Measurement, 11: 127–137, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366367.2013.841511

REJOINDER

Why Should We Assess the Goodness-of-Fit of IRT Models?

Alberto Maydeu-Olivares
Faculty of Psychology, University of Barcelona

In IRT measurement applications, the application of goodness-of-fit (GOF) methods informs us of the discrepancy between the model and the data being fitted (the room for improvement). By routinely reporting the GOF of our IRT models, together with the substantive results of the application of the fitted model, we will be able to learn "how bad is this fit for this purpose" and establish reasonable fit criteria, which are likely to depend on the intended use of the model. In psychological research, greater attention should be paid to modeling the process used by individuals to respond to test items. GOF methods provide an invaluable tool for this purpose as they often show that our models do not capture well the underlying response process.

Keywords: comparative fit index, close fit, test theory

Over the years, the term item response theory (IRT) has been used to mean various things, but loosely stated one can describe IRT as a set of models for how individuals may respond to educational, psychological, and so forth, test items. When I was a graduate student in the 1990s, one of the hot topics in Psychometrics was the proposal of new IRT models. I remember attending conferences and listening to one presentation after another describing the introduction of yet another IRT model and wondering, "Do we need this new model? Does it reproduce data better?" One of the highlights of IRT research during this period was the publication of the *Handbook of Modern Item Response Theory* (van der Linden & Hambleton, 1997). Each chapter of this edited volume describes an IRT model. The editors asked the authors to include a section in each chapter describing statistical methods to determine the fit of the model to data. In so doing, the reader could readily realize that goodness-of-fit methods lagged behind estimation methods for these models. Simply put, given two competing models described in the handbook, one

Correspondence should be addressed to Alberto Maydeu-Olivares, Faculty of Psychology, University of Barcelona, P. Vall d'Hebron, 171, 08035 Barcelona. E-mail: amaydeu@ub.edu

could not generally determine which of the two provided a better fit to the data, or if one could, whether the best fitting model provided a “good enough” fit. The overall Pearson’s X^2 statistic and the likelihood ratio G^2 statistic were described in most of the chapters, along with the usual warning that their *p*-values could not be trusted except for models involving just a handful of items. Also, several chapters suggested using the likelihood ratio test statistic to compare the relative fit of competing nested models, G^2_{dif}, rightfully pointing out that the asymptotic approximation to the sampling distribution of G^2_{dif} is less severely affected than for X^2 and G^2 by the sparseness of the data. However, in my opinion, when describing G^2_{dif} it was not emphasized enough that the adequacy of the asymptotic approximation relies on the largest model being correctly specified (Haberman, 1977) and years later a report was published (Maydeu-Olivares & Cai, 2006) to remind applied researchers of this fact. Chapters describing Rasch-type models described a variety of test statistics with known asymptotic distribution and good performance in small samples. Of particular importance in this area is the fundamental work of Dr. Cees A. W. Glas (see Glas, 1988, 1999, 2010; Glas & Suárez-Falcón, 2003; Glas & Verhelst, 1989, 1995),which extends beyond Rasch-type models, and toward which the Focus article does not do justice.

The Focus article describes an array of competing methods for the goodness-of-fit assessment of IRT models that have been developed by a number of researchers since van der Linden and Hambleton’s 1997 Handbook was published. To provide a historical context to these methods and to give proper credit to some of this work, I find it helpful to describe my personal journey through the field.

A PERSONAL JOURNEY

I obtained my PhD in Quantitative Psychology at the University of Illinois. While a graduate student there I was very privileged to join the Model Based Measurement Laboratory led by the late Dr. Michael Levine and by Dr. Fritz Drasgow. One of the areas of research in the lab was precisely to investigate how well existing IRT models were able to reproduce data from existing tests. We used Pearson’s X^2 statistic computed for every item, pair of items, and item triplets computed for every model under consideration. These statistics were summarized descriptively (Drasgow, Levine, Tsien, Williams, & Mead, 1995; Tay & Drasgow, 2011) to qualitatively judge the goodness-of-fit of IRT models. Although the approach can be successfully applied to compare the relative fit of competing models (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Maydeu-Olivares, 2005a), I wondered whether we could formalize the goodness-of-fit assessment process. More specifically, the asymptotic distribution of these X^2s was unknown and I wondered whether an alternative item-fit statistic with known asymptotic distribution could be used instead, and whether we could derive the joint distribution of such yet-to-be determined statistics.

