2

Unidimensional Logistic Response Models

Wim J. van der Linden

CONTENTS

2.1 Introduction

The family of models discussed in this chapter is for responses to dichotomous items. This family of models has a long tradition of successful applications in educational and psychological testing as well as several other areas of behavioral and cognitive measurement. In fact, it is by far the most frequently used family of models for these applications.

The most general member of the family, known as the three-parameter logistic (3PL) model, was introduced by Birnbaum (1968). It seems fair to ascribe the origins of a special version of it, known as the two-parameter logistic (2PL) model, to Lord (1952). Although he actually used the normal-ogive instead of the logistic response function, his model had a parameter structure identical to that of the 2PL model. The one-parameter logistic (1PL) model is a special case of the 2PL model. It can be shown to be equivalent to the Rasch model. Rasch (1960, Section 6.8) was aware of this option but invariably used the simpler representation of it with exponential versions of the difficulty and ability parameters in this chapter. However, as the Rasch model is an exponential family model from which it borrows special statistical properties, it deserves a separate review (Volume One, Chapter 3). In this chapter, we therefore review the 1PL model only to the extent that it shares statistical properties with the other two models.

2.2 Presentation of the Models

All three models exist as versions with fixed-effects and random-effects parameters. Historically, the introduction of the former preceded the latter. The main reason for the introduction of the latter was to overcome the computational issues associated with the fixed-effects models discussed later in this chapter.

2.2.1 Fixed-Effects Models

The distributions addressed by the fixed-effects versions of the three models are for the dichotomous responses $U_{pi} = 0,1$ by test takers $p = 1, \ldots, P$ on items $i = 1, \ldots, I$. The prime examples of this type of responses are items scored as correct or incorrect. The distributions are Bernoulli with probability functions

$$f(u_{pi}; \pi_{pi}) = \pi_{pi}^{u_{pi}} (1 - \pi_{pi})^{1-u_{pi}}, \quad p = 1, \ldots, P; \quad i = 1, \ldots, I, \tag{2.1}$$

where $\pi_{pi} \in [0,1]$ are the success parameters for the distributions; that is, the probabilities of a response $U_{pi} = 1$ by the test takers on each of the items (Casabianca and Junker, Volume Two, Chapter 2).

Making the usual assumption of independence between the responses by the same test taker ("local independence"), and assuming they all worked independently, the probability function of the joint distribution of a complete response matrix, $\mathbf{U} \equiv (U_{pi})$, is the product of each of these Bernoulli distributions:

$$f(\mathbf{u}; \pi) = \prod_{p=1}^{P} \prod_{i=1}^{I} \pi_{pi}^{u_{pi}} (1 - \pi_{pi})^{1-u_{pi}} \tag{2.2}$$

with parameter vector $\boldsymbol{\pi} \equiv (\pi_{11}, \ldots, \pi_{1I}, \ldots, \pi_{P1}, \ldots, \pi_{PI})$.

The 3PL model explains each π_{pi} as a function of parameters for the effects of the test taker's ability and the properties of the item. More specifically, let θ_p denote the parameters for the effects of the individual abilities of the test takers and a_i, b_i, and c_i the effects of the items generally interpreted as representing their difficulties, discriminating power, and success probabilities when guessing randomly on them, respectively. For a given response matrix, the 3PL model equations are

$$\pi_{pi} = c_i + (1 - c_i) \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}, \quad p = 1, \ldots, P; \quad i = 1, \ldots, I \tag{2.3}$$

with $\theta_p \in (-\infty, \infty)$, $a_i \in (0, \infty)$, $b_i \in (-\infty, \infty)$, and $c_i \in [0,1]$ as ranges for the values of their parameters. The model thus consists of $P \times I$ nonlinear equations, one for each of the success parameters. In other words, rather than a single-level probabilistic model, it is a system of second-level mathematical equations. (The statistical literature is somewhat ambiguous in its assignments of the number of levels to a model; some sources count the levels of parameters as we do here, while the others count the levels of randomness that are modeled.)

