Chapter 8
The "New Psychometrics" – Item Response Theory

Classical test theory is concerned with the reliability of a test and assumes that the items within the test are sampled at random from a domain of relevant items. Reliability is seen as a characteristic of the test and of the variance of the trait it measures. Items are treated as random replicates of each other and their characteristics, if examined at all, are expressed as correlations with total test score or as factor loadings on the putative latent variable(s) of interest. Characteristics of their properties are not analyzed in detail. This led Mellenbergh (1996) to the distinction between theories of tests (Lord and Novick, 1968) and a theories of items (Lord, 1952; Rasch, 1960). The so-called "*New Psychometrics*" (Embretson and Hershberger, 1999; Embretson and Reise, 2000; Van der Linden and Hambleton, 1997) is a theory of how people respond to items and is known as *Item Response Theory* or *IRT*. Over the past twenty years there has been explosive growth in programs that can do IRT, and within R there are at least four very powerful packages: **eRm** (Mair and Hatzinger, 2007), **ltm** Rizopoulos (2006), **lme4** (Doran et al., 2007) and **MiscPsycho**, (Doran, 2010). Additional packages include **mokken** (van der Ark, 2010) to do non-metric IRT and **plink** (Weeks, 2010) to link multiple groups together. More IRT packages are being added all of the time.

In the discussion of Coombs' *Theory of Data* the measurement of attitudes and abilities (2.9) were seen as examples of comparing an object (an item) to a person. The comparison was said to be one of order (for abilities) or of distance (for attitudes). The basic model was that for ability there is a latent value for each person, θ_i and a latent difficulty or location[1] δ_j for each item. The probability of a particular person getting a specific ability item correct was given in Equation 2.14 and is

$$prob(correct|\theta,\delta) = f(\theta - \delta) \tag{8.1}$$

while for an attitude, the probability of *item endorsement* is

$$prob(endorsement|\theta,\delta) = f(|\theta - \delta|) \tag{8.2}$$

and the question becomes what are the functions that best represents the data.

[1] The original derivation of IRT was in terms of measuring ability and thus the term *item difficulty* was used. In more recent work in measuring quality of life or personality traits some prefer the term *item location*. Although I will sometimes use both terms, most of the following discussion uses difficulty as the term used to describe the location parameter.

At the most general level, the probability of being correct on an item will be a *monotonically increasing* function of *ability* while the probability of endorsing an attitude item will be a *single peaked* function of the level of that *attitude*. Although the distinction between ability and attitude scaling is one of ordering versus one of distance seems clear, it is unclear which is the appropriate model for personality items. (Ability items are thought to reflect maximal competencies while personality items reflect average or routine thoughts, feelings and behaviors.) Typical analyses of personality items assume the ordering model (Equation 8.1) but as will be discussed later (8.5.2) there are some who recommend the distance model (Equation 8.2). The graphic representation of the probability of being correct or endorsing an item shows the *trace line* of the probability plotted as a function of the latent attribute (Lazarsfeld, 1955; Lord, 1952). Compare the hypothetical trace lines for ability items (monotonically increasing as in Figure 2.9) with that of attitude items (single peaked as in Figure 2.10). A trace line is also called an *item characteristic curve* or *icc* which should not be confused with an *Intra-Class Correlation* (*ICC*). Two requirements for the function should be that the trait (ability or attitude) can be unbounded (there is always someone higher than previously measured, there is always a more difficult item) and that the response probability is bounded (0,1). That is $-\infty < \theta < \infty$, $-\infty < \delta < \infty$ and $0 < p < 1$. The question remains, what are the best functions to use?

8.1 Dichotomous items and monotonic trace lines: the measurement of ability

An early solution to this question for the ability domain was proposed by Guttman (1950) and was a deterministic step function with no model of error (Equation 2.16). However, a person's response to an item is not a perfect measure of their underlying disposition and fluctuates slightly from moment to moment. That is, items do have error. Thus two common probabilistic models are the *cumulative normal* (2.17) and the *logistic* model (2.18). Although these models seem quite different, with the addition of a multiplicative constant (1.702) these two models appear to be almost identical over the range from -3 to 3 (Figure 2.8) and because the logistic function is easier to manipulate, many derivations have been done in terms of the logistic model. However, as Samejima (2000) has noted, even with identical item parameters, these two models produce somewhat different orderings of subjects. Even so, it is probably clearer to first discuss the logistic function and then consider some alternative models.

8.1.1 Rasch Modeling - one parameter IRT

If all items are assumed to equally good measures of the trait, but to differ only in their difficulty/location, then the *one parameter logistic* (*1PL*) *Rasch model* (Rasch, 1960) is the easiest to understand:

$$p(correct_{ij}|\theta_i, \delta_j) = \frac{1}{1 + e^{\delta_j - \theta_i}}. \tag{8.3}$$

That is, the probability of the i^{th} person being correct on (or endorsing) the j^{th} item is a logistic function of the difference between the person's ability (latent trait) (θ_i) and the item

difficulty (or location) (δ_j). The more the person's ability is greater than the item difficulty, the more likely the person is to get the item correct. To estimate a person's ability we need only know the probability of being correct on a set of items and the difficulty of those items. Similarly, to estimate item difficulty, we need only know the probability of being correct on an item and the ability of the people taking the item. Wright and Mok (2004) liken this to the problem of a comparing high jumpers to each other in terms of their ability to jump over fences of different heights. If one jumper is more likely to jump over a fence of a particular height than is another, or equally likely to clear a higher fence than the other, it is the ratio of likelihoods that is most useful in determining relative ability between people as well as comparing a person to an item.

The probability of missing an item, q, is just 1 - p(correct) and thus the *odds ratio* of being correct for a person with ability, θ_i, on an item with difficulty, δ_j is

$$OR_{ij} = \frac{p}{1-p} = \frac{p}{q} = \frac{\frac{1}{1+e^{\delta_j-\theta_i}}}{1-\frac{1}{1+e^{\delta_j-\theta_i}}} = \frac{\frac{1}{1+e^{\delta_j-\theta_i}}}{\frac{e^{\delta_j-\theta_i}}{1+e^{\delta_j-\theta_i}}} = \frac{1}{e^{\delta_j-\theta_i}} = e^{\theta_i-\delta_j}. \tag{8.4}$$

That is, the odds ratio will be a exponential function of the difference between a person's ability and the task difficulty. The odds of a particular pattern of rights and wrongs over n items will be the product of n odds ratios

$$OR_{i1}OR_{i2}\ldots OR_{in} = \prod_{j=1}^{n} e^{\theta_i-\delta_j} = e^{n\theta_i}e^{-\Sigma_{j=1}^{n}\delta_j}. \tag{8.5}$$

Substituting P for the pattern of correct responses and Q for the pattern of incorrect responses, and taking the logarithm of both sides of equation 8.5 leads to a much simpler form:

$$ln\frac{P}{Q} = n\theta_i + \sum_{j=1}^{n}\delta_j = n(\theta_i + \bar{\delta}). \tag{8.6}$$

That is, the log of the pattern of correct/incorrect for the i^{th} individual is a function of the number of items * (θ_i - the average difficulty). Specifying the average difficulty of an item as $\bar{\delta} = 0$ to set the scale, then θ_i is just the logarithm of P/Q divided by n or, conceptually, the average logarithm of the p/q

$$\theta_i = \frac{ln\frac{P}{Q}}{n}. \tag{8.7}$$

Similarly, the pattern of the odds of correct and incorrect responses across people for a particular item with difficulty δ_j will be

$$OR_{1j}OR_{2j}\ldots OR_{nj} = \frac{P}{Q} = \prod_{i=1}^{N} e^{\theta_i-\delta_j} = e^{\Sigma_{i=1}^{N}(\theta_i)-N\delta_j} \tag{8.8}$$

and taking logs of both sides leads to

$$ln\frac{P}{Q} = \sum_{i=1}^{N}(\theta_i) - N\delta_j. \tag{8.9}$$

Letting the average ability $\bar{\theta} = 0$ leads to the conclusion that the difficulty of an item for all subjects, δ_j, is the logarithm of Q/P divided by the number of subjects, N,

$$\delta_j = \frac{ln\frac{Q}{P}}{N}. \tag{8.10}$$

That is, the estimate of ability (Equation 8.7) for items with an average difficulty of 0 does not require knowing the difficulty of any particular item, but is just a function of the pattern of corrects and incorrects for a subject across all items. Similarly, the estimate of item difficulty across people ranging in ability, but with an average ability of 0 (Equation 8.10) is a function of the response pattern of all the subjects on that one item and does not depend upon knowing any one person's ability. The assumptions that average difficulty and average ability are 0 are merely to fix the scales. Replacing the average values with a non-zero value just adds a constant to the estimates.

The independence of ability from difficulty implied in equations 8.7 and 8.10 makes estimation of both values very straightforward. These two equations also have the important implication that the number correct ($n\bar{p}$ for a subject, $N\bar{p}$ for an item) is monotonically, but not linearly related to ability or to difficulty. That the estimated ability is independent of the pattern of rights and wrongs but just depends upon the total number correct is seen as both a strength and a weakness of the Rasch model. From the perspective of *fundamental measurement*, Rasch scoring provides an additive interval scale: for all people and items, if $\theta_i < \theta_j$ and $\delta_k < \delta_l$ then $p(x|\theta_i, \delta_k) < p(x|\theta_j, \delta_l)$. But this very additivity treats all patterns of scores with the same number correct as equal and ignores potential information in the pattern of responses (see the discussion of the normal ogive model at at 8.1.2).

