# Chapter 5 – Model Selection

The models described above were each introduced by their developers with a particular rationale in mind. Generally, little empirical evidence has been tendered to substantiate the rationales for each model, and in the case of the PCM, the original rationale for its development had been demonstrated to be invalid (see e.g., Molenaar, 1983; Verhelst & Verstralen, 1997, and the earlier discussion of item "steps"). Nevertheless, each model has come to acquire a life of its own, often finding potential application in contexts beyond the reasons for their original development. For example, each of these eight models can legitimately be applied to modeling Likert-type rating data, even though only three of them were explicitly developed to do so (RSM, SIM, & RS-GRM. The question remains however, as to how a measurement practitioner should choose from among the range of models available.

## General Criteria

Data characteristics will often simplify a practitioner's choice. For example, if the data is scored dichotomously, then the choice between an adjacent category and a cumulative boundaries model becomes moot. Similarly, continuous or nominally scored data precludes the use of a model for ordered categorical data – unless some data manipulation is first performed. Finally, if all items to be modeled do not have the same number of categories, then a rating scale model is ruled out.

Beyond data characteristics, models can be selected on the basis of measurement philosophy. Samejima (1995; 1996; 1997a) argues that the fundamental issue in measurement is whether or not a model faithfully reflects the psychological reality that produced the data, and she provides a set of specific mathematical criteria for testing a model. Rasch model proponents, in contrast, argue that the related principles of specific objectivity and fundamental measurement, which some Rasch models (but no non-Rasch models) meet,

should be the criteria by which models are chosen. Specific objectivity (Rasch, 1966) requires that comparisons among item parameters be independent of person parameter values, and vice versa. This is closely related to the mathematical principle underlying sufficient statistics (Rasch, 1977; van der Linden, 1994).

## Mathematical Approaches

Selecting IRT models based on mathematical criteria typically involves some method for assessing model-data fit, where the model that appears to best fit a set of data is selected. One such method is to graphically analyze fit. An exploratory heuristic (Kingston & Dorans, 1985) involves comparing model predicted response functions with empirical response functions (McKinley & Mills, 1985; 1989). A more formal approach is to graphically analyze residuals (Ludlow, 1985; 1986) where residuals are essentially the difference between a person's observed and expected response.

### Statistical Fit Tests

The most common method for attempting to assess model-data fit is to use some form of statistical fit test. These can be grouped into four general classes: (1) residual-based measures; (2) multinomial distribution based-tests; (3) response function based-tests; and (4) Guttman error based-tests.

Fit measures can also be classified in terms of the level of generality of their application. Fit can be assessed at the global level, in terms of the fit of an entire dataset from a complete measurement instrument. It can also be assessed in terms of the fit of specific groups of items from a test, if specific hypotheses about fit are to be tested. Finally, fit can be assessed in terms of the fit of individually modeled items to a set of data.

#### Residual-based measures

For direct residual-based measures, a simple response residual comprises the difference between observed and expected item responses. This can be standardized by dividing the

residual by the standard deviation of the observed score (Masters & Wright, 1997; Wright & Masters, 1982; Wu, 1997).

Response residuals can be summed over respondents to obtain an item fit measure. Generally, the accumulation is done with squared standardized residuals, which are then divided by the total number of respondents to obtain a mean square statistic. In this form, the statistic is referred to as an unweighted mean square (Masters & Wright, 1997; Wright & Masters, 1982) and has also come to be known as "outfit" (Smith, Shumacker & Bush, 1998; Wu, 1997) perhaps because it is highly sensitive to outlier responses (Adams & Khoo, 1996; Smith, et al., 1998; Wright & Masters, 1982).

A weighted version of this statistic was developed to counteract its sensitivity to outliers (Smith, 2000). In its weighted form the squared standardized residual is multiplied by the observed response variance and then divided by the sum of the item response variances. This is sometimes referred to as an information weighted mean square and has become known as "infit" (Smith, et al., 1998; Wu, 1997).

Since the statistic is approximately distributed as a mean square, however, it is not symmetrical about the mean (Smith, 2000). Partly in an effort to obtain a statistic that allowed a single critical value (Smith, et al., 1998) a cube root transformation of the mean square statistics was proposed. The resulting, standardized statistics are referred to as unweighted and weighted *t*-statistics, respectively. They have approximate unit normal distributions.

**Multinomial distribution-based tests**

In the context of general statistical theory, polytomous IRT models can be considered exponential models based on discrete data (Kelderman, 1996). Asymptotic goodness-of-fit statistics can be used with such data (Kelderman & Rijkes, 1994), with the most direct test of

fit based on the multinomial distribution of response patterns, given that the model holds true (Andersen, 1997).