My first attempt at addressing these questions was the statistic denoted as L_2 in Equation (3) of the Focus article. I introduced this statistic in my dissertation (Maydeu-Olivares, 1997). For item parameters estimated from tetrachorics (the focus of my dissertation) the statistic is not asymptotically chi-square (unless fully weighted least squares is used); yet, it often yields reasonable *p*-values in simulations. But there was no theory that supported its use for item parameters estimated by maximum likelihood (ML). Why did I focus on limited information testing methods? Partly, it was due to the influence of Dr. Roderick P. McDonald (see McDonald & Mok, 1995) also on the Illinois faculty at the time. But it was also the result of work by Reiser and VandenBerg (1994), who showed that the limited information test statistics used in limited information estimation methods were more accurate when data is sparse than the full information test statistics used with full information estimation methods (i.e., ML estimation). The implicit question raised by Reiser and VandenBerg's article was clear: How can we use limited information goodness-of-fit statistics with full information estimation methods?

After completing my dissertation, I discovered that Dr. Mark Reiser had already succeeded at addressing this question in a landmark article (Reiser, 1996) in which he proposed the first limited information overall goodness-of-fit statistic that could be used with ML estimates. This statistic is referred to as R_2 in the Focus article, see Equation (12). In subsequent work, he applied this statistic to latent class analysis as well (Reiser & Lin, 1999) and showed how to decompose it in terms of independent components (Reiser, 2008). However, there were two features of the R_2 statistic that left us unsatisfied: (1) degrees of freedom could not be determined a priori and, furthermore, they depended on the true parameter values (Reiser, 1996) and (2) the statistic could not be used with parameters estimated by methods other than ML (such as estimation from tetrachorics). In joint work with my colleague Dr. Harry Joe, we proposed (Maydeu-Olivares & Joe, 2005, 2006) a unified framework for limited and full information estimation and goodness-of-fit testing that tackled these issues. It is in these papers that we introduced the overall goodness-of-fit testing statistic M_2, which can be used with any consistent estimator and has the usual degrees of freedom (number of statistics minus number of estimated parameters). This statistic is inspired by an analogous statistic first proposed by Dr. Michael Browne in the context of covariance structure analysis (Browne, 1982). It differs from Browne's statistic in that raw moments are used (i.e., cross-products) instead of central moments (i.e., covariances) and in that the weight matrix is evaluated using parameter estimates (instead of sample moments).

Raw moments are used instead of central moments for convenience. With categorical data there is no need to use central moments, and the use of raw moments (which are marginal probabilities and proportions) facilitates deriving the asymptotic distribution of the statistics. Browne's test statistic is known to perform poorly in small sample sizes (e.g., Curran, West, & Finch, 1996). In contrast, M_2 works very well in small samples (Maydeu-Olivares & Joe, 2006). This is because in M_2 the weight matrix is evaluated under the model (using item parameter estimates), whereas in Browne's statistic the weight matrix is evaluated using sample moments. In covariance structure analysis the weight matrix cannot be computed under the model because a covariance structure model does not make any assumptions about the third and fourth order joint moments involved in the computation of the weight matrix. In contrast, because IRT models are models for the response patterns, the weight matrix can be computed under the model. This makes a big difference in the behavior of the statistics in small samples.

About this time, inspired by Satorra and Bentler (1994) I considered employing a different approach to assess the overall goodness of fit of IRT, a mean and variance correction to an easily computed limited information test statistic. I wrote a note on this idea and sent it to Dr. David Thissen at the University of North Carolina for comments. Several months later he replied to me, letting me know that 2 graduate students there, now Dr. Li Cai and Dr. Donna Coffman, had independently written a class project (!!) on precisely the same topic. We ended up publishing a report together on this approach (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006), which effectively corrects an error in Bartholomew and Leung (2002). This collaboration signaled the beginning of a very fruitful collaboration with Dr. Cai on the topic of goodness-of-fit assessment of IRT models.

