Chapter 8

Don't Tie Yourself to an Onion: Don’t Tie Yourself to Assumptions of Normality

Maarten Marsman, Gunter Maris and Timo Bechger

Abstract A structural measurement model (Adams, Wilson, & Wu, 1997) consists of an item response theory model for responses conditional on ability and a structural model that describes the distribution of ability in the population. As a rule, ability is assumed to be normally distributed in the population. However, there are situations where there is reason to assume that the distribution of ability is nonnormal. In this paper, we show that nonnormal ability distributions are easily modeled in a Bayesian framework.

Keywords: Bayes estimates, finite mixture, item response theory, Gibbs sampler, Markov chain Monte Carlo, one-parameter logistic model, plausible values

Introduction

A structural measurement model (Adams, Wilson, & Wu, 1997) consists of an item response theory (IRT) model for responses conditional on ability and a structural model that describes the distribution of ability in the population. At Cito, structural measurement models are used for test equating and to relate student characteristics to response behavior.

As a rule, we assume that ability is normally distributed in the population. We do this because it is easy and because we often do not have a clear alternative. However, there are situations where we have reason to assume that ability is not normally distributed in the population, for instance, when we know that the population consists of students who are selected from a larger, possibly normal population, based on one of our own tests. Thus, the onion in our title refers to the assumption of normality to which structural measurement models seem intimately tied. The most common form of nonnormality of the ability distribution found in Cito applications is skewness. This has enticed Molenaar (2007) and Verhelst (2008) to develop more general models that can handle skew ability distributions. They developed complex procedures to estimate these models using maximum likelihood methods, and the procedure developed by Verhelst is implemented in the Cito program SAUL.

We propose to use a (finite) mixture of normal distributions to model different forms of nonnormality, such as skewness, kurtosis, and multimodality. The structural measurement model is formulated in a Bayesian framework, and the Gibbs sampler is used to estimate the parameters. The Gibbs sampler requires a sample from the posterior distribution of ability, and the mixture plays the role of the prior distribution of ability. Draws from the posterior of ability are called plausible values (PVs; Marsman, Maris, Bechger, & Glas, 2011; Mislevy, 1991), and once they are obtained, estimating the parameters from the mixture becomes a routine exercise.

The structure of the paper is as follows. First, we introduce a real-data example to motivate our concerns about the assumption of normality. Then, we outline a Bayesian procedure that is then applied to the data of the motivating example. The paper ends with a discussion.

Motivating Example: Entreetoets Data

We use data of N = 136,495 students responding to k = 39 math items of the Cito Entreetoets. The measurement model is the one-parameter logistic model (OPLM; Verhelst & Glas, 1995). Since the OPLM is an exponential family (EF) IRT model, it can be fitted independently from the structural model using conditional likelihood methods. The parameters were estimated and showed reasonable fit.

We estimated the parameters of the structural model using the Gibbs sampler as described in Marsman et al. (2011) using noninformative priors. Marsman et al. (2011) developed several algorithms that use the method of composition (Tanner, 1993) to sample PVs from the posterior distribution of ability conditional on the response data. Because the OPLM is in the EF, the posterior of ability is characterized by its sufficient statistic, and the conditional composition algorithm for EF IRT models (the CC-EF algorithm) can be used. Furthermore, Marsman et al. show how recycling intermediate candidate values can increase efficiency when students come from few marginal distributions using the CC-EF-R algorithm.

The Markov chain Monte Carlo (MCMC) algorithm ran for 1,000 iterations, which is sufficient due to the low amount of autocorrelation and thus results in almost immediate convergence of the Markov chain. The expected a posteriori (EAP) estimates and posterior standard deviations are $\hat{\mu}$ = 0.1804185 (0.0004750) and $\hat{\sigma}$ = 0.1599017 (0.0003887).

To study the fit of the estimated model, we compared the observed (weighted sum) score distribution with the generated score distribution under the model; see Figure 1.

