A model-based approach to goodness-of-fit evaluation in item response theory

DL Oberski JK Vermunt
Dept of Methodology and Statistics, Tilburg University, The Netherlands

Author note

This work was supported by the Netherlands Organization for Scientific Research (NWO) [Vici grant 453-10-002]. Correspondence should be addressed to DL Oberski, Department of Methodology and Statistics, Faculty of Social Sciences, Tilburg University, Warandelaan 2, PO Box 90153, 5000LE, The Netherlands. E-mail: doberski@uvt.nl.

We congratulate Albert Maydeu on his lucid and timely overview of goodness-of-fit assessment in IRT models, a field to which he himself has contributed considerably in the form of limited-information statistics (Joe & Maydeu-Olivares, 2010; Maydeu-Olivares & Joe, 2005).

In this commentary we would like to focus on two aspects of model fit: 1) what causes there may be of misfit, and 2) what consequences misfit may have. We present our view on these topics in an integrated framework that is slightly different from that presented by Prof. Maydeu. In the following section we elaborate on these points. Subsequently we provide a short illustration using an IRT analysis and, finally, draw some conclusions.

A model-based approach to goodness-of-fit

Why does the model not fit?

Any IRT model with more than zero degrees of freedom can be seen as a restricted version of an alternative model. For example, the Rasch (1PL) model can be seen as a 2PL model in which all discrimination parameters have been set equal; the standard local independence model can be seen as a model in which bivariate dependencies between items have been set to zero; and an IRT model with covariates implicitly restricts direct effects of covariates and interactions between covariates and the latent trait to zero, respectively corresponding to uniform and non-uniform item biases (DIF). These restrictions can be seen as the cause of model misfit. The goal of goodness-of-fit assessment is then to detect which restrictions should be freed.

A tool from statistics to detect misspecified restrictions is the score test (Rao, 1948). The score test is also known as the "Lagrange multiplier test" and was introduced to IRT by Glas (1998, 1999) and van der Linden & Glas (2010). Denoting the restrictions corresponding to the fitted IRT model as $a_0 = 0$ and the sample parameter estimates under that model as $\hat{\theta}$, if we consider some alternative set of restrictions, $a_a = 0$, the "score" $s(\hat{\theta})$ is the vector of first derivatives of the likelihood with respect to the parameters of the *alternative* model, evaluated at the ML estimates of the *restricted* model, $s(\hat{\theta}) = \frac{\partial L}{\partial \theta}\Big|_{\theta=\hat{\theta}}$, where L is the likelihood. If a restriction is true in the population, then its $s(\hat{\theta})$ should equal zero; therefore a test of the null hypothesis that the restriction to be investigated holds in the population is H_0: $s(\hat{\theta}) = 0$, for which the test statistic T_a can be constructed as

$$T_a = A_a' \, s(\hat{\theta})' (A_a' \, I(\hat{\theta}) \, A_a \,)^{-1} s(\hat{\theta}) A_a \,,$$

where $A_a \ = \partial a_a / \partial \theta$, and I is the information matrix with respect to all possible parameters. Asymptotically, T_a, which is referred to as the "modification index" in structural equation modeling (Saris, Satorra, & Sörbom, 1987), will follow a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the alternative model and the model fitted.

The score test is well known to be asymptotically equivalent to the Wald test after freeing a set of parameters, but it does not require the estimation of the possibly large number of alternative models, and has in simulations and applications often been found more accurate (Agresti, 2002). It is also asymptotically equivalent to a likelihood ratio test of the restriction; again an important advantage is that the researcher need not estimate many alternative models when using score tests. In IRT modeling, in which large numbers of items are often modeled and numerical integration is usually required, this is an attractive feature. Thus, the advantages of the score test are that it provides a well-studied test of specific restrictions warranting direct interpretation in terms of an alternative model, while at the same time the many possible alternative models need not be estimated.

Oberski, Van Kollenburg, & Vermunt (2013) show that the score test and the *ad hoc* bivariate residual (BVR) tests discussed by Maydeu are in certain cases equivalent. For example, the BVR between pairs of indicators amounts to an uncorrected score test of the restriction that a local dependence parameter needs to be introduced. Maydeu's mean-and-variance corrected BVR should therefore closely approximate the score test of this restriction. For further details we refer to Oberski et al. (2013).

While the score test has several advantages, it shares an important disadvantage with other goodness-of-fit statistics such as the BVR: it only assesses statistical significance of a misspecification, not its substantive importance. A measure assessing the substantive size of the misspecification is the so-called "expected parameter change" (EPC_{self}), well-known in structural equation modeling (Saris et al., 1987), and introduced to latent class modeling by Oberski & Vermunt (forthcoming). The EPC_{self} is the change in a particular restricted parameter itself that can be expected if this restriction were freed. It can be defined as

$$\text{EPC}_{\text{self}} = P_{\text{self}}\big(A_a' \, I(\hat{\theta})A_a\,\big)^{-1} s(\hat{\theta})A_a\,,$$

where P_{self} selects the possibly misspecified parameter itself (Bentler & Chou, 1992). Like the score test and other goodness-of-fit measures, it can be calculated without estimating alternative models. Unlike the score test, however, it assesses the substantive size of the misspecification rather than merely the statistical significance.