In applications, IRT models are often applied to gigantic contingency tables (e.g., 5^{50}). The overall goodness-of-fit test statistics described above cannot be computed in models of this size. Simply, they are quadratic form statistics in univariate and bivariate margins and there are too many margins. To solve this problem, statistics that "condense" the information provided in the margins are needed. The necessary theory to construct such statistics was put forth in Joe and Maydeu-Olivares (2010) and was effectively applied by Cai and Hansen (2013). The resulting statistic M_{ord} (Equation 21 of the Focus article) can be successfully applied to test very large models, but only if the data are ordinal.

Yet, in such large models, it is unrealistic to expect any IRT model to fit exactly. In applications involving large models, a more fruitful avenue involves testing for approximate fit. Yet, in this case, a question immediately arises, "What cut-off to use?" And in recent work with Dr. Harry Joe, summarized in the Focus article, we have attempted to address this question by introducing RMSEAs for IRT modeling, again drawing on work from the covariance structure literature (see also McDonald & Mok, 1995).

In Maydeu-Olivares and Joe (2006), we also introduced a statistic for piece-wise diagnostics, referred to as M_{ij} in the Focus article (Equation 38), that attempted to come full circle with this personal journey. It is effectively a correction to the Pearson's X^2 statistic used in the Illinois Model Based Measurement Lab so that the resulting statistic has an asymptotic chi-square distribution. However, while on sabbatical at the University of North Carolina, a graduate student, Mr. Yang Liu, performed simulations to investigate the small sample performance of M_{ij}. The simulations revealed that while the statistic has excellent empirical Type I errors, it lacks power. This brought about the development of a full array of alternatives to M_{ij}, led by Mr. Liu, also summarized in the Focus article.

What else is to be done? A whole lot, not least is to investigate how well the different alternative statistics behave in applications, as well as in simulations. Back in the Illinois lab where I grew up as a researcher, Dr. Levine and Dr. Drasgow taught me to be demanding. An IRT model should not only fit the data well where the model is calibrated, but it should also fit well in "fresh" cross-validation samples as well (holding the parameters fixed at the values estimated in the calibration sample). When comparing the overall fit of an IRT model in calibration and cross-validation samples using X^2 (or G^2 for that matter), fit substantially worsens in the cross-validation samples. This led researchers to question the suitability of existing parametric IRT models. It turns out that this is a statistical artifact. Pearson's X^2 should not be used in cross-validation samples because it does not take into account that there are 2 sources of sampling variability (calibration and cross-validation sample) and as a result it does not follow an asymptotic chi-square distribution. In Joe and Maydeu-Olivares (2006) we introduced an alternative to X^2, referred to as X^2_{xval}, suitable for this type of cross-validation. We showed that when X^2_{xval} is used, IRT models may fit cross-validation data as well as they fit calibration data. Of course, due to data sparseness, X^2_{xval} is well approximated by asymptotic methods only in small models (just as X^2 in calibration samples). We have recently developed limited information statistics for cross-validation samples. These will enable applied researchers to test models whose fit cannot be assessed with existing methods, due to the lack of degrees of freedom. This work in progress will enable, for instance, testing the fit of nonparametric IRT models such as those (Drasgow, Levine, Williams, McLaughlin, & Candell, 1989) whose fit we tested using averages of X^2s for single items, and averages of X^2s for pairs of items when I was a graduate student at Illinois. At the personal level, this will complete a journey that has lasted 25 years.

A REJOINDER

Edwards (this issue) puts forth a very important question: Do we have in IRT modeling "production level" goodness-of-fit (GOF) statistics as we have in structural equation modeling (SEM)? The answer is an outright yes. In SEM the standard overall GOF statistics are the likelihood ratio (LR) test statistic (often referred to simply as the chi-square statistic) when data are assumed to be normality distributed, and Satorra-Bentler mean (or mean and variance) corrections to the LR statistic. The source of misfit is usually assessed using modification indices (aka score/Lagrange multiplier tests) although sometimes z statistics for residual means and covariances are also used (these are often referred to as standardized residuals). The LR test statistic is so widely used because of convenience: it is a side product of the estimation process. Therefore, one would only use alternative test statistics if they were much better than the LR statistic, which is not the case. When using ML, the limited information test statistics described in the Focus article are not a side product of the model estimation process. They are to be computed in addition to estimating the item parameters. Therefore, we can choose which test statistic to use. Based on existing research, I recommend using M_{ord} for routine evaluation of IRT models for ordinal data such as those obtained when administering Likert-type items, and also for binary data (M_{ord} reduces to M_2 in the binary case). M_{ord} can assess the fit of models of any size and it is already implemented in FlexMirt (Cai, 2012). For assessing the source of misfit I recommend using z statistics for residual means and cross-products. *R* code for computing all the statistics for piece-wise assessment reported in this paper is given in Liu and Maydeu-Olivares (2013).

