists available at ScienceDirect

Journal of School Psychology

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

Introduction to bifactor polytomous item response theory analysis☆

Michael D. Toland [a,*], Isabella Sulis [b], Francesca Giambona [b], Mariano Porcu [b], Jonathan M. Campbell [a]

[a] Department of Educational, School, and Counseling Psychology, University of Kentucky, United States
[b] Department of Social Sciences and Institutions, University of Cagliari Italy

ARTICLE INFO

Article history:
Received 8 January 2016
Received in revised form 22 November 2016
Accepted 22 November 2016
Available online 29 December 2016

Keywords:
Bifactor
Item response theory
Graded response model
flexMIRT
IRTPRO
R
Mplus
STATA

ABSTRACT

A bifactor item response theory model can be used to aid in the interpretation of the dimensionality of a multifaceted questionnaire that assumes continuous latent variables underlying the propensity to respond to items. This model can be used to describe the locations of people on a general continuous latent variable as well as on continuous orthogonal specific traits that characterize responses to groups of items. The bifactor graded response (bifac-GR) model is presented in contrast to a correlated traits (or multidimensional GR model) and unidimensional GR model. Bifac-GR model specification, assumptions, estimation, and interpretation are demonstrated with a reanalysis of data (Campbell, 2008) on the Shared Activities Questionnaire. We also show the importance of marginalizing the slopes for interpretation purposes and we extend the concept to the interpretation of the information function. To go along with the illustrative example analyses, we have made available supplementary files that include command file (syntax) examples and outputs from flexMIRT, IRTPRO, R, M*plus*, and STATA. Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jsp.2016.11.001. Data needed to reproduce analyses in this article are available as supplemental materials (online only) in the Appendix of this article.

1. Introduction

A range of psychometric techniques are available to determine whether a batch of manifest categorical items reflects a unidimensional or multidimensional latent construct. When a construct is presumed to be unidimensional (e.g., Autism Spectrum Disorder, social skills, self-esteem, depression, school belongingness) all items are supposed to reflect a single continuous latent variable, whereas a multidimensional construct reflects multiple continuous latent variables (factors or dimensions) that explain associations among the items and factors extracted or specified (e.g., student engagement [academic, behavior, cognitive, psychological], Appleton, Christenson, Kim, & Reschly, 2006; bullying [physical, verbal, social, electronic, racial, sexual, sexual preference], Beran, Stanton, Hetherington, Mishna, & Shariff, 2012). A popular psychometric method used for modeling unidimensional and multidimensional constructs is item response theory (IRT; De Ayala, 2009; Embretson & Reise, 2000; Lord, 1959; Lord &

☆ Data needed to reproduce analyses in this article are available as supplemental materials (online only) in the Appendix of this article.
* Corresponding author at: Department of Educational, School, and Counseling Psychology, University of Kentucky, Lexington, KY, United States.
E-mail address: toland.md@uky.edu (M.D. Toland).
Action Editor: James Peugh

http://dx.doi.org/10.1016/j.jsp.2016.11.001
0022-4405/

Novick, 1968). In general, parametric IRT models make basic assumptions regarding appropriate dimensionality, local (conditional) independence, and functional form (monotonicity).

First, it is assumed that responses to the questionnaire items are solely a function of the number of continuous latent traits specified in a given IRT model (De Ayala, 2009). So, whenever a construct is solely unidimensional all items are expected to load on a single continuous latent trait to explain the item response patterns. For example, in an early childhood reading test a single latent trait, reading literacy, could be assumed to be the basis of test performance. This assumption could be violated to varying degrees for different reasons. For instance, some items may focus on reading skills in English, whereas other items may focus on listening skills in English and then produce a multidimensional solution. Or, for another example, a unidimensional questionnaire could be developed to measure elementary students' attitude towards school belongingness with items written to tap into this heterogonous concept based on different facets (e.g., academic, social), but this could produce a multidimensional solution related to the different facets. As a result, it is common practice for researchers in school psychology to conduct dimensionality analyses to determine if the unidimensionality assumption is tenable when using a unidimensional IRT model. If multidimensionality exists, then the degree of the violation should be assessed in order to verify if the unidimensional IRT model provides a reasonable representation of the data or if a more complex multidimensional model should be used.