It is common to introduce the 3PL model graphically instead of as the system of equations in Equation 2.3, emphasizing the shape of the success probabilities π_{pi} as a function of a mathematical variable θ. Figure 2.1 shows these response functions for 40 arithmetic items estimated under the 3PL model. For each of these functions, the c_i parameter represents the height of the lower asymptote to it. More formally, these parameters are defined as

$$c_i = \lim_{\theta_p \to -\infty} \pi_{pi}. \tag{2.4}$$

Naively, for a multiple-choice item with A alternatives, one might expect to find $c_i = 1/A$. But in practice, guessing turns out to be somewhat more complicated than a test taker just picking one of the alternatives completely at random; empirical estimates of the c_i parameters typically appear to be slightly lower than $1/A$. The b_i parameters represent the location of the curves on the scale for θ. Formally, they are the values of θ with success probability

$$\pi_{pi} = (1 + c_i)/2, \tag{2.5}$$

that is, the probability halfway between their maximum of 1 and minimum of c_i. Finally, the α_i parameters can be shown to be proportional to the slope of the response functions at $\theta = b_i$, at which point the slopes take the value

$$\frac{\partial \pi_{pi}}{\partial \theta_p} = 25 a_i (1 - c_i). \tag{2.6}$$

Although graphs of response functions as in Figure 2.1 definitely add to our understanding of the 3PL model, they are also potentially misleading. This happens especially

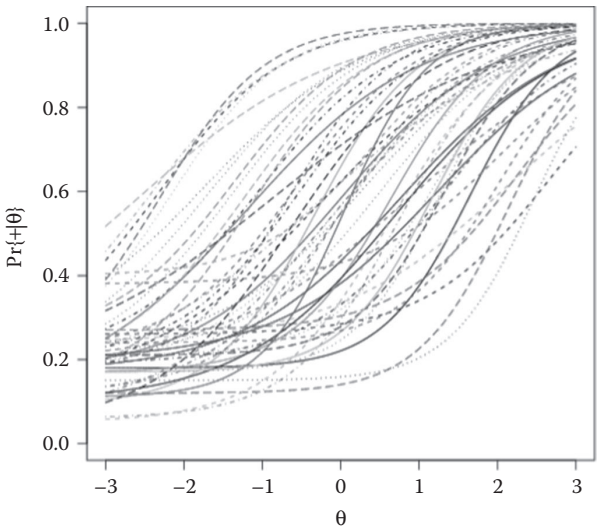

FIGURE 2.1
Response functions for a set of 40 arithmetic items estimated under the 3PL model.

when, mainly for didactic reasons, the model is introduced by explaining the graph of the response function for one fictitious item only. For the case of one item, the system of equations is not identified. Consequently, as its parameters cannot be estimated, the model does not have any empirical meaning (see below). For the case of multiple items, as in Figure 2.1, the graph nicely reveals the relative values of the parameters of each of the I items in Equation 2.3 but still hides those of the P ability parameters.

Use of Equation 2.3 as the representation of the model also reveals that, in spite of its wide acceptance, the adoption of the qualifier "3PL" in the name of the model is potentially confusing as well. It suggests a restriction of the parameter count to the parameters of one item only, ignoring both those of all other items and the ability parameters. However, replacing it by the more appropriate qualifier "$(P + 3I)$PL" would still leave us with a misnomer. The shapes of the response functions shown in Figure 2.1 are *not* those of a logistic function. Only the 2PL and 1PL models below have logistic response functions.

Finally, it is still not uncommon to find a version of the model with a scale constant $D = 1.7$ added to it, giving it $1.7a_i(\theta_p - b_i)$ as its core structure. The original purpose of this practice was to bring the shape of the response functions close to that of the normal-ogive model, generally believed to be the "true model" that would give us a scale with "equal units of measurement" in the early days of test theory (Volume One, Chapter 1). However, as discussed below, the scale for the θ_p parameters is arbitrary up to a nonmonotonic transformation, and the practice of bringing it close to the scale of a normal ogive is a relict from a distant past, bound to create communication problems between users of item response theory (IRT) rather than add to their understanding of it.