Consider the case of 1000 subjects taking the Law School Admissions Exam (the LSAT). This example is taken from a data set supplied by Lord to Bock and Lieberman (1970), which has been used by McDonald (1999) and is included in both the **ltm** and the **psych** package. The original data table from Bock and Lieberman (1970) reports the response patterns and frequencies of five items (numbers 11 - 15) of sections 6 and 7. Section 6 were highly homogeneous Figure Classification items and Section 7 were items on Debate. This table has been converted to two data sets: `lsat6` and `lsat7` using the `table2df` function. Using the `describe` function for descriptive statistics and then the `rasch` function from **ltm** gives the statistics shown in Table 8.1. The presumed item characteristic functions and the estimates of ability using the model are seen in Figure 8.1. Simulated *Rasch* data can be generated using the `sim.rasch` function in **psych** or the `mvlogis` function in **ltm**.

What is clear from Table 8.1 is that these items are all very easy (percent correct endorsements range from .55 to .92 and the Rasch scaled values range from -1.93 to -.20). Figure 8.1 shows the five parallel trace lines generated from these parameters as well as the distribution of subjects on an underlying normal metric. What is clear when comparing the distribution of ability estimates with the *item information* functions is that the test is most accurate in ordering those participants with the lowest scores.

Because the person and item parameters are continuous, but the response patterns are discrete, neither person nor item will fit the Rasch model perfectly. The residual for the pairing of a particular person with ability θ_i for a particular item with difficulty δ_j will be the difference between the observed response, x_{ij}, and the modeled response, p_{ij}

$$x_{ij} - p_{ij} = x_{ij} - \frac{1}{1 + e^{\theta_i - \delta_j}}.$$

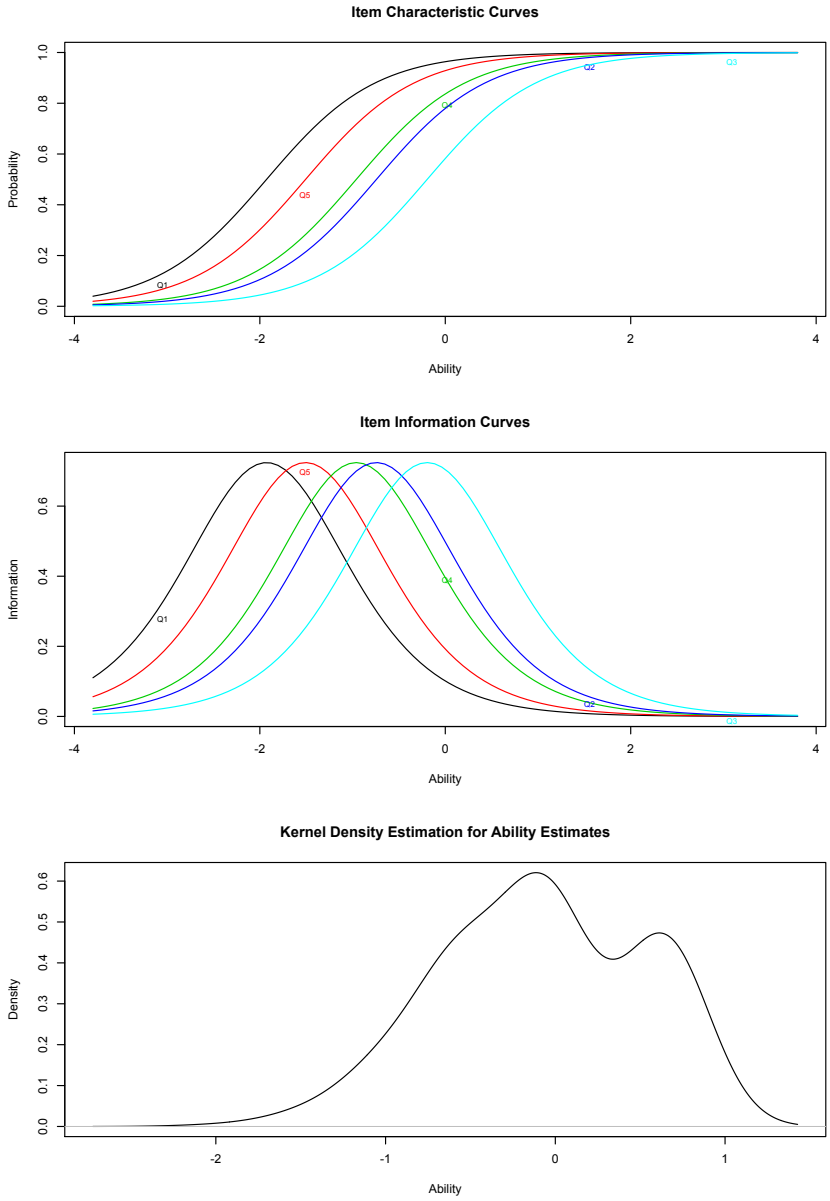

Fig. 8.1 The `rasch` function estimates item characteristic functions, item information functions, as well as the distribution of ability for the various response patterns in the `lsat6` data set. All five items are very easy for this group. Graphs generated by `plot.rasch` and `plot.fscores` in the **ltm** package. Because the Rasch model assumes all items are equally discriminating, the trace lines and the resulting item information functions are all identically shaped. The item information is just the first derivative of the item characteristic curve.

Table 8.1 Given the 1000 subjects in the LSAT data set (taken from Bock and Lieberman (1970), the item difficulties may be found using the Rasch model in the **ltm** package. Compare these difficulties with the item means as well as the item thresholds, τ. The items have been sorted into ascending order of item difficulty using the `order` and `colMeans` functions.

```
data(bock)
> ord <- order(colMeans(lsat6),decreasing=TRUE)
> lsat6.sorted <- lsat6[,ord]
> describe(lsat6.sorted)
> Tau <- round(-qnorm(colMeans(lsat6.sorted)),2)  #tau = estimates of threshold
> rasch(lsat6.sorted,constraint=cbind(ncol(lsat6.sorted)+1,1.702))
```

```
   var    n mean   sd median trimmed mad min max range  skew kurtosis   se
Q1   1 1000 0.92 0.27      1    1.00   0   0   1     1 -3.20     8.22 0.01
Q5   2 1000 0.87 0.34      1    0.96   0   0   1     1 -2.20     2.83 0.01
Q4   3 1000 0.76 0.43      1    0.83   0   0   1     1 -1.24    -0.48 0.01
Q2   4 1000 0.71 0.45      1    0.76   0   0   1     1 -0.92    -1.16 0.01
Q3   5 1000 0.55 0.50      1    0.57   0   0   1     1 -0.21    -1.96 0.02

> Tau
   Q1    Q5    Q4    Q2    Q3
-1.43 -1.13 -0.72 -0.55 -0.13

Call:
rasch(data = lsat6.sorted, constraint = cbind(ncol(lsat6.sorted) +
    1, 1.702))

Coefficients:
Dffclt.Q1  Dffclt.Q5  Dffclt.Q4  Dffclt.Q2  Dffclt.Q3     Dscrmn
   -1.927     -1.507     -0.960     -0.742     -0.195      1.702
```

Because p_{ij} is a binomial probability it will have variance $p_{ij}(1-p_{ij})$. Thus, the residual expressed as a z score will be

$$z_{ij} = \frac{x_{ij} - p_{ij}}{\sqrt{p_{ij}(1-p_{ij})}}$$

and the square of this will be a χ^2 with one degree of freedom. Summing this over all n items for a subject and dividing by n yields a goodness of fit statistic, *outfit*, which represents the "outlier sensitive mean square residual goodness of fit statistic" for the subject, the *person outfit*, (Wright and Mok, 2004, p 13)

$$u_i = \sum_{j=1}^{n} z_{ij}^2/n \tag{8.11}$$

with the equivalent *item outfit* based upon the sum of misfits across people for a particular item

$$u_j = \sum_{i=1}^{N} z_{ij}^2/N. \tag{8.12}$$

Because the *outfit* will be most responsive to unexpected deviations (missing an easy item for a very able person, or passing a difficult item for a less able person), the *infit* statistic is the "information weighed mean square residual goodness of fit statistic" (Wright and Mok,

2004, p 13). The weight, W_{ij} is the variance $W_{ij} = p_{ij}(1-p_{ij})$ and the *person infit* is

$$v_i = \frac{\sum_{j=1}^{n} z_{ij}^2 W_{ij}}{\sum_{j=1}^{n} W_{ij}} \tag{8.13}$$

(Wright and Mok, 2004). The *infit* and *outfit* statistics are available in the *eRm* package by Mair and Hatzinger (2009).