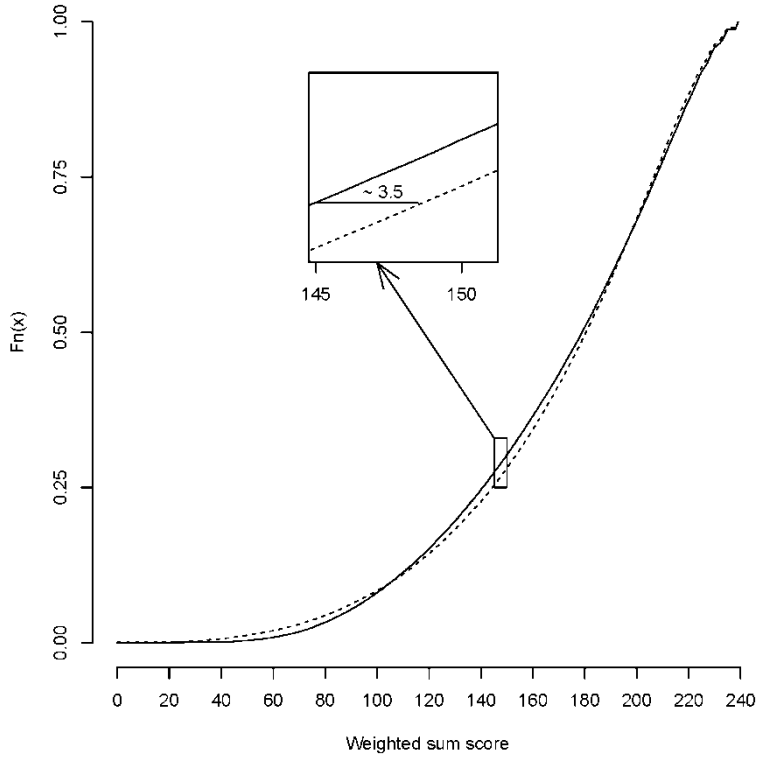


Figure 1 Observed (solid) and replicated (dashed) weighted score distributions

There were discrepancies between the observed and generated data. The magnified section in Figure 1 shows a section of the score distributions where they differ in approximately 3.5 score points. The geometric mean of the item weights in the OPLM model was set at 6, so a difference of 3.5 points would be a little less than 0.6 raw score points.

Table 1 Equating the Score Distributions

Score	Observed	Generated with normal	Generated with mixture	Difference with normal	Difference with mixture
0	0	8	0	8	0
25	11	282	45	271	34
50	455	1,602	700	1,147	245
75	3,309	4,981	3,514	1,672	205
100	11,000	11,358	10,730	358	270
125	23,671	22,080	23,672	1,591	1
150	41,325	38,513	41,757	2,812	432
175	64,197	62,108	64,184	2,089	13
200	92,872	93,487	92,882	615	10
225	125,543	126,655	125,473	1,112	70

The observed differences may have large implications in test equating. This is illustrated in Table 1, which contains six columns. The first column is a score used as a possible cut-off point in an equating procedure. The number of persons who received that score or lower as observed in the sample or generated with a normal density are given in the second and third columns, respectively. In the fifth column, we look at the difference between the number of persons we observe with what we expect under the model. These differences are not anywhere near zero, illustrating that model misfit can have large implications.

