6

The Three-Parameter Model

In this chapter we present a model for addressing chance success on an item. This chance success is reflected in an IRF with a nonzero lower asymptote. To model this lower asymptote, we extend the 2PL model to produce the three-parameter model. Parallel to the structure of the chapters discussing the 1PL and 2PL models, we present examples of a three-parameter model calibration using the mathematics data set introduced in Chapter 2.

Through the previous chapters we have developed a "toolbox" of model-fit techniques. This toolbox includes methods for assessing the tenability of various assumptions. To summarize these approaches, the unidimensionality assumption can be assessed using nonlinear factor analysis, linear factor analysis, and structural equation modeling. We can assess the tenability of the functional form assumption by examining the empirical IRFs. Moreover, model–data fit can be assessed through fit statistics (e.g., `INFIT`, `OUTFIT`, M2), comparing the predicted and empirical IRFs, as well as by obtaining evidence of item parameter estimate invariance through the use of several statistics (e.g., correlations, RMSD, UA_{22}). We have also examined person fit through fit statistics.

In this chapter we add to our toolbox. Specifically, (1) we introduce the likelihood ratio, AIC, and BIC statistics for making model comparisons, (2) we use Q_3 for assessing the tenability of the conditional independence assumption, and (3) we discuss the appropriateness of a person's estimated location as a measure of their true location. Although for pedagogical reasons we present the model-fit techniques separately, in practice they would be used collectively. The last topic we cover in this chapter is the handling of missing data.

Conceptual Development of the Three-Parameter Model

Individuals at the lower end of the latent continuum may be expected to have a high probability of providing a response of 0. For example, examinees who have low mathematics proficiency may be expected to incorrectly respond to, say, a topology question

on a mathematics examination. If this mathematics examination uses a multiple-choice item format, then some of these low-proficiency individuals may select the correct option simply by guessing. Similarly, people low in neuroticism who are administered a neuroticism inventory using a true/false response format may be expected to respond "False" to a question depicting a neurotic behavior. However, owing to inattention or fatigue, some of these individuals may respond "True" to the question. In these cases, the item's response function has a lower asymptote that may not be asymptotic with 0.0 but may be with some nonzero value. The three-parameter model addresses this nonzero lower asymptote.

To develop the three-parameter model, we need to be concerned with two cases. The first case is, "What is the probability of a response of 1 on an item when an individual responds consistent with their location θ?" Our answer is that the probability of a response of 1 is modeled by the 2PL model. Conversely, the probability of a response of 0 (i.e., $p(x_j = 0 \mid \theta, \delta)$) when an individual responds consistent with their location θ is given by $(1 - p_j)$; Figure 2.12 depicts these two functions. The $p(x_j = 0 \mid \theta, \delta)$ response function has a lower asymptote of 1 and an upper asymptote of 0. That is, as θ approaches $-\infty$, the event "a response of 0" is almost certain to occur.

The second case to consider is, "What should be the probability of a response of 1 on an item due to chance alone?" To answer this question, let us symbolize this probability as χ_j. In other words, when a person can be successful on item j regardless of the person's location, then the corresponding probability is given by χ_j. To determine the pseudo-random guessing response function, we need to consider χ_j and the probability of a response of 0 given the 2PL model (i.e., $p(x_j = 0 \mid \theta, \delta) = [1 - p_j]$). Noting that the event "a response of 1 due to chance alone" is independent of the event "a response of 0 given θ" allows us to apply the multiplication rule. That is, when a person can be successful on item j on the basis of chance alone the probability is given by the pseudo-random guessing response, $\chi_j[1 - p_j]$. Multiplying by $[1 - p_j]$ transforms the lower asymptote of $p(x_j = 0 \mid \theta, \delta)$ to equal χ_j. Thus, as θ goes to $-\infty$, p_j approaches 0.0 and $\chi_j[1 - p_j]$ simplifies to χ_j. Conversely, as θ goes to ∞, p_j approaches 1.0 and $\chi_j[1 - p_j]$ approaches 0.0. Thus, the probability of a response of 1 for an individual with an infinitely low location is χ_j.

Putting these two (mutually exclusive) cases together, we obtain the probability of a response of 1

$$p_j^* = p_j + \chi_j(1 - p_j), \tag{6.1}$$

where p_j is given by the 2PL model. Equation 6.1 may be rearranged to be

$$p_j^* = \chi_j + (1 - \chi_j)p_j. \tag{6.2}$$

By substitution of the 2PL model for p_j, we obtain the *three-parameter logistic* (3PL) model

$$p(x_j = 1 \mid \theta, \alpha_j, \delta_j, \chi_j) = \chi_j + (1 - \chi_j)\frac{e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}}. \tag{6.3}$$

One can view Equation 6.3 from a slightly different perspective than above and explain why Equation 6.2 is not simply $\chi_j + p_j$. The effect of the term $(1 - \chi_j)$ compresses the 2PL model's IRF to range from zero to $(1 - \chi_j)$. By adding χ_j to this compressed IRF (i.e., Equation 6.3), we transform the IRF to have a range from χ_j to 1.0. One implication of this compression is that it effectively reduces the IRF's slope.

Although, strictly speaking, Equation 6.3 is not in logistic form, it is referred to as a logistic model. (Because there is a normal ogive version of the three-parameter model, Equation 6.3 is sometimes presented, incorporating the scaling factor *D*.) As is the case with the 1PL and 2PL models, δ_j represents item *j*'s location and α_j reflects its discrimination parameter. The additional parameter, χ_j, is referred to as the item's *pseudo-guessing* or *pseudo-chance* parameter and equals the probability of a response of 1 when θ approaches $-\infty$ (i.e., $\chi_j = p(x_j = 0 \mid \theta, \rightarrow -\infty)$. As such, χ_j represents the IRF's lower bound or asymptote. With the 3PL model, there are three parameters characterizing the item *j* (i.e., $\alpha_j, \delta_j, \chi_j$) plus a person parameter.

The 3PL model is based on the same assumptions discussed in Chapter 2 with the 1PL model. Recall that these assumptions are a unidimensional latent space, conditional independence, and a specific functional form. For brevity we use p_j instead of $p(x_j = 1 \mid \theta, \alpha_j, \delta_j, \chi_j)$ in the following.

Examples of the 3PL model's IRF are given in Figure 6.1. The two items shown have the same discrimination and location parameters, but they have different χ_js. For item 1 $\chi_1 = 0.1$ and for item 2 $\chi_2 = 0.05$. We see that the IRFs have nonzero lower asymptotes

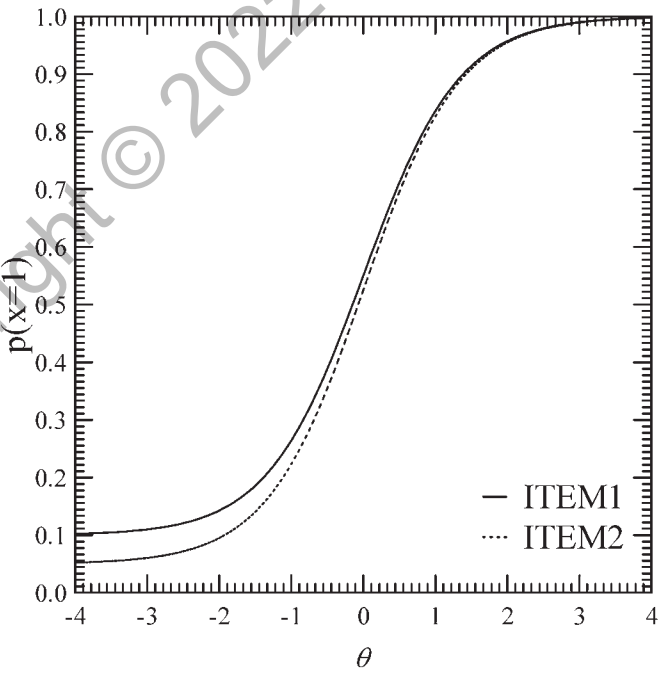

FIGURE 6.1. 3PL model IRFs for two items with $\alpha_1 = 1.5, \delta_1 = 0.0, \chi_1 = 0.1$, and $\alpha_2 = 1.5, \delta_2 = 0.0, \chi_2 = 0.05$.

and that each IRF is asymptotic with its corresponding χ_j value. In addition, we see that item 1 with the larger χ_j has the higher IRF. In general, as χ_j increases, so does p_j, all other things being equal. In the context of proficiency assessment, this means that items with larger χ_js are easier than those with smaller χ_js. The figure shows that the valid range for χ_j is 0.0 to 1.0.

As is the case with the 1PL and 2PL models, the IRF's slope is at a maximum at the item *j*'s location. This point of inflexion occurs midway between the lower and upper asymptotes. The lower asymptote is the floor of the IRF and represents the smallest probability for a response of 1, whereas the upper asymptote is the ceiling for the IRF and reflects the largest probability of a response of 1. If we let Υ_j denote item *j*'s upper asymptote, then a general expression for determining the midpoint (i.e., the probability at δ_j) for any of our dichotomous models is $(\Upsilon_j + \chi_j)/2$. For example, with the 1PL and 2PL models, the lower asymptote is 0 and the upper asymptote is 1. Therefore, for the 1PL and 2PL models we have that $\chi_j = 0.0$, $\Upsilon_j = 1.0$, and the probability of a response of 1 at δ_j is $(1 + 0.0)/2 = 0.50$. For the 3PL model, if $\chi_j > 0.0$, then the probability of a response of 1 at δ_j is greater than 0.50. For example, if $\chi_j = 0.20$ and $\Upsilon_j = 1.0$, then the probability of a response of 1 at δ_j is $(1 + \chi_j)/2 = (1 + 0.2)/2 = 0.6$.[1] Moreover, as is true with the 1PL and 2PL models, the 3PL model's discrimination parameter is proportional to the slope at the inflexion point. However, the relationship between α_j and the slope now involves χ_j. Specifically, the slope for the 3PL model is $0.25\alpha_j(1-\chi_j)$.[2] Therefore, an item's discriminatory effectiveness is affected by the magnitude of χ_j. Specifically, as χ_j increases, an item's discriminatory effectiveness decreases, all other things being equal. For example, we see from Figure 6.1 that item 1's discriminatory effectiveness (reflected in its IRF's slope) is less than that of item 2.

Additional Comments about the Pseudo-Guessing Parameter, χ_j

Our first comment is about the different labels used for χ_j. Originally, χ_j was referred to as the item's guessing parameter (e.g., Lord, 1980, p. 12). However, because χ_j is typically lower than what would be predicted by a random guessing model (i.e., the reciprocal of the number of multiple-choice options), χ_j is now referred to as the pseudo-guessing parameter. This difference between χ_j and the random guessing model prediction is due to differential option attractiveness. That is, the random guessing model assumes that all options are equally attractive. Yet we know from traditional item analyses that item alternatives vary in their degree of attractiveness to persons. For instance, using keywords in alternatives is a typical tactic to increase the attractiveness of alternatives. Moreover, test-taking preparation instructs examinees who do not know the answer to a question to select the longest option because it is usually the correct response. As such, the random guessing model's assumption is not reflected in the response data.

Our second comment concerns the nature of χ_j. As mentioned earlier, χ_j's function is to reflect that some individuals with infinitely low locations may obtain a response of 1 when, according to the 2PL model, they should not. These responses are a manifesta-

tion of the interaction between person and item characteristics (including item format). In the case of proficiency instruments, person characteristics include not only a person's θ, but also their test-wiseness and "risk-taking" tendencies. These last two factors are tangential latent person variables. Therefore, although χ_j is considered to be an item parameter, it may be more reflective of a *person* characteristic (i.e., another person parameter) than of an item characteristic or, at least, an interaction between person and item characteristics.

Our final comments concern the implicit assumption made by the use of χ_j and the effect of χ_j on estimation. In regard to the former, we see from Equation 6.3 that the presence of χ_j in the model assumes that, regardless of a person's location, their propensity to "guess" is constant across the continuum (i.e., χ_j does not vary as a function of θ). This assumption may or may not be reasonable in all situations. With respect to effects, nonzero χ_js lower the estimate of a person's location (Wainer, 1983) and reduce the amount of item information.[3] Thus, although we are modeling nonzero χ_js, it is very desirable that our χ_js be close to zero. Of course, in this case the 2PL model may provide a sufficiently reasonable representation of the data.

Conceptual Parameter Estimation for the 3PL Model

The estimation of item parameters proceeds as discussed in previous chapters. However, unlike the 1PL and 2PL models, the 3PL model does not have sufficient statistics for parameter estimation (Baker, 1992; Lord, 1980). The log likelihood surface for an item with three item parameters would require four dimensions to graphically represent it. However, the general idea can be represented as a series of static multiple surfaces similar to the one presented in Figure 5.3, but with each surface slightly different from the others and associated with a particular value of χ_j (e.g., 0.0, 0.01, 0.02). (Obviously, the discrete nature of this series of surfaces does not accurately reflect the continuous nature of χ_j.) The essence of the estimation process would be to identify across these "multiple surfaces" the values of α_j, δ_j, and χ_j that maximize the log likelihood for an item.[4]

In some cases, distinguishing between these multiple surfaces may be problematic. For instance, if there are insufficient data at the lower end of the continuum, then there may be multiple sets of α_j, δ_j, and χ_j that account for the data. As such, the corresponding IRFs are similar to one another in this region (cf. Mislevy, 1986a). As an example, assume that in a given calibration sample everyone is located above –1. As a result, there is insufficient data to estimate the lower asymptote. Figure 6.2 presents two IRFs that can account for empirical data. One IRF is based on $\alpha = 0.8$, $\delta = -0.05$, and $\chi = 0.435$, whereas the other has the item parameter values of $\alpha = 0.56$, $\delta = -1.8$, and $\chi = 0.0$. As can be seen, these two IRFs are very similar to one another above –1 and, in fact, differ by less than 0.01 in the θ range –1 to 1 and by less than 0.018 in the range –1 to 3. Without additional information (e.g., persons located around –3, or prior information), it is not possible to determine whether χ_j should be 0.435 or 0. In terms of our "multiple surfaces" analogy, this means that we cannot distinguish between the log likelihood

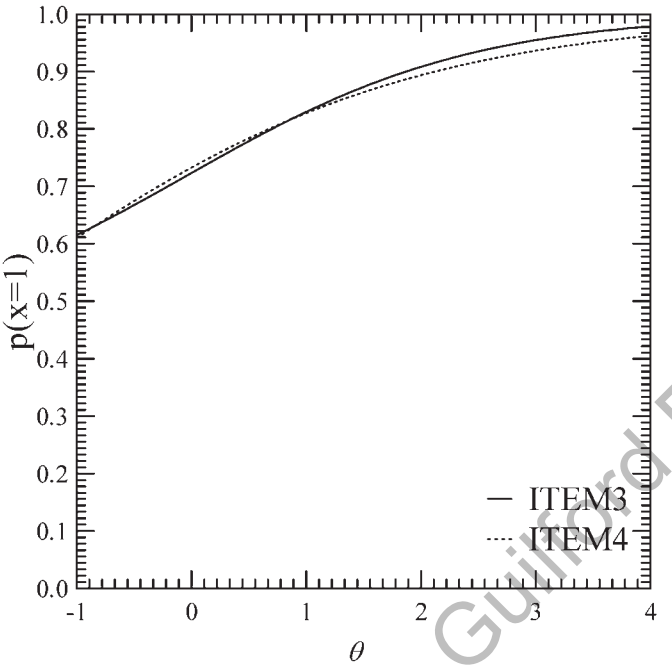

FIGURE 6.2. 3PL model IRFs when $\alpha_1 = 0.8$, $\delta_1 = -0.05$, $\chi_1 = 0.435$ and when $\alpha_2 = 0.56$, $\delta_2 = -1.8$, $\chi_2 = 0.0$.

surface associated with $\chi = 0.435$ and the one when $\chi = 0.0$. Therefore, if the respondents are located above –1, it is difficult to determine which of these two sets of item parameter estimates is "best," and so we have difficulty obtaining a converged solution for the item.[5]

In general, the estimation of χ_j may be problematic for some items because of the paucity of persons at the lower end of the continuum; because the items are located at the lower end of the continuum (e.g., very easy items); and/or because the items have low estimated discrimination parameters. Problems in estimating χ_j can influence the estimation of the item's other parameters. In these situations, a criterion may be used to determine whether χ_j should be estimated. For instance, LOGIST used the "stability" criterion of $(\delta_j - 2/\alpha_j)$. Specifically, χ_j is estimated only when $(\delta_j - 2/\alpha_j) > -2.5$; –2.5 is the default value and may be changed. The stability criterion is the location on the θ continuum "at which the proportion of correct responses is only about 0.03 above the lower asymptote" (Wingersky et al., 1982, p. 21). Alternative strategies are to fix χ_j to a specific value or to impose a prior distribution. With respect to the former, the selection of a constant (common) value for χ may be done arbitrarily (e.g., LOGIST's [1/m – 0.05] where m is the number of item options), by averaging the nonproblematic $\hat{\chi}_j$s, by averaging the $\hat{\chi}_j$s for items located at the lower end of the continuum, or by fixing the lower asymptote to some nonzero value determined by inspecting the lower asymptote of empirical IRFs.

We may also use a prior distribution with χ_j. de Gruijter (1984) has demonstrated that the use of a prior distribution for estimation of χ_j can lead to reasonable parameter estimates for the model. The regression toward the mean phenomenon that typically occurs when using a prior distribution is not as problematic in estimating χ_j as it is when estimating person and item location parameters (Lord, 1986). In general, we recommend use of a prior on the $\hat{\chi}_j$s as the first strategy to facilitate estimating the lower asymptote.

In regard to the item's other parameters, empirical data calibration has shown that the $\hat{\alpha}_j$s and $\hat{\delta}$s are nonlinearly related and, typically, have a positive correlation (Lord, 1975). In addition, Lord found that items with $\hat{\delta}$s less than about –0.5 almost never have $\hat{\alpha}$s greater than 1 and that items located above 0.5 almost always have $\hat{\alpha}$s greater than 1.0. In this regard, we examined the calibration results from the reading and mathematics tests from the National Education Longitudinal Study, 1988 (NELS: 88; Ingels, Scott, Rock, Pollack, & Rasinski, 1994) base year, and found the correlation between the $\hat{\alpha}$s and the $\hat{\delta}$s is 0.25 for the reading test and 0.59 for the mathematics test, the 3PL model calibration used LOGIST; also see Yen (1987). Baker and Kim (2004) present the mathematics for estimating the three item parameters, and a Bayesian estimation procedure is presented in Swaminathan and Gifford (1986).

So far we have been concerned with item parameter estimation. We now turn our attention to person parameter estimation. Any of the methods that were previously discussed, such as MLE or EAP, could be used. However, in some cases the use of unrestricted MLE for person location estimation may encounter problems. For example, Samejima (1973a) showed that there is not a unique solution for θ for every possible response pattern under the three-parameter model. For these problematic response patterns, the likelihood function may have more than one maximum. For example, assume we have a two-item instrument with $\alpha_1 = 2.0$, $\delta_1 = 0.0$, $\chi_1 = 0.25$ for the first item and $\alpha_2 = 1.0$, $\delta_2 = -0.5$, $\chi_2 = 0.0$ for the second item. On these two items, assume that a person has a response of 1 on item 1 and a response of 0 on item 2 (Samejima, 1973a). Assuming a proficiency testing situation, then this response pattern reflects a person correctly answering the "harder/more discriminating" item (possibly by guessing) and incorrectly answering the "easier/less discriminating" item. The corresponding likelihood function is presented in Figure 6.3.

As we see, the likelihood function has a *local* maximum at approximately –0.05, and as θ becomes progressively smaller, the likelihood function begins to approach an asymptote of 0.25. Therefore, this likelihood function is asymptotic with a value of 0.25 and without a unique person location estimate. Stated another way, these types of response vectors do not have a *global* maximum and have multiple maxima. In these cases, the use of standard MLE with the three-parameter model may yield a $\hat{\theta}$ that turns out to be a local, not a global, maximum. When a local maximum is suspected, then using a different starting/provisional estimate for the MLE algorithm (see Appendix A) might produce a different $\hat{\theta}$. (In fact, the presence of multiple solutions for a given response vector is evidence that one or more of the solutions represent local maxima.)

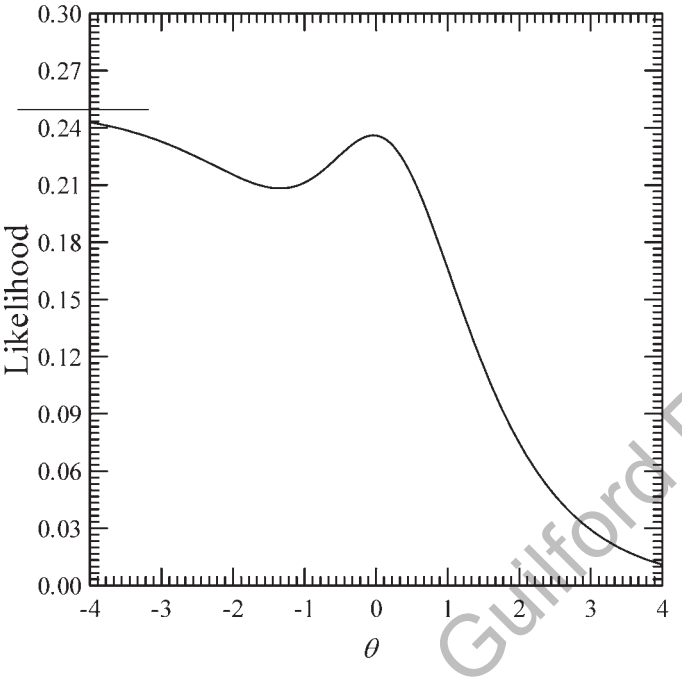

FIGURE 6.3. Likelihood function for a two-item instrument with no unique maximum in which the first response is correct and the second is incorrect (i.e., $\underline{\mathbf{x}}' = 10$).