In addition to the `lsat6` and `lsat7` data sets, a sample data set that has been very well documented includes Bond's Logical Operations Test (*BLOT*) (Bond, 1995) and the Piagetian Reasoning Task (*PRTIII* by Shayer et al. (1976)) the data for which may be found online at `http://homes.jcu.edu.au/~edtgb/book/data/Bond87.txt` or in the introductory text by Bond and Fox (2007). By copying the data into the clipboard and using the `read.clipboard.fwf` function to read a *fixed width formatted* file, we are able to compare the results from the *ltm* and *eRm* packages with some of the commercial packages such as *WINSTEPS* (Linacre, 2005). It is important to note that the estimates of these three progams are not identical, but are rather linear transforms of each other. Thus, the `qnorm` of the `colMeans`, and the `rasch` estimates from *ltm* will match the `RM` estimates from *eRm* if the slopes are constrained to be one, and an additive constant is added. The `RM` estimates differ from the *WINSTEPS* merely in their sign, in that `RM` reports *item easiness* estimates while *WINSTEPS* reports *item difficulty*. Finally, `RM` by default forces the mean difficulty to be 0.0, while `rasch` does not. This agreement is also true for the person estimates, which are just inverse logistic, or `logit`, transforms of the total score. This is seen by some the power of the Rasch model in that it is monotonic with total score. In fact, the θ estimate is just a `logit` transformation of total score expressed as a percent correct, $\bar{P}$, and the percent incorrect, $\bar{Q} = 1 - \bar{P}$

$$\theta_i = -ln(\frac{1}{\bar{P}_i} - 1) = ln(\frac{\bar{P}_i}{\bar{Q}_i}). \tag{8.14}$$

That is, for complete data sets, Rasch scores are easy to find without the complexity of various packages. But the power of the Rasch (and other IRT models) is not for complete data, but when data are missing or when items are tailored to the subject (see **??**). Additionally, by expressing person scores on the same metric as item difficulties, it is easy to discover when the set of items are too difficult (e.g., the PRT) or too easy (e.g., the BLOT) for the subjects (Figure 8.2).

Good introductions to the *Rasch model* include a chapter by Wright and Mok (2004), and texts by Andrich (1988) and Bond and Fox (2007). Examples of applied use of the *Rasch model* for non-ability items include developing short forms for assessing depression (Cole et al., 2004) and the impairment in one's daily life associated with problems in vision (Denny et al., 2007). For depression, an item that is easy to endorse is "I felt lonely" while "People were unfriendly" is somewhat harder to endorse, and "I felt my life had been a failure" is much harder to endorse (Cole et al., 2004). For problems with one's vision, many people will have trouble reading normal size newsprint but fewer will have difficulty identifying money from a wallet, and fewer yet will have problems cutting up food on a plate (Denny et al., 2007).

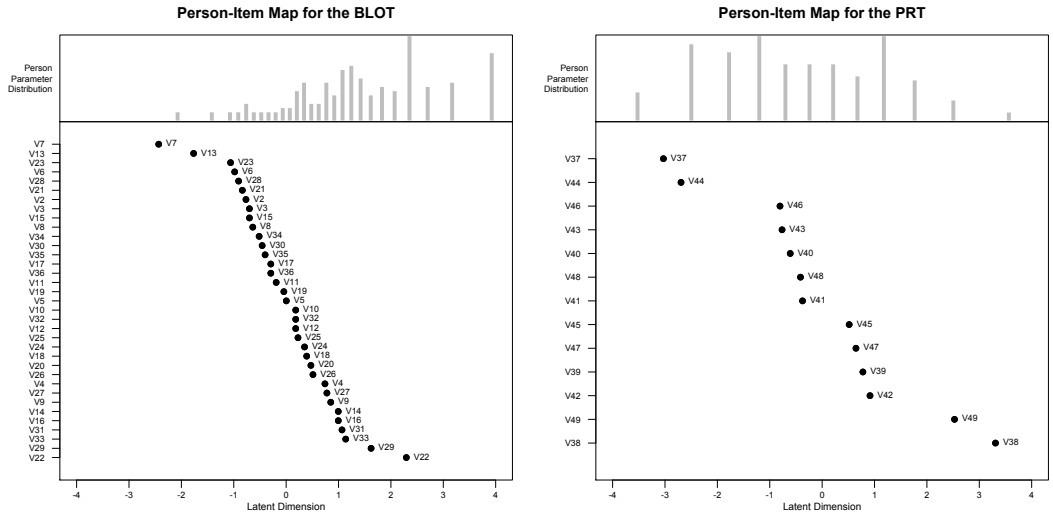

Fig. 8.2 The `plotPImap` function in the *eRm* package plots the distribution of the the person scores on the same metric as the item difficulties. While the hardest BLOT item is easier than the ability of 22% of subjects, the PRT items are much harder and all except two items are above 30% of the subjects. In that the subjects are the same for both tests, that the distribution of latent scores is not equivalent suggests problems of precision.

8.1.2 The normal ogive – another one parameter models

The Rasch model is based upon a logistic function rather than the cumulative normal. As shown earlier, with the multiplication of a constant, these functions are practically identical, and the logistic is relatively easy to calculate. However, with modern computational power, ease of computation does not compensate for difficulties of interpretation. Most people would prefer to give more credit to getting a difficult item correct than getting an easy item correct. But the logistic solution will give equal credit to any response pattern with the same number of correct items. (To the proponents of the Rasch model, this is one of its strengths, in that Rasch solutions are monotonic with total score.) The *cumulative normal* model (also known as the *normal ogive* or the one parameter normal, *1PN*, model) gives more credit for passing higher items:

$$p(correct|\theta,\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\inf}^{\theta-\delta} e^{-\frac{u^2}{2}} du \tag{8.15}$$

where $u = \theta - \delta$. By weighting the squared difference between a person and an item, greater weight is applied to missing extremely easy and passing extremely hard items than missing or passing more intermediate items. This will lead to scores that depend upon the particular pattern of responses rather than just total score. The difference between these two models may be seen in Figure 8.3 which compares logistic and normal ogive models for the 32 possible response patterns of five items ranging in difficulty from -3 to 3 (adapted from Samejima (2000). Examine in particular the response patterns 2-7 where the score estimate increases even though the number correct remains 1.

Another strength of the cumulative normal or normal ogive model is that it corresponds directly to estimates derived from factor analysis of the tetrachoric correlation matrix (McDonald, 1999). This is more relevant when two parameter models are discussed, for then the factor loadings can be translated into the item discrimination parameter (8.3).

The cumulative normal model, however, has an interesting asymmetry, in that for a pattern of all but one wrong item, the difficulty of the one passed item determines the ability estimate, while for a set of items that are all passed except for one failure, the difficulty of the failed item determines the ability estimate. This leads to the non-intuitive observation that getting the hardest four of five items correct shows less ability than getting all but the hardest item correct (Samejima, 2000). Consider five items with difficulties of -3, -1.5, 0, 1.5 and 3. These five items will lead to the 32 response patterns although scores can only be found for the cases of at least one right or one wrong. IRT estimates of score using the cumulative normal model range from -2.28 to 2.28 and correlate .93 with total correct. IRT estimates using the logistic function also range from -2.28 to 2.28 and correlate 1.0 with total correct. However, the logistic does not discriminate between differences in pattern within the same total score while the cumulative normal does (data in Figure 8.3 adapted from Samejima (2000)), To Rasch enthusiasts, this is a strength of the Rasch model, but to others, a weakness. An alternative, discussed by Samejima (1995, 2000), is to add an *acceleration parameter* to the logistic function which weights harder items more than easy items. This generalization is thus one of the many two and three parameter models that have been proposed.

8.1.3 Parameter estimation

The parameters of classical test theory, total score and item-whole correlations are all easy to find. The parameters of IRT, however, require iterative estimation, most typically using the principal of *maximimum likelihood* and the assumption of *local independence*. That is, for a pattern of responses, v, made up of individual item responses u_j with values of 0 or 1, the probability, P, of passing an item and the probability, $1-P=Q$, of failing an item has a probability distribution function of U_j of a response on the j^{th} item that depends upon the subject's ability, θ_i, and characteristics of the item difficulty, δ_j:

$$f_j(u_j|\theta) = P_j(\theta_i)^{u_j} Q_j(\theta_i)^{1-u_j} \tag{8.16}$$

then, with the assumption of local independence, the probability of data pattern u of successes and failures is the product of the individual probabilities

$$L(f) = \prod_{j=1}^{n} P_j(\theta_i)^{u_j} Q_j(\theta_i)^{1-u_j} \tag{8.17}$$

(Lord and Novick, 1968) and, by taking logarithms

$$lnL(f) = \sum_{j=1}^{n} [x_j lnP_j + (1-x_j) ln(Q_j)] \tag{8.18}$$

(McDonald, 1999, p 281).

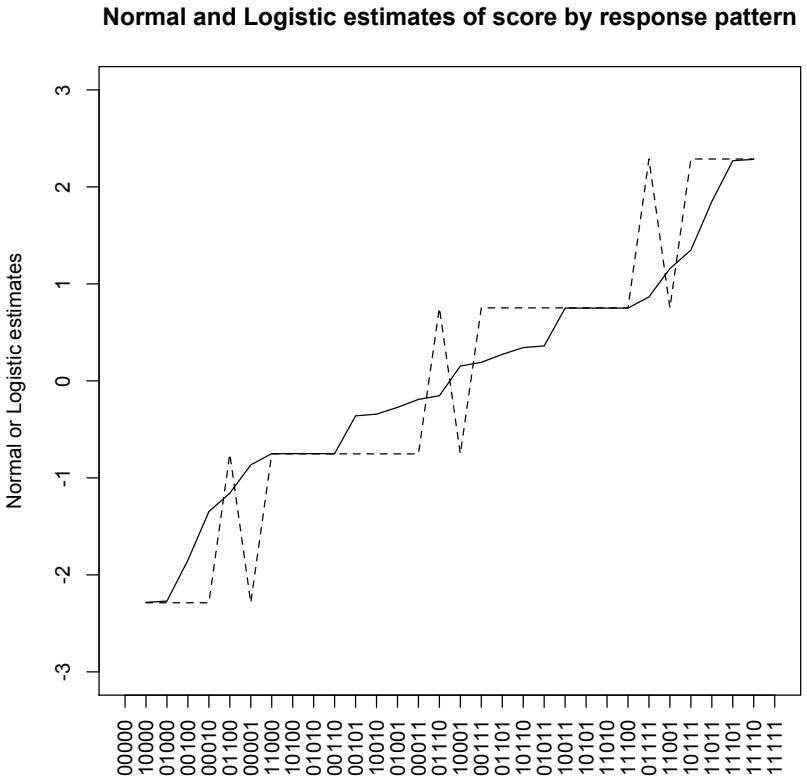

Fig. 8.3 For 32 response patterns of correct (1) or incorrect (0), the estimated scores from the logistic function do not discriminate between different response patterns with the same total score. The cumulative normal model does, but has some problematic assymetries (Samejima, 2000). The solid line represents the cumulative normal scores, the dashed line the logistic estimates. Neither model can find estimates when the total is either 0 or 5 (all wrong or all correct).