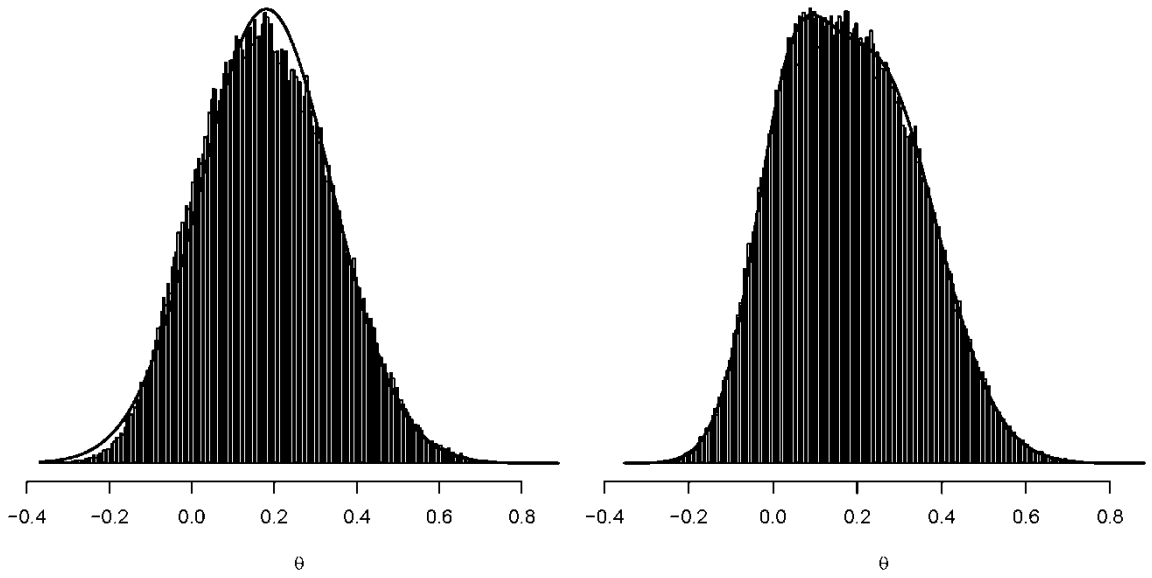
Using Plausible Values

To estimate the parameters from the structural model, we used PVs, which are draws from the posterior of ability. We want the PVs to have the same distribution as ability in the population.

However, since we do not know how ability is distributed, we introduce a structural model and use it as a prior distribution. The posterior distribution now has two ingredients:

1. The likelihood (IRT model): we can add more items, which will make the likelihood dominate the prior distribution so that the posterior converges to the true posterior distribution.
2. The prior distribution: we can adjust the prior to match information in the likelihood, so that the likelihood more easily dominates the prior and the posterior converges to the true posterior distribution.

When the measurement model is firmly established, it is guaranteed that the distribution of PVs is closer to the true posterior distribution than the estimated structural model. We illustrate this in Figure 2(a), where a histogram of generated PVs with a normal distribution are given along with the estimated normal density. Clearly, the measurement model makes the PV distribution negatively skewed and as a result does not match the prior distribution.



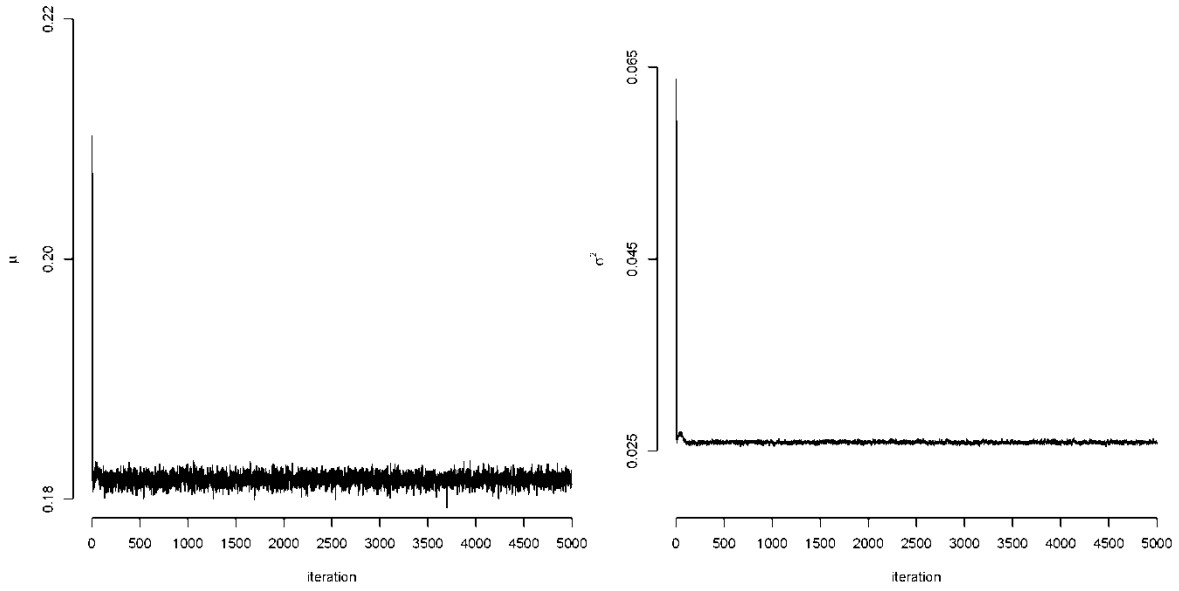
(a) Normal population model. (b) Mixture population model.