Although the example only uses two items, Samejima (1973a) speculated that "the likelihood function may be more complicated, with possibly more than one local maximum in addition to a terminal maximum" (p. 225). We have empirical support for the occurrence of multimodal likelihood functions from the work of Yen, Burket, and Sykes (1991). Specifically, in an analysis of 14 empirical data sets they found that as many as 3.1% of the examinees had response vectors whose likelihood functions had multiple maxima (cf. Fischer, 1981).

The multimodal likelihood function seen in Figure 6.3 is due to the specific $\underline{\mathbf{x}}$ and the particular relationship among the α_j, δ_j, and χ_j. If $\chi_j = 0$ for both items (i.e., the 1PL and 2PL models), then the likelihood function has a unique solution. Therefore, one possibility of addressing these multimodal likelihood function situations is to use the *truncated 2PL model* (Samejima, 2001) for person parameter estimation. The truncated 2PL model capitalizes on the fact that the 2PL model's IRF (with appropriate item parameter values) is virtually indistinguishable from that of the 3PL model above a critical value, θ_g. Below θ_g the probability of a response of 1 is 0 for the truncated 2PL model (i.e., the IRF is truncated at θ_g). Therefore, for the truncated 2PL model there are two conditions: (1) for $-\infty < \theta < \theta_g$ where we have that $p_j = 0$; and (2) for $\theta_g < \theta < \infty$; p_j is given by the 2PL model (Equation 5.1). Samejima (1973a) shows that $\theta_g = 0.5 \ln(\chi_j) + \delta_j$. An alternative approach for handling multimodal likelihood functions is to use a Bayesian person estimation technique (e.g., EAP).

How Large a Calibration Sample?

The answer to the question of how large a sample is needed depends, in part, on the estimation procedure, instrument characteristics (e.g., the distribution of item parameter estimates, instrument length, etc.), response data characteristics (e.g., amount of missing data), and person distribution. In general, attempts at answering this question have involved conducting simulation studies where the parameter estimates can be compared, directly or indirectly, with the corresponding parameters.

For example, Yen (1987) investigated the parameter recovery of MMLE and JMLE as implemented in BILOG and LOGIST, respectively, using a fixed sample size of 1,000. In this Monte Carlo study, she investigated three different instrument lengths (10, 20, and 40 items) and different θ distributions: normal (0, 1), negatively skewed (skew = −0.4/kurtosis = −0.1), positively skewed (skew = 0.4/kurtosis = −0.1), and platykurtic (skew = 0.1/kurtosis = −0.4). Generally speaking, she found that MMLE estimates were more accurate than those of JMLE, particularly at the 10-item instrument length. With respect to MMLE, the item discrimination estimation results using the 20- and 40-item instruments were comparable to one another in terms of their RMSD, with values ranging from 0.09 to 0.20. Moreover, the correlations between $\hat{\alpha}$ and α ($r_{\alpha\hat{\alpha}}$) ranged from 0.88 to 0.94, regardless of the normality or non-normality of the θ distribution. For the 10-item length, the RMSD doubled to 0.48 and $r_{\alpha\hat{\alpha}}$ decreased to 0.84. In terms of estimating item locations, the 20- and 40-item instruments had correlations ($r_{\delta\hat{\delta}}$) from 0.97 to 0.99 with RMSDs of 0.07 to 0.16; the 10-item length had an $r_{\delta\hat{\delta}}$ of 1.00 and RMSDs of 0.18. In general, item location is better estimated than item discrimination. The lower asymptote showed $r_{\chi\hat{\chi}}$ s between 0.11 and 0.54, with RMSDs of 0.03 to 0.08 across the various instrument lengths and irrespective of the nature of the θ distribution. Although not a formal parameter recovery study, Mislevy (1986a) presents results indicating that BILOG does a reasonably good job in recovering item parameters with a sample size of 1,000 and a 20-item instrument.

This research appears to indicate that for MMLE a sample of 1,000 persons may lead to reasonably accurate item parameter estimates with the 3PL model under favorable conditions (e.g., a symmetric θ distribution, an instrument length of 20 items). This rough guideline assumes the use of prior distributions for χ_j and α_j. However, it is strongly recommended that calibration sample sizes exceed 1,000 to mitigate the convergence problems that sometimes plague 3PL model calibrations. In fact, Thissen and Wainer (1982) suggest trying to avoid estimating χ_j if possible under unrestricted MLE, and they also suggest that the use of a prior distribution when estimating χ_j seems "to offer some hope" (p. 410). In cases where one has a smallish sample size and/or one experiences difficulty in estimating the item parameters with the 3PL model, then fixing the lower asymptote to a reasonable nonzero value for some or all the items may help. In addition, some convergence problems (e.g., –2ln*L* values that oscillate across iterations) may sometimes be rectified by using the RIDGE subcommand available in BILOG and PARSCALE. The calibration sample size caveats and considerations previously mentioned in Chapters 3 and 5, such as model–data misfit tolerance, ancillary

technique sample size requirements, the amount of missing data, and so on are also applicable to the three-parameter model.[6]

Assessing Conditional Independence

In Chapter 2, we stated that one assumption underlying IRT models is that the responses to one item are not related to those on any other item conditional on θ(s). This assumption is the conditional (or local) independence assumption. When this assumption is violated, then the accuracy of our item parameter estimates is affected and the total instrument information is overestimated (Chen & Thissen, 1997; Oshima, 1994; Sireci, Wainer, & Thissen, 1991; Thissen, Steinberg, & Mooney, 1989; Yen, 1993). As such, any subsequent use of the item parameter estimates for, say, equating (see Chapter 11) will be potentially adversely affected. In the following, we discuss some causes of item *dependence*, some ways to handle this dependence, and then a statistic for identifying local dependent items post administration.

Violation of the conditional independence assumption may occur for various reasons, such as structural dependence among items, content clues, instrument length, insufficient allotted time to complete an instrument (i.e., speededness), and/or an insufficient number of latent variables in the IRT model. Examples of items with structural dependence are a set of survey questions that all refer to the same, say, life-changing event (e.g., a diagnosis of cancer, contracting HIV), comprehension questions that use the same reading passage, or trigonometry problems based on a common figure. In all of these cases, one may see local dependence. In addition, when there is insufficient time to respond to all the items on an instrument, the items affected by the lack of time may exhibit dependence. As a consequence, their corresponding parameter estimates are adversely affected. Conversely, when there is sufficient time to respond to an instrument but the instrument is very long, one may observe local dependence due to fatigue or diminished motivation. Practice effects may also lead to local dependence.

For some of these causes, it is possible to identify the items that may be prone to local dependence prior to administering the instrument. In general, an instrument should be inspected for connections between the items. This inspection involves looking for similarity in the questions' text, an item providing one or more cues as to how to respond to another item, the items sharing grammatical inconsistencies or common information (e.g., a passage or a figure), the items sharing a nesting/hierarchical relationship, and so on. Depending on the outcome of this inspection, rewriting the items may be sufficient to address the anticipated dependency. In other cases, the items cannot be rewritten because they need to be logically related or structurally dependent. In these cases the dependent items may be combined to form an item cluster.

An item cluster (also known as an item bundle or testlet [Thissen, Steinberg, & Mooney, 1989b; Wainer & Kiely, 1987; Wainer & Lewis, 1990]) is a group of interdependent items that may be created pre- or post administration. There are at least two ways to score an item cluster. In one approach, each item in the item cluster provides an item score and the score on the item cluster is, for example, the sum of these item

scores. For instance, if an item cluster consists of three 1-point items, then the possible scores on the item cluster would be 0, 1, 2, or 3. In effect, the item cluster is treated as a single "item" for estimating a person's location. One way of utilizing this item cluster score is to use a model that can handle both dichotomous and polytomous responses (e.g., see Yen, 1993). Models that can address not only polytomous responses, but also dichotomous responses, are presented in the following chapters.

In the foregoing polytomous model approach to handling item clusters, there is some loss of information. For example, an item cluster score does not say anything about the response pattern that produced the score. Whether this is an important issue is context-specific. However, if the loss of this information is important, then an alternative approach to scoring an item cluster is to use a model that incorporates a parameter that reflects the dependency among items within the item cluster. Bradlow, Wainer, and Wang (1999) developed such a model by augmenting the 2PL model. The augmentation is a random effect parameter that reflects a person-specific *testlet* effect.[7] The Bradlow et al. model may be applied to both items that are independent and those in testlets; one- and three-parameter models also exist (see Wang & Wilson, 2005; Wainer, Bradlow, & Du, 2000). This *testlet model* and its variants form the basis of testlet response theory (Wainer, Bradlow, & Wang, 2007).

Various indices have been developed for identifying local dependence. A review of some of these indices and an examination of which index works best may be found in Kim, de Ayala, Ferdous, and Nering (2011); also see Glas (1999), Glas and Falcón (2003), Orlando and Thissen (2000), and Rosenbaum (1984) for related indices. One of these indices is Yen's (1984) Q_3 index. Although no index may be considered to be the best in terms of combining high power to detect conditional item dependence with low false positive rates, the Q_3 index works reasonably well (e.g., see Kim et al., 2011). Because of Q_3's simplicity and its comparative good performance, we use it to demonstrate evaluating the conditional independence assumption with our mathematics data example.

Q_3 is the correlation between the residuals for a pair of items. The residual for an item is the difference between an individual's observed response and their expected item response. Therefore, after fitting the model, the Pearson correlation coefficient is used to examine the linear relationship between pairs of residuals. In the current context, the observed response is either a 1 or a 0 and the expected response is the probability according to the 3PL model. Symbolically, the residual for person i on item j is

$$d_{ij} = x_{ij} - p_j(\hat{\theta}_i)$$

and for item k it is

$$d_{ik} = x_{ik} - p_k(\hat{\theta}_i).$$

Q_3 is the correlation between d_{ij} and d_{ik} across respondents

$$Q_{3(j,k)} = r_{d_j d_k} \tag{6.4}$$

If $|Q_3|$ equals 1.0, then the two items are perfectly interdependent. In contrast, a Q_3

of 0.0 is a necessary, but not sufficient, condition for independence because a $Q_3 = 0$ can be obtained when the items in the pair are independent of one another *or* because they exhibit a nonlinear relationship. Therefore, Q_3 is useful for identifying items that exhibit item *dependence*. Under conditional independence Q_3 should have an expected value of $-1/(L-1)$ (Yen, 1993).

As has been mentioned, in some cases one can explain item dependence in terms of multidimensionality. That is, the dependency between two items is due to a common additional latent variable such as test-wiseness. If two items are independent, then their interrelationship is completely explained by the latent structure of the model. If one applies a unidimensional model when two dimensions are called for, then the items that are influenced by both latent variables show a negative local dependence, and items that are affected by only one of the two latent variables show a positive local dependence (Yen, 1984). However, if only one of the latent variables is used, then the items that are influenced only by that underlying variable show a slight negative local dependence. To obtain a large Q_3 value for an item pair, we need to have similarity of parameters for the items in question and the items need to share one or more unique dimensions. Therefore, similarity of parameters is a necessary, but not sufficient, condition for obtaining a large Q_3 value.

Some research (e.g., Yen, 1984) has found that although the value of Q_3 is not as much influenced by the sample size as other measures, it is affected by the instrument's length. This may be due to item scores being involved in both x_{ij} and x_{ik} as well as (implicitly) in $p_j(\hat{\theta})$. As a result, Q_3 may tend to be slightly negative due to part–whole contamination (Yen, 1984). The implication is that one would expect to see substantially more negative Q_3s for short instruments than for longer instruments. In this case, these negative values may be artifacts due to the instrument's length.

There are various ways to use Q_3 to identify locally dependent items. First, we can use Q_3 in a statistical *z*-test (Yen, 1984). This would require that Q_3 be transformed by the Fisher *r*-to-$\dot{z}$ transformation ($\dot{z}_{Q_3}$) and then used in a *z*-test,

$$z = \dot{z}_{Q_3} \Big/ \sqrt{1/(N-3)}$$

$\dot{z}_{Q_3}$ has a mean of 0.0 and a variance of 1/(N – 3). The standard unit normal table is used to provide critical values for identifying items with $\dot{z}_{Q_3}$ values that are unlikely to be observed owing to chance alone. However, because the typical calibration sample size will result in a test with a great deal of power, we will most likely reject the null hypothesis of independence for trivially small correlations. An additional issue is that the sampling distribution of Q_3 may not be symmetric (Chen & Thissen, 1997). That is, because the Q_3 sampling distribution may not approximate the standard normal very well, the critical values would be inappropriate. As a result, the Q_3 empirical Type I error rates do not match the nominal significance level that one would expect under normal curve theory. Moreover, Marais (2013) states that the "sampling properties of the correlations among residuals are unknown. It is therefore not possible to use these statistics for for-

mal tests of local independence" (p. 121). Therefore, rather than using the critical values from the standard unit normal table in a statistical inferential fashion, it is preferable to use them as guidelines/screening values for informed judgment.

A second way of using Q_3 is to take advantage of the fact that Q_3 is a correlation. Specifically, because Q_3 is a correlation coefficient, the square of Q_3 (Q_3^2) may be interpreted as a measure of the amount of residual variance shared by an item pair. Therefore, item pairs with a large proportion (e.g., 5% or greater) of shared variability would indicate dependent items.

Alternatively, one could compare Q_3 to a cutpoint. That is, an observed Q_3 that is larger than the cutpoint would indicate item dependence. Yen (1993) suggests one such cutpoint in the context of instruments with a minimum of 17 items. Specifically, a Q_3 screening value of 0.2 was suggested to identify items exhibiting dependence (i.e., $|Q_3| > 0.2$ indicates local item dependence). Although this cutpoint has been found to produce small Type I error rates, it also leads to comparatively lower power than other detection methods (Chen & Thissen, 1997).

To address some of these issues, Christensen, Makransky, and Horton (2017) conducted a study to arrive at empirically based critical values. Using simulation in conjunction with empirical estimates, they found, for example, that critical values of 0.19 and 0.24 cut off 5% and 1%, respectively, of $\max(Q_{3(j,k)})$'s empirical distribution above them with a 9-item instrument; $\max(Q_{3(j,k)})$ is the largest observed $Q_{3(j,k)}$. Although the study focused on the Rasch model, their results may be considered indicative that no single critical value can be used in all situations regardless of the IRT model used. They concur with Marais's (2013) conclusion that Q_3's evaluation should take into account all of an instrument's Q_3s (cf. Marais, 2013, p. 121). Specifically, a given $Q_{3(j,k)}$ is compared to "$\max(Q_{3(j,k)}) - \bar{Q}_3$", where Christensen et al. (2017) define the average Q_3 as

$$\bar{Q}_3 = 2\sum_{j>k} Q_{3_{jk}} / (\mathrm{L}(\mathrm{L}-1)).$$

As Equation 6.4 shows, Q_3 is the zero-order correlation for item j's and k's respective residuals. (In this paragraph, all references to items are to the items' residuals.) As such, unless all item pairs are independent of one another, the correlation between item j and k will contain information from the other items to varying degrees. For example, assume that we have a three-item instrument and the correlation between items 1 and 2 is –0.181. If the correlation between items 1 and 3 is zero and the correlation between items 2 and 3 is also zero, then the correlation between items 1 and 2 is the zero-order correlation (e.g., $Q_3 = r_{d_1 d_2} = -0.181$). However, if the correlation between items 1 and 3 is –0.098 and if the correlation between items 2 and 3 is –0.304, then our zero-order correlation's magnitude is affected by the linear relationships item 3 has with items 1 and 2. To obtain an accurate assessment of the linear relationship between items 1 and 2, we should remove the linear influences of item 3 on items 1 and 2. Thus, we introduce a modified Q_3 statistic in which we calculate the (L – 2)-order partial correlation for items j and k

$$Q_3^P = r_{d_j d_k \cdot \underline{\mathbf{z}}}, \tag{6.5}$$

where $\underline{\mathbf{z}}$ represents all the instrument's items except items j and k. For our example, the first-order partial correlation between items 1 and 2 removing the linear effects of item 3 is $Q_3^P = -0.222$. Comparing Q_3^P and Q_3 shows item 3's linear effect on Q_3. In short, in this case Q_3 shows less item dependency between items 1 and 2 than does Q_3^P. As is the case with Q_3, large values of Q_3^P reflect item dependence, with values around 0 indicating either no linear relationship between items j and k or item independence.

Example: Application of the 3PL Model to the Mathematics Data, MMLE, BILOG-MG

A number of programs perform 3PL model calibration, including, but not limited to, BILOG-MG, XCALIBRE, `mirt`, NOHARM, SAS `proc irt`, and `TAM`. For comparison with our previous calibrations we use BILOG and then `mirt`.

Table 6.1 shows the command file for performing the calibration. As can be seen, we estimate both item and person parameters in a single run and save both the item estimates (`PARm = 'MATH3PL.PAR'`) and person location estimates (`SCOre = 'MATH3PL_EAP.SCO'`) using the `SAVE` subcommand on the `GLOBAL` command line and the `SAVE` command.

Table 6.2 contains the corresponding abridged Phases 1 and 2 output. The echo of the program parameters indicates that the intended model (`3 PARAMETER LOGISTIC`) and the logistic metric (`LOGIT METRIC`) are being used. The echo of the Phase 2 program parameters shows the maxima of 50 EM and 20 Newton cycles (`CALIB` line) as well as the default convergence criterion of 0.01. Moreover, the output indicates the use of prior distributions for the estimation of α_j and χ_j (i.e., `CONSTRAINT DISTRIBUTION ON SLOPES` and `CONSTRAINT DISTRIBUTION ON ASYMPTOTES`, respectively).

