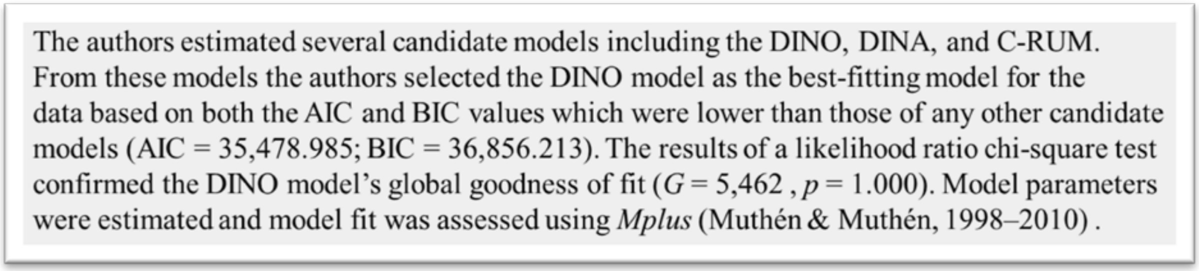
Questions and Sample Answers for Chapter 12

Question 1

What are the problems with using the χ^2-based goodness-of-fit statistics that are provided by *Mplus* as a default to assess the model fit of DCMs?

Question 2

While conducting a literature review on research to model reading comprehension, you find a recently authored paper that applies DCMs to data from a large-scale administration of a dichotomously scored 25-item standardized reading assessment (*N*=1,674). From the abstract, the article appears to be a perfect fit for your research needs. Scanning the paper, you read the following excerpt which describes the authors' assessment of data-model fit:



The authors estimated several candidate models including the DINO, DINA, and C-RUM. From these models the authors selected the DINO model as the best-fitting model for the data based on both the AIC and BIC values which were lower than those of any other candidate models (AIC = 35,478.985; BIC = 36,856.213). The results of a likelihood ratio chi-square test confirmed the DINO model's global goodness of fit (G = 5,462 , p = 1.000). Model parameters were estimated and model fit was assessed using *Mplus* (Muthén & Muthén, 1998–2010) .

What do you make of these conclusions? What information, if any, could the authors have provided about model-data fit to strengthen their argument for the DINO model being an appropriate model for the data under consideration?

Question 3

Describe the difference between absolute, relative, item, and person fit statistics. How should they be used jointly to make a decision about model-data fit?

Question 4

Discuss how posterior predictive model checking (PPMC) is similar to, and different from, other types of fit assessment as described in this chapter.

Section 2 – Sample Answers

Question 1

In categorical data analyses, χ^2 statistics are used to compare how similar the observed proportions of respondents are to the expected proportions of respondents under the assumption of statistical independence. A significant χ^2 test would lead us to reject the hypothesis that variables of interest are independent. An alternate test statistic that is frequently used to evaluate the distribution of categorical data is the G-statistic. The G-statistic is a maximum-likelihood or likelihood-ratio-based test statistic that is very similar to χ^2 statistic. One important caveat for using χ^2 test statistic is that contingency tables should have a sample size that is large enough to ensure a sufficient minimum number of observations in each cell. For example, Agresti and Finlay (1997) recommend that the expected frequency should exceed 5 in each cell.

The problems with using general goodness-of-fit statistics for DCMs are as follows. In a diagnostic assessment, the number of possible response patterns is contingent upon the number of items; specifically, *I* items on the assessment will result in 2^I potential response patterns that result in a large global contingency table for differences between observed and expected response vectors. Consequently, the χ^2 statistic for a statistical test of residual independence among item responses, which is what the DCM tries to achieve, would have to be computed across all response patterns. This is computationally impossible in almost all applications, which is why programs like *Mplus* attempt to simplify the problem by deleting sparse cells. This results in the loss of a large amount of statistical information, however, and results of different global tests can sometimes be contradictory.