What if the data are polytomous unordered? In this case M_{ord} and z statistics cannot be used as they are suitable only for ordinal (or binary) data. There are no proposals that can assess the overall fit of a model to a 5^{50} table. At most, only models with about 20 items can be tested. At present I would use M_2, which is implemented in FlexMirt and also in IRTPro (Cai, du Toit, & Thissen, 2011) and mean-and-variance corrected X^2 statistics to assess these models, but further research in this area is needed.

What if the model involves lower asymptote parameters, as the 3-parameter logistic (3PL) model, a question raised by Edwards (2013)? This issue also remains to be investigated. In principle, M_2 can be used to test this model, but it may be that Reiser's R_2 is more suitable to assess the GOF of this model.

As I have pointed out, score test statistics are more widely used in SEM for piece-wise model-fit assessment than z statistics, and Oberski and Vermunt (this issue) raise the very interesting question of whether they should be the method of choice in IRT modeling as well (instead of z statistics or other residual-based test statistics described in the Focus article). Score tests have 2 clear advantages over residual-based statistics: (a) they suggest a way to modify the model and an estimate of the value of the parameter if this were added to the model and (b) they are most powerful if the alternative model used to specify the score test is correctly specified. However, the drawback of score tests is that an alternative model needs to be specified. SEM focuses on linear models, and the alternative model used to compute the score tests (modification indices) is simply another linear model with additional linear relationships among the variables being modeled. But in IRT, nonlinear models are used and applied researchers may be interested in detecting (a) whether the shape of the nonlinear function (usually logistic) is correctly specified, (b) whether the distribution of the latent traits (usually normal) is correctly specified, and (c) whether asymptote parameters should be added to the response function. What alternative model should be used in this case?

Consider the PROMIS anxiety example described in the Focus article. A unidimensional graded-response model (Samejima, 1969) was fitted to that data. I could have fitted Muraki's (1992) generalized partial credit model instead. Both models have the same number of parameters but they are not equivalent and the difference in fit is not great (Maydeu-Olivares, Drasgow, & Mead, 1994). In this case, Bock's (1972) nominal model can be used as an alternative model. Whereas for each item, both the graded and generalized partial credit model have (loosely speaking) a common slope but a different intercept associated to each category, Bock's model has (subject to identification constraints) a different slope and intercept associated to each response category. Thus, Bock's model is a suitable alternative model to be used in a score test of Muraki's model. Such a score test would inform us if, for any given item, different slopes should be used. Better yet, Thissen and Steinberg's (1984) model could be used instead of Bock's model, as their model adds a lower asymptote parameter to each response category. Thus, the use of Thissen and Steinberg's model as an alternative model in a score test would inform us of the suitability of lower asymptote parameters and/or different slope parameters per item.

But for the graded model, what alternative (less restricted) model should be used in computing score tests? The only alternatives that I am familiar with are a threshold-drift model (Glas, 1999) and a bifactor model (Liu & Thissen, 2013), and Liu and Thissen (2013) show that different results are obtained depending on the alternative model used. How useful are score tests to detect departures from the model they were not intended to detect (nonlogistic shape functions or nonnormal latent trait distributions)? This is an open question that can only be addressed by simulations. Currently, I believe that score tests are most useful in applications for which there is a clear alternative model in mind. One such example is the one described by Oberski and Vermunt (this issue). Another example is when multidimensionality is suspected and a bifactor alternative can be employed. Yet, another example involves the investigation of differential item functioning or when there is a clear alternative model (such as Muraki's model being nested within Bock's model).