The Phase 2 results show convergence in 14 EM cycles, and 3 Newton cycles were executed. The item parameter estimates from the converged solution are presented in Table 6.3. The item discrimination, location parameter, and pseudo-guessing parameter estimates are obtained from the `SLOPE`, `THRESHOLD`, and `ASYMPTOTE` columns,

TABLE 6.1. BILOG Command File for the 3PL Model Item Calibration

```
Example 3PL Calibration w/ person scoring

>GLOBAL DFName = 'C: \Math.dat', NPArm = 3, LOGistic, SAVe;
>SAVE PARm = 'MATH3PL.PAR', SCOre = 'MATH3PL_EAP.SCO';
>LENGTH NITems = (5);
>INPUT NTOtal = 5, NIDchar = 10;
>ITEMS ;
>TEST1 TNAme = 'TEST0001', INUmber = (1(1)5);
(10A1, T1, 5(1X,A1))
>CALIB CYCLES=50, NEWtON=20, PLOt = 1.0000, ACCel = 1.0000,
       CHIsquare = (5, 8);
>SCORE ;
```

TABLE 6.2. BILOG Output: Phases 1 and 2 (Abridged)

<Phase 1 results>

```
                :
 >GLOBAL DFNAME='MATHPAT.DAT', NPARM=3, NWGHT=3, LOG, SAVE;

 FILE ASSIGNMENT AND DISPOSITION
 ===============================
 SUBJECT DATA INPUT FILE     C:\MATH.DAT
 BILOG-MG MASTER DATA FILE   MF.DAT
                                        WILL BE CREATED FROM DATA FILE

 CALIBRATION DATA FILE       CF.DAT
                                        WILL BE CREATED FROM DATA FILE

 ITEM PARAMETERS FILE        IF.DAT
                                        WILL BE CREATED THIS RUN

 CASE SCALE-SCORE FILE       SF.DAT
 CASE WEIGHTING                         NONE EMPLOYED

 ITEM RESPONSE MODEL                    3 PARAMETER LOGISTIC
                                        LOGIT  METRIC (I.E., D = 1.0)          :
     19601 OBSERVATIONS READ FROM FILE:   C:\MATH.DAT
     19601 OBSERVATIONS WRITTEN TO FILE:  MF.DAT


 ITEM STATISTICS FOR SUBTEST TEST0001

                                                          ITEM*TEST CORRELATION
 ITEM   NAME        #TRIED    #RIGHT   PCT      LOGIT    PEARSON  BISERIAL
 ------------------------------------------------------------------------------
    1   ITEM0001  19601.0   17395.0    88.7    -2.07     0.246    0.407
    2   ITEM0002  19601.0   12624.0    64.4    -0.59     0.439    0.564
    3   ITEM0003  19601.0   11094.0    56.6    -0.27     0.416    0.524
    4   ITEM0004  19601.0    8369.0    42.7     0.29     0.405    0.511
    5   ITEM0005  19601.0    7592.0    38.7     0.46     0.312    0.397
 ------------------------------------------------------------------------------
```

<Phase 2 results begin>

```
                :
CALIBRATION PARAMETERS
 ======================
MAXIMUM NUMBER OF EM CYCLES:                50
MAXIMUM NUMBER OF NEWTON CYCLES:            20
 CONVERGENCE CRITERION:                      0.0100
 ACCELERATION CONSTANT:                      1.0000
LATENT DISTRIBUTION:                       NORMAL PRIOR FOR EACH GROUP
 PLOT EMPIRICAL VS.  FITTED ICC'S:          YES, FOR ITEMS WITH FIT PROBABILITY
                                            LESS THAN 1.00000
 DATA HANDLING:                            DATA ON SCRATCH FILE
 CONSTRAINT DISTRIBUTION ON ASYMPTOTES:    YES
 CONSTRAINT DISTRIBUTION ON SLOPES:        YES
 CONSTRAINT DISTRIBUTION ON THRESHOLDS:    NO
 SOURCE OF ITEM CONSTRAINT DISTIBUTION
       MEANS AND STANDARD DEVIATIONS:      PROGRAM DEFAULTS          :
                :
METHOD OF SOLUTION:

 EM CYCLES (MAXIMUM OF    50)
 FOLLOWED BY NEWTON-RAPHSON STEPS (MAXIMUM OF  20)
                :
```

(continued)

TABLE 6.2. *(continued)*

```
 [EM STEP]

-2 LOG LIKELIHOOD =     111772.062
CYCLE      1;   LARGEST CHANGE=   0.35458              :
          -2 LOG LIKELIHOOD =    110088.005

-2 LOG LIKELIHOOD =     110066.473
CYCLE     14;   LARGEST CHANGE=   0.00782

 [FULL NEWTON CYCLES]
-2 LOG LIKELIHOOD:       110065.8456
CYCLE    15;   LARGEST CHANGE=   0.10845
               :

-2 LOG LIKELIHOOD:       110066.0225
CYCLE    17;   LARGEST CHANGE=   0.00327
               :
```

respectively. For instance, for item 1 the item discrimination estimate ($\hat{\alpha}_1$) is 1.608, and the item location estimate ($\hat{\delta}_1$) is –1.561, with a pseudo-guessing parameter estimate ($\hat{\chi}_1$) of 0.228. By and large, the $\hat{\chi}_j$s are acceptable.[8]

As part of our model–data fit analysis, we compare the empirical and predicted IRFs for each of our items. Figure 6.4 shows item 4's empirical and predicted IRFs. The estimate of the item's pseudo-guessing parameter is identified by the symbol *c* instead of

TABLE 6.3. BILOG Output: Phase 2 (Abridged)

```
                :
ITEM      INTERCEPT     SLOPE     THRESHOLD    LOADING    ASYMPTOTE     CHISQ      DF
             S.E.        S.E.        S.E.        S.E.        S.E.       (PROB)
 ------------------------------------------------------------------------------------
 ITEM0001 |   2.510  |  1.608  |  -1.561  |   0.849  |   0.228  |    693.1   4.0
          |   0.160* |  0.092* |   0.154* |  0.049*  |  0.096*  |  (0.0000)
          |          |         |          |          |          |
 ITEM0002 |   0.661  |  2.802  |  -0.236  |   0.942  |   0.156  |    826.5   3.0
          |   0.100* |  0.217* |   0.050* |  0.073*  |  0.029*  |  (0.0000)
          |          |         |          |          |          |
 ITEM0003 |  -0.328  |  2.397  |   0.137  |   0.923  |   0.202  |    665.4   4.0
          |   0.121* |  0.183* |   0.041* |  0.071*  |  0.020*  |  (0.0000)
          |          |         |          |          |          |
 ITEM0004 |  -1.482  |  2.788  |   0.532  |   0.941  |   0.147  |    495.5   5.0
          |   0.177* |  0.254* |   0.022* |  0.086*  |  0.011*  |  (0.0000)
          |          |         |          |          |          |
 ITEM0005 |  -1.420  |  1.608  |   0.883  |   0.849  |   0.156  |    611.5   5.0
          |   0.153* |  0.134* |   0.033* |  0.071*  |  0.016*  |  (0.0000)
 ------------------------------------------------------------------------------------
                                                                  * STANDARD ERROR

    LARGEST CHANGE =    0.003266                                 3291.9  21.0
                                                                 (0.0000)
                :
```

χ and the item's location by b. As can be seen, there is a close correspondence between the empirical and predicted IRFs. This correspondence provides evidence of data fit for this item. Figure 6.4 is typical of the other empirical versus predicted IRFs plots. Item 4's IRF and information function are presented in the left and right panels, respectively, of Figure 6.5. From the right panel we see that the item's information function is not quite centered about the item's location. This is true for all items calibrated with the 3PL model and is due to the influence of a nonzero lower asymptote. The actual location of the maximum of the item information is discussed below.

Fit Assessment: Conditional Independence Assessment

We use Q_3^P for evaluating the conditional independence assumption; Appendix G "Conditional Independence Using Q_3" shows the analysis using Q_3 and a simulation approach for identifying a screening value. With a five-item instrument, there are 10 Q_3^P values to calculate (i.e., L(L – 1)/2). To calculate Q_3^P we need to have person location estimates to calculate the expected responses, p_js. Our (EAP) estimates are obtained from the `MATH3PL _ EAP.SCO` file that we created in our calibration. These $\hat{\theta}$s, the response data, and the item parameter estimates are used to calculate the p_js as well as the residuals ($x_{ij} - p_j(\hat{\theta}_i)$). Using the residuals, we calculate the 10 third-order partial correlations (i.e., Q_3^P). Table 6.4 shows the Q_3^Ps for the mathematics data example; the scatterplots

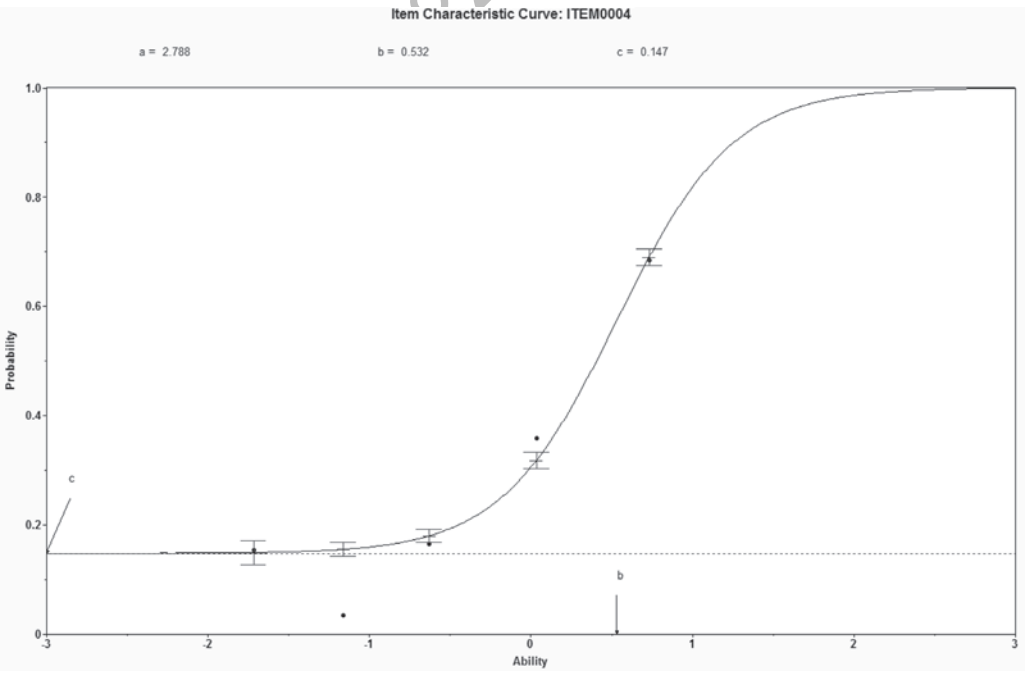

FIGURE 6.4. Empirical and predicted IRFs for item 4.

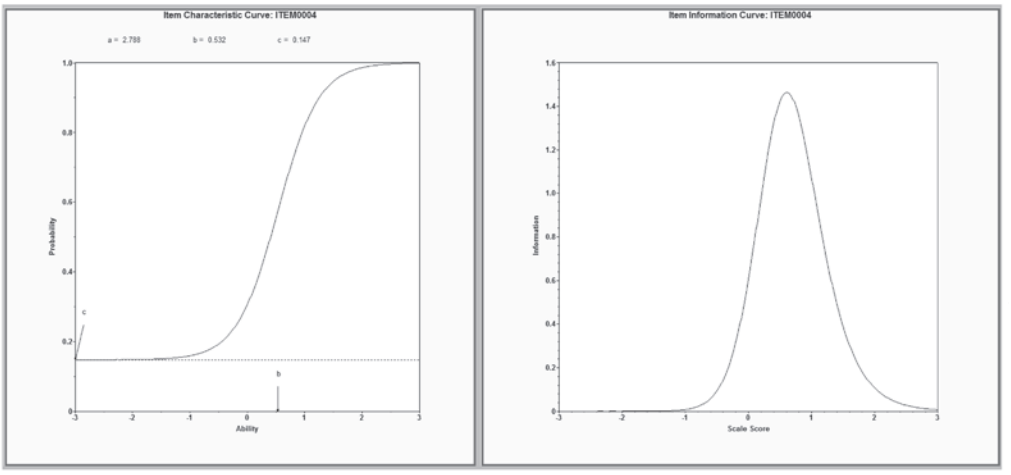

FIGURE 6.5. Item response and item information functions for item 4.

(not presented) corresponding to these values were inspected for anomalies, but none were found.

Figure 6.6 contains a dot density plot of our Q_3^Ps with the location of the mean Q_3^P ($\bar{Q}_3^P$ = –0.29789). To obtain $\bar{Q}_3^P$ we use the Fisher r to $\dot{z}$ transformation to convert each correlation to $\dot{z}$ and then calculate the average $\dot{z}$. This average $\dot{z}$ is transformed back to correlation. That is,

$$Q_3^P = \tanh\left(\sum_{j}^{L-1}\sum_{k=j+1}^{L} \operatorname{arctanh}(Q_{3(j,k)}^P) / {}_{L}C_r\right), \tag{6.6}$$

where ${}_{L}C_r = \frac{(L(L-1))}{2}$.

To identify values that reflect item dependence, we use a "gap" approach informed by $\bar{Q}_3^P$ in which an item pair that is substantially separated from the item pair cluster

TABLE 6.4. Q_3^P Statistics for the Math Data Set; $(Q_3^P)^2$ Are in Parentheses

	Items			
	1	2	3	4
2	-0.30300 (0.09181)			
3	-0.24700 (0.06101)	-0.44300 (0.19625)		
4	-0.23100 (0.05336)	-0.38500 (0.14823)	-0.37000 (0.13690)	
5	-0.16900 (0.02856)	-0.28100 (0.07896)	-0.25100 (0.06300)	-0.27800 (0.07728)

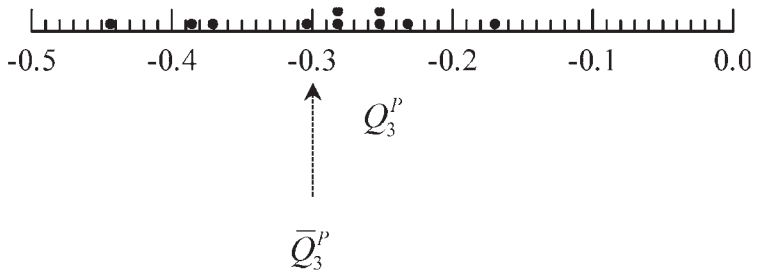

FIGURE 6.6. Dot density plot for Q_3^P.

reflects a potentially item dependent pair. For instance, we see that we have a cluster of values in the –0.30 to –0.22 range that contains $\bar{Q}_3^P$. These item pairs around the mean Q_3^P do not show item dependence. However, below this range there is a gap (~0.15) to get to the lowest value Q_3^P of –0.443. This item pair (items 2 and 3) has almost 20% of their variability in common (Table 6.4). Conversely, above this range we have a gap (~0.07) to reach our rightmost point Q_3^P = –0.169 for item pair 1–5. However, this item pair has only about 3% of shared variability. As such, we do not consider this item pair to be exhibiting item dependence. Additionally, one might consider the two values, Q_3^P = –0.37 and Q_3^P = –0.385 for item pairs 3–4 and 2–4, respectively, with a gap of approximately 0.075 from the cluster's lowest value as potentially exhibiting dependence. These two pairs, items 3–4 and items 2–4, have 13.7% and 14.8%, respectively, of their variance in common. Although these three item pairs may be considered to be exhibiting item dependence, evidence of conditional dependence in the remaining seven pairs is absent. Our analysis shows that after fitting the unidimensional 3PL model to the data, the items in these three item pairs (i.e., items 2, 3, 4) had more than 13% of their residual variability in common.[9] (These item pairs may or may not be found to be exhibiting item dependence with either the 1PL or 2PL models.)

How one deals with items that are considered sufficiently dependent to be problematic post administration is contingent on what one believes is the cause of the dependency. Again, inspection of the items exhibiting local dependence may be useful (the local dependence may be related to the locations of the items in the instrument, their text, dimensionality, and so on). In some cases where there is a great deal of dependence, it may be necessary to remove one of the dependent items for pragmatic reasons and because it is not clear as to why there is a dependency between the items. Because highly dependent items are in a sense redundant, the removal of one of the dependent items may not be problematic. In other cases, the items may be combined to form a testlet or the items combined to form an item parcel that is scored polytomously. In either case the instrument would need to be recalibrated.

For our example, the local dependence exhibited by the two item pairs could be addressed by forming a parcel for each pair. Parcel 1 would consist of items 1 and 4, whereas parcel 2 would involve items 2 and 5. Each parcel would have possible scores of 0 through 2, and the instrument would consist of three "items" (i.e., parcel 1, parcel 2,

and item 3). The corresponding response data could be calibrated using, for example, the polytomous partial credit model discussed in Chapter 7.

Fit Assessment: Model Comparison

In this and in previous chapters, our focus has been on whether a particular model is exhibiting model–data fit. We now present three model–data fit statistics that can be used for making model comparisons and selection. These complementary procedures should be used after obtaining evidence supporting model–data fit. The first of these is based on the likelihood ratio (G^2) test statistic for comparing the relative fit of hierarchically related models. The second statistic is analogous to the use of R^2 for comparing various regression models, whereas the third is based on an information criterion.

The change in G^2 across models can be used to determine whether two hierarchically related models significantly differ from one another. For instance, the 2PL model can be considered to be nested within the 3PL model because constraining the 3PL model's χ_js to be 0 yields the 2PL model. Similarly, imposing the constraint that all items have the same discrimination parameter on the 2PL model produces the 1PL model. If we impose the constraints that all the items have a common item discrimination parameter and χ_js equals 0, then the 3PL model reduces to the 1PL model. As such, the 1PL model is nested within both the 2PL and 3PL models. In the following discussion, we refer to the more complex (or less constrained) model as the *full* model and the less complex/simpler (or more constrained) model as the *reduced* model. The likelihood ratio test is the difference between two deviance statistics

$$\Delta G^2 = (-2\ln L_R) - (-2\ln L_F) = G_R^2 - G_F^2, \tag{6.7}$$

where L_R is the maximum of the likelihood for the reduced model and L_F is the maximum of the likelihood for the full model. The degrees of freedom (*df*) for evaluating the significance of ΔG^2 is the difference in the number of parameters between the full model and the reduced model.[10] This statistic is distributed as a χ^2 when the sample size is large and the full (nesting) model holds for the data. A nonsignificant statistic indicates that the additional complexity of the nesting model is not necessary. For instance, if a comparison of the 2PL model with the 3PL model is not significant, then the additional estimation of the pseudo-guessing parameters (i.e., the increased model complexity) is not necessary to improve model–data fit over and above that obtained with the 2PL model.

Table 6.5 contains the values of the –2 log likelihoods (i.e., –2ln*L*) for the 1PL, 2PL, and 3PL models from our BILOG calibrations. The –2ln*L* is the last entry from the converged solution's iteration history and is labeled `–2 LOG LIKELIHOOD:` in the output. As can be seen, as the models increase in complexity, the corresponding G^2s decrease. The difference between the 1PL and 2PL models is

$$\Delta G^2 = (-2\ln L_R) - (-2\ln L_F) = 110{,}774.295 - 110{,}397.103 = 377.191$$

with 4 *df*. Therefore, at the instrument level the 2PL model represents a significant (at the 5% level) improvement in fit over the 1PL model. An analogous comparison between the 2PL and 3PL models also shows a significant improvement in fit by the 3PL model over the 2PL model. Therefore, the 3PL model fits significantly better than either the 2PL or 1PL model.

Our second model comparison statistic is complementary to ΔG^2. This approach uses G^2 in a manner analogous to comparing various regression models' R^2s. That is, the change in R^2s may be used for assessing the relative improvement in the proportion of variability accounted for by one model over another model. In the current context, our strategy is to calculate the relative reduction in G^2s (Haberman, 1978). For instance, for the comparison of the 2PL model (G_F^2) with the 1PL model (G_R^2) we would calculate

$$R_\Delta^2 = \frac{G_R^2 - G_F^2}{G_R^2} = \frac{110774.295 - 110397.103}{110774.295} = 0.0034$$

This R_Δ^2 indicates that the 2PL model results in a 0.34% improvement in fit over the 1PL model. Comparing the 3PL and 2PL models we have

$$R_\Delta^2 = \frac{110397.103 - 110066.023}{110397.103} = 0.0030$$

Therefore, using the more complex 3PL model results in an improvement of fit of 0.3% over the 2PL model. We do not consider this to be a meaningful improvement in fit vis-à-vis the increase in model complexity. Summarizing the results so far, we have that the 3PL model fits significantly better than the 2PL and 1PL models, but it does not result in a *meaningful* improvement of fit of over either model.

Table 6.5 shows the AIC and BIC values for the three models. As is the case with ΔG^2 above, we see that even taking the 3PL model's additional complexity into account (i.e., relative to the 1PL and 2PL models), these statistics indicate that it is the best fitting of these three models.

Although our triangulation with ΔG^2, AIC, and BIC shows that the 3PL model is the best-fitting model of the three considered, our R_Δ^2 shows that the differences are slight. In fact, the correlation between the 1PL model-based $\hat{\theta}$s and those of the 2PL model is 0.9908, and for the $\hat{\theta}$s from the 2PL and 3PL models the correlation is 0.9869; the lowest

TABLE 6.5. Model Fit Statistics

Model	-2ln*L*	*df*	Relative Change	Number of Parameters	AIC	BIC
1PL[a]	110,774.295	25		6	110,786.295	110,833.595
2PL	110,397.103	21	0.0034	10	110,417.103	110,495.937
3PL	110,066.023	16	0.0030	15	110,096.023	110,214.273

[a]Five item locations plus a common α.

correlation is between the 1PL and 3PL models' $\hat{\theta}$s, $r = 0.9778$. That is, although the 3PL model is the best fitting of the three models, we have a high degree of linear agreement in the ordering of individuals across the three models. Based solely on the R^2_{Δ}, the magnitude of the $\hat{\theta}$ intercorrelations, the variability in the $\hat{\alpha}$s (both this chapter and Chapter 5), and the axiom "Make everything as simple as possible, but not simpler" (Albert Einstein), we would select the 2PL model for modeling these data. (However, we believe that a reasonable argument can be made for selecting the 1PL model.) Additional points to consider in model selection are presented below in the section "Issues to Consider in Selecting among the 1PL, 2PL, and 3PL Models."

Example: Application of the 3PL Model to the Mathematics Data, MMLE, `mirt`

As in Chapter 5, we assume the data and the relevant libraries are loaded into our R workspace. To perform our calibration, we specify the `3PL` model in our call to the `mirt` function (`ThreePL = mirt(mathdata,1, '3PL', SE = T, SE.type = 'Fisher')`). Our calibration required 58 iterations to obtain convergence (Table 6.6).

Examining our item parameter estimates, we notice that our first item has a large $\hat{\chi}_1$ of 0.609 and a comparatively good discrimination ($\hat{\alpha}_1 = 2.289$); the item is located at $\hat{\delta}_1 = -0.703$. The large $\hat{\chi}_1$ coupled with good discrimination is somewhat counterintuitive. Our traditional item statistics corroborate the item's easiness, with almost 89% of the respondents providing a correct response (P-value = 0.8875); this item did not do as well as the other items in differentiating among observed scores (corrected $r_{1,NC} = 0.246$). Our observed score frequency distribution shows that only 3.5% of our sample have a $X = 0$, and our empirical IRFs for item 1 (Figure 6.7) show there is little observed data below approximately –1. In toto, we conjecture that there is not enough information at the lower end of the continuum to "properly" estimate item 1's lower asymptote.