For $n$ items, each with $m$ categories, there are $n^m$ possible response patterns. The joint distribution of all response patterns has the same likelihood as the multinomial distribution (Andersen, 1997). A fitted model can be tested against the general multinomial model by comparing observed and expected frequencies for the response patterns. A number of test statistics that are asymptotically distributed as $\chi^2$ are available to test differences between observed and expected frequencies (Kelderman, 1996). These include Pearson's $\chi^2$, the log-likelihood ratio statistic ($\chi^2$), the Freeman-Tukey statistic, and the Neyman modified $\chi^2$. These are themselves all special cases of a general statistic for the fit of observed to expected frequencies (Cressie & Read, 1984). Cressie and Read also provided a compromise statistic for use when the type of alternative hypothesis being tested is unknown, as is usually the case in model testing. This class of fit tests is particularly relevant in the context of MLE (Baker, 1992) which is prevalent in IRT model estimation.

**Response function-based tests**

This type of fit assessment was initially proposed as a means for studying person fit. In this approach, instead of observed and expected frequencies of response patterns, observed and expected log-likelihoods of individual item responses are subtracted and standardized (Rost & von Davier, 1994).

**Guttman error-based tests**

An alternative to the response function-based tests is to take a non-parametric approach consisting of counting the number of Guttman response errors across item pairs. A more comprehensive approach is to weight the Guttman errors. The $Q_i$ statistic (Rost & von Davier, 1994) is such a statistic and is essentially a ratio of log-likelihood ratios (Karabatsos,

2000). The test is a simple function of conditional item response patterns, Guttman response patterns, and anti-Guttman response patterns.

**Fit Statistic Problems**

Each of the four general approaches to statistically testing fit has its own problems. The residual measures are based on unknown distributional properties (Masters & Wright, 1997) which some consider to be dubious (Rogers & Hattie, 1987). The lack of known distributional properties makes it difficult to justify critical values for the test statistic, with the result that several different values have been proposed (Ludlow & Haley, 1992; Smith, et al., 1998; Wu, 1997).

The major difficulty with any of the multinomial distribution-based $\chi^2$ tests is that they require impossibly large response samples when more than a dozen items with three or four response categories each are used. Otherwise, the expected frequencies of response patterns become small and the test statistics' approximations to a $\chi^2$ distribution becomes poor (Kelderman, 1996).

While the response function-based approach is promising, very little research has been conducted using it in the context of polytomous IRT models. Concerns remain regarding the asymptotic properties of the statistics, particularly with estimated model parameters and in the face of model violations. Similarly, the main drawback of the $Q_i$ statistic is that there has been very little research to determine whether it operates as intended.

Two general problems arise with tests of model fit. First, the power of all the fit tests described above strongly depends on the variance of the statistic (Rost & von Davier, 1994). That is, when item trait locations are close to respondents' $\theta$ levels, there is no good way to know whether there is lack of fit. This is a well recognized problem with no obvious solution. Furthermore, as with all inferential test statistics, these fit tests are sensitive to

sample size. Given sufficiently large sample sizes, even small deviations from model fit will be identified as statistically significant.

An important theoretical concern also threatens the validity of fit tests. Baker (1992) notes that, in practice, IRT modeling involves the application of curve-fitting techniques to the observed proportions of category responses in the hope that fit is sufficiently adequate to justify faith in the model being fitted. However, Garcia-Pérez and Frary (1991) point out that testing fit under this approach involves the fundamental contradiction that fit is measured after parameters are estimated to fit the model as well as possible. In other words, it is not possible to test the fit of IRT models against data because there is no way to test the adequacy of modeled response functions independently of the parameter estimation process (Garcia-Pérez & Frary, 1991).

A more sophisticated approach to choosing a model than solely considering fit statistics, is to simultaneously take into account the goodness-of-fit of a model and the number of parameters modeled to achieve that fit (Sclove, 1987). This typically involves a penalty term that increases with the number of parameters in the fitted models. Akaike's information criterion (AIC: Akaike, 1977) implements a form of model selection along these lines, and has occasionally been used in IRT (e.g., Wilson, 1992; Wilson & Adams, 1993; 1995). Some research suggests that the AIC is not asymptotically consistent (see Sclove, 1987) and it is not clear that the AIC is appropriate for comparing models with different types of parameters, such as adjacent category and cumulative boundary models.