An alternative to the use of score statistics is to simply use residual-based statistics such as a mean-and-variance-adjusted X^2, or a z statistic. These statistics simply inform us of which items or pairs of items do not fit well, but they do not suggest ways to modify the model (except when multidimensionality is suspect or when we may remove items from the test). But if we feel we need to modify the model (because it does not even provide a close fit to the data), we can look at the data using fit plots for the items flagged as misfitting (Chernyshenko et al., 2001; Drasgow et al., 1995; Maydeu-Olivares, Morera, & D'Zurilla, 1999; Stark, Chernyshenko, Drasgow, & Williams, 2006). Clearly, future research should compare the relative performance of score statistics to residual-based statistics in detecting the source of misfit.

As we have seen, there are strong analogies between GOF developments in IRT modeling and in SEM (see also Maydeu-Olivares, 2005b), but research on GOF testing in SEM is ahead of similar research in IRT, and Cai and Monroe (this issue) suggest adapting the most useful tools in the SEM GOF toolbox to incorporate them into the IRT GOF toolbox. One of the most useful of the SEM GOF tools is undoubtedly the Tucker-Lewis goodness-of-fit index. Hence, Lee and Cai's (2012) counterpart is a most welcome addition to the IRT GOF toolbox. The Tucker-Lewis (1973) index, as well as other similar indices such as the Comparative Fit Index (CFI: Bentler, 1990), compare the fitted model to a baseline model. My main concern with regard to comparative fit indices in general is the choice of baseline model. In the factor analysis literature, an independence model is almost invariably used as the baseline model. Lee and Cai also use an independence model as baseline for what they refer to as a zero-factor model). But, if we believe that the items are independent, why do we fit a factor analysis or an IRT model? To put it differently, if we fit a factor analysis or an IRT model because we try to account for the observed associations in the data, why do we use an independence model as baseline to assess its fit? I firmly believe that these indices are useful to gauge the relative performance of competing models using the least parameterized model as baseline. For instance, they can be used to compare a bifactor model to a 1-dimensional model using the latter as baseline. But these indices should not be used to compare a bifactor model to an independence model, and a 1-dimensional model to an independence model to help choosing between the bifactor model and the 1-dimensional model. I believe that these indices can be used to gauge model-data fit as well, but to do so, we need to use a meaningful baseline model other than the independence model.

ASSESSING THE GOODNESS-OF-FIT OF IRT MODELS

Why should we assess the goodness-of-fit of IRT models? Because IRT modeling is about identifying a plausible process that individuals may have used to respond to items. Consequently, we must assess how well the models we are fitting are doing their job. It is true as Edwards (this issue) suggests that some of the models that we use are so simplistic that they are unrealistic. My reply to this concern is twofold.

The first part of my reply is let's make them more realistic, that is, let's improve our models. Much of my substantive research has focused on measuring psychological constructs using ratings. Our models for these data assume that the answer to the second item being administered depends only on the traits being measured, and it does not depend on the response to the first item. I do not think this is a reasonable assumption in all testing situations, but it is an assumption we invariably make. Also, we now strongly suspect that individuals respond differently to items depending on their direction (positively worded or negatively worded). Our current models struggle when both positively and negatively worded items measuring the same construct are included in a questionnaire. They often lead us to believe that we are measuring 2 constructs (such as optimism and pessimism). I just think that we are using models that are too simplistic. Currently, IRT modeling focuses mainly on how well we measure individuals and, to a lesser extent, on the underlying decision process, As a psychologist, I am keenly interested in the response process itself, in modeling that process, and therefore in assessing the fit of the model. For instance, in a landmark article, Thissen and Steinberg (1986) classified parametric IRT models into difference models—the graded response model— and divide-by-total models, which includes Bock's nominal model and Muraki's generalized partial credit model. In my experience modeling rating data, the graded response model always fits better than Muraki's model (and Bock's model rarely outperforms the graded model). For an illustration of this point using descriptive methods, see Maydeu-Olivares (2005a). This suggests to me that the process used by individuals to respond to these items is more closely approximated by the graded-response model.

