Measurement 90 (2016) 549–559

Contents lists available at ScienceDirect

Measurement

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

Assessing item fit: A comparative study of frequentist and Bayesian frameworks ☆

Muhammad Naveed Khalid, Cees A.W. Glas *

Department of Research Methodology, Measurement, and Data Analysis, Faculty of Behavioral Science, University of Twente, P.O. Box 217 7500, AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 6 March 2014
Received in revised form 4 May 2016
Accepted 6 May 2016
Available online 7 May 2016

Keywords:
Differential item functioning
Item fit
Item characteristics curve
Lagrange multiplier test
Local independence
Posterior predictive checks
Power
Type I error rate

ABSTRACT

Goodness of fit for item response theory (IRT) models in a frequentist and Bayesian framework are evaluated. The assumptions that are targeted are differential item functioning (DIF), local independence (LI), and the form of the item characteristics curve (ICC) in the one-, two-, and three parameter logistic models. It is shown that a Lagrange multiplier (LM) test, which is a frequentist based approach, can be defined in such a way that the statistics are based on the residuals, that is, differences between observations and their expectations under the model. In a Bayesian framework, identical residuals are used in posterior predictive checks. In a Bayesian framework, it proves convenient to use normal ogive representation of IRT models. For comparability of the two frameworks, the LM statistics are adapted from the usual logistic representation to normal ogive representation. Power and Type I error rates are evaluated using a number of simulation studies. Results show that Type I error rates are conservative in the Bayesian framework and that there is more power for the fit indices in a frequentist framework. An empirical data example is presented to show how the frameworks compare in practice.

1. Introduction

Psychometric theory is the mathematical framework for measurement in many fields of psychology and education. These measurements may concern abilities, personality traits, attitudes, opinions, and achievement. Item response theory (IRT) models play a prominent role in psychometric theory. In these models, the properties of a measurement instrument are completely described in terms of the properties of the items, and the responses are modeled as functions of item and person parameters. While many of the technical challenges that arise when applying IRT models have been resolved (e.g., model parameter estimation), the assessment of model fit remains a major hurdle for effective IRT model implementation [23].

Model checking, or assessing the fit of a model, is an important part of any data modeling process. Before using the model to make inferences regarding the data, it is crucial to establish that the model fits the data well enough according to some criteria. In particular, the model should explain aspects of the data that influence the inferences made using the IRT model. Otherwise, the conclusions obtained using the model might not be relevant. IRT models are based on a number of explicit assumptions, so the method for the evaluation of model fit focus on these assumptions. The most important assumptions underlying these models are subpopulation invariance (DIF), the form of the ICC, local stochastic independence, and item score pattern. Researchers have proposed a significant number of fit statistics for assessing fit of IRT models. These statistics developed to be sensitive to specific model violations [4,15,16,20,21,19,25,26,28–31,36,35,50,52,53]. An essential feature of these statistics is that they are based on information that is aggregated over persons; therefore they will refer to as aggregate test statistics.

To date most of the research to IRT model fit procedures has been done in a frequentist framework. Chi-square statistics are natural tests of the discrepancy between the observed and expected frequencies or proportions computed in a residual analysis. Both Pearson and likelihood ratio statistics have been proposed; these statistics have been standard tools for assessing model fit since the earliest applications of IRT.

A number of problems arise in using chi-square statistics as tests of model data fit in the IRT context. Principal among them is whether the statistics have the chi-square distribution claimed and if so, whether the degrees of freedom are correctly determined. Glas and Suárez Falcón [19] note that the standard theory for chi-square statistics does not hold in the IRT context because the

☆ Note: This research was conducted when first author was a PhD student at the University of Twente, The Netherlands.

* Corresponding author.

E-mail address: c.a.w.glas@utwente.nl (C.A.W. Glas).

http://dx.doi.org/10.1016/j.measurement.2016.05.020
0263-2241/© 2016 Elsevier Ltd. All rights reserved.

observations on which the statistics are based do not have a multinomial or Poisson distribution. Simulation studies [56,32,37,38] have shown that the fit statistics in common use do generally appear to have an approximate chi-square distribution; however, the number of degrees of freedom remains at issue. Orlando and Thissen [37] argued that because the definition of the observed proportions correct are based on model-dependent trait estimates, the degrees of freedom may not be as claimed. Stone and Zhang [51] agreed with the assessment of Orlando and Thissen [37] and further noted that when the expected frequencies depend on unknown item and ability parameters, and when these are replaced by their estimates, the distribution of the chi-square statistic is grossly affected. Glas and Suárez Falcón [19] have also criticized these procedures along the same lines for failing to take into account the stochastic nature of the item parameter estimates. The model fit indices which are based on the likelihood ratio and Wald statistics are also problematic (computational intensive) because every alternative model for every model violation for every person and each item would have to be estimated [16].