An extension of the AIC that is based on the likelihood ratio of two comparison models, rather than a single baseline model (Maydeu-Olivares, Drasgow & Mead, 1994), is the ideal observer index (IOI: Levine, Drasgow, Williams, McCusker & Thomasson, 1992). This index is more appropriate than the AIC when comparing models with different types of parameters. However, the asymptotic statistical properties of this index have not been well

investigated and there is some indication that it is sensitive to both test length and sample size (Maydeu-Olivares et al., 1994). Furthermore, computation of the IOI is not straightforward, in part because the index is estimated in a simulation study that involves a high burden of time and effort (van Engelenburg, 1997). Ultimately, this index is more suitable as a research tool than a practical method for choosing among models in a practical testing context.

**An Example**

A simple demonstration of the use of fit tests with the data used in the PCM, RSM, GPCM and GRM practical examples from the previous two chapters is provided here. The estimated item parameter values for the four models are provided again in Table 5.1. This helps to remind us that the RSM is the most restrictive model of the four, while the GPCM and GRM are the least restrictive models by virtue of the number and type of parameters estimated. Table 5.1 also provides *p*-values for individual item fit statistics for the item that was modeled to provide these parameter values.

PARSCALE (Muraki & Bock, 1999) provides a multinomial distribution-based $\chi^2$ test of item fit for each model. The *p*-values in Table 5.1 show that this item best fits the GPCM; fits the GRM somewhat less well and fits the PCM and RSM least well with no differentiation among the last two models in terms of probability of fit. By comparison, WinMira (von Davier, 2000) provides a Guttman error-based standardized $Q_i$ statistic but only for the PCM and RSM. This program does not estimate GPCM or GRM parameters. Unlike the PARSCALE fit test, the $Q_i$ statistic is able to distinguish between the PCM and RSM, in probability terms, showing that the more restrictive RSM has a distinctly poorer probability of fitting this item.

Tests of fit were also provided by each program for the entire 12-item questionnaire from which this example item was taken. However, even with almost 1,000 responses to the 12 items these data were still too sparse to reliably estimate the overall fit statistics.

## Differences in Modeled Outcome

Underlying most methods for choosing among different IRT models is the implicit assumption that the models produce different results. However, there is little demonstrated evidence that different polytomous IRT models do produce substantially different measurement outcomes when applied to the same data. Some of the few comparative studies of polytomous IRT models suggest that it is of little consequence which model is used (e.g., Dodd, 1984; Maydeu-Olivares et al., 1994; Ostini, 2001; van Engelenburg, 1997; Verhelst, Glas & de Vries, 1997).

In the context of a discussion about the consequences of choosing the "wrong" model Wainer and Thissen (1987) note that inaccuracies in the estimation of item parameters becomes error in the $\theta$ estimates. However, in their research with dichotomous models they found that the bias in trait estimation that resulted from using the wrong response model was small and could be corrected, to some extent, by using robust estimation techniques. Yen (1981), on the other hand, listed potential problems with fitting the wrong dichotomous model to a given data set. Similar research does not yet exist for polytomous IRT models.

Unfortunately, common software routines for implementing different polytomous models typically employ different fit tests, which can produce widely different indications of model fit (Ostini, 2001). The result is that reliance on model/software specific fit indicators can result in different data being excluded from analyses, which subsequently results in the appearance of differences in model functioning, even though the difference is primarily the effect of fit test variations.

## Conclusion

Considering the fundamental structural differences between the two major types of polytomous IRT models, it is certainly a matter of some interest as to whether they produce demonstrable, meaningful differences in measurement outcome. It is not clear that this

question can be answered by recourse to statistical procedures, either graphical or goodness-of-fit tests. A definitive conclusion will also not always be provided by considerations of the different measurement philosophies implicit in the two approaches. In part, this is because the importance of specific objectivity is not universally agreed upon. However, it is also partly due to the fact that there are models based on adjacent categories that do not meet the requirements for models to provide specific objectivity. In the case of polytomous models therefore, differences in measurement philosophy do not boil down to the question of specific objectivity. Similarly, Samejima's criteria do not automatically rule out adjacent category models. So even philosophical criteria may not differentiate polytomous IRT models.

Ultimately, the best approach to model selection should begin with a consideration of data characteristics. It may then be influenced by measurement philosophy considerations. Tests of model-data it would then ideally involve input from multiple sources including tests of specific model assumptions, as well as goodness-of-fit tests, either graphical or statistical, or both, together with sound judgement and substantive expertise.