As Figure 6.7 shows, the leftmost empirical point ($\theta \cong -1.12$) has a proportion correct of about 0.33 that is not quite being captured by the IRF. With 6 fractiles we see a smoother empirical pattern that appears to indicate that the IRF should continue further down to reflect the leftmost empirical point. Because we are modeling the data, we decide to impose priors on our $\hat{\chi}$s to enhance the correspondence between our $\hat{\chi}$s and our data. This two-step process begins with determining the item parameter number for the parameter of interest and then specifying the prior using the item parameter number. To determine the item parameter number, we execute `mirt` using the `pars = 'values'` argument. The corresponding output object (`modThreePL`) is shown in Table 6.7, with the leftmost column containing the item parameter numbers and the column labeled `item` and `name` specifying the item and corresponding parameter label, respectively. For instance, item 1's information is on the first four lines, with item parameter number 1 being used for α_1 (labeled `a1`), number 2 for δ_1 (labeled `d`), number 3 for χ_1 (labeled `g`), and number 4 for Υ_1 (labeled `u`), respectively. We can identify the item parameter numbers for $\chi_1, \chi_2, \chi_3, \chi_4$, and χ_5 from the appropriate line in the `modThreePL` display.

TABLE 6.6. `mirt` Session for the 3PL Calibration of the Mathematics Data (No Prior)

```
> # read data, load mirt, etc.

    > print((ThreePL = mirt(mathdata,1,'3PL',SE=T,SE.type='Fisher')))
    Iteration: 58, Log-Lik: -55028.684, Max-Change: 0.00008

    Calculating information matrix...

    Call:
    mirt(data = mathdata, model = 1, itemtype = "3PL", SE = T, SE.type = "Fisher")

    Full-information item factor analysis with 1 factor(s).
    Converged within 1e-04 tolerance after 58 EM iterations.
    mirt version: 1.30
    M-step optimizer: BFGS
    EM acceleration: Ramsay
    Number of rectangular quadrature: 61
    Latent density type: Gaussian

    Information matrix estimated with method: Fisher
    Condition number of information matrix = 1897.529
    Second-order test: model is a possible local maximum

    Log-likelihood = -55028.68
    Estimated parameters: 15
    AIC = 110087.4; AICc = 110087.4
    BIC = 110205.6; SABIC = 110157.9
    G2 (16) = 55.71, p = 0
    RMSEA = 0.011, CFI = NaN, TLI = NaN


> coef(ThreePL,simplify=TRUE,IRTpars=TRUE)
    $items
           a      b     g u
    I1 2.289 -0.703 0.609 1
    I2 2.640 -0.306 0.108 1
    I3 2.523  0.154 0.212 1
    I4 2.736  0.519 0.143 1
    I5 1.618  0.867 0.154 1

> # get proportion correct
> summary(mathdata)
          I1                I2               I3               I4               I5
    Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.0000
    1st Qu.:1.0000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000
    Median :1.0000   Median :1.000   Median :1.000   Median :0.000   Median :0.0000
    Mean   :0.8875   Mean   :0.644   Mean   :0.566   Mean   :0.427   Mean   :0.3873
    3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.0000
    Max.   :1.0000   Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.0000

> # obtain corrected item total correlations & item total correlations
> NC=c(rep(-1,19601))                                          # create & initialize NC

> for (i in 1:19601){NC[i]=sum(mathdata[i,])}                  # calculate NC
```

(continued)

TABLE 6.6. *(continued)*

```
> table(NC)                                                    # frequency distribution
    NC
       0    1    2    3    4    5
     691 3099 4269 4116 4041 3385

> # calculate corrected point biserial & point biserial for each item
> for (j in 1:5){
+ print(j)
+ print(cor((NC-mathdata[,j]),mathdata[,j])); print(cor(NC,mathdata[,j])) }
    [1] 1
    [1] 0.2460252
    [1] 0.4473804
    [1] 2
    [1] 0.4389591
    [1] 0.6882788
    [1] 3
    [1] 0.4156754
    [1] 0.6804917
    [1] 4
    [1] 0.4051335
    [1] 0.6727627
    [1] 5
    [1] 0.3117375
    [1] 0.6017258

> itemfit(ThreePL,S_X2.tables=T,empirical.table=1)   # item 1
    $`theta = -1.1215`
          Observed  Expected z.Residual
    cat_0     1309  554.1609   32.06538
    cat_1      651 1405.8391  -20.13198

    $`theta = -0.936`
          Observed  Expected z.Residual
    cat_0      331  483.2472  -6.925715
    cat_1     1629 1476.7528   3.961826
    :

> itemfit(ThreePL,group.bins=10,empirical.plot=1,empirical.CI=0)   # produces Figure 6.7
> itemfit(ThreePL,group.bins=6,empirical.plot=1,empirical.CI=0)    # produces Figure 6.7
```


Alternatively, we can extract the parameter numbers by using a Boolean expression to select only the parameter numbers from the `parnum` variable when the variable name is set to `g (with(modThreePL,parnum[name = = 'g'])`; we use the `with` function to minimize typing the output object name. The corresponding item parameter numbers for our $\chi_1, \chi_2, \chi_3, \chi_4$, and χ_5 are 3, 7, 11, 15, and 19, respectively.

To impose the prior, we specify the type (`'prior.type'`) and its parameters (`'prior _ 1', 'prior _ 2'`). The prior has a location value of –1.5 and a scale value of 0.5. We use a generic `for` loop to implement the three assignments (`'prior.type', 'prior _ 1', 'prior _ 2'`) for each item. Our `for` loop executes for each of our items; the number of items is stored in the output object `ThreePL` in the variable `@Data$nitems` (see Table 6.6 for the creation of `ThreePL`). The body of the loop

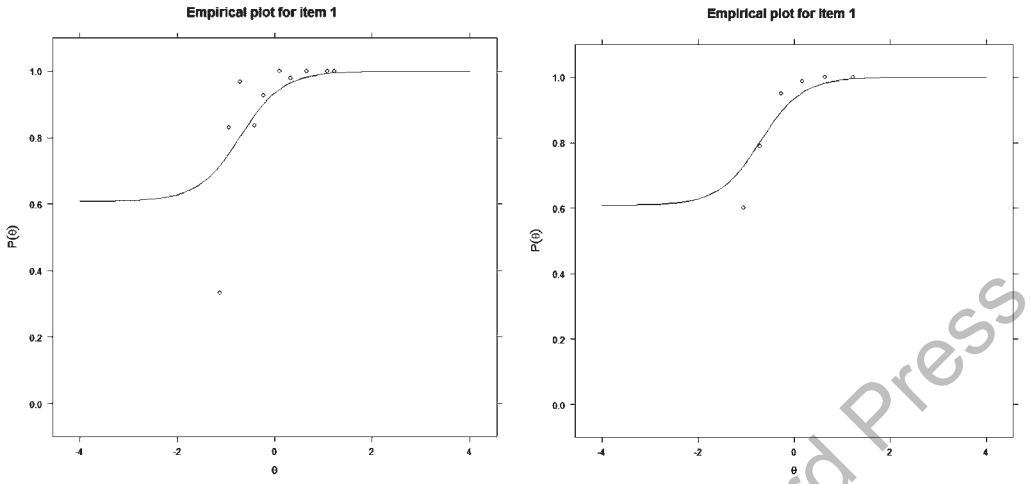

FIGURE 6.7. IRF for item 1 with observed proportions (left: 10 fractiles; right: 6 fractiles).

contains three assignments, with the indexing of the item parameter done by using the variable `itm`. We initialize `itm` to our first item parameter number 3. Each iteration of the `for` loop increments `itm` by the difference between successive parameter numbers (i.e., 4).

In our call to `mirt`, we pass our modified `modThreePL` by using the `pars` argument (`pars = modThreePL`). Our calibration required fewer iterations (i.e., 37) than when we did not impose priors on the estimation of the $\hat{\chi}$s (i.e., 58 iterations). In contrast to our previous estimates for item 1, we now have $\hat{\alpha}_1 = 1.683$, $\hat{\delta}_1 = -1.542$, and $\hat{\chi}_1$ of 0.221. Our new $\hat{\chi}_1$ is closer to the data. Figure 6.8 shows the corresponding IRFs. Both figures show a better agreement with the data than is seen in Figure 6.7. Comparing items 2–5's $\hat{\chi}$s with and without use of the prior shows a difference on the order of 0.007 or less for items 3–5 and 0.058 for item 2. The items' corresponding IRFs and item information are shown in Figure 6.9. The nonzero lower asymptotes are evident in the IRFs. Similarly, the varying discriminations is seen in both the slope of the IRFs and the heights of the item information functions. As in previous chapters, our `mirt` item parameter estimates show correlations of 0.991 or higher with those of BILOG.

Because our fit analysis at both the model and item levels proceeds as demonstrated in Chapter 5, we do not repeat it here. However, we note that according to the information criteria (e.g., `AIC, BIC`), the 3PL model is found to fit better than the 2PL model (i.e., `anova(TwoPL, ThreePL)`). Moreover, we show how to examine conditional dependence using Q_3 (`residuals(ThreePL,type = "Q3")`); Q_3 is one of the local dependency statistics provided by `mirt`. See Appendix G, "Conditional Independence Using Q_3," for more information on using Q_3.

We examine item parameter invariance using multiple-group analysis. We begin by creating two random samples, along with a binary group indicator variable (`grp`), followed by concatenating the two samples to create our data frame `mathdatagrp`. (Because we use Chapter 5's seed (`set.seed(99999)`), these samples are the same as those in Chapter 5.) To impose priors on the $\hat{\chi}$s, we use the `pars = 'values'`

TABLE 6.7. `mirt` Session for the 3PL Calibration of the Mathematics Data (Prior)

```
> # This is a continuation of the session from Table 6.6

> # go get item parameter numbers (3, 7, 11, 15, 19)
> print((modThreePL = mirt(mathdata,1,'3PL',SE=T,SE.type='Fisher',pars='values')))
   group  item  class  name parnum      value lbound ubound   est prior.type prior_1
1    all    I1   dich    a1      1  0.8510000   -Inf    Inf  TRUE       none     NaN
2    all    I1   dich     d      2  2.3841111   -Inf    Inf  TRUE       none     NaN
3    all    I1   dich     g      3  0.1500000  0e+00      1  TRUE       none     NaN
4    all    I1   dich     u      4  1.0000000  0e+00      1 FALSE       none     NaN
5    all    I2   dich    a1      5  0.8510000   -Inf    Inf  TRUE       none     NaN
6    all    I2   dich     d      6  0.7257898   -Inf    Inf  TRUE       none     NaN
7    all    I2   dich     g      7  0.1500000  0e+00      1  TRUE       none     NaN
:
18   all    I5   dich     d     18 -0.5626501   -Inf    Inf  TRUE       none     NaN
19   all    I5   dich     g     19  0.1500000  0e+00      1  TRUE       none     NaN
20   all    I5   dich     u     20  1.0000000  0e+00      1 FALSE       none     NaN
21   all GROUP GroupPars MEAN_1     21  0.0000000   -Inf    Inf FALSE   none     NaN
22   all GROUP GroupPars COV_11     22  1.0000000  1e-04    Inf FALSE   none     NaN

   prior_2
1      NaN
2      NaN
:
21     NaN
22     NaN

> with(modThreePL,parnum[name == 'g'])
    [1]  3  7 11 15 19

> ThreePL@Data$nitems                          # ThreePL was created in Table 6.6
    [1] 5

> itm=3
> for(j in 1:ThreePL_a@Data$nitems[1]){
+      modThreePL[itm,'prior.type']='norm'
+      modThreePL[itm,'prior_1']=-1.5
+      modThreePL[itm,'prior_2'] =0.5
+ itm=itm+4
+ }   # end for j loop

> modThreePL                                   # checking that prior information was correctly  imposed
   group  item  class  name parnum      value lbound ubound   est prior.type prior_1
1    all    I1   dich    a1      1  0.8510000   -Inf    Inf  TRUE       none     NaN
2    all    I1   dich     d      2  2.3841111   -Inf    Inf  TRUE       none     NaN
3    all    I1   dich     g      3  0.1500000  0e+00      1  TRUE       norm    -1.5
4    all    I1   dich     u      4  1.0000000  0e+00      1 FALSE       none     NaN
5    all    I2   dich    a1      5  0.8510000   -Inf    Inf  TRUE       none     NaN
6    all    I2   dich     d      6  0.7257898   -Inf    Inf  TRUE       none     NaN
7    all    I2   dich     g      7  0.1500000  0e+00      1  TRUE       norm    -1.5
 :
19   all    I5   dich     g     19  0.1500000  0e+00      1  TRUE       norm    -1.5
:

> # use prior information with 'pars=' argument
> print((ThreePL = mirt(mathdata,1,'3PL',SE=T,SE.type='Fisher',pars=modThreePL)))
    Iteration: 37, Log-Lik: -55033.198, Max-Change: 0.00008
```

(continued)

TABLE 6.7. *(continued)*

```
    Calculating information matrix...
    Warning message:
    In ESTIMATION(data = data, model = model, group = rep("all", nrow(data)),  :
      Information matrix with the Fisher method does not
      account for prior parameter distribution information

    Call:
    mirt(data = mathdata, model = 1, itemtype = "3PL", SE = T, SE.type = "Fisher",
        pars = modThreePL)

    Full-information item factor analysis with 1 factor(s).
    Converged within 1e-04 tolerance after 37 EM iterations.
    mirt version: 1.30
    M-step optimizer: nlminb
    EM acceleration: Ramsay
    Number of rectangular quadrature: 61
    Latent density type: Gaussian

    Information matrix estimated with method: Fisher
    Condition number of information matrix = 50656.81
    Second-order test: model is a possible local maximum

    Log-posterior = -55033.2
    Estimated parameters: 15
    DIC = 110096.4
    G2 (16) = 61.17, p = 0
    RMSEA = 0.012, CFI = NaN, TLI = NaN

> print(coef(ThreePL,IRTpars=TRUE,printSE=T),digits=5)
    $I1
               a        b       g  u
    par 1.64259 -1.54157 0.22136  1
    SE  0.11009  0.14091 0.02691 NA

    $I2
               a        b       g  u
    par 2.92857 -0.21561 0.16569  1
    SE  0.27594  0.05758 0.05430 NA

    $I3
               a       b       g  u
    par 2.45749 0.14175 0.20510  1
    SE  0.22644 0.04933 0.01449 NA

    $I4
               a       b       g  u
    par 2.88706 0.52929 0.15035  1
    SE  0.32700 0.02762 0.04145 NA

    $I5
               a       b       g  u
    par 1.64094 0.87442 0.15803  1
    SE  0.17635 0.04386 0.04594 NA

    $GroupPars
        MEAN_1 COV_11
    par      0      1
    SE      NA     NA
```

(continued)

TABLE 6.7. *(continued)*

```
> anova(TwoPL,ThreePL)                                   # The object TwoPL was created in Chapter 5

    Model 1: mirt(data = mathdata, model = 1, itemtype = "2PL", SE = T,
        SE.type = "Fisher")
    Model 2: mirt(data = mathdata, model = 1, itemtype = "3PL", SE = T,
        SE.type = "Fisher", pars = modThreePL)

          AIC     AICc    SABIC       HQ      BIC      DIC    logLik   logPost  df
    1 110417.0 110417.0 110464.0 110442.8 110495.8 110417.0 -55198.50 -55198.50 NaN
    2 110093.1 110093.1 110163.7 110131.8 110211.3 110096.1 -55031.55 -55033.03   5
      Bayes_Factor
    1           NA
    2            0

> itemfit(ThreePL,group.bins=10,empirical.plot=1,empirical.CI=0)  # produces Figure 6.8 (left)
> itemfit(ThreePL,group.bins=6,empirical.plot=1,empirical.CI=0)   # produces Figure 6.8 (right)

> residuals(ThreePL,type="Q3")                                    # Yen's Q3
    Q3 matrix:
           I1     I2     I3     I4     I5
    I1  1.000 -0.185 -0.100 -0.092 -0.059
    I2 -0.185  1.000 -0.300 -0.212 -0.137
    I3 -0.100 -0.300  1.000 -0.208 -0.097
    I4 -0.092 -0.212 -0.208  1.000 -0.158
    I5 -0.059 -0.137 -0.097 -0.158  1.000


> marginal_rxx(ThreePL)
    [1] 0.6104937

> # empirical reliability
> ThreePLrxx=fscores(ThreePL,method="EAP",full.scores=T,full.scores.SE=T,returnER=T)
> ThreePLrxx
           F1
    0.6294784

> # examination of invariance ---------  use of priors
> set.seed(99999)
> caseU=runif(19601)
> sortmathdata=mathdata
> sortmathdata$unif=caseU
> sortmathdata=sortmathdata[order(sortmathdata$unif),]
> mathdata1=sortmathdata[1:9800,]       ;  mathdata2=sortmathdata[9801:19601,]
> mathdata1=within(mathdata1,rm(unif))  ;  mathdata2=within(mathdata2,rm(unif))
> names(mathdata1) = c(paste0("I",1:5)) ;  names(mathdata2) = c(paste0("I",1:5))

> mathdata1$grp='0'         ;  mathdata2$grp='1'        # create (0,1) group variable

> mathdatagrp=rbind(mathdata1, mathdata2)               # concatenate the two randomized groups
> grpvar=mathdatagrp$grp                                # extract group variable

> mathdatagrp=within(mathdatagrp,rm(grp))
> ThreePLgrp=multipleGroup(mathdatagrp,1,itemtype='3PL',group=grpvar, pars='values')
> ThreePLgrp

> ThreePLgrp
   group  item   class  name parnum     value lbound ubound   est prior.type prior_1
1      0   I1    dich    a1      1 0.8510000   -Inf    Inf  TRUE       none      NaN
2      0   I1    dich     d      2 2.3841111   -Inf    Inf  TRUE       none      NaN
3      0   I1    dich     g      3 0.1500000  0e+00      1  TRUE       none      NaN
4      0   I1    dich     u      4 1.0000000  0e+00      1 FALSE       none      NaN
```

(continued)

TABLE 6.7. *(continued)*

```
5      0   I2    dich    a1      5  0.8510000   -Inf    Inf  TRUE       none      NaN
6      0   I2    dich     d      6  0.7257898   -Inf    Inf  TRUE       none      NaN
:
41     NaN
42     NaN
43     NaN
44     NaN

> itm=with(ThreePLgrp,parnum[name == 'g'])                    #obtain & store item numbers for 'g'
> itm
    [1]  3  7 11 15 19 25 29 33 37 41
> ngrps=2                                                     # number of groups

> # imposing priors on group "0" and group "1"
> for(j in 1:(ThreePL@Data$nitems[1]*ngrps)){
> + ThreePLgrp[itm[j],'prior.type']='norm'; ThreePLgrp[itm[j],'prior_1']=-1.5;
     ThreePLgrp[itm[j],'prior_2']=0.5   }

> ThreePLgrp                                          # checking that prior information was correctly imposed
   group  item   class  name parnum     value lbound ubound  est prior.type prior_1
1      0   I1    dich    a1      1  0.8510000   -Inf    Inf  TRUE       none      NaN
2      0   I1    dich     d      2  2.3841111   -Inf    Inf  TRUE       none      NaN
3      0   I1    dich     g      3  0.1500000  0e+00      1  TRUE       norm     -1.5
4      0   I1    dich     u      4  1.0000000  0e+00      1 FALSE       none      NaN
5      0   I2    dich    a1      5  0.8510000   -Inf    Inf  TRUE       none      NaN
6      0   I2    dich     d      6  0.7257898   -Inf    Inf  TRUE       none      NaN
7      0   I2    dich     g      7  0.1500000  0e+00      1  TRUE       norm     -1.5
8      0   I2    dich     u      8  1.0000000  0e+00      1 FALSE       none      NaN
:
41     0.5
42     NaN
43     NaN
44     NaN

> ThreePLgrp=multipleGroup(mathdatagrp,1,itemtype='3PL',group=grpvar,pars=ThreePLgrp)
    Iteration: 40, Log-Lik: -55030.224, Max-Change: 0.00010

> ThreePLgrp
    Call:
    multipleGroup(data = mathdatagrp, model = 1, group = grpvar,
        itemtype = "3PL", pars = ThreePLgrp)

    Full-information item factor analysis with 1 factor(s).
    Converged within 1e-04 tolerance after 40 EM iterations.
    mirt version: 1.30
    M-step optimizer: nlminb
    EM acceleration: Ramsay
    Number of rectangular quadrature: 61
    Latent density type: Gaussian

    Log-posterior = -55030.22
    Estimated parameters: 30                                  ← Nparm*L*2 groups
    DIC = 110120.4
    G2 (1) = 79.67, p = 0
    RMSEA = 0.063, CFI = NaN, TLI = NaN
```

(continued)