To address the above mentioned issues Glas [16] has proposed procedures based on the Lagrange multiplier (LM) statistic by Aitchison and Silvey [1]. The LM statistics estimate the IRT model only once and produce a number of tables of residuals that are informative with respect to specific model violations. An advantage of the use of LM tests is the necessity to formulate specific parametric alternatives to the assumptions targeted by test statistics. Glas have sketched the approach of LM test in the marginal maximum likelihood (MML) frame work (see, for instance, [5,34] which is the standard procedure for parameter estimation in IRT. However, MML frame work may be less efficient (in terms of computation) for multilevel and multidimensional psychometric models [12,8] due to complex dependency structures of models and require the evaluation of multiple integrals to solve the estimation equations for parameters.

These computational problems are avoided in a fully Bayesian framework and now-a-days this framework is widely used for parameter estimation in complex psychometric IRT models. When comparing the fully Bayesian framework with the MML framework the following considerations play a role. First, a fully Bayesian procedure supports definition of a full probability model for quantifying uncertainty in statistical inferences (see, for instance, [16, p. 3]). This does involve the definition of priors, which creates some degree of bias, but this can be minimized using of non-informative priors. Second, estimates of model parameters that might otherwise be poorly determined by the data can be enhanced by imposing restrictions on these parameters via their prior distributions. However this can also be done in a Bayes modal framework, which is closely related to the MML framework [34].

Recently Sinharay [47] and Sinharay et al. [48] have applied the popular Bayesian approach of posterior predictive checks (PPCs) to the assessment of model violations in unidimensional IRT models. However, PPCs is not free from criticism. Bayarri and Berger [6] have showed that PPCs also comes with problems due to twice use of data as a result posterior *p*-value were conservative (i.e., often failed to detect model misfit) and inadequate behavior of posterior *p*-values [7]. Robins et al. [41] showed that PPP values need not be uniformly distributed under null conditions, even asymptotically. Rather, the distribution is centered at .5 but is less dispersed than a uniform distribution [33,41].

The advantage of a Bayesian approach, particularly when implemented through Markov chain Monte Carlo (MCMC) sampling from the posterior distribution, is the easy calculation of the posterior distribution of any function of the estimates. However, the frequentist approach has a long standing, more rigorously developed tradition of statistical test for model fit. The purpose of this study is to introduce analogous frequentist procedures (LM test) and Bayesian procedures (PPCs) and to compare their Type I error rate and power.

This article is organized as follows. First, the model violations, assumptions targeted by item fit statistics, that examined in this study are presented. The second section introduces the description of LM statistics and PPCs. The third section outlines the design of simulation studies. Next, results from a simulation study comparing empirical Type I error rates and power for the above frameworks are presented. Then, both frameworks are applied in the context of an empirical example. Finally, some conclusions are drawn, and some suggestions for further research are given.

2. Fit to IRT models

The fit of a model, or the correspondence between model predictions and observed data, is generally regarded as an important property of model-based procedures like IRT. When a model does not fit the data, valid use of estimated parameters is compromised. IRT models are based on a number of explicit assumptions which can be viewed from two perspectives: the items and respondents. In the first case, for every item, residuals (differences between predictions from the estimated model and observations) and item fit statistics are computed to assess whether item violates the model. In the second case, residuals and person fit statistics are computed for every person to assess whether the responses to the items follow the model.

For unidimensional IRT models, a number of item fit statistics may be of interest, depending on the context of the problem. These models assume item parameters invariance, a specific shape of the ICC, local independence, fit of response pattern, and normality of the ability distribution, and each of these assumptions should be checked using suitable fit measures. The first assumption entails that the item responses can be described by the same parameters in all possible subpopulations. The shape of ICC describes the relation between the latent variable and the observable responses to items. Evaluation is usually done by comparing observed and expected item response frequencies given some measure of latent trait level. The third assumption, local independence, assumes that responses to different items are independent given the latent trait variable value. The important assumption evaluated from the perspective of person fit is the invariance of the ability parameter over sub-tests.

In a Bayesian framework, the normal-ogive representation of IRT models has a number of important computational advantages [2]. Since the objective of this article is to compare the Bayesian and the frequentist likelihood-based framework, we adopt the normal-ogive representation and also apply it to the likelihood-based framework. In the 1–2-, and 3-parameter models, it is assumed that the proficiency level of a respondent (indexed n) can be represented by one dimensional proficiency parameter θ_n. In the 3PNO model the probability of correct response to item i, denoted by $X_{ni} = 1$, as a function of θ_n is given by

$$\begin{aligned} P(X_{ni} = 1|\theta_n) &= P_i(\theta_n) \\ &= c_i + (1 - c_i)\int_{-\infty}^{a_i(\theta_n - b_i)} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-t^2}{2}\right]\partial t \\ &= c_i + (1 - c_i)\Phi(a_i(\theta_n - b_i)). \end{aligned} \tag{1}$$

Note that $\Phi(\cdot)$ is the cumulative standard normal distribution. The three item parameters a_i, b_i, c_i are called the discrimination, difficulty and guessing parameter, respectively. The 2PNO model follows upon setting the guessing parameter c_i equal to zero, and 1PNO model follows upon introducing the additional constraint $a_i = 1$.