The 2PL model follows from Equation 2.3 upon the choice of $c_i = 0$ for all items. It has

$$\pi_{pi} = \frac{\exp[a_i(\theta_p - b_i)]}{1 + \exp[a_i(\theta_p - b_i)]}, \quad p = 1, \ldots, P; \quad i = 1, \ldots, I \tag{2.7}$$

as its system of model equations. Observe that this model does assume the logistic function

$$\Psi(x) = \frac{\exp(x)}{1 + \exp(x)}$$

as response function. For this choice, location parameters b_i simplify to the point on the θ scale with success probability $\pi_{pi} = 0.5$ for each item, which now also is the point at which the slope of the response functions in Equation 2.6 specializes to $0.25a_i$. A more substantive assumption underlying the 2PL model is thus absence of any guessing—an assumption that may hold when there is nothing at stake for the test takers. However, the model is sometimes applied in cases where the assumption is clearly untenable, typically with the claim that the c_i parameters are generally difficult to estimate. But even if this claim were valid (it is not), it would empirically be more defensible to set these parameters equal to a value somewhat larger than zero to better account for the effects of guessing.

The 1PL model follows upon the additional assumption of $a_i = 1$ for all items, yielding

$$\pi_{pi} = \frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)}, \quad p = 1, \ldots, P; \quad i = 1, \ldots, I \tag{2.8}$$

as its system of model equations. As already alluded to, reparameterizing the model through substitution of

$$\begin{aligned} \theta_p &= \ln(\tau_p), \\ b_i &= \ln(\xi_i),\ \tau_p, \xi_i > 0 \end{aligned} \tag{2.9}$$

into it gives us

$$\pi_{pi} = \frac{\tau_p}{\tau_p + \xi_i}, \quad p = 1,\ldots,P; \quad i = 1,\ldots,I, \tag{2.10}$$

which is the version typically used as representation of the Rasch model (Volume One, Chapter 3).

Attempts have been made to add an upper asymptote to the model in Equation 2.3 to make it robust against careless response behavior in addition to guessing (Barton and Lord, 1981), but this "4PL" model just has not made it. An empirically more successful model may be the Rasch model extended with the structure for the guessing parameters in Equation 2.3 added to it; for details on this 1PL-G model, see San Martín et al. (2006). Finally, the success of the family of logistic models in this chapter has led to its generalization to items with polytomous scoring and multidimensional ability parameters; examples of such generalizations are reviewed by Masters (Volume One, Chapter 7), Muraki and Muraki (Volume One, Chapter 8), Tutz (Volume One, Chapter 9), and Reckase (Volume One, Chapter 12).

2.2.2 Random-Effects Models

In spite of the convincing nature of the logistic models, their initial number of applications was extremely low. For the Rasch model, the favorite method of parameter estimation was maximum conditional likelihood (CML) estimation (Volume One, Chapter 3), but due to the limited computational power available in the 1960–1970s, it appeared impossible to deal with its elementary symmetric functions for longer tests, even for Fischer's efficient summation algorithm (van der Linden, Volume Two, Chapter 6). The typical estimation methods in use for the 3PL and 2PL models were maximum joint likelihood (JML) meth-ods. It is still not fully clear why these methods occasionally failed to show convergence, even for larger datasets and shorter test lengths. Although lack of convergence may be caused by lack of model identifiability, it is now clear that the major computer programs imposed effective additional restrictions to prevent this from happening (see below). In hindsight, these disturbing results might have been just due to limited numerical precision and/or the presence of the ability parameters as incidental parameters during item calibra-tion (Volume Two, Chapter 9).

One of the first attempts to neutralize the possible role of incidental parameters was Bock and Lieberman's (1970) reformulation of the fixed-effects three-parameter normal-ogive model as a model with random ability parameters. Their reformulation applies equally well to the logistic models.