TABLE 6.7. *(continued)*

```
> coef(ThreePLgrp,simplify=TRUE,IRTpars=TRUE)
    $`0`                                                    ← Group 0 item parameter estimates
    $items
           a      b     g u
    I1 1.629 -1.584 0.198 1
    I2 2.977 -0.208 0.185 1
    I3 2.474  0.130 0.212 1
    I4 3.184  0.527 0.158 1
    I5 1.661  0.875 0.162 1

    $means
    F1
     0

    $cov
       F1
    F1  1


    $`1`          Group 1 item parameter estimates
    $items
           a      b     g u
    I1 1.637 -1.565 0.198 1
    I2 2.931 -0.210 0.155 1
    I3 2.429  0.151 0.197 1
    I4 2.644  0.532 0.143 1
    I5 1.622  0.875 0.155 1

    $means
    F1
     0

    $cov
       F1
    F1  1

> plot(ThreePLgrp, type = 'trace', theta_lim=c(-4,4))          # Figure 6.10
> plot(ThreePLgrp,type="score",theta_lim=c(-4,4))              # Figure 6.11
>
> # end of examination of invariance ---------

> plot(ThreePL, type = 'trace', theta_lim=c(-4,4))             # produces Figure 6.9 (top)
> plot(ThreePL, type = 'infotrace', theta_lim=c(-4,4))         # produces Figure 6.9c (top)
```


argument with the `multipleGroup` function. However, we store the item parameter numbers for the two groups in `itm` to directly access them in our `for` loop (`itm = with(ThreePLgrp,parnum[name == 'g'])`). Group 0's item parameter numbers are the same as above, whereas for group 1 we have 25, 29, 33, 37, and 41. We impose the priors on our two groups using a for loop in which the appropriate item numbers are indexed by *j* (i.e., `itm[j]`). We call `multipleGroup` a second time and pass to it our data frame containing the prior information (`pars = ThreePLgrp`). Convergence was achieved in 40 iterations. Comparison of group 0's item parameter estimates with those of group 1's shows close correspondence. Our groups' IRFs for each item are shown in Figure 6.10. At the item level, the agreement between the two sets of IRFs for

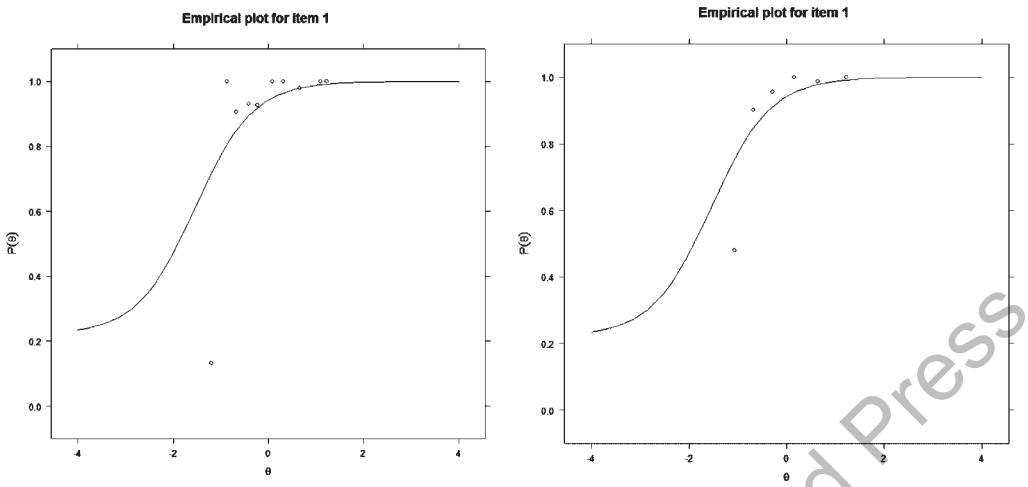

FIGURE 6.8. IRF for item 1 with observed proportions (left: 10 fractiles; right: 6 fractiles).

each item provides us with evidence of invariance. (Because our item parameter point estimates show estimation error, we interpret the minor discrepancies between IRFs as being within this margin of error.) Additionally, at the "model level," our groups' total characteristic curves (TCCs) show strong agreement with one another and provide us with additional invariance evidence (Figure 6.11).

Assessing Person Fit: Appropriateness Measurement

Various person fit measures have been previously discussed. From one perspective, these measures are trying to determine whether the person is behaving in a fashion consistent with the model. Alternatively, one may ask, what is the *appropriateness* of a person's estimated location, $\hat{\theta}$, as a measure of their true location (θ)? For instance, imagine that a person has a response pattern of missing easy items and correctly answering more difficult items. Did this pattern arise from the person's correctly guessing on some difficult items and incorrectly responding to easier items, or does this reflect a person who was able to copy the answers on some items? Various statistically based indices have been developed to measure the degree to which an individual's response pattern is unusual or is inconsistent with the model used for characterizing their performance. These indices of person fit are examples of *appropriateness measurement* (e.g., Levine & Drasgow, 1983; Meijer & Sijtsma, 2001).

One index, l_z, has been found to perform better than other person fit measures (e.g., Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, & Williams, 1985). This index is based on a standardization of the person log likelihood function to address the interaction of ln$L$ and θ. As such, this standardization of log likelihood allows us to compare individuals at different θ levels on the basis of their l_z values.

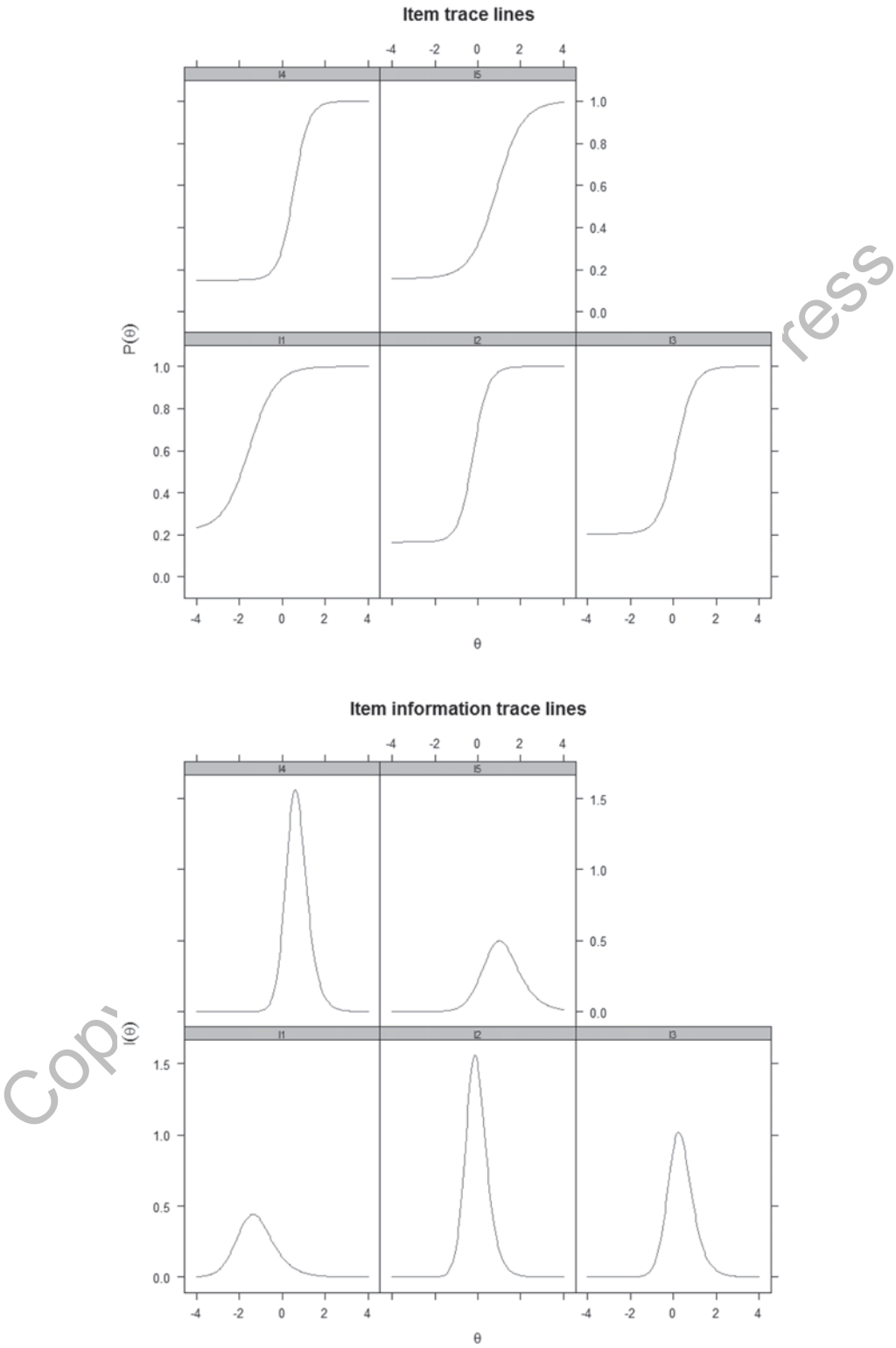

FIGURE 6.9. IRFs and item information for all items.

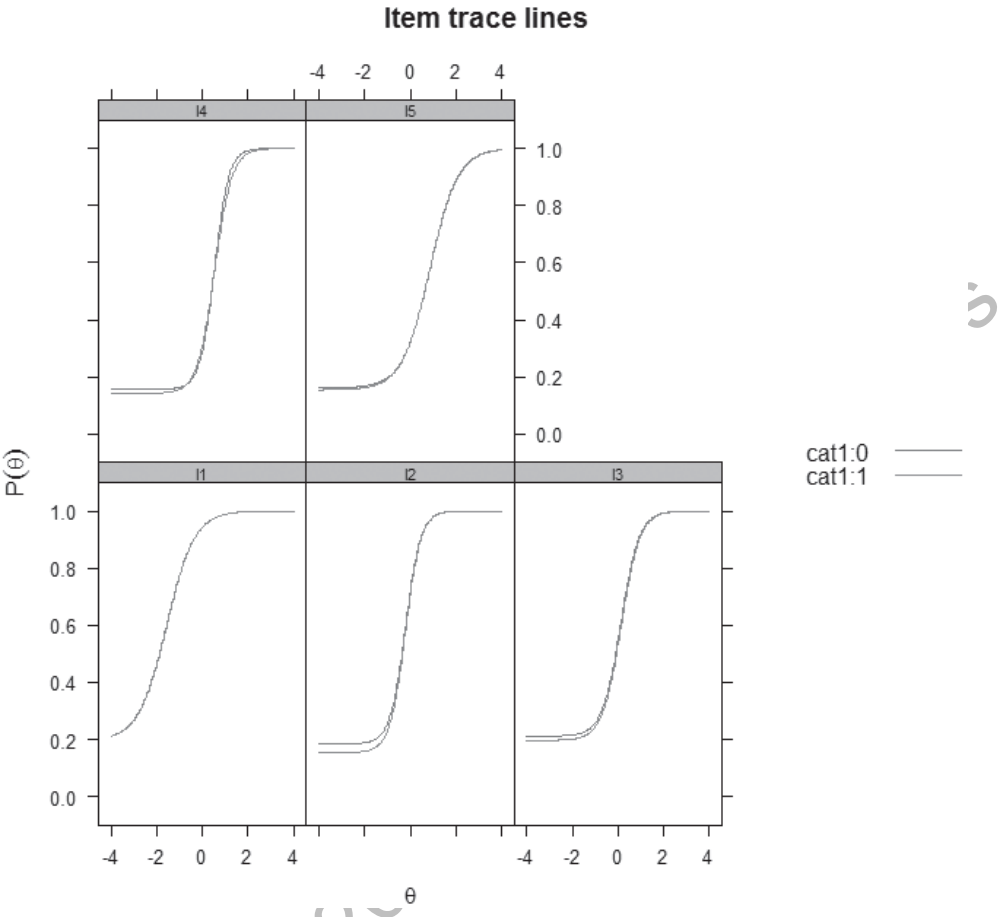

FIGURE 6.10. IRFs for two-group analysis.

To present *l*z we start with the log likelihood function for a person *i*'s response vector

$$\ln L(\underline{\mathbf{x}}_i, \theta, \underline{\boldsymbol{\alpha}}, \underline{\boldsymbol{\delta}}, \underline{\boldsymbol{\chi}}) = \sum_{j=1}^{\mathrm{L}} [x_{ij} \ln(p_j) + (1 - x_{ij}) \ln(1 - p_j)] \tag{6.8}$$

To standardize ln*L* we need both its variance and expected value. The expected value of the ln*L* is given by

$$\mathcal{E}(\ln L) = \sum_{j=1}^{\mathrm{L}} [p_j \ln(p_j) + (1 - p_j) \ln(1 - p_j)] \tag{6.9}$$

and its variance by

$$Var(\ln L) = \sum_{j=1}^{\mathrm{L}} \left\{ p_j (1 - p_j) \left[\ln \frac{p_j}{1 - p_j} \right]^2 \right\} \tag{6.10}$$

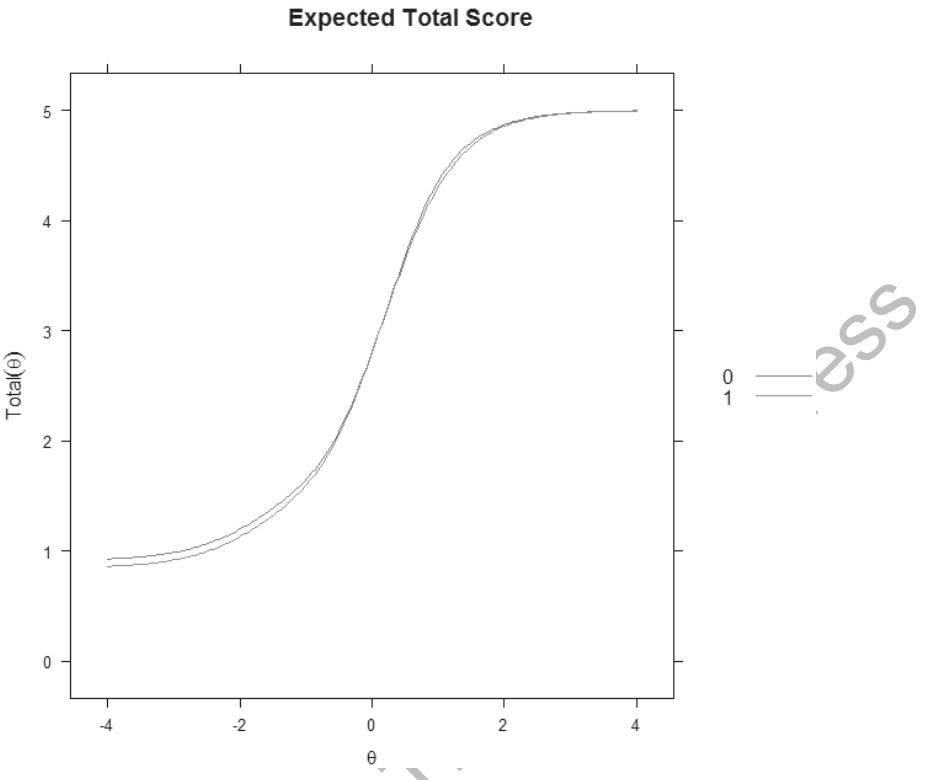

FIGURE 6.11. Total characteristic curves for two-group analysis.

Using Equations 6.8–6.10 and the *z*-score formula, we obtain

$$l_z = \frac{\ln L - \mathcal{E}(\ln L)}{\sqrt{Var(\ln L)}} \tag{6.11}$$

In practice, we use estimates in lieu of parameters in the calculation of p_j (e.g., $\hat{\theta}$ for θ).

Although l_z is purported to have a unit normal distribution, this has not necessarily been true for instruments of different lengths (Drasgow et al., 1985, 1987; Levine & Drasgow, 1983). Moreover, because the l_z uses person parameter estimates in its calculation, it is not asymptotic normal. To this end, Nering (1995) found l_z's detection accuracy approaching the significance level is adversely affected by how well the person locations are estimated. Therefore, using the standard normal curve for hypothesis testing with l_z may be inadvisable in some situations. Nevertheless, various guidelines exist for using l_z for informed judgment. In general, a "good" l_z is one around 0.0. A negative l_z reflects a relatively unlikely response vector (i.e., inconsistent responses), whereas a positive value indicates a comparatively more likely response vector than would be expected on the basis of the model (i.e., hyperconsistent responses). Also see

Appendix G, "The Person Response Function," for a graphical approach that can be used for detecting aberrant response vectors.

Snijders (2001) proposed an alternative to l_z that addresses the use of person estimates in its calculation. Thus, Snijders modifies l_z by incorporating a set of modification weights ($\tilde{w}_j$) in calculating l_z^*

$$l_z^* = \frac{\ln L - \mathcal{E}(\ln L) + c_{\mathrm{L}} r_0}{\sqrt{Var^*(\ln L)}}, \tag{6.12}$$

where $Var^*(\ln L) = \sum_{j=1}^{\mathrm{L}} \left\{ p_j (1-p_j) w_j^{*2} \right\}$, $w_j^* = w_j - c_n r_j$, $c_{\mathrm{L}} = \sum_{j=1}^{\mathrm{L}} p'_j w_j \Big/ \sum_{j=1}^{\mathrm{L}} p'_j r_j$

(p'_j is the first derivative of the model), $w_j = \ln[p_j/1 - p_j]$, r_j depends on the model, and r_0 depends on the ability estimation technique and model. For example, for the 1PL model $r_j = 1$, for the 2PL model $r_j = \alpha_j$, and the 3PL $r_j = [\alpha_j \exp(\theta - \delta_j)]/[\chi_j + \alpha_j \exp(\theta - \delta_j)]$ with $r_0 = 0$ for MLE and $r_0 = -\theta$ for MAP and a θ distribution that is $N(0,1)$ (Magis, Raîche, & Béland, 2012).

The R package `PerFit` (Tendeiro, Meijer, & Niessen, 2016, 2018) can be used to calculate l_z and l_z^* as well as other person fit statistics. Table 6.8 shows our R session for obtaining l_z^* for our math data calibration using `mirt`. We begin by extracting our item parameter estimates into the object `itests` and then convert our person estimates to a vector `PersonEst`. Both are passed to the `lzstar` function to calculate each person's l_z^* with the results stored in the object `lzstar_stat`. We then use the `cutoff` function to obtain a screening value for the 5% level (`Blvl = .05`); `cutoff` uses 1000 bootstraps to generate an empirical sampling distribution and determine the value that cuts off 5% of the distribution. Passing the l_z^* results (`lzstar_stat`), and the `cutoff` function's output object (`lzstarcut_05`) to the `flagged.resp` function allows us to create an object that contains those cases whose $|l_z^*|$ values exceed the absolute value of screening point (`$Cutoff` = –1.8083). For our example, 562 cases (2.87% of our sample; `Prop.flagged`) are identified as having $|l_z^*|$ values exceeding the |screening point| value.

Figure 6.12 contains the distribution of l_z^* values, with the vertical line indicating the screening value's location (–1.808) along with its confidence band (CB) on the abscissa. Thus, cases to the left of the vertical line are potentially misfitting persons. Alternatively, the lower bound of the CB could be used if one wishes to take into estimation error in identifying individuals for further examination. In this latter case, the tick marks on the top of the graph identify these potentially misfitting persons. Below we examine some cases from this distribution.

From above we know that items 1–5's P-values are 0.887, 0.644, 0.566, 0.427, and 0.387, respectively. Thus, for the first case identified (`FlaggedID` = 20–the 20th line in the data file), they incorrectly answered the easiest and hardest items as well as an item of moderate difficulty ($\underline{\mathbf{x}}$ = 01010). In contrast, the case with the `FlaggedID` of 19306 incorrectly answered the easiest item, but correctly answered the progressively more

TABLE 6.8. PerFit Session to Obtain l_z^* for the 3PL Calibration of the Mathematics Data (Prior)[a]

```
> library(PerFit)
> packageVersion("PerFit")
    [1] '1.4.3'

> itests=coef(ThreePL,simplify=TRUE,IRTpars=TRUE)$items
         [,c('a','b','g')]                                          # extract estimates
> PersonEst=as.vector(peopleThreePL[,1])
> lzstar_stat=lzstar(mathdata, IRT.PModel = "3PL",Ability=PersonEst,IP=itests)

> # determine screening value for the 5% level
> lzstarcut_05=cutoff(lzstar_stat,ModelFit="Parametric", Blvl=.05)
> FlgdCase_lzstar = flagged.resp(lzstar_stat,cutoff.obj=lzstarcut_05,scores=T)
> FlgdCases=FlgdCase_lzstar$Scores
> FlgdCase_lzstar$Cutoff
    $Cutoff
    [1] -1.8083

    $Cutoff.SE
    [1] 0.1751

    $Prop.flagged
    [1] 0.0287

    $Tail
    [1] "lower"

    $Cutoff.CI
       2.5%   97.5%
    -1.9985 -1.4599

    attr(,"class")
    [1] "PerFit.cutoff"

> FlgdCases=FlgdCase_lzstar$Scores
> head(FlgdCases,6)
         FlaggedID It1 It2 It3 It4 It5 PFscores
    [1,]        20   0   1   0   1   0  -2.4340
    [2,]        35   0   1   1   0   0  -1.7903
    [3,]        41   1   0   1   1   1  -1.6627
    [4,]        45   0   1   1   1   0  -2.9155
    [5,]        55   0   1   1   1   0  -2.9155
    [6,]       114   1   0   1   1   1  -1.6627

> tail(FlgdCases,4)
          FlaggedID It1 It2 It3 It4 It5 PFscores
    [559,]     19306   0   1   1   1   1  -4.0532
    [560,]     19354   0   1   1   0   1  -2.7950
    [561,]     19525   0   1   1   0   1  -2.7950
    [562,]     19533   0   1   1   1   0  -2.9155

> # for consistency with the above results we pass the cutoff object to the
> # plot function otherwise it will recompute the screening value
> plot(lzstar_stat,cutoff.obj=lzstarcut_05,Type="Histogram")      # produces Figure 6.12

> PRFplot(mathdata,respID=19601,IP=itests,Ability=PersonEst)      # produces Figure 6.13
    Respondent 19601: Press ENTER.

> PRFplot(mathdata,respID=19306,IP=itests,Ability=PersonEst)      # produces Figure 6.13
    Respondent 19306: Press ENTER.
```

[a]The cutoff function uses a bootstrap to determine the screening value. The bootstrap uses a random number generator. The seed used is 88888.