Figure 2 Histograms of PVs and estimated density

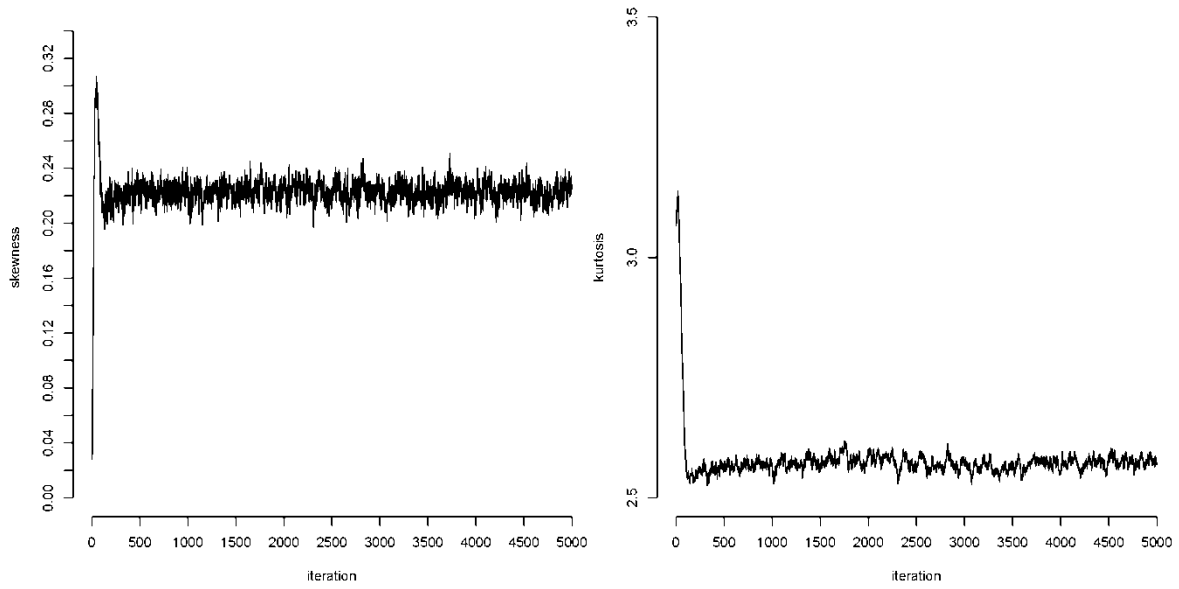
Motivating Example: Reanalysis

We can speed up convergence of the PV distribution to the true posterior distribution by improving the fit of the structural model and adjusting it to better match the information in the likelihood. We do this by using a mixture of two normal distributions. Standard methods for estimating a normal mixture using Gibbs samplers are readily available, and we refer the interested reader to Congdon (2010, Chapter 3) or Fox (2010, Chapter 6).

The assignment to a component was modeled as a binomial random variable and given a flat Beta prior, uniform over the range [0,1]. The parameters from the structural model were assigned noninformative priors as in the previous example. This is not convenient when there is a risk of (almost) empty mixture components, in which case informative priors can be assigned. The parameters from the mixture distribution were estimated using the Gibbs sampler, which ran for 5,000 iterations. Convergence of the estimated distribution was relatively fast as can be seen by inspecting the traceplots of the mean, variance, skewness, and kurtosis of the mixture model, shown in Figure 3.



(a) Mean of mixture model. (b) Variance of mixture model.



(c) Skewness of mixture model. (d) Kurtosis of mixture model.

Figure 3 Trace plots of the mean, variance, skewness, and kurtosis of the mixture distribution

Because of more freely estimating the population model, the distribution of PVs converged to the true posterior ability distribution. This can be seen in Figure 2(b), where a histogram of generated PVs with the mixture distribution are given along with the estimated density. The distribution of PVs and the estimated structural model are aligned, confirming that it has converged to the true posterior distribution. Since PVs are now a sample from the structural model, we can use them for other purposes. For instance, in large-scale educational surveys, such as the Program for International Student Assessment (PISA) and the European Survey on Language Competences (ESCL), PVs are provided for secondary analyses.