The best estimate for θ_i is then that value that has the maximum likelihood which may be found by iterative procedures using the `optim` function or IRT packages such as *ltm* or *eRm*.

8.1.4 Item information

When forming a test and evaluating the items within a test, the most useful items are the ones that give the most information about a person's score. In classic test theory, *item information* is the reciprocal of the squared *standard error* for the item or for a one factor test, the ratio of the item communality to its uniqueness:

$$I_j = \frac{1}{\sigma^2_{e_j}} = \frac{h^2_j}{1-h^2_j}.$$

When estimating ability using IRT, the information for an item is a function of the first derivative of the likelihood function and is maximized at the inflection point of the *icc*. The information function for an item is

$$I(f,x_j) = \frac{[P'_j(f)]^2}{P_j(f)Q_j(f)} \tag{8.19}$$

(McDonald, 1999, p 285). For the 1PL model, P', the first derivative of the probability function $P_j(f) = \frac{1}{1+e^{\delta-\theta}}$ is

$$P' = \frac{e^{\delta-\theta}}{(1+e^{\delta-\theta})^2} \tag{8.20}$$

which is just P_jQ_j and thus the information for an item is

$$I_j = P_jQ_j. \tag{8.21}$$

That is, information is maximized when the probability of getting an item correct is the same as getting it wrong, or, in other words, the best estimate for an item's difficulty is that value where half of the subjects pass the item. Since the test information is just the sum of the item information across items, a test can be designed to provide maximum information (and the smallest *standard error of measurement*) at a particular point by having items of a particular difficulty, or it can be designed to have relatively uniform information across a range of ability by having items of different difficulties.

8.1.5 Two parameter models

Both the *one parameter logistic* (*1PL*) and the *one parameter normal* ogive (*1PN*) models assume that items differ only in their difficulty. Given what we know about the factor structure of items, this is unrealistic. Items differ not only in how hard they are to answer, they also differ in how well they assess the latent trait. This leads to the addition of a *discrimination* parameter, α, which has the effect of magnifying the importance of the difference between the subject's ability and the item difficulty. In the *two parameter logistic* (*2PL*) model this leads to the probability of being correct as

$$p(correct_{ij}|\theta_i,\alpha_j,\delta_j) = \frac{1}{1+e^{\alpha_i(\delta_j-\theta_i)}} \tag{8.22}$$

while in the *two parameter normal* ogive (*2PN*) model this is

$$p(correct|\theta,\alpha_j,\delta) = \frac{1}{\sqrt{2\pi}}\int_{-\inf}^{\alpha(\theta-\delta)} e^{-\frac{u^2}{2}}du \tag{8.23}$$

where $u = \alpha(\theta-\delta)$.

The information function for a two parameter model reflects the item discrimination parameter, α,

$$I_j = \alpha^2 P_j Q_j \tag{8.24}$$

which, for a 2PL model is

$$I_j = \alpha_j^2 P_j Q_j = \frac{\alpha_j^2}{(1+e^{\alpha_j(\delta_j-\theta_j)})^2}. \tag{8.25}$$

The addition of the *discrimination* parameter leads to both a better fit to the data but also leads to non-parallel trace lines. Thus, an item can be both harder at low levels of ability and easier at high levels of ability (Figure 8.4). Indeed, the trace lines for two estimates of θ will cross over when

$$\theta_{ij} = \frac{\alpha_i \delta_i - \alpha_j \delta_j}{\alpha_i - \alpha_j} \tag{8.26}$$

(Sijtsma and Hemker, 2000). This leads to a logical problem. By improving the intuitive nature of the theory (that is to say, allowing items to differ in their discriminability) we have broken the simple additivity of the model. In other words, with the *1PL* model, if one person is more likely to get an item with a specific level of difficulty correct than is another person, then this holds true for items with any level of difficulty. That is, items and people have an additive structure. But, with the addition of the *discrimination* parameter, rank orders of the probability of endorsing an item as a function of ability will vary as a function of the item difficulty. It is this failure of *additivity* that leads some (e.g., Cliff (1992); Wright (1977); Wright and Mok (2004)) to reject any generalizations beyond the *Rasch* model. For, by violating additivity, the basic principles of fundamental *measurement theory* are violated. In a comparison of 20 different IRT models (10 of dichotomous responses, 10 for polytomous responses) Sijtsma and Hemker (2000) show that only the 1PL and 1PN models for dichotomous and the RSM and PCM models for polytomous items have this important property of not having item functions intersect each other. The RSM and PCM models are discussed below.

The `ltm` function in *ltm* estimates the two parameter model and produces the associated *icc* and *item information curve*, *iic*. Compare Figure 8.5 with the same set of graphs for the Rasch model 8.1. Although the items have roughly equal slopes and do not intersect, they do differ in the information they provide.

8.1.6 Three parameter models

The original work on IRT was developed for items where there was no guessing, e.g., getting the right answer when adding up a column of numbers (Lord, 1952). But, when taking a multiple choice ability test with n alternatives it is possible to get some items correct by guessing. Knowing nothing about the material one should be able to get at least 1/n% correct by random guessing. But guessing is not, in fact, random. Items can differ in how much they attract guesses for people who know very little about the material and thus a third item parameter can be introduced, the guessing parameter, γ:

$$p(correct_{ij}|\theta_i, \alpha_j, \delta_j, \gamma) = \gamma + \frac{1-\gamma}{1+e^{\alpha_i(\delta_j-\theta_i)}} \tag{8.27}$$

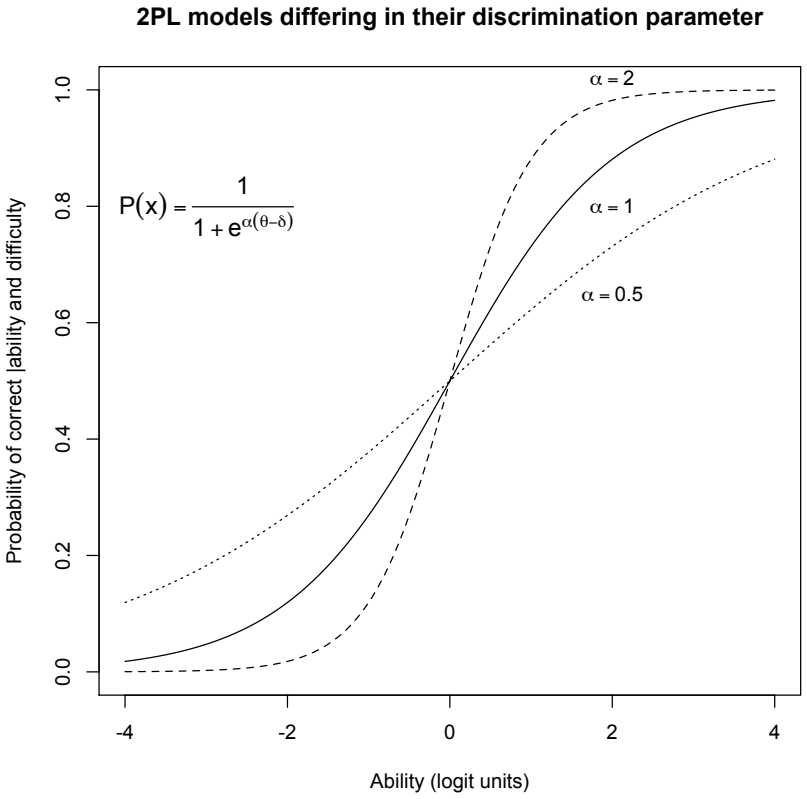

Fig. 8.4 By adding a discrimination parameter, the simple additivity of the 1PL model is lost. An item can both be harder for those with low ability and easier with those with high ability (b=2) than less discriminating items (b = 1, .5). Lines drawn with e.g., `curve(logistic(x,b=2))`.

(Figure 8.6 upper panel).

Unfortunately, the addition of the guessing parameter increases the likelihood that the trace lines will intersect and thus increases the non-additivity of the item functioning.