At this point we would like to comment on Maydeu's statement that "the goodness of approximation of an IRT model should be regarded as the effect size of its misfit". While we think the idea behind this statement takes us in the right direction, it is, in our opinion, not completely correct. As shown by Saris, Satorra, & Van der Veld (2009), fit indices, including the RMSEA, can be highly sensitive to certain misspecifications and not others due to aspects of the model that are unrelated to the size of the misspecification itself. In IRT, for example, misspecifications related to an item with low marginal information will be less easily detected than those for items with high marginal information, *even when the misspecifications are exactly the same size*. In this sense, therefore, we think that Maydeu's statement is a good starting point, but somewhat too simplistic when applied to RMSEA and similar measures. The score test and expected parameter change statistics indeed provide "the effect size of misfit", at least the effect of misfit on the misspecified parameter itself.

What difference does it make that the model does not fit?

Any IRT analysis is performed with some goal in mind. Another perspective on the "effect of misfit" is therefore that it should refer to the consequence of misspecification for certain parameters of interest. For example, an IRT analysis may be performed to examine differences in ability between boys and girls. In that case the parameter of interest is the regression coefficient of the covariate "gender" on the latent ability (indicating the difference in means between boys and girls). This regression coefficient will be biased if boys and girls do not react equivalently to some item, i.e. when there is item bias or differential item functioning. From this perspective the main question is not whether there is some nonzero item bias or not, but rather what the effect of such item bias might be on the parameter of interest. This point was made in the context of invariance testing by Borsboom (2006), and in the context of selection in educational testing by Millsap (1995, 1997, 1998, 2007). Recently, Oberski (2013) suggested using a measure that assesses the change in the parameter of interest when freeing a possibly misspecified restriction, the "EPC-interest". It can be defined as

$$\text{EPC}_{\text{interest}} = P_{\text{interest}}\left(A'_a\, I(\hat{\theta})A_a\,\right)^{-1} s(\hat{\theta})A_a\,.$$

Like EPC$_{\text{self}}$, this yields an estimate of the parameter change if a particular restriction were freed. The only difference between EPC$_{\text{interest}}$ and the more widely known EPC$_{\text{self}}$ is that instead of the change in the parameter itself, the selection matrix P_{interest} selects the parameter(s) of interest. It directly provides the effect of misspecification on the goals of the IRT analysis being conducted.

Illustration with an IRT model with covariates

We now demonstrate the use of the score test, EPC-self, and EPC-interest in an IRT analysis of an 18-item math test taken from 2156 eight grade pupils, a data set collected by Doolaard (1999). The goal of this analysis was to assess the effects of four covariates – whether the school participates in the national school leaving Cito examination (Cito; 0=no, 1=yes), nonverbal intelligence (ISI; standardized), socioeconomic status (SES, standardized), and gender (0=male, 1=female) – on a pupil's latent ability. A 2PL model was estimated, including effects of the covariates on the latent ability. This yielded effect size estimates for the covariates Cito, ISI, SES, and gender of 0.631 (s.e. 0.073), 0.654 (s.e. 0.046), 0.323 (s.e. 0.034), and -0.292 (s.e. 0.058), respectively.

Misfit may potentially affect the effect size estimates of the covariates, particularly when there is item bias, i.e. residual relationships between the items and the covariates. Therefore these bivariate residuals, shown in Table 1, are examined. The Table shows that the bivariate Pearson residuals between items and covariates do indicate some misfit, if 4 is taken as a cutoff value as suggested by Vermunt & Magidson (2005). However, though larger values do indicate possible item bias, these BVRs do not asymptotically follow a chi-square distribution (Tay, Vermunt, & Wang, 2013). One of the corrections suggested by Maydeu could be used. These corrected statistics should provide very similar values to the score tests for freeing direct effect parameters from covariates to items, shown in the subsequent columns. It can be seen that these score tests for uniform item bias are generally larger than the bivariate residuals. They do follow the same pattern however: certain direct effects appear to be statistically significant. The last four columns of Table 1 reports the EPC-self values indicating whether the direct effects of the covariates are also substantively large. Indeed this appears to be the case for some of them. Particularly the effect of gender on item y11 is statistically significant and rather large on a logit scale (0.610). Therefore this item appears to have considerable item bias, something that in principle could potentially affect the parameters of interest.

Inclusion of this direct effect in the model indeed yields a statistically significant and substantively large value (0.627), close to its EPC-self value. However, the resulting change in the parameters of interest after estimating this alternative model turns out to be rather small, namely -0.002, -0.003, -0.002, and -0.032, for the effects of Cito, ISI, SES, and gender, respectively. This could also have been seen without estimating the alternative model by examining the EPC-interest values: these are -0.002, -0.002, -0.002, and -0.030, respectively. The fact that these are also the largest EPC-interest values indicates that although there are several misspecifications (statistically significant item biases which are ignored), none of these strongly affects the covariate effects on the latent trait that are the parameters of interest.