In addition, DCMs are designed to perfectly replicate the marginal proportion correct for each item, thus making goodness-of-fit tests for individual items useless. For item pairs it can be possible to compute χ^2 statistics but those would have to be visually displayed or otherwise aggregated to allow for an overall evaluation of the global goodness-of-fit of a given model. This can be cumbersome in practice, but is certainly no different from similar challenges within the area of pair-wise item fit assessment using a posterior predictive model checking (PPMC) approach under a Bayesian estimation framework, for example.

Question 2

The conclusions drawn in this paper are somewhat problematic because the authors do not provide sufficient information to allow a critical reader to judge the degree of model-data fit at a finer level of analytic grainsize. Specifically, AIC and BIC indices only provide information about the relative goodness-of-fit amongst candidate models. To their credit the authors include multiple measures of relative fit, as these information criteria can suggest different models as being the best model for the data. Yet, even with multiple relative fit indices one can have the situation that any of the candidate models fits poorly in an absolute sense.

The authors do provide the likelihood-ratio statistic (*G*) calculated by *MPlus*, but it is not a reliable index of global model fit. An assessment with 25 dichotomously scored items yields 33,554,432 possible response patterns and with only 1,674 respondents the statistic is calculated using a sparse contingency table. It is worth noting that the authors did not report the results of the global χ^2 test as appears in the *MPlus* output as it is likely to have given results contradictory to the likelihood-ratio test. In order to demonstrate adequate absolute fit, the authors could have utilized resampling methods to generate an empirical distribution of full information statistics such as the chi-squared statistic, or they could have used limited information statistics which compare the expected and observed dependencies among pairs or even triplets of items. If the authors were working within a fully Bayesian framework, they could have used posterior predictive model checking (PPMC) methods to examine both global and item-level model-data fit.

Question 3

The assessment of model-data fit at different levels of the model is a crucial step for any real-data application of DCMs. Researchers generally distinguish between relative, absolute, item, and person fit even though the latter three are all variations on the idea that model-predicted response patterns are strikingly different from observed response patterns.

Relative model-data fit is typically one of the first steps in any model-data fit assessment enterprise when competing candidate models have been estimated as one, or perhaps a few, candidate models can be chosen. For nested models, comparisons can be done using likelihood-ratio test statistics and for non-nested models – as well as nested models technically – information-based indices such as AIC, BIC, or DIC can be used. Absolute model-data fit is an assessment of whether a particular model fits well when information about all items and respondents is considered jointly – or if information from individual items or respondents is aggregated. This is often the second step in model-data fit assessment.

If some indication of poor model-data fit is present using absolute indices, researchers are often interested in determining where the problems come from. Are there specific items or respondents that appear to misfit? Are there groups of items or groups of respondents that appear to misfit? A solid model-data fit assessment enterprise tries to identify and remedy these issues, which often takes on the form of (a) understanding the relationship between the issues and patterns of fit statistics and (b) performing sensitivity analyses that involve deleting particular items or respondents from the data set. Research on item and person fit statistics has evolved since the publication of the book and additional technical articles are now available. Nevertheless, in practice, researchers are just beginning to pay closer attention to fit assessment at different levels.

Question 4

PPMC is a framework for the specification, estimation, and interpretation of fit statistics, so-called discrepancy measures, under a Bayesian estimation framework. Literature in this area has highlighted that researchers should define or identify fit statistics that are sensitive to the particular kinds of misfit that they expect to be present in their data set.

PPMC is grounded in a fully Bayesian estimation approach and technically involved the computation of predicted response data from the posterior predictive distribution, which involves the posterior distribution of the model parameters and the likelihood of the data. As always with fit statistics, observed and predicted response patterns at some level of aggregation can then be compared.

Interestingly, the idea of PPMC can also be applied to frequentist estimation approaches because maximum-likelihood estimators are asymptotically normally distributed. Thus, if one is willing to make this asymptotic assumption in any given context, one could use the estimated standard errors for the parameter estimates and draw from normal distributions that are specified with them in order to then generate predicted data that can be compared to observed data. In other words, incorporating uncertainty about parameter estimates to compute fit statistics at some level of definition is possible within frequentist and Bayesian estimation frameworks even if the field of PPMC has popularized this notion with a Bayesian estimation framework specifically.