8.1.7 Four parameter models

Some items are so difficult that even with extreme levels of a trait not everyone will respond to the item. That is, the upper asymptote of the item is not 1. Although Reise and Waller (2009) have shown that including both a lower (γ) and upper (ζ) bound to the item response will improve the fit of the model and can be argued for in clinical assessment of disorders resulting in extremely rare behaviors, it would seem that that the continued addition of parameters at some point leads to an overly complicated (but well fitting) model (Figure 8.6 lower panel). The model is

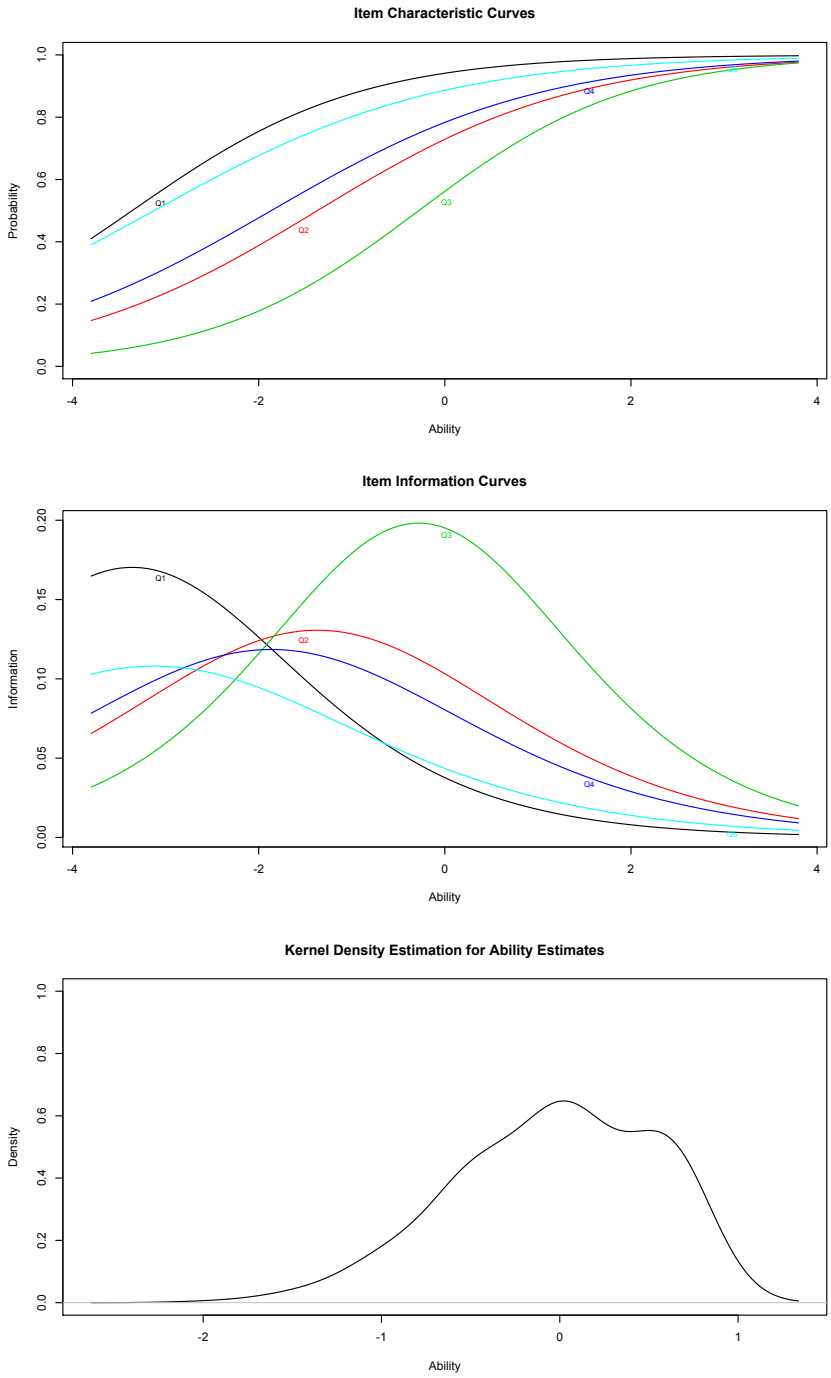

Fig. 8.5 The two parameter solution to the `lsat6` data set shows that the items differ somewhat in their discriminability, and hence in the information they provide. Compare this to the one parameter (Rasch) solution in Figure 8.1. The item information curve emphasizes that the items are maximally informative at different parts of the difficulty dimension.

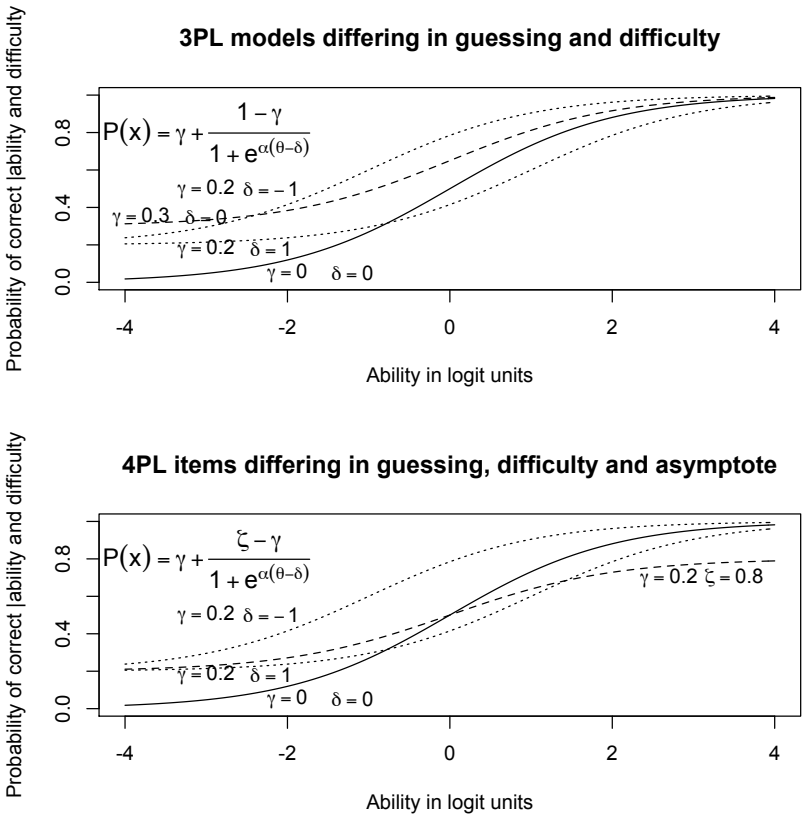

Fig. 8.6 The 3PL model adds a lower asymptote (γ) to model the effect of guessing. The 4PL model adds yet another parameter (ζ) to reflect the tendency to never respond to an item. As the number of parameters in the IRT model increases, the fit tends to be better, but at the cost of item by ability interactions. It is more difficult to make inferences from an item response if the probability of passing or endorsing an item is not an additive function of ability and item difficulty.

$$P(x|\theta_i, \delta_j, \gamma_j, \zeta_j) = \gamma_j + \frac{\zeta_j - \gamma_j}{1 + e^{\alpha_j(\delta_j - \theta_i}}. \tag{8.28}$$

When considering the advantages of these more complicated IRT models as compared to classic test theory Reise and Waller (2009) note that only with proper IRT scoring can we detect the large differences between individuals who differ only slightly on total score. This is the case for extremely high or low trait scores that are not in the normal range. But this is an argument in favor of all IRT models and merely implies that items should be "difficult" enough for the levels of the trait of interest.

8.2 Polytomous items

Although ability items are usually scored right or wrong, personality items frequently have multiple response categories with a natural order. In addition, there is information that can be obtained from the pattern of incorrect responses in multiple choice ability items. Techniques addressing these type of items were introduced by Samejima (1969) and discussed in great detail by Ostini and Nering (2006); Reise et al. (1993); Samejima (1996); Thissen and Steinberg (1986).

8.2.1 Ordered response categories

A typical personality item might ask "How much do you enjoy a lively party" with a five point response scale ranging from "1: not at all" to "5: a great deal" with a neutral category at 3. The assumption is that the more sociable one is, the higher the response alternative chosen. The probability of endorsing a 1 will increase monotonically the less sociable one is, the probability of endorsing a 5 will increase monotonically the more sociable one is. However, to give a 2 will be a function of being above some *threshold* between 1 and 2 and below some threshold between 2 and 3. A possible response model may be seen in Figure 8.7-left panel. This *graded response model* is based upon the early work of Samejima (1969) who discussed how a n alternative scale reflects n-*1* thresholds for normal IRT responses (Figure 8.7-right panel). The probability of giving the lowest response is just the probability of not passing the first threshold. The probability of the second response is the probability of passing the first threshold but not the second one, etc. For the 1PL or 2PL logistic model the probability of endorsing the k^{th} response is a function of ability, item thresholds, and the discrimination parameter and is

$$P(r=k|\theta_i,\delta_k,\delta_{k-1},\alpha_k)=P(r|\theta_i,\delta_{k-1},\alpha_k)-P(r|\theta_i,\delta_k,\alpha_k)=\frac{1}{1+e^{\alpha_k(\delta_{k-1}-\theta_i)}}-\frac{1}{1+e^{\alpha s_k(\delta_k-\theta_i)}} \tag{8.29}$$

where all b_k are set to $b_k=1$ in the 1PL Rasch case.

Because the probability of a response in a particular category is the difference in probabilities of responding to a lower and upper threshold, the *graded response model* is said to be a *difference model* (Thissen and Steinberg, 1986, 1997) which is distinguished from a family of models known as *divide by total* models. These model include the *rating scale model* in which the probability of response is the ratio of two sums:

$$P(X_i=x|\theta)=\frac{e^{\sum_{s=1}^{x}(\theta-\delta_i-\tau_s)}}{\sum_{q=0}^{m}e^{\sum_{s=1}^{q}(\theta-\delta_i-\tau_s)}} \tag{8.30}$$

where τ_s is the the difficulty associated with the s response alternatives and the total number of alternatives to the item is m (Sijtsma and Hemker, 2000).