The second part of my reply is that we must assess the GOF of our IRT models to determine how well we are doing when measuring unobservable constructs (how much room for improvement—statistically speaking—there is). Engelhard and Perkins (this issue) describe the 2 prevalent traditions to psychological measurement. Within the Rasch modeling tradition, assessing goodness-of-fit comes naturally and suitable GOF statistics have existed for quite some time. Items are carefully constructed and selected to fit the desired model. And as the second example of the Focus article reveals, we have become very good at it. We are able to construct tests such that the selected model cannot be rejected. But outside a Rasch modeling tradition, items are also constructed and selected, and a model is fitted. Within this model-data tradition (using Engelhard and Perkins's terminology) we are also keenly interested in assessing the GOF of the fitted model, for if the intended model is rejected, we wish to determine how far away we are from the data we are fitting, and whether the piece-wise GOF statistics suggest an easy fix.

Common sense should be used when judging the results of a GOF assessment because in principle it is easier to model a 2^5 contingency table that a 5^{50} one. In the first case, one should strive to find a model that cannot be rejected, whereas in the second case it is not realistic to expect to find such a model. Also, it is easier to find a model that cannot be rejected with 300 observations (we do not have much power) than with 3,000. In the latter, if the model is rejected, researchers should check the magnitude of the misfit (using for instance residual correlations if the data are ordinal).

A common concern runs through most of the comments about the Focus article (Cai & Monroe; Edwards; and Thissen (all in this issue). Paraphrasing Cai and Monroe, we see some SEM practitioners being unnecessarily obsessed with the GOF of their models, failing to devote the necessary time to investigate the usefulness of their model and its substantive interpretation. The concern is that the introduction of these new IRT GOF methods brings about a similar phenomenon to IRT research. This is a well-founded concern that I share. Does this mean that we should not assess the fit of IRT models? Certainly not. The overall GOF of a model informs us of the discrepancy between the model and the data. By routinely reporting the GOF of our IRT models, together with the substantive implications of the application of the fitted model, we will be able to learn "how bad is this fit for this purpose" and establish reasonable criteria, which will depend on the intended use of the model. Surely, a model that shows a substantial degree of misfit may prove useful for purpose A (but not necessarily for purposes B and C). But certainly, there are limits. What those limits are is what we ought to determine. And we need GOF statistics to do that.

Thissen (this issue) point out that sometimes tests are constructed to serve a purpose and that in these cases one should use a specific test statistic that checks whether the purpose is indeed served. I agree. The statistics described in the Focus article are all-purpose. There is no guarantee that they have power against all possible alternatives (intended purposes of the test). Different test statistics have different power we regard to different alternatives. This is why the piece-wise statistics in the anxiety PROMIS example do not completely agree on what items are misfitting. Cai and Monroe (this issue) report a test statistic that has higher power than M_2 (an all-purpose statistic) to detect an IRT model with a multimodal distribution. When interest lies in specific hypotheses, we may need specific tests (see also Thissen, this issue). But the converse may also be true—that all-purpose statistics such as M_2 are as powerful in detecting specific departures from the model as specific-purpose statistics (Maydeu-Olivares & Montaño, 2013).

By using a common GOF metric in our IRT studies together with the substantive results obtained, we will be able to establish cut-off criteria for different purposes. This is, in my view, how we can overcome the danger of being over-zealous about the GOF of our IRT models. For instance, if the intended purpose of the test is test linking, by reporting a common GOF metric, possibly complemented by a specific purpose statistic (Thissen, this issue), we may learn that a cut-off value of X on the common GOF metric denotes a close enough fit for test linking, but based on previous studies that used the common GOF metric we know that X is not enough to make reliable inferences about the dimensionality of the construct being measured. This is where I would like us to go.

REFERENCES

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*(1), 1–15. doi:10.1348/000711002159617

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–46.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. doi:10.1007/BF02291411

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141). Cambridge, United Kingdom: Cambridge University Press.

Cai, L. (2012). FlexMIRT: A numerical engine for multilevel item factor analysis and test scoring. [Computer program]. Seattle, WA: Vector Psychometric Group.

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modelling. [Computer software]. Chicago, IL: Scientific Software International.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245–276. doi:10.1111/j.2044-8317.2012.02050.x

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2 tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173–194. doi:10.1348/000711005X66419

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562. doi:10.1207/S15327906MBR3604_03

Curran, P., West, S., & Finch, J. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(l), 16–29.