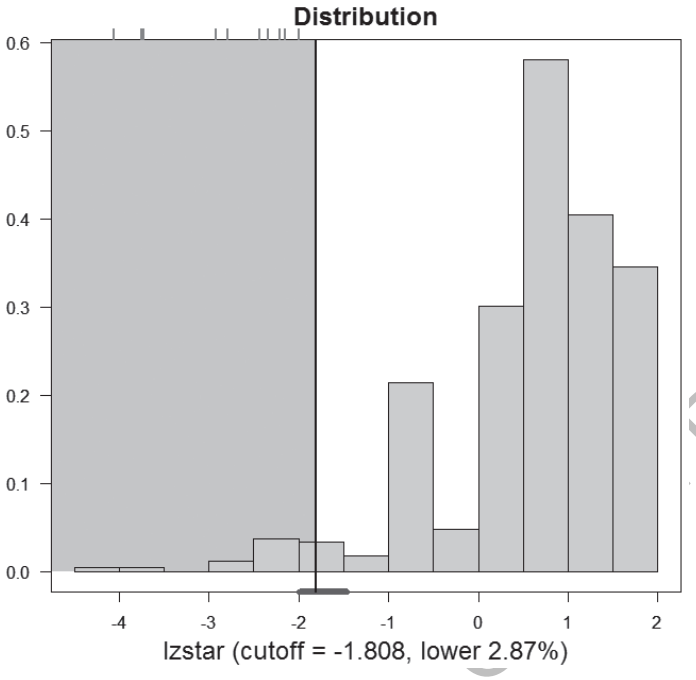

FIGURE 6.12. Distribution of person fit scores.

difficult items. Both of these individuals are not behaving consistently with expectations. As such, their $\hat{\theta}$s may be inappropriate for them. Of course, with only five items we have a small item sample and insufficient information to fully determine if these are a problem. For example, if we had 20 items, we might very well find that 19,306 would behave consistent with expectations.

With five items we can easily look at the response pattern. However, with a longer instrument a graphical approach may be more instructive. Therefore, to demonstrate a graphical approach, we use the `PRFplot` function to obtain the person response function (PRF). As discussed in Appendix G, the PRF relates the probability of a response of 1 to item location. In Figure 6.13, we show the PRFs for two persons. The left side is for a person (#19,601) who was not identified as potentially misfitting. This person's response pattern of 11110 (i.e., correctly answering all items except the most difficult) and PRF are consistent with what is expected. That is, as items become more difficult relative to this person's $\hat{\theta}$, the probability of a correct response decreases; $\hat{\theta}_{19601} = 0.720$. In contrast, person #19,306 ($\underline{\mathbf{x}} = 01111$; $\hat{\theta}_{19306} = 0.404$) has a PRF that is inconsistent with what one would expect using the 3PL model. One possible explanation of this $\underline{\mathbf{x}}$ is that this person's initial inattentiveness/carelessness may have led to their incorrect response to the first item, although they had the ability to correctly answer the easiest item. Thus, this person should have had a $\underline{\mathbf{x}} = 11111$ with a commensurate $\hat{\theta}$ of 1.233. Alternatively, it may be that the person guessed and/or copied some or all of their responses to items 2–5. In this case, their $\underline{\mathbf{x}}$ should be something along the lines of 00000, with an $\hat{\theta} = -1.396$. It is also possible that there is something unique about

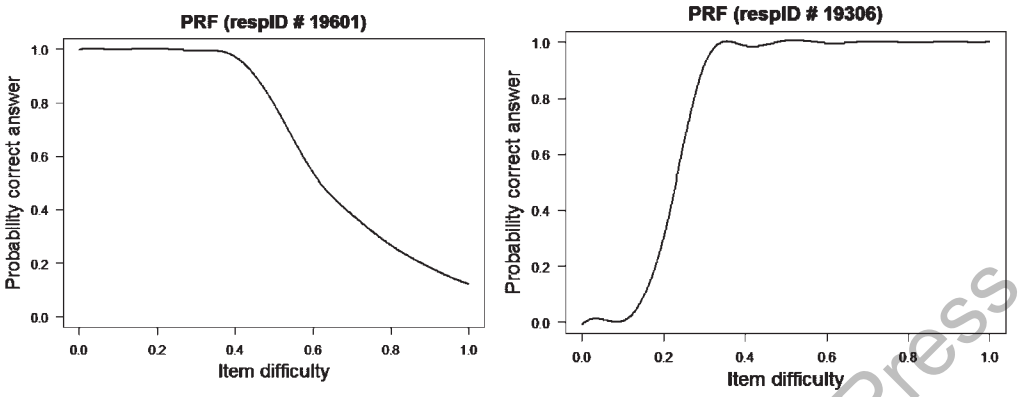

FIGURE 6.13. Person fit plots for fitting person (left) and misfitting person (right).

item 1 that led this person to respond incorrectly. In other words, the $\hat{\theta}_{19306}$ of 0.404 is an inappropriate estimate of this person's math ability. Of course, we have insufficient information to determine the cause of this person's response pattern.

Information for the Three-Parameter Model

The amount of information an item provides for estimating θ under the 3PL model is

$$I_j(\theta) = \alpha_j^2 \left[\frac{(p_j - \chi_j)^2}{(1-\chi_j)^2} \right] \left[\frac{1-p_j}{p_j} \right]. \tag{6.13}$$

Because guessing behavior reflects "noise," it may be intuited that one effect of a nonzero χ_j is to reduce the amount of information available for locating people on the θ continuum.[10,11] Equation 6.13 shows that this is indeed the case. For a given α_j and δ_j, an item provides more information for person estimation when $\chi_j = 0$ than when it is nonzero. Therefore, for the 3PL model the upper limit of $I_j(\theta)$ is given by the more restrictive 2PL model. If one sets $\chi_j = 0$ and simplifies, then Equation 6.13 reduces to Equation 5.4.

In contrast to the 1PL and 2PL models with their maximum item information at δ_j, Figure 6.5 shows that for the 3PL model the peak of the item information does not occur at δ_j but slightly above it. This offset from δ_j is given by[12]

$$\frac{\ln\left[\frac{1}{2} + \frac{\sqrt{1+8\chi_j}}{2}\right]}{\alpha_j}.$$

At this location, the maximum item information value is (Lord, 1980)

$$\text{Max}(I_j(\theta)) = \frac{\alpha_j^2}{8(1-\chi_j)^2}\left[1-20\chi_j-8\chi_j^2+(1+8\chi_j)^{1.5}\right]. \quad (6.14)$$

As has previously been the case, the total information for an instrument is the sum of the item information

$$I(\theta) = \frac{1}{\sigma_e^2(\theta)} = \sum_{j=1}^{L} I_j(\theta). \quad (6.15)$$

In the foregoing, we have focused on the amount of information an item provides for estimating a person's location.[13,14] However, we can also look at how much information the calibration sample provides for estimating a particular item parameter. The information for estimating α_j, δ_j, and χ_j is, respectively (Lord, 1980).

$$I_{\alpha_j} = \frac{1}{(1-\chi_j)^2}\sum_{i}^{N}\left((\theta_i-\delta_j)^2(p_j-\chi_j)^2\frac{(1-p_j)}{p_j}\right), \quad (6.16)$$

$$I_{\delta_j} = \frac{\alpha_j^2}{(1-\chi_j)^2}\sum_{i}^{N}\left((p_j-\chi_j)^2\frac{(1-p_j)}{p_j}\right), \quad (6.17)$$

and

$$I_{\chi_j} = \frac{1}{(1-\chi_j)^2}\sum_{i}^{N}\frac{(1-p_j)}{p_j}. \quad (6.18)$$

In Figure 6.14, we present the information function (dash line) for estimating α_5 with item 5's IRF (solid line) overlaid.[15] As can be seen, the information function is bimodal with different maxima. These different maxima reflect that we have a nonzero χ_5. As χ_j increases, the left maximum decreases and shifts its location, whereas the right maximum increases in value and stays at the same θ location. The modes are located in the θ neighborhood of the IRF beginning its trajectory toward becoming asymptotic. It is also apparent that the modes occur on opposite sides of the item's location, with the leftmost mode always less than the rightmost mode. This characteristic is a reflection of positive α_5 (i.e., if $\alpha_5 < 0$, then the leftmost mode would be greater than the rightmost mode). As α_j decreases, the distance between the modes increases, the maxima values increase, and function broadens. The location of the minimum (i.e., 0) of the information function between the two modes corresponds to δs. In other words, persons located at the item's location do not contribute information for estimating the item's discrimination.

Figure 6.15 shows that the information function for estimating δ_j is unimodal, with the mode located at the item's δ_j. Therefore, individuals around the item's location provide the greatest information for estimating δ. As is the case with Figure 6.14, the different heights of the modes across the items is a reflection of the interaction among the item's parameters as well as their different values across items. In short, poor estimation of one or more of the parameters (e.g., χ_j) affects the estimation of the item's other parameter(s).

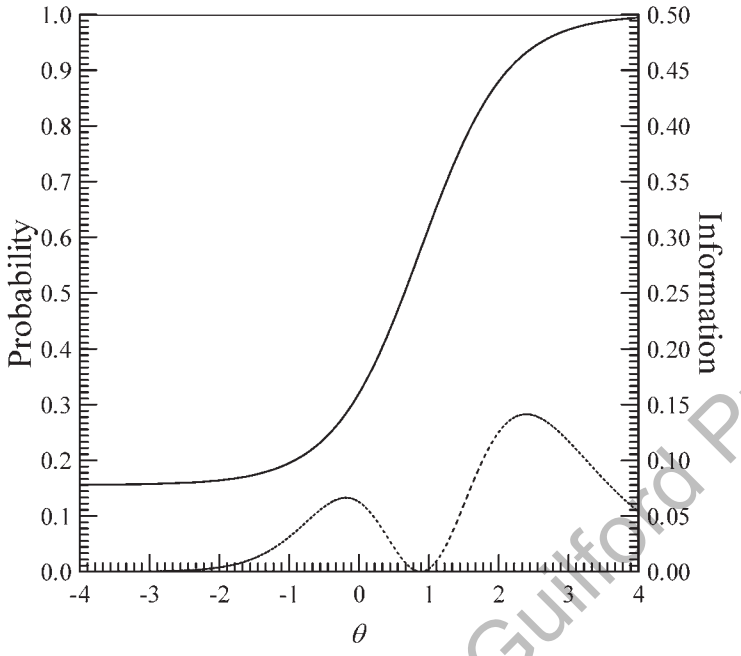

FIGURE 6.14. Information for estimating α_j as a function of θ for item 5 ($\hat{\alpha}_5 = 1.608$, $\hat{\delta}_5 = 0.883$, $\hat{\chi}_5 = 0.156$).

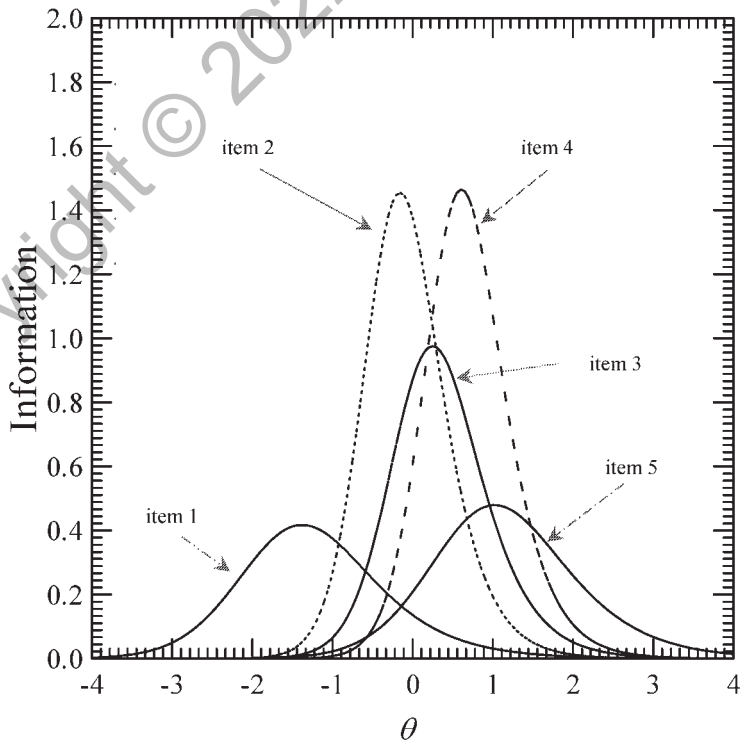

FIGURE 6.15. Information for estimating δ_j as a function of θ for each of five items.

With respect to χ_j one sees (Figure 6.16) that most information for estimating χ_j comes from the lower end of the θ continuum. Depending on the particular item, there is virtually no useful information for estimating χ_j from individuals located above 2.0. However, the information functions' plateaus show that even at the lower end of the θ continuum there is a finite amount of information available for estimating χ_j. Moreover, the larger the χ_j, the greater the shift in the beginning of this plateau toward the lower end of the continuum than when χ_j is smaller. We also see that the larger the χ_j, the lower the plateau, indicating less information for estimating these large χ_j values than for estimating smaller χ_j values.

For completeness, we now discuss the information functions for the 1PL and 2PL models. If we plot the information functions for estimating δ_j for the 1PL model, we find that across items the corresponding information functions have a constant height, with the location of the modes corresponding to the items' δ_js. In addition, the information functions for estimating a common α across items are bimodal, but unlike the 3PL model case, the functions have a constant height across modes and across items. The minima of the information functions are zero and occur between the two modes at the items' δ_js.

For the 2PL model, the information function for estimating the δ_j is also unimodal. Its height across items varies as a direct function of the items' α_js, with the location of the modes corresponding to an item's δ_j. With respect to item discrimination, the information function for estimating α_j is bimodal, with a constant height across the modes for an item and equidistant from δ_j. However, the modes vary across the items as an

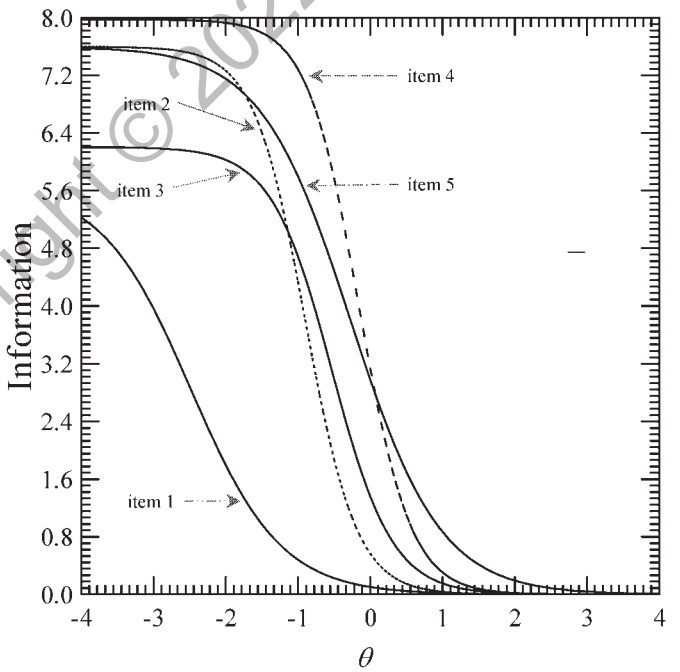

FIGURE 6.16. Information for estimating χ_j as a function of θ for each of five items.

indirect function of the items' α_js. As is the case with the 1PL and 3PL models, the location of the minimum of the information function between the two modes corresponds to the item's δ_j and has a value of 0.

Metric Transformation, 3PL Model

Linear rescaling of α_j and δ_j (or their estimates) is accomplished as performed with the 2PL model. Because the pseudo-guessing parameter is on the probability scale, it does not have an indeterminacy in its scale and there is no need to rescale χ_j. Person location parameters (or their estimates) are transformed by $\theta^* = \zeta(\theta) + \kappa$.

The total characteristic curve for the 3PL model is determined as shown, for example, in Chapter 4. As is the case with the 1PL and 2PL models, all individuals with the same location, θ, obtain the same expected trait score, T. Furthermore, neither θ nor T depends on the distribution of persons. However, unlike the case with the previous models, with the 3PL the TCC lower asymptote is asymptotic with $\sum_j \chi_j$. As an example, the expected trait score for individuals with a $\hat{\theta}$ of 1.1746 on our mathematics test would be

$$\mathrm{T} = \sum^{\mathrm{L}} p_j = 0.9927 + \ldots + 0.6813 = 4.4952.$$

Therefore, a person with an estimated location of 1.1746 would be expected to obtain almost 4.5 correct answers on the mathematics test. Figure 6.17 contains the TCC with the transformation of $\hat{\theta} = 1.1746$ to its corresponding T identified. Comparing this figure with the TCC for the 1PL model (Chapter 4, Figure 4.9) shows that it is steeper than the 1PL model's. The steepness of the TCC is a function of not only the discrimination parameter estimates (for the 3PL model the mean α is 2.3778 and for the 1PL model the common α is 1.421), but also the variability of the δ_js as well as the magnitude of the χ_js. As is seen, the lower asymptote of the TCC approaches the $\sum \chi_j = 0.889$, and its upper asymptote is the instrument's length because $\Upsilon_j = 1$ for all IRFs.[16]

Handling Missing Responses

From the preceding discussion, we know that IRT models are concerned with modeling *observed* responses. However, in working with empirical data, one will, at times, encounter situations where some items do not have responses from all individuals in the calibration sample. Some of these missing data may be considered to be missing by design or may be structurally missing. For example, one may administer an instrument to one group of people and an alternate form of the instrument to another group. If these two forms have some items in common, then the calibration sample can consist of both groups. As a result, our data contain individuals who have not responded to all items. Figure 11.1 in Chapter 11 contains a graphical depiction of this. In situations where the nonresponses are missing by design, these missing data may be ignored because of the IRT properties of person and item parameter invariance. However, when non-

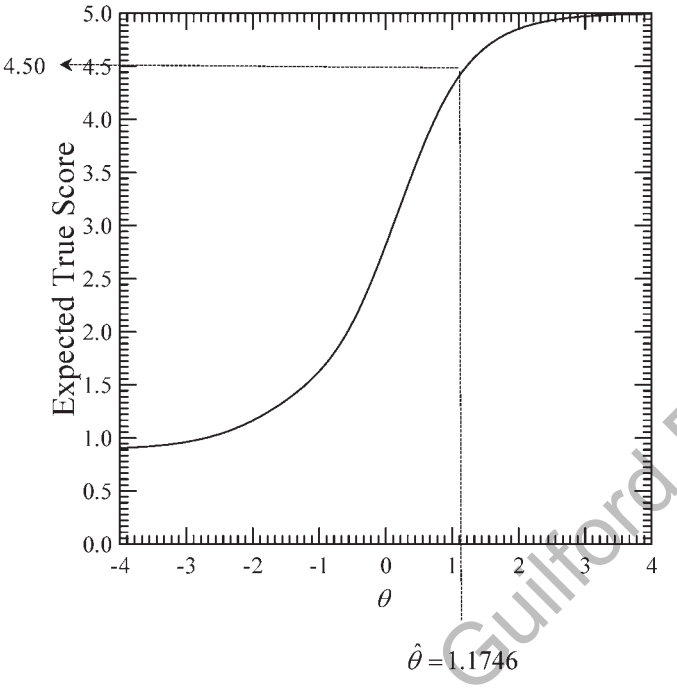

FIGURE 6.17. TCC for the five-item mathematics instrument calibrated with the 3PL model.

responses are not structurally missing, then one needs to consider how to treat these nonresponses. We begin with a brief overview of a taxonomy for missing data and then address handling missing data in the IRT context.

In general, missing data (e.g., omitted responses) may be classified in terms of the mechanism that generated the missing values. According to Little and Rubin (1987), missing data may be classified as *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR; a.k.a., NMAR: not missing at random). MCAR refers to data in which the missing values are statistically independent of the values that could have been observed, as well as other variables. In contrast, when data are MAR, then the missing values are conditionally independent of one or more variable(s). In both of these cases, the data are missing at random either unconditionally (MCAR) or conditionally on one or more variables (MAR). If the data are neither MCAR nor MAR, then the missing values are considered to be MNAR and are *nonignorable*. Nonignorable missing values are data for which the probability of omission is related to what the response would be if the person had responded.