Estimation of the *graded response model* may be done with the `grm` function in the *ltm* package and of the *rating scale model* with the `RSM` function in the *eRm* package. Consider the Computer Anxiety INdex or *CAIN* inventory discussed by Bond and Fox (2007) and available either from that text or an online repository at `http://homes.jcu.edu.au/~edtgb/book/`

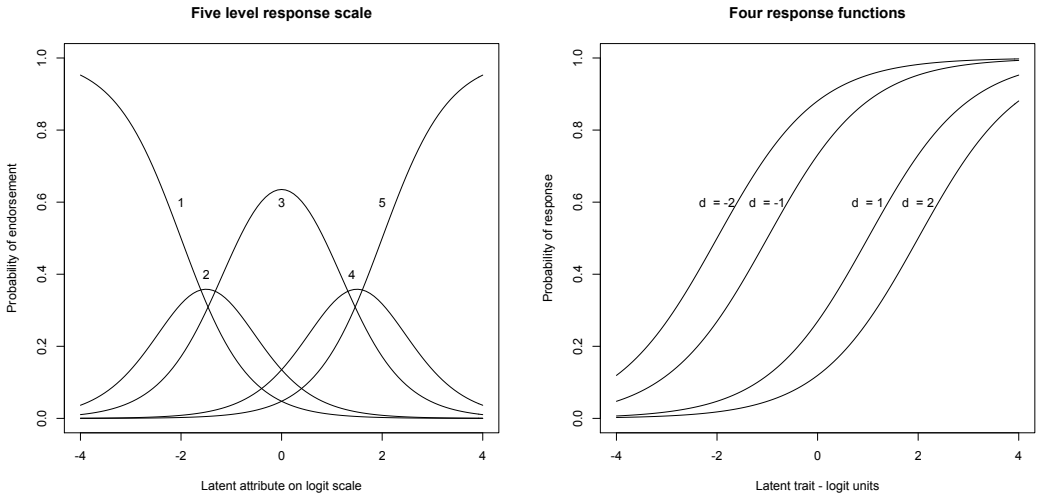

Fig. 8.7 The response probability to the five alternatives of ordered categories will be monotonically decreasing (for the lowest alternative), single peaked (for the middle alternatives, and monotonically increasing for the highest alternative (left panel). These five response patterns are thought to reflect four response functions (right panel). The difficulty thresholds are set arbitrarily to -2, -1, 1, and 2.

`data/Cain.dat.txt`. Some of the CAIN items as presented need to be reversed scored before analysis can proceed. Doing so using the `reverse.code` function allows an analysis using the `grm` function.

8.2.2 Multiple choice ability items

An analysis of the response probabilities for all the alternatives of a multiple choice test allows for an examination of whether some distractors are behaving differently from others. This is particularly useful when examining how each item works in an item bank. With the very large samples available to professional testing organizations, this can be done empirically without model fitting (Wainer, 1989). But with smaller samples this can be done using the item parameters (Thissen et al., 1989). In Figure 8.8 distractor D1 seems to be equally attractive across all levels of ability, D2 and D3 decrease in their probability of response as ability increases, and distractor D4 appeals to people with slightly less than average ability. The correct response, while reflecting guessing at the lowest levels of ability increases in the probability of responding as ability increases. This example, although hypothetical, is based upon real items discussed by Thissen et al. (1989) and Wainer (1989).

Table 8.2 The `grm` function in the the *ltm* package may be used to analyze the CAIN data set from Bond and Fox (2007) using a *graded response model*. The data are made available as a supplement to Bond and Fox (2007) and may be retrieved from the web at http://homes.jcu.edu.au/~edtgb/book/data/Cain.dat.txt. Some of the items need to be reverse coded before starting. The model was constrained so that all items had an equal discrimination parameter. The `coef` function extracts the cutpoints from the analysis, the `order` function allows for sorting the items based upon their first threshold and `headtail` reports the first and last four lines.

```
keys <- c(-1,-1,-1,-1,1,-1,1,-1,1,-1,1,1,-1,1,-1,-1,-1,1,1,1,1,1,1,1,1,-1)
rev.cain <- reverse.code(keys,cain,mini=rep(1,26),maxi=rep(6,26))
cain.grm <- grm(rev.cain,constrained=TRUE)
cain.coef <- coef(cain.grm)
sorted.coef <- cain.coef[order(cain.coef[,1]),]
headtail(sorted.coef)
```

```
       Extrmt1 Extrmt2 Extrmt3 Extrmt4 Extrmt5 Dscrmn
Item 21   -5.02   -3.26   -2.71   -1.91   -0.42   1.15
Item 1    -4.96   -2.69   -1.93   -0.35     1.7   1.15
Item 13   -4.81   -3.91   -3.01   -1.99   -0.48   1.15
Item 9    -4.48   -3.57   -2.69   -2.13   -1.07   1.15
...         ...     ...     ...     ...     ...    ...
Item 14   -3.15    -2.3   -1.39   -0.52    0.89   1.15
Item 23   -2.94   -1.96    -1.1   -0.21    0.85   1.15
Item 15   -2.36   -1.43    -0.6    0.35    1.41   1.15
Item 25   -1.02   -0.12    0.58    1.17    1.69   1.15
```

8.3 IRT and factor analysis of items

At first glance, the concepts of *factor analysis* as discussed in Chapter 6 would seem to be very different from the concepts of Item Response Theory. This is not the case. Both are models of the latent variable(s) associated with observed responses. Consider first the case of one latent variable. If the responses are dichotomous functions of the latent variable then the observed correlation between any two responses (items) is a poor estimate of their underlying latent correlation. The observed ϕ correlation is attenuated both because of the dichotomization, but also if there are any differences in mean level for items. With the assumption of normally distributed latent variables, the observed two x two pattern of responses may be used to estimate the underlying, latent, correlations using the *tetrachoric correlation* (4.5.1.4).

If the correlations of all of the items reflect one underlying latent variable, then factor analysis of the matrix of tetrachoric correlations should allow for the identification of the regression slopes (α) of the items on the latent variable. These regressions are, of course just the factor loadings. Item difficulty, δ_j and item discrimination, α_j may be found from factor analysis of the tetrachoric correlations where λ_j is just the factor loading on the first factor and τ_j is the normal threshold reported by the `tetrachoric` function (McDonald, 1999; Lord and Novick, 1968; Takane and de Leeuw, 1987).

$$\delta_j = \frac{D\tau}{\sqrt{1-\lambda_j^2}}, \qquad \alpha_j = \frac{\lambda_j}{\sqrt{1-\lambda_j^2}} \tag{8.31}$$

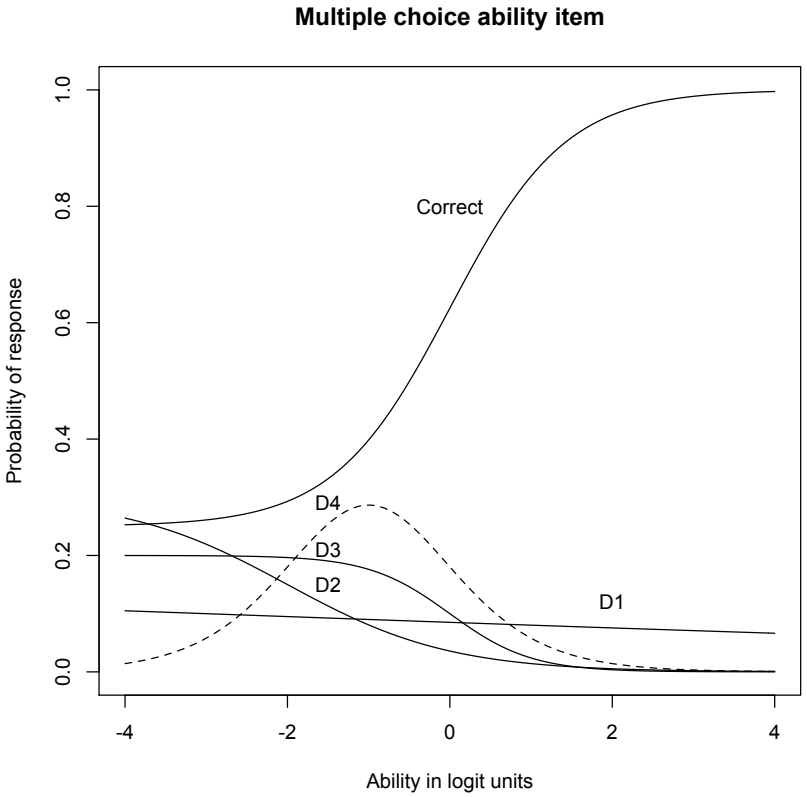

Fig. 8.8 IRT analyses allow for the detection of poorly functioning distractors. Although the correct response is monotonically increasing as a function of ability, and the probability of responding to distractors D1...D3 decreases monotonically with ability, the distractor D4 (dotted line) seems to appeal to people with slightly below average ability. This item should be examined more closely.

where D is a scaling factor used when converting to the parameterization of *logistic* model and is 1.702 in that case and 1 in the case of the normal ogive model. Thus, in the case of the normal model, factor loadings (λ_j) and item medians (τ) are just

$$\lambda_j = \frac{\alpha_j}{\sqrt{1+\alpha_j^2}}, \qquad \tau_j = \frac{\delta_j}{\sqrt{1+\alpha_j^2}}.$$

Consider the item data discussed in 6.6. These were generated for a normal model with difficulties ranging from -2.5 to 2.5 and with equal slopes of 1. Applying the equations for δ and α (8.31) to the τ and λ estimates from the `tetrachoric` and `fa` functions results in difficulty estimates ranging from -2.45 to 2.31 and that correlate $> .995$ with the theoretical values and with those found by either the `ltm` or `rasch` functions in the *ltm* package. Similarly, although the estimated slopes are not all identical to the correct value of 1.0, they have a mean of 1.04 and range from .9 to 1.18. (Table 8.3).