The second assumption, local independence, states that the probability of selecting a specific category response for an item is independent of the category selected on any other items on the questionnaire for people who have the same level of the latent trait. This assumption is violated when there still remains an association across item responses beyond the latent trait. For example, (i) a test is developed to measure children's knowledge of autism, but some items inform children on how to respond to other items. Or, (ii) some children work together as a group when responding to a set of items. Or, (iii) at some schools teachers are teaching classrooms of children about autism while other teachers are not. As another example, (iv) a questionnaire is developed to measure youth's behavioral intentions towards a youth with Attention-Deficit/Hyperactivity Disorder (ADHD) for three domains (social, academic and recreational), but items asking about a similar domain demonstrate a residual correlation even after accounting for the general trait. If any of these situations occur, then the local independence assumption is violated (conditional dependence exists) and alternative psychometric models are recommended (Gibbons & Hedeker, 1992; Reise, Morizot, & Hays, 2007). In the first and fourth examples the violation is related to the assessment (i.e., construct irrelevant variance), whereas the other two examples are related to violations due to the data structure (i.e., group effect). The methodology presented herein focuses on how to handle local item dependence issues when the first type of violation is present.

The third assumption, functional form or monotonicity, states that the appropriate IRT model has been chosen to represent the functional association between the latent trait and item response behavior. Logistic IRT models assume a monotonically increasing function represents the empirical data. This assumption is violated when people with higher levels of the latent trait tend to have lower probabilities of endorsing a higher category. For example, a school psychologist may want to measure kids' aggressiveness using a 4-category response format (e.g., Not at all, Sometimes, Often, A lot) that is assumed ordered. This assumption could be violated if for any given item kids with higher levels of aggressiveness at any interval along the latent trait (aggressiveness) have a lower tendency to actually endorse higher categories than kids with lower levels of the latent trait. This happens when kids have confusion about the order of the response options, or use response options interchangeably. Typically, functional form is checked using model-data fit statistics. We discuss model-data fit in our empirical example below. Alternatively, one could use a non-parametric IRT approach (i.e., Mokken scaling, Mokken, 1971) to assess monotonicity. There are several well-written books (Sijtsma & Molenaar, 2002) and papers (Meijer & Baneke, 2004; Sijtsma & Hemker, 2000; van der Ark, 2007) on this topic. If this assumption is not found to be tenable, then researchers may consider Mokken scaling as an alternative to parametric IRT modeling.

In general, many methods have been developed to test the tenability of IRT model assumptions. For instance, exploratory (i.e., theory developing) and confirmatory (i.e., theory testing) factor analysis techniques are commonly used to evaluate the dimensionality assumption. Similar to factor analysis, multidimensional IRT (MIRT) models have been advanced to handle multiple dimensions or latent traits (Embretson & Reise, 2000; Reckase, 2009). MIRT models are often classified as either exploratory or confirmatory (see De Boeck & Wilson, 2004); Embretson & Reise, 2000). Herein, we focus on confirmatory IRT models.

As aforementioned, dimensionality is an important assumption of IRT modeling. As described by Reise et al. (2007), Reise, Moore, and Haviland (2010), and Reise (2012), the dimensionality pattern of a particular questionnaire can be analyzed:

- using a standard unidimensional model where all item responses reflect the same common general latent trait (see Fig. 1 – Model A);
- considering more than one dimension and fitting a unidimensional model to each subscale separately, thus ignoring any correlations among subscales (Fig. 1 – Model B);
- considering more than one correlated dimensions (correlated traits or multidimensional item response theory [MIRT] model; Fig. 1 – Model C);
- through a higher-order model where dimensions (namely first-order traits) are correlated with a higher-order trait (the second-order trait; Fig. 1 – Model D);
- by specifying a special case of the MIRT model, a bifactor model, which includes a general (primary) trait that explains the items while simultaneously taking into account for the so called specific traits (Fig. 1 – Model E).

Reise et al. (2007) noted that the bifactor model is a useful compliment to the unidimensional model because it enables researchers to examine distortions in the unidimensional model when it is fit to multifaceted data and the utility of subscales. Notably, Reise (2012) indicates that the correlated traits model (see Fig. 1 – C) can be problematic because the general and specific