Various approaches for handling missing data have been developed. Some of these approaches share the goal of creating "complete data," so standard analysis techniques may be applied. For instance, complete data may be created by deleting either the case that contains the missing value(s) either in its entirety or some subset of the case, or by replacing the missing value(s) by estimate(s) of what the missing value could have been. The replacement of the missing values by estimates is, in general, known as imputation. There are a number of single imputation methods (e.g., cold-deck imputation, hot-

deck imputation, mean substitution) as well as multiple imputation methods. Multiple imputation (MI) methods differ from single imputation methods by creating multiple (imputed) complete data sets to model the uncertainty in sampling from a population, whereas only one complete data set is created with single imputation. Other missing data methods are maximum likelihood-based. For greater detail, see C. H. Brown (1983); R. L. Brown (1994); Dillman, Eltinge, Groves, and Little (2002); Enders (2001, 2003); and Roth (1994).

Returning to the IRT context, there are various reasons why an individual's response vector may not contain responses to each item. We present three conditions that lead to missing data. The first condition is mentioned above. In the *missing by design* case (e.g., *not-presented* items), such as in adaptive testing (Appendix D) or simultaneous calibration (see Chapter 11), the nonresponses represent conditions in which the missingness process may be ignored for purposes of person location estimation (Mislevy & Wu, 1988, 1996). Therefore, the estimation is based only on the observed responses.

A second situation that produces missing data occurs when an individual has insufficient time to answer the item(s). These *not-reached* items are (typically) identified as collectively occurring at the end of an instrument (this assumes the individual responds to the test items in a serial fashion) and represent *speededness*. (Of course, the absence of not-reached items does not mean that speededness did not occur because respondents may randomly guess on items.) Although IRT should be applied to unspeeded tests, Lord (1980) stated that if we knew which items the examinee did not have time to consider, then these not-reached items may be ignored for person location estimation because they contain no readily quantifiable information about the individual's location (e.g., their proficiency). Therefore, when one has (some) missing data due to not-reached items, then the person's location is estimated using only the observed responses. However, this should not be interpreted as indicating that one should apply IRT to speeded instruments nor that these not-reached items are unaffected by being speeded. Speeded situations may lead to violation of the unidimensionality assumption and biased item parameter estimates. Research has shown that the speeded items' α_js and δ_js are overestimated and the χ_js underestimated (Oshima, 1994). Because of the overestimation of α_j, the corresponding item information and, therefore, the instrument's total information becomes inflated. Identifying the speeded items as not-reached within BILOG mitigates the bias in item parameter estimation. (See Goegebeur, De Boeck, Wollack, and Cohen [2008] for a gradual process change model that models speededness as a person-specific effect.)

The third situation that produces missing data occurs when an examinee intentionally chooses to not respond to a question for which they do not know the answer. These *omitted responses* represent nonignorable missing data (Lord, 1980; Mislevy & Wu, 1988, 1996). Again, assuming that an individual responds in a serial fashion to an instrument, omitted responses may be distinguished from not-reached items because omits appear throughout the response vector and not just at the end of the vector. Lord (1980) has argued that omitted responses should not be ignored because an individual could obtain as high a proficiency estimate as they wished by simply answering only

those items they had confidence in answering correctly. This idea has found some support in Wang, Wainer, and Thissen's (1995) study on examinee item choice.

The effect of omitted responses on EAP person location estimates has been studied (de Ayala, Plake, & Impara, 2001; de Ayala, 2006; Finch, 2008; Glas & Pimentel, 2008; Rose, von Davier, & Xu, 2010). Results show that for dichotomous data, omits should not be treated as incorrect, nor should they be ignored; also see Lord (1974a, 1983c). However, using a fractional value of 0.5 in place of omitted values leads to improved person location estimation, compared with treating the omits as incorrect or using a fractional value equal to the reciprocal of the number of item options (i.e., $1/m$ where m is the number of response categories). (The $1/m$ approach assumes that an individual responds randomly to a multiple-choice item format and was suggested by Lord [1974a, 1980].) The results also seem to indicate that this would be true for MLE person location estimation. By using this fractional value, one is simply imputing a response for a binomial variable and thereby "smoothing" irregularities in the likelihood function. Although this research was conducted using the 3PL model, it appears that the results would apply to both the 1PL and 2PL models.

An alternative approach that may be fruitful in some situations is to treat omission as its own response category and apply a polytomous model such as the multiple-choice model or the nominal response model; both models are discussed in Chapter 9. Additionally, Holman and Glas (2005) present a "multiple" model approach that uses an IRT model to model the missing-data process and an IRT model for the observations. Missingness can also be addressed using MI. Several MI routines are available, including SAS `proc mi`, SPSS's `multiple imputation` (or `EM` from `missing value analysis`) (SPSS Incorporated, 2019), `Missing Value Analysis` (SYSTAT, 2017), or, for example, the `R` package `mice` (van Buuren & Groothuis-Oudshoorn, 2011, 2019). These routines assume that missing data are MAR. After imputation of omitted responses these complete data may then be calibrated.

The practitioner should be aware of several issues in the treatment of omits. For instance, in the context of proficiency assessment, all imputation procedures that produce complete data for analysis are, in effect, giving partial credit for an omitted response. For example, Lord's (1974a, 1980) suggested use of $1/m$ gives an individual partial credit worth, say 0.2 (i.e., $m = 5$), for having omitted an item. A second issue to be aware of is that using the same imputed value for all omits assumes that individuals located at different points can all be treated the same. These issues are raised so that the practitioner understands the assumptions that are being made with some of the missing data approaches discussed. However, these may or may not be of concern to a particular practitioner. For example, when IRT is used in personality testing or with attitude or interest inventories, these may be nonissues. A third issue to be noted is that omits tend to be associated with personality characteristics, demographic variables, and proficiency level (Mislevy & Wu, 1988; Stocking, Eignor, & Cook, 1988). Thus, in those situations where information on these variables is available, one may wish to use this information as covariates in the imputation process. Use of these covariate(s) may or may not have any meaningful impact on the person location estimates.

When calibrating a data set, it is good practice to identify items without responses

by some code. For instance, in the data file, not-reached items may be identified by a code of, say, 9, not-presented items by a code of 8, omitted items by a code of 7. With certain calibration programs (e.g., BILOG-MG), any ASCII character may be used (e.g., the letters "R" for not-reached, "P" for not-presented, and "O" for omit). In these cases, the code used must be identified for the program. With BILOG one would use the `KFName`, `NFName`, and/or `OFName` subcommands on the `GLOBAL` or `INPUT` command line, depending on the version of BILOG one is using. For BILOG, omitted responses must be identified as such, whereas with other programs any response code encountered in the data file that is not identified as a valid response is considered to reflect an omitted item. Omitted responses that have been identified by an omitted response code are, by default, treated as incorrect by BILOG.

Issues to Consider in Selecting among the 1PL, 2PL, and 3PL Models

The issues to be considered in selecting among the 1PL, 2PL, and 3PL models involve, in part, one's philosophy of whether the data should fit the model or vice versa (see Chapter 2), as well as the application context (e.g., sample size, instrument characteristics and considerations, assumption tenability, political realities). Given that the 1PL model is the most restrictive of the three models, there have been a number of studies that have investigated use of the 1PL model when it misfits. For instance, Forsyth, Saisangjan, and Gilmer (1981) investigated the robustness of the Rasch model when the dimensionality and constant α assumptions are violated. Because their empirical data came from an examination using a multiple-choice item format, it was assumed that some examinees would engage in guessing. Forsyth et al. concluded that "the Rasch model does yield reasonably invariant item parameter and ability estimates . . . even though the assumptions of the model are not met" (p. 185). Similar results were obtained by Dinero and Haertel (1977) using simulation data.

Wainer and Wright (1980) stated, "It seems that the Rasch model yields rather good estimates of ability and difficulty even when its assumption of equal slopes is only roughly approximated" (p. 373). Furthermore, Lord and Novick (1968) stated, "It appears that if the number of items is very large, then inferences about an examinee's ability based on his total test score will be very much the same whether" (p. 492) the Rasch model or the 3PL model is used. In this regard, recall that for the mathematics data example the Pearson correlation between the $\hat{\theta}$s based on the 1PL and the 3PL models' $\hat{\theta}$s for the example's data is 0.9764. For the other model combinations, we have a correlation of 0.9907 for the 1PL and the 2PL models' $\hat{\theta}$s, and for the 2PL and the 3PL models' $\hat{\theta}$s the correlation is 0.9859. Although these are all reasonably strong correlations, the correlations among the standard errors, $s_e(\hat{\theta})$s, for the various model combinations paint a different picture. The correlation between the 1PL model estimated standard errors and those of the 2PL model is 0.9721, between the 1PL model and the 3PL model the correlation is 0.3318, and for the 2PL and the 3PL models' estimated standard errors it is 0.2275. Therefore, in situations where confidence bands about $\hat{\theta}$ are used for

classification decisions, the same individual would be classified differently depending on the model used. Presumably, using longer instruments would allow for greater agreement among the standard errors. Moreover, the magnitude of the correlations between the 1PL, 2PL, and 3PL models' $\hat{\theta}$s would be affected by the correlation between α_j and δ_j (Yen, 1981).

For samples of 200 or fewer, Lord (1983a) found that the Rasch model was slightly superior to the 2PL model in terms of person estimation. As previously mentioned, Thissen and Wainer (1982) studied the asymptotic standard errors of the one-, two-, and three-parameter models. They suggested fitting the 1PL model first and examining its model–data fit. If only a few items misfit and they could be omitted without adversely affecting the instrument (e.g., the validity of the $\hat{\theta}$s), then one should consider removing them. However, if the omission of these misfitting items is problematic, then one should increase the sample size and try to fit the 2PL model (presumably the item[s] misfit is due to varying item discrimination). In contrast, Gustafsson (1980) suggested grouping the items into homogeneous subsets rather than removing them from the instrument. For instance, looking at the mathematics 2PL model calibration example, we see that in terms of the $\hat{\alpha}_j$s there are three groupings of items. Items 3 and 4 are very similar in terms of their $\hat{\alpha}_j$s, items 1 and 5 are somewhat similar to one another, and item 2 is substantially different from the other four items. Therefore, three subsets could be created for the mathematics data example. Assuming item misfit is due to varying item discrimination, we can alternatively use the OPLM model approach in which the item locations are estimated but the item discrimination(s) are imputed (Verhelst & Glas, 1995; Verhelst et al., 1995). The use of mixture models (see Appendix F, "Mixture Models"), as well as some of the models presented in von Davier and Carstensen (2007), may also provide additional solutions. (It should be recalled that the desirable properties of the Rasch model [e.g., specific objectivity] hold only when one has model–data fit.)

Yen (1981) advocates a process of first fitting all three models (i.e., 1PL, 2PL, 3PL) to the empirical data set of interest. Subsequently, simulation data sets are generated based on item parameter distributions that are similar to those found with the calibration of the empirical data set. For example, we would generate a data set using the 1PL model, another with the 2PL model, and so on. The final step involves comparing the fit analyses across models in conjunction with the fit analysis of the empirical data to facilitate model selection.

In a simulation study, Yen (1981) generated different data sets based on various models and compared the fit of the 1PL, 2PL, and 3PL models to these data. When she used the 3PL model for data generation, she found that the 2PL model fitted the data almost as well as the 3PL model did, although the item parameters estimates were not the same across the two models. She noted that when an item was difficult and had a moderate to high discrimination, it was difficult for the 2PL to model a nonzero lower asymptote. She concluded that although the 2PL model performed almost as well as the 3PL model in modeling the response vectors, one might observe sample dependency when difficult items have their discrimination parameters estimated with low-proficiency-level examinees.

As may be inferred from the above, there are variants of the dichotomous models.

For instance, it is possible to constrain the 3PL model to produce modified versions, such as constraining the α_js to a constant value as well as the χ_js to a nonzero value. This model is sometimes referred to as a modified 1PL model (Cressie & Holland, 1983; also see Kubinger & Draxler, 2007). Furthermore, one may use the 3PL model with the χ_js for certain problematic items fixed to a constant nonzero value, whereas χ_j is estimated for other items. In general, for those situations where one is not holding χ_js fixed, it would be prudent (as done above) to use a prior distribution on the χ_js when estimating the lower asymptotes. In addition, with some data, one may obtain unreasonably large estimates of α_j (e.g., greater than 3). For these situations, use of a prior distribution on the α_js may be in order.

As discussed in this chapter and the preceding chapters, it is the validity of the person location estimates that is paramount. From a pragmatic perspective, if convincing validity evidence can be accrued for person location estimates using a particular model in a particular application, then it would seem that the above arguments, though interesting in their own right, are somewhat irrelevant.[18]

Summary

The 3PL model attempts to obtain useful information from a response pattern over and above that contained in the response vector's observed score. To achieve this objective, the 3PL model consists of parameters that reflect the item's location and discrimination as well as the lower asymptote of the IRF. As is true with the 2PL model's IRFs, the 3PL model's IRFs may potentially cross because the 3PL model allows for varying discrimination. With the 3PL model, item discrimination is proportional to the slope of the IRF at the point of inflexion and is equal to $0.25\alpha_j(1 - \chi_j)$. In addition, the 3PL model's IRFs may cross because the model allows for the lower asymptote parameters, χ_js, to vary across items. The lower asymptote parameter is restricted to the range 0 to 1 (inclusive) and reflects the probability of obtaining a response of 1 by individuals who are extremely low on the latent variable continuum. The lower asymptote parameter is typically referred to as the pseudo-guessing parameter.

In previous chapters, fit is investigated in terms of item statistics, empirical and predicted IRFs, and examination of the invariance of item parameter estimates across random calibration subsamples. In this chapter, we also used ΔR^2_Δ and ΔG^2 for assessing relative model–data fit. Moreover, we introduced an appropriateness index to gauge person fit and the Q^P_3 and Q_3 statistics for assessing the tenability of the conditional independence assumption. The Q^P_3 and Q_3 statistics may be useful for identifying sets of items that are exhibiting item dependence. When items are found to be interdependent, it may make sense to bundle them together and obtain an item score for the item parcel. The resulting item score is polytomous and ordinal in nature (i.e., a larger item score reflects more of the latent variable than does a smaller value). The analysis of these data can be accomplished through a polytomous model.

Chapter 7 introduces polytomous models that are derived from the Rasch model. These models, the partial credit and rating scale models, are appropriate for ordinal

polytomous data. These models assume that an instrument's items are equally effective in discriminating among individuals. As the models' names imply, the partial credit model can be used with data that reflect degrees of response correctness, whereas the rating scale model can be used with data from response formats, such as the Likert or summated ratings format. In actuality, both models are applicable to data that reflect degrees of response endorsement, but they differ from one another in their respective simplifying assumptions. In Chapter 8, use of polytomous models for ordinal data continues, but with models that are not based on the Rasch model.

Notes

1. Although the three-parameter model allows for the possibility that the lower asymptote is nonzero, the upper asymptote is still 1.0. That is, as θ approaches positive infinity, the probability of a response of 1 is 1.0 or, symbolically, $p(x = 1|\theta \rightarrow \infty) \rightarrow 1$. An alternative model, the *four-parameter logistic* model (Barton & Lord, 1981), extends the three-parameter model to allow for the possibility that persons with very large θs may still not have a success probability equal to 1 (see McDonald, 1967). The motivation behind the model's development was to improve person location estimation. For instance, if a person with a very large θ makes a clerical error on an easy item, then their estimate would be more drastically lowered using a model with an upper asymptote of 1 than when this asymptote was less than 1 (Barton & Lord, 1981). To address this situation, Barton and Lord (1981) introduced a parameter that reflected the IRF's upper asymptote (Υ_j) into the 3PL model. As a consequence, as θ goes to ∞ the probability of a response of 1 is Υ_j or, symbolically, $p(x = 1|\theta \rightarrow \infty) \rightarrow \Upsilon$. The *four-parameter logistic* (4PL) model is

$$p(x_j = 1|\theta, \alpha_j, \delta_j, \chi_j, \Upsilon_j) = \chi_j + (\Upsilon_j - \chi_j)\frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}} \tag{6.19}$$

   Barton and Lord (1981) compared the model in Equation 6.19 to the 3PL model using empirical data. They found that the 3PL model did as well or better than the 4PL model. Barton and Lord concluded that "there is no compelling reason to urge the use of this <4PL> model" (p. 6). However, it should be noted that although the α_js, δ_js, and χ_js were estimated (using JMLE), the Υ_js were *not* estimated. Rather, the Υ_js were held fixed at either 0.98 or 0.99. Given the study's design decisions, it is doubtful that this one study should be considered definitive. In contrast, Loken and Rulison (2010) using WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2004) obtained promising estimation results in the estimation of all four item parameters. `mirt` and SAS `proc irt` can be used to estimate the 4PL model.

2. The first derivative of the 3PL model is

$$p'_j = \alpha_j(1-p_j)\frac{(p_j-\chi_j)}{(1-\chi_j)},$$

where

$$(1-p_j)=1-\left[\chi_j+(1-\chi_j)\frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}}\right]=1-\left[\chi_j+\frac{(1-\chi_j)}{1+e^{-\alpha_j(\theta-\delta_j)}}\right].$$

Because by definition α_j is defined at $\theta = \delta_j$, p_j simplifies to

$$p_j=\chi_j+(1-\chi_j)\frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}}=\chi_j+\frac{(1-\chi_j)}{1+e^0}=\chi_j+\frac{(1-\chi_j)}{2}=\frac{2\chi_j+1-\chi_j}{2}=\frac{1+\chi_j}{2}$$

and $(1 - p_j)$ simplifies to

$$(1-p_j)=1-\left[\chi_j+\frac{1-\chi_j}{1+e^{-\alpha_j(\theta-\delta_j)}}\right]=1-\frac{1+\chi_j}{2}=\frac{2-(1+\chi_j)}{2}=\frac{1-\chi_j}{2}.$$

By substitution for p_j in p'_j we obtain

$$p'_j=\alpha_j(1-p_j)\frac{(p_j-\chi_j)}{(1-\chi_j)}=\alpha_j\left[\frac{1-\chi_j}{2}\right]\left[\frac{\left(\frac{1+\chi_j}{2}-\chi_j\right)}{1-\chi_j}\right]=\alpha_j\left[\frac{\left(\frac{1+\chi_j}{2}-\chi_j\right)}{2}\right]=0.25\alpha_j(1-\chi_j)$$

When the *D* scaling constant is used, then the slope for the 3PL model is

$$0.25D\alpha_j(1-\chi_j)=0.425\alpha_j(1-\chi_j).$$

3. For example, assume we have a two-item instrument with $\alpha_1 = 2.0, \delta_1 = 0.0, \chi_1 = 0.25$ for the first item and $\alpha_2 = 1.0, \delta_2 = -0.5, \chi_2 = 0.0$ for the second item. According to the 3PL model, a person with the response vector $\underline{\mathbf{x}} = 01$ will have an $\hat{\theta}$ of –0.55. However, if we use the 2PL model (i.e., $\chi_1 = \chi_2 = 0.0$), then our $\hat{\theta}$ is –0.1558. For the Rasch model (i.e., $\alpha_1 = \alpha_2 = 1.0$ and $\chi_1 = \chi_2 = 0.0$), our $\hat{\theta}$ is approximately –0.25. Comparing these $\hat{\theta}$s shows that one effect of including a nonzero χ_j in our model is to reduce the $\hat{\theta}$s relative to not including χ_j.
4. Some users of the Rasch model have argued that the item discrimination parameter cannot be estimated as is done with the 2PL and 3PL models (e.g., see Wright, 1977b). According to Gustafsson (1980), when one has unequal discriminations, the item locations are related to the calibration sample's characteristics on the latent variable (e.g., a high- or low-proficiency group). In fact, he states that "it is difficult to make a distinction between the assumption of unidimensionality and the assumption of homogeneous item discrimination" (p. 208). Lumsden (1978) expresses a similar opinion: "Test scaling methods are self-contradictory if they assert both unidimensionality and different slopes for the ICC. . . . If the unidimensionality requirement is met, the Rasch (1960) one-parameter model will be realized" (p. 22). (Lumsden also suggested abandoning the two- and three-parameter normal ogives.) Gustafsson (1980) suggests that it may be prudent to investigate the robustness of the Rasch model in the face of varying item discriminations for specific applications.

5. According to Holland (1990a), there can be at most two parameters per item, and "models that contain three or more parameters per item can only estimate these parameters successfully for one of two reasons; either they are not applied to a large enough item set or the test is not unidimensional" (p. 17); also see Cressie and Holland (1983) and Holland (1990b). As such, there appear to be more parameters in the 3PL model than can be supported by a unidimensional test.

6. As is true with the two-parameter model, JMLE no longer seems to be used for parameter estimation with the three-parameter model. However, for completeness, we describe some of the past research in this area. The Hulin et al. (1982) study of JMLE presented in Chapter 5 also examined parameter estimation accuracy for two models (2PL, 3PL); this study had the additional factors of sample sizes (200, 500, 1000, 2000) and instrument length (15, 30, 60 items). They found that for a given condition the 2PL model results were better than those for the 3PL. However, for both models, and not surprisingly, the larger the sample size and the longer the instrument, the more accurate the estimates. In addition, the average error (i.e., root mean squared) in recovering the true IRFs for both models and using at least 30 items was no greater than 0.05 for a sample size of 1000 and less than 0.07 with 500 cases. In general, increasing the instrument's length for a given sample size resulted in more accurate estimates.