Table 8.3 Three different functions to estimate irt parameters. `irt.fa` factor analyzes the tetrachoric correlation matrix of the items. `ltm` applies a two parameter model, as does `rasch`. The slope parameter in the `rasch` solution is constrained to be equal 1. Although the estimated values are different, the three estimates of the item difficulty correlate $> |.995|$ with each other and with the actual item difficulties (a). The slope parameters in the population were all equal 1.0, and the differences in the estimates of slopes between the models reflect random error. Note that the ltm and rasch difficulties are reversed in sign from the a values and irt.fa values.

```
set.seed(17)
items <- sim.npn(9,1000,low=-2.5,high=2.5)$items
p.fa <- irt.fa(items)$coefficients[1:2]
p.ltm <- ltm(items~z1)$coefficients
p.ra <- rasch(items, constraint = cbind(ncol(items) + 1, 1))$coefficients
a <- seq(-2.5,2.5,5/8)
p.df <- data.frame(a,p.fa,p.ltm,p.ra)
round(p.df,2)
```

```
           a Difficulty Discrimination X.Intercept.   z1 beta.i beta
Item 1 -2.50      -2.45           1.03         5.42 2.61   3.64    1
Item 2 -1.88      -1.84           1.00         3.35 1.88   2.70    1
Item 3 -1.25      -1.22           1.04         2.09 1.77   1.73    1
Item 4 -0.62      -0.69           1.03         1.17 1.71   0.98    1
Item 5  0.00      -0.03           1.18         0.04 1.94   0.03    1
Item 6  0.62       0.63           1.05        -1.05 1.68  -0.88    1
Item 7  1.25       1.43           1.10        -2.47 1.90  -1.97    1
Item 8  1.88       1.85           1.01        -3.75 2.27  -2.71    1
Item 9  2.50       2.31           0.90        -5.03 2.31  -3.66    1
```

Simulations of IRT data with 1 to 4 parameter models and various distributions of the underlying latent may be done using the `sim.irt` function. The `irt.fa` function may be used to factor analyze dichotomous items (using tetrachoric correlations) and to express the results in terms of IRT parameters of difficulty (δ) and discrimination (α). Plotting the subsequent results shows the idealized two parameter *item characteristic curves* or *iccs*. In a very helpful review of the equivalence of the Item Response and factor analytic approaches, Wirth and Edwards (2007) review various problems of bias in estimating the correlation matrix and hence the factor loadings and compare the results of various *Confirmatory Categorical Factor Analyses* (*CCFA*s). Comparing their results to those of `irt.fa` suggests that the simple approach works quite adequately.

8.4 Test bias and Differential Item Functioning

If items differ in their parameters between different groups this may be seen as *test bias* or more generally as *Differential Item Functioning* (*DIF*). (Test bias implies that the items differentially favor one group over another, whereas DIF includes the case where an item is either easier or harder, or more or less sensitive to the underlying trait for different groups). Consider the case of sex differences in depression. Items measuring depression (e.g., "In the past week I have felt downhearted or blue" or "In the past week I felt hopeless about the future" have roughly equal endorsement characteristics for males and females. But the item

"In the past week I have cried easily or felt like crying" has a much higher threshold for men than for women (Schaeffer, 1988; Steinberg and Thissen, 2006). This example of using IRT to detect DIF may be seen clearly in Figure 8.9. As pointed out by Steinberg and Thissen (2006) "For many purposes, graphical displays of trace lines and/or their differences are easier to interpret than the parameters themselves" (p 406).

There are a number of ways to test for DIF, including differences in item difficulty as well as differences in item sensitivity. Of course, when using a one parameter model, only differences in difficulty can be detected while the sort of difference discussed by Steinberg and Thissen (2006) requires at least a two parameter model. In their development of items for short scales to assess physical and emotional well being Lai et al. (2005) consider items that differ more than .5 logistic units to be suspect.

Sometimes tests will have items that have compensatory DIF parameters. That is, a test that includes some items that are easier for one group (the reference group) than another (the focal group), and then other items that are harder for the reference group than the focal group. It is thus logically possible to have unbiased tests (no *differential test functioning* made up of items that themselves show DIF (Raju et al., 1995).

Steinberg and Thissen (2006)

Reeve and Fayers (2005)

8.5 Non-monotone trace lines – the measurement of attitudes

8.5.1 Unfolding theory revisited

8.5.2 Political choice

Van Schuur and Kruijtbosch (1995) Chernyshenko et al. (2007) Borsboom and Mellenbergh (2002)

8.6 IRT and adaptive testing

Whether comparing school children across grade levels, or assessing the health of medical patients, the usefulness of IRT techniques becomes most clear. For classical test theory increases reliability by aggregating items. This has the unfortunate tendency of having many items that are either too easy or too difficult for any one person to answer and requires tests that are much too long for practical purposes. The IRT alternative is to give items that are maximally informative for the person at hand, rather than people in general. If items are tailored to the test taker, then people are given items that they will endorse (pass) with a 50% likelihood. (The analogy to high jumping is most appropriate here: Rather than ask an olympic athlete to jump bars of .5, .6, ...,1, ..., 1.5, ... 2.35,2.4,2.45, 2.5 meters, the first jump will typically be 2.2 or 2.3 meters. Similarly, for elementary school students, jumps might be .2, .. 1 meters, with no one being given a 2.4 meter bar). Thus, the person's score is not how many they passed, but rather a function of how difficult were the items they passed or failed.

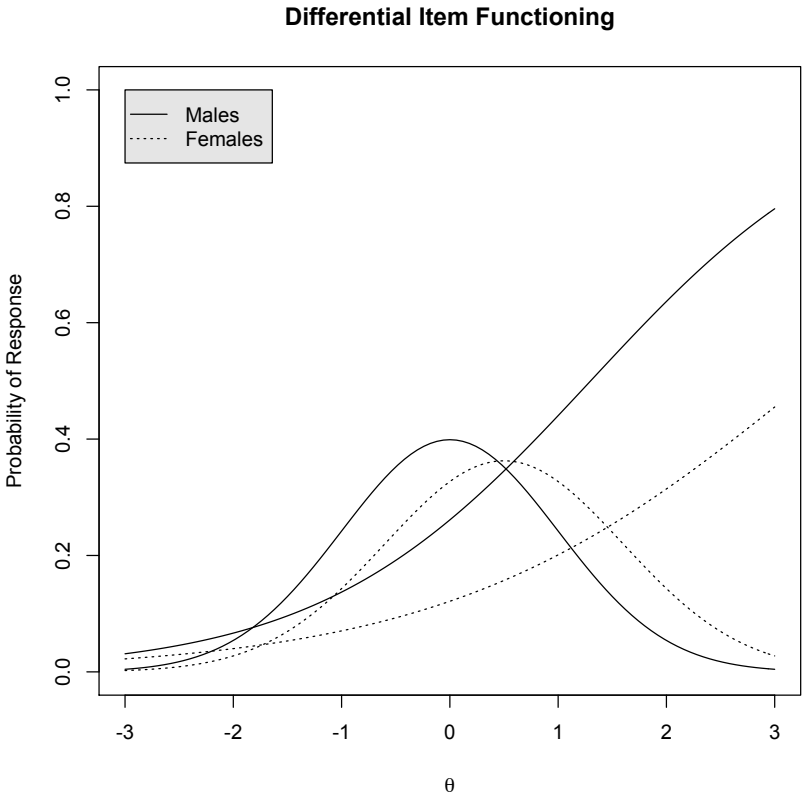

Fig. 8.9 When measuring depression, the item "In the past week I have cried easily or felt like crying" has a lower threshold ($\delta = 1.3$) and is slightly more sensitive to the latent trait ($\alpha = .8$ for women than it is for men ($\delta = 3.3, \alpha = .6$)). The figure combines item characteristic curves (montonically rising) with the distribution of participant scores (normal shaped curves). Data from Schaeffer (1988), figure adapted from Steinberg and Thissen (2006)).

In an adaptive testing framework, the first few items that are given are of average difficulty. People passing the majority of those items are then given harder items. If they pass those as well, then they are given yet harder items. But, if they pass the first set and then fail most of the items in the second set, the third set will be made up of items in between the first and second set. The end result of this procedure is that everyone could end up with the same overall passing rate (50%) but drastically different scores (Cella et al., 2007; Gershon, 2005). Unfortunately, this efficiency of testing has a potential bias in the case of high stakes testing (e.g., the Graduate Record Exam or a nursing qualification exam). For people who are higher on anxiety tend to reduce their effort following failure while those lower on anxiety increase their effort following failure. Given that the optimal adaptive test will produce failure 50% of the time, adaptive testing will lead to an underestimate of the ability of the more anxious participants as they will tend to reduce their effort as the test proceeds (Gershon, 1992).

8.7 Item banking and item comparison

When doing adaptive testing or developing short forms of instruments that are optimized for a particular target population, it is necessary to have a bank of items of various difficulties in order to choose the most discriminating ones. This is not much different from the normal process of scale development (to be discussed in Chapter 16) but focuses more upon choosing items to represent the full range of item location (difficulty). Developing an item bank requires understanding the construct of interest, writing items at multiple levels of difficulty and then validating the items to make sure that they represent the relevant construct. Items are chosen based upon their discriminability (fit with the dimension) and location (difficulty). Additional items are then developed to fit into gaps along the difficulty dimension. Items need to be shown to be equivalent across various sub groups and not to show differential item functioning.