Conclusion

This commentary expanded on Maydeu's overview of goodness-of-fit testing in item response theory by providing a model-based framework for GOF assessment. Within this framework, the score test, EPC-self, and EPC-interest are useful tools for GOF evaluation that aid the researcher in determining 1) in what sense the model fits the data and in what sense it does not; and 2) whether possible misfit is important or not. Many unresolved issues remain with the use of these measures, such as the problem of equivalent models, multiple testing, dealing with dependencies between parameters, and a shift of responsibility for deciding what may cause misfit and what is “of interest” towards the practitioner. These issues are, however, not absent in other goodness-of-fit assessment procedures, but merely less explicit. Therefore the model-based perspective appears to us to be a fruitful area of future research.

References

Agresti, A. (2002). *Categorical data analysis*. New York: John Wiley & Sons.

Bentler, P. M., & Chou, C.-P. (1992). Some new covariance structure model improvement statistics. *Sociological Methods & Research*, *21*(2), 259–282.

Borsboom, D. (2006). When does measurement invariance matter? *Medical care*, *44*(11), S176–S181.

Doolaard, S. (1999). *Schools in Change Or Schools in Chains? The Development of Educational Effectiveness in a Longitudinal Perspective*. Twente University Press.

Glas, C. A. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica*, *8*(3), 647–667.

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

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*(3), 393–419.

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

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*(4), 577–605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*(3), 248.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research*, *33*(3), 403–424.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461–473.

Oberski, D. (2013). Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models. *Political Analysis*.

Oberski, D., Van Kollenburg, G., & Vermunt, J. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, *7*(3).

Oberski, D., & Vermunt, J. (forthcoming). The Expected Parameter Change (EPC) for Local Dependence Assessment in Binary Data Latent Class Models.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Proceedings of the Cambridge Philosophical Society* (Vol. 44, pp. 50–57).

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological methodology*, *17*, 105–129.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*(4), 561–582.

Tay, L., Vermunt, J. K., & Wang, C. (2013). Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning. *International Journal of Testing*, *13*(3), 201–222.

Van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*(1), 120–139.

Vermunt, J. K., & Magidson, J. (2005). Technical guide for Latent GOLD 4.0: Basic and advanced. *Belmont Massachusetts: Statistical Innovations Inc*.

Table 1

CITO example 2PL model. Item bias (DIF) analysis: goodness-of-fit assessment of the restrictions that the direct effects of covariates on the indicators are zero.

	Bivariate residuals (BVR's)				Score statistics (MI's)				Expected parameter changes (EPC's)			
	cito	isi	ses	sex	cito	isi	ses	sex	cito	isi	ses	sex
y1	11.0	9.5	0.3	0.0	11.8	13.5	0.3	0.0	-0.44	-0.24	0.03	-0.02
y2	2.8	1.5	1.4	5.7	3.0	2.2	1.7	6.1	0.24	-0.11	0.09	0.30
y3	0.3	0.2	0.1	5.6	0.3	0.4	0.1	6.0	0.07	-0.04	0.02	-0.26
y4	1.1	1.0	2.0	0.2	1.4	1.6	2.4	0.2	-0.18	-0.09	-0.10	-0.06
y5	1.0	0.0	0.5	1.8	1.1	0.0	0.5	1.9	0.13	0.00	0.04	-0.14
y6	4.2	0.6	4.4	0.0	4.5	0.9	5.0	0.0	-0.25	0.06	-0.11	0.00
y7	1.0	0.1	0.0	1.8	1.0	0.1	0.0	1.9	-0.13	-0.02	0.00	-0.15
y8	7.0	1.2	0.9	0.1	7.6	2.0	1.1	0.1	-0.44	0.12	-0.08	0.04
y9	0.0	0.9	0.4	0.4	0.0	1.4	0.4	0.5	0.01	0.08	-0.03	0.07
y10	0.4	0.1	2.1	0.2	0.4	0.1	2.5	0.2	0.09	-0.02	0.09	-0.05
y11	1.9	0.6	4.2	**26.2**	2.0	0.9	4.9	**27.5**	0.19	0.07	-0.14	**0.61**
y12	0.1	3.1	0.5	0.1	0.1	4.8	0.6	0.1	-0.03	0.13	0.04	-0.03
y13	0.6	1.7	0.3	11.6	0.7	2.9	0.4	12.2	-0.16	-0.18	-0.06	0.60
y14	0.2	1.0	0.0	1.7	0.2	1.5	0.0	1.8	0.08	0.10	-0.01	-0.18
y15	11.4	2.9	0.2	0.5	12.2	4.3	0.2	0.5	0.46	-0.14	0.03	-0.08
y16	2.2	11.1	0.3	0.6	2.3	17.3	0.4	0.7	0.18	0.27	-0.03	-0.09
y17	0.6	2.4	2.6	1.5	0.6	4.1	3.2	1.6	-0.14	0.19	0.15	0.20
y18	0.7	0.3	0.7	3.5	0.7	0.5	0.8	3.8	0.11	-0.04	0.05	-0.21