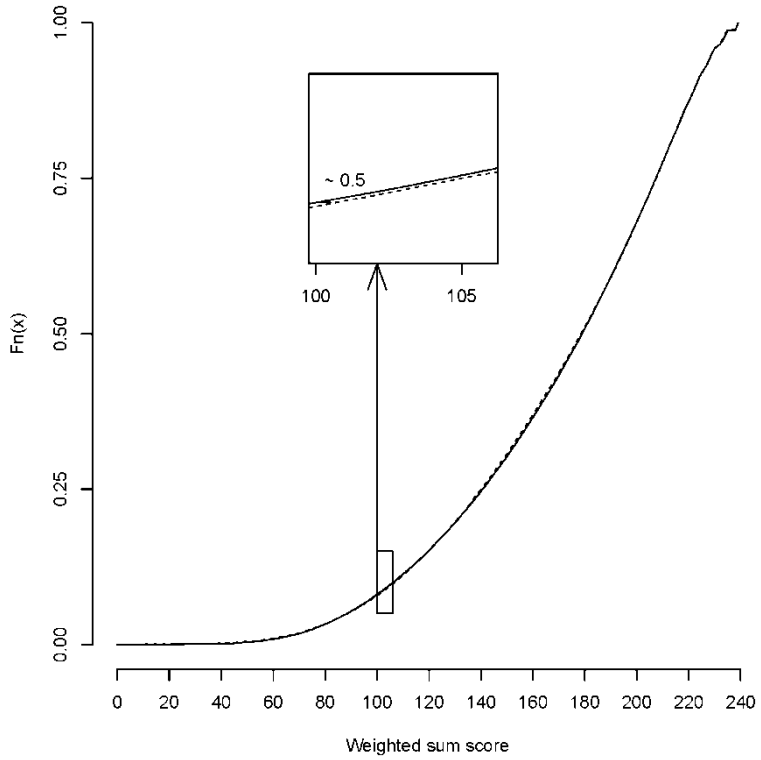


Figure 4 Observed (solid) and replicated (dashed) weighted score distributions at iteration 2,000

The fit of the structural measurement model was assessed via a comparison of observed and replicated score distributions as before; see Figure 4. The mixture distribution provides a better fit to the data than the normal distribution as provided.

The magnified section in Figure 4 shows that there are still discrepancies, although they are less severe than those found in Figure 1. The observed difference of less than of 0.5 points on the weighted score scale refers to less than 0.1 raw score points, approximately.
This can still have a large effect in test equating. The fourth column in Table 1 shows the number of persons who received the cut-off score or lower as generated with a mixture. The sixth column shows the difference between the number of persons we observed with what we expected under the mixture model. Clearly these numbers are still substantial. The question is: do we have to explain the results to the parents of 2,812 or of 432 children, which is large either way, but substantially larger under the normality assumption.

Discussion

As a rule, assumptions of normality (the onion) are introduced in test equating, although situations exist where there is reason to believe that the ability distribution is not normal. In this paper, we showed that we can easily model deviations from normality in a Bayesian

framework by using PVs and a mixture distribution as a prior. Compared to regular applications of mixture distributions, we are not interested in its components but use it merely for curve-fitting. As an illustration, we applied a mixture of two normal distributions to Entreetoets data. It is easy to extend the mixture to include more components at a low cost in terms of additional parameters. The mixture is flexible and can model many deviations from normality, such as skewness, kurtosis, and multimodality.

In addition, ignoring deviations from normality can have serious effects on test equating. Furthermore, if the ability distribution shows similar deviations in repeated assessments and these deviations are ignored, the effects can add up in the equating procedure. As a result, the projected norm can drift away from the originally proposed norm. Thus, it is very important to correctly model the ability distribution to provide valid inference in test equating. The mixture solution is easily applied in test equating, and the Entreetoets example shows that it can improve model fit and consequently provide better predictions.

References

Adams, R., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*,*22* ,47–76.

Congdon, P. (2010). *Applied Bayesian hierarchical methods*. Boca Raton, FL: Chapman & Hall/CRC Press.

Fox, J. (2010). *Bayesian item response modeling.* New York, NY: Springer.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2011). *A conditional composition algorithm for latent regression.* (Measurement and Research Department Report No. 11-2). Cito.

Mislevy, R. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.

Molenaar, D. (2007). *Accounting for non-normality in latent regression models using a cumulative normal selection function.* (Measurement and Research Department Report No. 07-3). Cito.

Tanner, M. (1996). *Tools for statistical inference* (3rd ed.). New York, NY: Springer.

Verhelst, N. (2008). *Untitled document containing updated theory and a short addition to the manual for the structural analysis of a univariate latent variable (SAUL) program*. Cito.

Verhelst, N., & Glas, C. (1995). Rasch models; foundations, recent developments, and applications. In G. Fischer & I. Molenaar (Eds.), *One Parameter logistic model* (pp. 215-238). New York, NY: Springer-Verlag.