   Skaggs and Stevenson (1989) report a similar finding using LOGIST. They also found that the average error in recovering the true IRFs for the 15-item instrument was about 0.07, and for the 30-item length the average error was slightly below 0.055 when using a sample size of 500. These average errors decreased to about 0.05 and about 0.037 for the 15- and 30-item lengths, respectively, when the sample was quadrupled to 2000 cases. Lord (1968) suggests that the sample size be greater than 1000 and that instruments be at least 50 items long when using LOGIST. However, Swaminathan and Gifford (1983) found that reasonably good estimates can be obtained with a 1000-person sample and a 20-item instrument. Therefore, it appears that samples of a 1000 or more with instruments of at least 20 items, and preferably longer, should be used with JMLE as implemented in LOGIST. However, work by Thissen and Wainer (1982) calls this sample size suggestion into question. For example, applying their observations to the 3PL model for an item with α_j of 1.5, $\delta_j = 2$ (or $\delta_j = -2$), and $\chi_j = 0.1$ would require 97,220, 22,142, and 46,743 individuals to estimate the item's δ_j, α_j, and χ_j, respectively, with an accuracy of one-tenth. Therefore, the calibration sample size would be 97,220.

7. The testlet model is equivalent to a second-order model or a restricted bifactor model (Li, Bolt, & Fu, 2006; Rijmen, 2010).

8. Although the same calibration sample is used for the 1PL, 2PL, and 3PL model calibrations, the different models produced different estimates. The mean item location estimate for the 1PL, 2PL, and 3PL models are –0.403, –0.400, and 0.036, respectively. Moreover, the mean discrimination estimate of 2.342 for the 3PL model is substantially greater than the common $\hat{\alpha} = 1.421$ found with the 1PL model or the 2PL model's mean discrimination estimate of 1.459. This is due to the nonzero lower

asymptote as well as to differences in metrics. With respect to the former explanation, we see from a comparison of the 2PL and 3PL models' $\hat{\alpha}_j$s that the 2PL model accommodates the nonzero asymptote by decreasing $\hat{\alpha}$ relative to what is obtained when we estimate the lower asymptote; for the 2PL model $\hat{\alpha}_1$ is 1.226 and for the 3PL model $\hat{\alpha}_1$ = 1.921. In fact, for all the items the 2PL model's $\hat{\alpha}_j$s are less than the corresponding 3PL model's $\hat{\alpha}_j$s. These lower 2PL model $\hat{\alpha}_j$s are associated with a metric that, relative to the 3PL model's $\hat{\alpha}_j$s, is stretched out and located lower than that of the 3PL model. In short, we have different metrics for the different model calibrations of the data. As such, the differences in the estimates across models for corresponding item parameters are partly due to a difference in metrics. Therefore, strictly speaking, we need to link the various metrics before directly comparing individual item parameter estimates across models.

9. The screening value of –0.2935 obtained in Appendix G, "Conditional Independence using Q_3," can be used for evaluating Q_3^P. As mentioned in Appendix G, the generated data are conditionally independent. Theoretically, when the data are conditionally independent, there is no linearity in the residuals to partial out and the zero-order correlation Q_3 is equivalent to Q_3^P. Because our generated data contain a random error component, it is possible that the intercorrelations among one or more item pairs will not be equal to zero but will be very close to zero; sample size will also affect the equivalence of the partial and zero-order correlations. However, any difference from zero will not be meaningful and should not affect our conclusions. Consequently, we see that the item pairs identified using the gap approach (i.e., 2–3, 3–4, and 2–4) are also identified using the screening value.

10. A complementary approach for determining the *df* for evaluating ΔG^2 is to use the difference in the model's *df*s. The *df* for a model is given by 2^L – (number of item parameters) – 1, where L is the number of items on the instrument and the number of item parameters is based on the model and the number of items. For example, for the 3PL model there are three item parameters (α_j, δ_j, and χ_j), and for a, say, five-item instrument the number of items parameter is 3 × 5 = 15. Therefore, for the 3PL model the *df* = 32 – 15 – 1 = 16. For the 2PL model there are two item parameters (α_j and δ_j), and with a five-item instrument the *df* = 32 – 10 – 1 = 21. With the 1PL model, each item has a location (δ_j) and all items have a common α. Therefore, with a five-item instrument there are six parameters that are estimated and the model's *df* = 32 – 6 – 1 = 25. With BILOG, if one uses the keyword `RASch`, the program performs a 1PL estimation and then rescales the common α to be 1 and adjusts all the δ_js accordingly; how this is done is demonstrated in Chapter 4. Therefore, with BILOG there are six, not five, item parameters estimated with the Rasch model. In contrast, a program like BIGSTEPS (or WINSTEPS) does not estimate a common α, and, as a result, there are only five δ_js estimated; that is, the *df* = 32 – 5 – 1 = 26.

11. From Lord and Novick (1968) we have that the area under the item information function is

$$\int_{-\infty}^{\infty} I_j(\theta) = \alpha_j \frac{\chi_j \ln(\chi_j) + 1 - \chi_j}{1 - \chi_j}.$$

With the use of the *D* scaling constant in the 3PL model, the item information is

$$I_j(\theta) = \frac{D^2 \alpha^2 (1 - p_j)(p_j - \chi_j)^2}{(1 - \chi_j)^2 p_j} \tag{6.20}$$

and the corresponding area under the item information function is equal to

$$\int_{-\infty}^{\infty} I_j(\theta) = D\alpha_j \frac{\chi_j \ln(\chi_j) + 1 - \chi_j}{1 - \chi_j}.$$

12. To determine where an item has its maximum information, recall that α_j is proportional to the slope of the IRF at δ_j (i.e., the slope at δ_j is $0.25\alpha_j(1 - \chi_j)$). The offset from δ_j to where an item has its maximum information is obtained from the item information equation. By substitution into Equation 6.13 and rearranging terms, we have

$$I_j(\theta) = \alpha_j^2 \frac{e^{-\alpha_j(\theta - \delta_j)}}{1 + e^{-\alpha_j(\theta - \delta_j)}} (1 - \chi_j) \frac{e^{\alpha_j(\theta - \delta_j) - \ln(\chi_j)}}{1 + e^{\alpha_j(\theta - \delta_j) - \ln(\chi_j)}}.$$

Following Lord and Novick (1968) and maximizing $I_j(\theta)$ with respect to $\alpha_j(\theta - \delta_j)$ leads to

$$\frac{\partial}{\partial \alpha(\theta - \delta)} \ln I_j(\theta) = \frac{\partial}{\partial \alpha(\theta - \delta)} \left[\ln \left(\frac{e^{-\alpha_j(\theta - \delta_j)}}{1 + e^{-\alpha_j(\theta - \delta_j)}} \right) + \ln \left(\frac{e^{\alpha_j(\theta - \delta_j) - \ln(\chi_j)}}{1 + e^{\alpha_j(\theta - \delta_j) - \ln(\chi_j)}} \right) \right]$$

$$= 2 \left(\frac{e^{-\alpha_j(\theta - \delta_j)}}{1 + e^{-\alpha_j(\theta - \delta_j)}} \right) - 1 + \left(\frac{e^{-\alpha_j(\theta - \delta_j) + \ln(\chi_j)}}{1 + e^{-\alpha_j(\theta - \delta_j) - \ln(\chi_j)}} \right)$$

$$= \frac{2}{1 + e^{-\alpha_j(\theta - \delta_j)}} - 1 + \frac{1}{1 + e^{-\alpha_j(\theta - \delta_j)/\chi_j}}$$

$$= \frac{1 - e^{\alpha_j(\theta - \delta_j)}}{1 + e^{\alpha_j(\theta - \delta_j)}} + \frac{\chi_j}{\chi_j + e^{\alpha_j(\theta - \delta_j)}}$$

$$= \frac{2\chi_j + e^{\alpha_j(\theta - \delta_j)} - e^{2\alpha_j(\theta - \delta_j)}}{(\chi_j + e^{\alpha_j(\theta - \delta_j)})(1 + e^{\alpha_j(\theta - \delta_j)})}.$$

To find the maximum of this last equation, its derivative is set to 0 and we solve for $\alpha_j(\theta - \delta_j)$

$$\frac{2\chi_j + e^{\alpha_j(\theta-\delta_j)} - e^{2\alpha_j(\theta-\delta_j)}}{(\chi_j + e^{\alpha_j(\theta-\delta_j)})(1+e^{\alpha_j(\theta-\delta_j)})} = 0.$$

Because this equation is equal to 0.0, when its numerator equals zero we only need to be concerned with the numerator

$$2\chi_j + e^{\alpha_j(\theta-\delta_j)} - e^{2\alpha_j(\theta-\delta_j)} = 0.$$

This last equation is in the form of a quadratic (i.e., $f(x) = ax^2 + bx + c$, where a, b, and c are real constants and $x = e^t$; therefore, $2c + e^t + e^{2t}$). We can solve this last equation by using the quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

with $a = -1$, $b = 1$, and $c = 2\chi_j$. Because in this case, $a < 0$, we have two solutions: $1 + 4(2c) > 0$ and $-1/2(-1) = 0.5$. Using the quadratic formula, we obtain by substituting the values for a, b, and c

$$x = \frac{-1 \pm \sqrt{(-1)^2 - 4(-1)(2c)}}{2(-1)} = \frac{-1 \pm \sqrt{1+8c}}{-2}.$$

The solutions are

$$x = \frac{-1+\sqrt{1+8c}}{-2} = \frac{1}{2} - \frac{\sqrt{1+8c}}{2} \text{ and } x = \frac{-1-\sqrt{1+8c}}{-2} = \frac{1}{2} + \frac{\sqrt{1+8c}}{2}$$

We can eliminate

$$\frac{1}{2} - \frac{\sqrt{1+8c}}{2}$$

because it leads to having to take the log of a negative number. Therefore, we have $x = e^{\alpha_j(\theta-\delta_j)}$ and $\ln(x) = \alpha_j(\theta - \delta_j)$. By substitution

$$\ln\left(\frac{1}{2} + \frac{\sqrt{1+8c}}{2}\right) = \alpha_j(\theta - \delta_j) \quad \frac{\ln\left(\frac{1+\sqrt{1+8c}}{2}\right)}{\alpha_j} = (\theta - \delta_j)$$

The item has the location of its maximum information at

$$\frac{\ln\left(\frac{1+\sqrt{1+8\chi_j}}{2}\right)}{\alpha_j} + \delta_j$$

and the offset is

$$\frac{\ln\left(\frac{1+\sqrt{1+8\chi_j}}{2}\right)}{\alpha_j}$$

That is, an item provides its maximum information at a location slightly higher than its δ_j. When $\chi_j = 0$, the offset equals 0.

13. The standard error for the person location estimate under the 3PL is

$$s_e(\hat{\theta}_i) = \sqrt{\sum_{j=1}^{L}\left[\frac{p_j(1-\chi_j)^2}{\alpha_j^2(1-p_j)p_j(1-\chi_j)^2}\right]} \tag{6.21}$$

where p_j is conditional on $\hat{\theta}_i$.

14. As mentioned in Chapter 5, the maximum information attainable by any scoring method is given by the total information function. Therefore, the optimal scoring weight for an item j is given by Equation 5.20

$$w_j(\theta) = \frac{p_j'}{p_j(1-p_j)}$$

Given that the first derivative for the 3PL model is

$$p_j' = \frac{\alpha_j(p_j-\chi_j)(1-p_j)}{(1-\chi_j)} \tag{6.22}$$

we have by substitution of Equation 6.22 into Equation 5.20 that the optimal scoring weight for the 3PL model is (Lord, 1980)

$$w_j(\theta) = \frac{\alpha_j(p_j-\chi_j)}{p_j(1-\chi_j)} = \frac{\alpha_j}{1+\chi_j e^{-\alpha_j(\theta-\delta_j)}}. \tag{6.23}$$

Therefore, the optimal weight is a function of not only the item parameters, but also the person's location. As a result, with the 3PL model it is not possible to know the optimal scoring weight for an individual. Equation 6.23 shows that when $\chi_j = 0$, then $w_j(\theta) = \alpha_j$. Similarly, whenever θ is very large (i.e., $\theta \to \infty$), then the item's optimal weight approaches its discrimination (i.e., $w_j(\theta) \to \alpha_j$). In contrast, whenever θ is very small (i.e., $\theta \to -\infty$), then $p_j \to \chi_j$ and $w_j(\theta) \to 0$. In this latter condition, the respondent's location makes the item ineffective. With the scaling constant, D, Equation 6.23 becomes

$$w_j(\theta) = \frac{D\alpha_j}{1+\chi_j e^{-D\alpha_j(\theta-\delta_j)}}$$

15. Equations 6.16–6.18 and the following equations are for maximum likelihood estimation. In addition to the information for each item parameter, we have information for the interrelationships among α_j, δ_j, and χ_j. Following Lord (1980) we have

$$I_{\alpha\delta_{(j)}} = \frac{\alpha_j}{(1-\chi_j)^2}\sum_i^N\left((\theta_i-\delta_j)(p_j-\chi_j)^2\frac{(1-p_j)}{p_j}\right), \tag{6.24}$$

$$I_{\alpha\chi_{(j)}} = \frac{1}{(1-\chi_j)^2}\sum_i^N\left((\theta_i-\delta_j)(p_j-\chi_j)^2\frac{(1-p_j)}{p_j}\right), \tag{6.25}$$

and

$$I_{\delta\chi_{(j)}} = \frac{\alpha_j}{(1-\chi_j)^2}\sum_i^N\left((p_j-\chi_j)^2\frac{(1-p_j)}{p_j}\right). \tag{6.26}$$

Collectively, Equations 6.16–6.18 and 6.24–6.26 form the information matrix ($\mathbf{I}_j$) for item *j*

$$\mathbf{I}_j = \begin{bmatrix} \text{Eq. 6.16} & & \\ \text{Eq. 6.24} & \text{Eq. 6.17} & \\ \text{Eq. 6.25} & \text{Eq. 6.26} & \text{Eq. 6.18} \end{bmatrix}. \tag{6.27}$$

The reciprocals of the square root of the main diagonal elements are the estimates of the standard errors of α_j, δ_j, and χ_j. On the normal metric, the corresponding item parameter information formulas are (Lord, 1980)

$$I_{\alpha_j} = \frac{D^2}{(1-\chi_j)^2}\sum_i^N\left((\theta_i-\delta_j)^2(p_j-\chi_j)^2\frac{(1-p_j)}{p_j}\right), \tag{6.28}$$

$$I_{\delta_j} = \frac{D^2\alpha_j^2}{(1-\chi_j)^2}\sum_i^N\left((p_j-\chi_j)^2\frac{(1-p_j)}{p_j}\right), \tag{6.29}$$

$$I_{\chi_j} = \frac{1}{(1-\chi_j)^2}\sum_i^N\frac{(1-p_j)}{p_j}, \tag{6.30}$$

$$I_{\alpha\delta_{(j)}} = \frac{D^2\alpha_j}{(1-\chi_j)^2}\sum_i^N\left((\theta_i-\delta_j)(p_j-\chi_j)^2\frac{(1-p_j)}{p_j}\right), \tag{6.31}$$

$$I_{\alpha\chi_{(j)}} = \frac{D}{(1-\chi_j)^2}\sum_i^N\left((\theta_i-\delta_j)(p_j-\chi_j)\frac{(1-p_j)}{p_j}\right), \tag{6.32}$$

and

$$I_{\delta\chi_{(j)}} = -\frac{D\alpha_j}{(1-\chi_j)^2}\sum_i^N\left((p_j-\chi_j)\frac{(1-p_j)}{p_j}\right). \tag{6.33}$$

16. For completeness, the information functions for estimating α_j for all five items are shown in Figure 6.18. As can be seen, the bimodal pattern exhibited in Figure 6.14 is true for all items. The different modal values reflect the magnitude of the nonzero

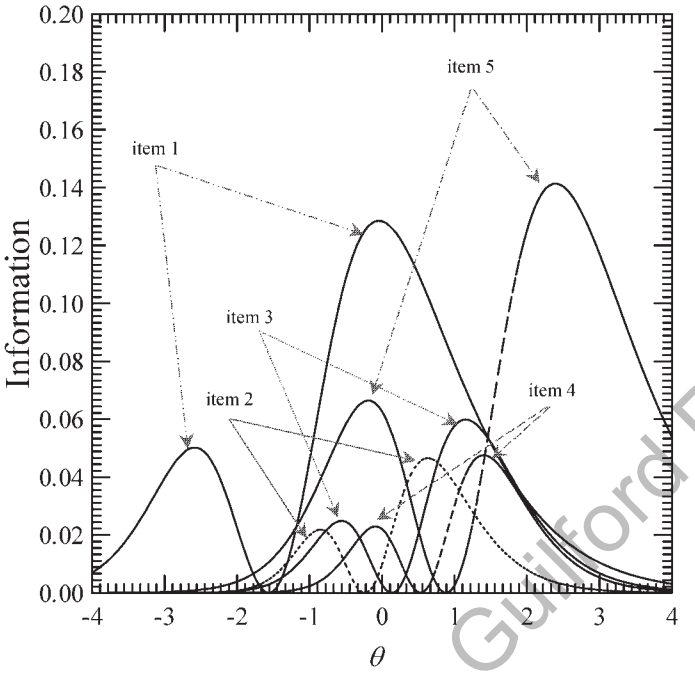

FIGURE 6.18. Information for estimating α_j as a function of θ for each of the five math items.

χ_js. It is also apparent that the modes occur on opposite sides of the item's location, with the leftmost mode always less than the rightmost mode. This characteristic is a reflection of positive α_js (i.e., if the α_js are negative, then the leftmost mode would be greater than the rightmost mode). The location of the minimum of the information function between the two modes corresponds to the item's δ_j; this minimum information is 0.

17. Typically, the TCC is depicted as ogival shaped and as resembling an IRF. However, the TCC's shape is a function of not only the number of items, but also the calibration model and the distribution/characteristics of the item parameter estimates. For example, if our $\hat{\delta}_j$s are more widely spaced than those used in Figure 6.17, the TCC's shape would change. Figure 6.19 contains the TCC for a five-item set that uses the same $\hat{\alpha}_j$s and $\hat{\chi}_j$s as in Figure 6.17, but with $\hat{\delta}_1 = -3.0$, $\hat{\delta}_2 = -2$, $\hat{\delta}_3 = 0.0$, $\hat{\delta}_4 = 2$, and $\hat{\delta}_5 = 3.0$. Clearly, this TCC is still monotonically nondecreasing, but it also contains ridges. (One needs to extend the abscissa to see that the TCC is asymptotic with $\sum \chi_j$.)

18. Based on the work of Yen (1981), it appears that whenever one applies an inappropriate model to a data set, one may obtain *sample-dependent* estimates (i.e., a contradiction to one of IRT's potential advantages). Therefore, adopting a model that expresses one's intentions and does not simply describe the data appears to be a prudent strategy. From a philosophical perspective, because all models are false, then this begs the question as to whether one may obtain sample-independent estimates

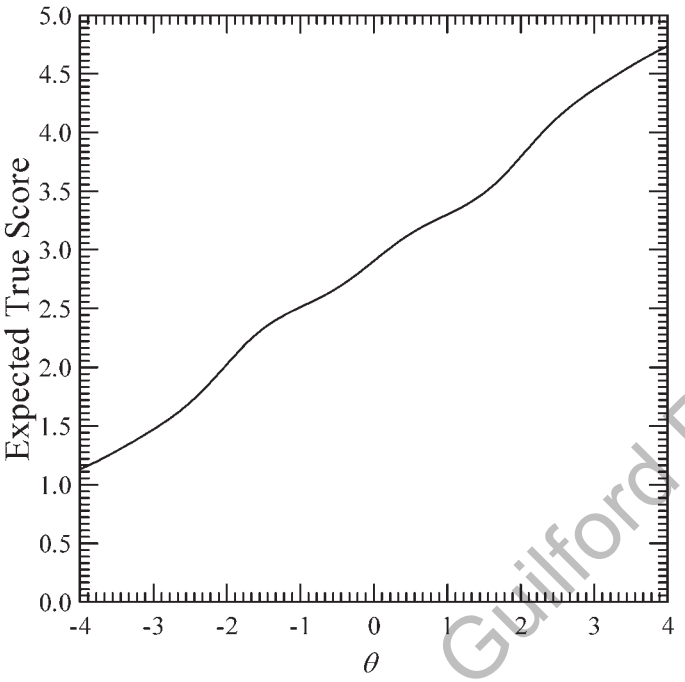

FIGURE 6.19. TCC for widely spaced δs.

in any truly absolute fashion. It is conjectured that, most likely, the best that one may be able to achieve is sample-independent estimates for a particular range of data (as demonstrated in Chapter 3). If these data represent the situations in which one is primarily interested, then whether one may obtain sample independent estimates in an absolute fashion may be academic.

Guilford Publications
370 Seventh Avenue
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com