Drasgow, F., Levine, M. V, Tsien, S., Williams, B., & Mead, A. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143–165.

Drasgow, F., Levine, M. V, Williams, B., McLaughlin, M. E., & Candell, G. L. (1989). Modeling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement*, *13*, 285–299.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*(4), 525–546.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*(3), 273–294. doi:10.1007/BF02294296

Glas, C. A. W. (2010). Item parameter estimation and item fit analysis. In Wim J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 269–288). New York, NY: Springer. doi:10.1007/978-0-387-85461-8

Glas, C. A. W., & Suárez-Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106. doi:10.1177/0146621602250530

Glas, C. A. W., & Verhelst, N. (1989). Extensions of the partial credit model. *Psychometrika*, *54*(4), 635–659.

Glas, C. A. W., & Verhelst, N. (1995). Testing the Rasch model. In G. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 69–96). New York, NY: Springer.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics*, *5*, 1148–1169.

Joe, H., & Maydeu-Olivares, A. (2006). On the asymptotic distribution of Pearson's X2 in cross-validation samples. *Psychometrika*, *71*(3), 587–592. doi:10.1007/s11336-005-1284-z

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*(3), 393–419. doi:10.1007/s11336-010-9165-5

Lee, T., & Cai, L. (2012, July). *A note on a Tucker-Lewis index for item response theory modeling*. Paper presented at the 2012 International Meeting of the Psychometric Society, Lincoln, NE.

Liu, Y., & Maydeu-Olivares, A. (in press). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*.

Liu, Y., & Thissen, D. (2013). Local dependence score tests for the graded response model. Unpublished manuscript.

Maydeu-Olivares, A. (1997). *Structural equation modeling of binary preference data. Dissertation Abstracts International: Section B*. University of Illinois.

Maydeu-Olivares, A. (2005a). Further empirical results on parametric versus non-parametric irt modeling of Likert-type personality data. *Multivariate Behavioral Research*, *40*(2), 261–279. doi:10.1207/s15327906mbr4002_5

Maydeu-Olivares, A. (2005b). Linear IRT, non-linear IRT, and factor analysis: A unified framework. In A. Maydeu-Olivares & J. J. Mcardle (Eds.), *Contemporary Psychometrics. A Festchrift for Roderick P. McDonald* (pp. 73–100). Mahwah, NJ: Erlbaum.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G 2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research*, *41*(1), 55–64. doi:10.1207/s15327906mbr4101_4

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement*, *18*(3), 245–256. doi:10.1177/014662169401800305

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2 n contingency tables. *Journal of the American Statistical Association*, *100*(471), 1009–1020. doi:10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. doi:10.1007/s11336-005-1295-9

Maydeu-Olivares, A., & Montaño, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*, *1*, 116–133.

Maydeu-Olivares, A., Morera, O., & D'Zurilla, T. J. (1999). Using graphical methods in assessing measurement invariance in inventory data. *Multivariate Behavioral Research*, *34*(3), 397–420. doi:10.1207/S15327906MBR3403_5

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of Fit in Item Response Models. *Multivariate Behavioral Research*, *30*(1), 23–40. doi:10.1207/s15327906mbr3001_2

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. doi:10.1177/014662169201600206

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika*, *61*(September), 509–528.

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*, *61*(Pt2), 331–360.

Reiser, M., & Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. *Sociological Methodology*, *29*(1), 81–111. doi:10.1111/0081-1750.00061

Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, *47*(1), 85–107. doi:10.1111/j.2044-8317.1994.tb01026.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 17*.

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. Von Eye & C. C. Clogg (Eds.), *Latent variable analysis. Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring?, *Journal of Applied Psychology*, *91*(1), 25–39. doi:10.1037/0021-9010.91.1.25

Tay, L., & Drasgow, F. (2011). Adjusting the a2/df ratio statistic for dichotomous item response theory analyses: Does the model fit? *Educational and Psychological Measurement*, *72*(3), 510–528. doi:10.1177/0013164411416976

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501–519. doi:10.1007/BF02302588

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.