A very nice example of the use of item banking in applied testing of various aspects of health is the "Patient-Reported Outcomes Measurement Information System" (*PROMIS*) developed by the National Institutes of Health. Combining thousands of different items assessing symptoms and outcomes as diverse as breathing difficulties associated with chronic lung disease to pain symptoms to satisfaction with life, the PROMIS item bank is used for *Computerized Adaptive Testing* (*CAT*) presenting much shorter and precise questionnaires than would have been feasible using conventional test (Cella et al., 2007).

Although most of the PROMIS work has been done using commercial software, more recent developments have used IRT functions Linking solutions together across group may be done using the **plink** (Weeks, 2010).

8.8 Non-parametric IRT

Mokken and Lewis (1982) but see Roskam et al. (1986) and a response Mokken et al. (1986)

8.9 Classical versus IRT models – does it make a difference?

Perhaps the greatest contribution of IRT methods is that person scores are in the same metric as the items. That is, rather than saying that someone is one standard deviation higher than another, it is possible to express what that means in probability of response. With items with a mean probability of .5, being one logit unit higher than the mean implies a probability of correct response of .73, while being two logit units higher implies a probability of correct response of .88. This is much more informative about the individual than saying that someone is one or two standard deviations above the mean score which without knowing the sample standard deviation tells us nothing about how likely the person is to get an item correct. In addition, rather than saying that a test has a reliability of r_{xx}, and therefore the *standard error of measurement* for any particular person's score is estimated to be $\sigma_x\sqrt{1-r_{xx}}$, it is possible to say that a particular score has an error of estimation of $\frac{1}{\sqrt{information}}$. That is, some scores can be measured with more precision than others. In addition, by recognizing that the relationship between the underlying attribute and the observed score is monotonic

(in the case of ability) but not linear, it is less likely that errors of inference due to scaling artifacts (e.g., 3.6) will be made.

In addition, to the practitioner the obvious advantage of IRT techniques is the possibility of item banking and adaptive testing (Cella et al., 2007). For it makes little sense to give items that are much too easy or much to hard to measure a particular attribute. The savings in tester and testee time is an important benefit of adaptive testing.

Why then do people still use classical techniques? To some dedicated IRT developers, using classical test theory is an anachronism that will fade with better training in measurement (Embretson, 1996). But to others, the benefits of combining multivariate models such as factor analysis or structural equation modeling with a theory of tests that is just a different parameterization of IRT values outweighs the benefits of IRT (McDonald, 1999). The supposed clear superiority of IRT is seen as an "urban legend" and rather IRT and CTT techniques each have their advantage (Zickar and Broadfoot, 2009). Because the correlation between CTT scores and IRT based scores tend to be greater than .97 for normal samples (Fan, 1998), the simple addition of items following principles of CTT seem quite adequate. In a very balanced discussion of how to construct quality of life inventories, Reeve and Fayers (2005) consider the benefits of both CTT and IRT. The great benefit of IRT techniques is the emphasis in characteristics of the items (e.g., in identifying DIF), and the ability to select items from item banks to optimally measure any particular level of a trait.

The graphical displays of item information as a function of item location and item discrimination make it easier to see and explain why some items were chosen and how and why to choose items ranging in their difficulty/location. The appeal of multi-parameter IRT models to the theorist concerned with best fitting complex data sets with elegant models needs to be contrasted with the simple addition of item responses for the practitioner. The use of IRT for selecting items from a larger item pool is a clear advantage for the test developer, but of less concern for the user who just wants to rank order individuals on some construct and have an estimate of the confidence interval for any particular score. But for the theorist or the practitioner, R includes multiple packages that provide powerful alternatives to the commercial IRT programs.

Chapter 9
Validity

Reliability is how compact a flight of arrows are, validity is whether you hit the target. Oxygen titrations can be reliable, but if the chemist is color blind, they are not valid measures of the oxygen level when compared to someone elses measures.

9.1 Types of validity

Does a test measure what it supposed to test? How do we know? Shadish et al. (2001), Borsboom and Mellenbergh (2004)

9.1.1 Face or Faith

Does the content appear reasonable? Is this important?

9.1.2 Concurrent and Predictive

Do tests correlate with alternative measures of the same construct right now and do they allow future predictions?

9.1.3 Construct

What is the location of our measure in our nomological network. Cronbach and Meehl (1955)

9.1.3.1 convergent

Do measures correlate with what they should correlate with given the theory?

9.1.3.2 discriminant

Do measures not correlate with what the theory says they should not correlate with? What is the meaning of a hyperplane? Measuring what something isn't is just as important as knowing what something is.

9.1.3.3 incremental

Does it make any difference if we add a test to a battery? Werner Wittmann and the principle of Brunswickian Symmetry Wittmann and Matt (1986) Also, the notion of higher order versus lower order predictions.

9.2 Validity and cross validation

9.2.1 Cross validation

9.2.2 Resampling and cross validation

Grucza and Goldberg (2007)

9.3 Validity: an modern perspective

Borsboom and Mellenbergh (2004) The ontology vs. epistemology of validity.

9.4 Validity for what?

9.4.1 Validity for the institution

9.4.2 validity for the individual

9.5 Validity and decision making

Wiggins (1973)

Chapter 10
Reliability + Validity = Structural Equation Models

10.1 Generating simulated data structures

10.2 Measures of fit

As has been seen in the previous sections, the use of fit statistics does not guarantee meaningful models. If we do not specify the model correctly, either because we do not include the correct variables or because we fail to use the appropriate measurement model, we will lead to incorrect conclusions. Widaman and Thompson (2003) MacCallum et al. (2006) Marsh et al. (2005)

Even if we have a very good fit, we are unable to determine causal structure from the model, even if we bother to add time into the model.

10.2.1 χ^2

As we saw in the previous chapter, χ^2 is very sensitive to many sources of error in our model specification. χ^2 is sensitive to failures of our distributional assumptions (continuous, multivariate normal) as well as to our failures to correctly specify the structure.

10.2.2 GFI, NFI, ...

10.2.3 RMSEA

10.3 Reliability (Measurement) models

10.3.1 One factor — congeneric measurement model

10.3.1.1 Generating congeneric data structures

10.3.1.2 Testing for Tau equivalent and congeneric structures

10.3.2 Two (perhaps correlated) factors

10.3.2.1 Generating multiple factorial data

10.3.2.2 Confirmatory factoring using sem

10.3.3 Hierarchical measurement models

10.3.3.1 Generating the data for three correlated factors

10.3.3.2 Testing hierarchical models

10.4 Reliability + Validity = Structural Equation Models

10.4.1 Factorial invariance

10.4.2 Multiple group models

10.5 Evaluating goodness of fit

10.5.1 Model misspecification: Item quality

10.5.1.1 Continuous, ordinal, and dichotomous data

Most advice on the use of latent variable models discusses the assumption of multivariate normality in the data. Further discussions include the need for continuous measures of the observed variables. But how does this relate to the frequent use of SEM techniques in analysis of personality or social psychological items rather than scales? In this chapter we consider typical problems in personality where we are interested in the structure of self reports of personality, emotion, or attitude. Using simulation techniques, we consider the effects of

normally distributed items, ordinal items with 6 or 4 or 2 levels, and then the effect of skew on these results. We use simulations to show the results more clearly. For a discussion of real data with some of these problems, see Rafaeli and Revelle (2006).

10.5.1.2 Simple structure versus circumplex structure

Most personality scales are created to have "simple structure" where items load on one and only one factor Revelle and Rocklin (1979); Thurstone (1947). The conventional estimate for the reliability and general factor saturation of such a test is Cronbach's coefficient α (Cronbach, 1951) Variations of this model include hierarchical structures where all items load on a general factor, g, and then groups of items load on separate, group, factors Carroll (1993); Jensen and Weng (1994). Estimates of the amount of general factor saturation for such hierarchical structures may be found using the ω coefficient discussed by (McDonald, 1999) and (Zinbarg et al., 2005).

An alternative structure, particularly popular in the study of affect as well as studies of interpersonal behavior is a "circumplex structure" where items are thought to be more complex and to load on at most two factors.

> "A number of elementary requirements can be teased out of the idea of circumplex structure. First, circumplex structure implies minimally that variables are interrelated; random noise does not a circumplex make. Second, circumplex structure implies that the domain in question is optimally represented by two and only two dimensions. Third, circumplex structure implies that variables do not group or clump along the two axes, as in simple structure, but rather that there are always interstitial variables between any orthogonal pair of axes Saucier (1992). In the ideal case, this quality will be reflected in equal spacing of variables along the circumference of the circle Gurtman (1994)(Gurtman, 1994; Wiggins, Steiger, & Gaelick, 1981). Fourth, circumplex structure implies that variables have a constant radius from the center of the circle, which implies that all variables have equal communality on the two circumplex dimensions (Fisher, 1997; Gurtman, 1994). Fifth, circumplex structure implies that all rotations are equally good representations of the domain (Conte & Plutchik, 1981; Larsen & Diener, 1992)." (Acton and Revelle, 2004).

Variations of this model in personality assessment include the case where items load on two factors but the entire space is made up of more factors. The Abridged Big Five Circumplex Structure (AB5C) of (Hofstee et al., 1992b) is an example of such a structure. That is, the AB5C items are of complexity one or two but are embedded in a five dimensional space.

10.5.2 Model misspecification: failure to include variables

10.5.3 Model misspecification: incorrect structure

10.6 What does it mean to fit a model