The reformulation involves the adoption of a different probability experiment. Instead of a set of P fixed test takers, it assumes these test takers to be randomly and independently sampled from the same probability distribution. Let Θ be the random ability parameter assumed to have a common density $f(\theta)$ for the test takers and $\pi_i(\theta)$ be the probability of a correct response to item i by a test taker with realization $\Theta = \theta$.

The focus is on the random response vector to be observed for each random test taker (observe that we now have two levels of randomness). We use $\mathbf{u}_v \equiv (u_{v1}, \ldots, u_{vI})$, $v = 1, \ldots, 2^I$ to denote each possible realization of this vector. Because of local independence, the conditional probability of observing $\mathbf{U}_v = \mathbf{u}_v$ given $\Theta = \theta$ is equal to

$$\pi_v(\theta) = \prod_{i=1}^{I} \pi_i(\theta)^{u_{vi}} [1 - \pi_i(\theta)]^{1-u_{vi}}. \tag{2.11}$$

Continuing the argument, the marginal probabilities of observing $\mathbf{U}_v = \mathbf{u}_v$ are equal to

$$\pi_v = \int \prod_{i=1}^{I} \pi_i(\theta)^{u_{vi}} [1 - \pi_i(\theta)]^{1-u_{vi}} f(\theta) d\theta, \quad v = 1, \ldots, 2^I. \tag{2.12}$$

A random sample of P test takers amounts to an equal number of independent draws from the space of all possible response vector. Let X_v denote the number of times vector v is observed, with $\sum_v X_v = P$. The probability model for this experiment is a multinomial distribution with probability function

$$f(\mathbf{x}; \boldsymbol{\pi}) = \frac{P!}{x_1! \ldots x_{2^I}!} \prod_{v=1}^{2^I} \pi_v^{x_v}, \tag{2.13}$$

where $\mathbf{x} \equiv (x_v)$ and now $\boldsymbol{\pi} \equiv (\pi_v)$.

Treating $p = 1, \ldots, P$ as index for the order in which the test takers are sampled, and using $\mathbf{U} \equiv (U_{pi})$ again to denote the response matrix, the function can also be written as

$$f(\mathbf{u}; \boldsymbol{\pi}) = \prod_{p=1}^{P} \int \prod_{i=1}^{I} \pi_{pi}^{u_{pi}} (1 - \pi_{pi})^{1-u_{pi}} f(\theta_p) d\theta_p \tag{2.14}$$

with π_{pi} given by Equation 2.3, which is the form usually met in the IRT literature. This form hides the crucial difference between the fixed-effects and random-effects models as a consequence of their different sampling distributions, though. The former is the second-level system of $P \times I$ equations for the Bernoulli parameters π_{pi} in Equation 2.3. The latter is a system at the same level but now with the 2^I equations for the multinomial parameters π_v in Equation 2.12.

Somewhat later, Bock and Aitkin (1981) introduced an application of the expectation–maximization (EM) algorithm for the estimation of the item parameters in the 3PL model based on the marginal likelihood associated with Equation 2.14, which has become one of the standard procedures for item calibration in IRT. A description of the logic underlying the EM algorithm is provided by Aitkin (Volume Two, Chapter 12). For an application of the algorithm, it is not necessary for the test takers to actually be randomly sampled; in principle, a statistical model can produce good results even if not all of its assumptions are met. And extensive parameter recovery studies of item calibration with maximum marginal likelihood (MML) estimation with the EM algorithm have certainly proven the random-effects models to do so, in spite of their imposition of a mostly arbitrary common

density $f(\theta)$ on the ability parameters. But for the issue of model identifiability addressed in the next section, it is crucial to be aware of the precise sampling model underlying it.

It is not necessary to treat the ability parameters as the only random parameters in the model, leaving the item parameters fixed. Several advantages may be related to the adoption of random-effects parameters for the items as well (De Boeck, 2008). A natural application of random-item IRT is modeling of the distributions of the item parameters in the different families of items produced by different settings of a rule-based item generator; for details, see Glas, van der Linden and Geerlings (Volume One, Chapter 26).

2.2.3 Model Identifiability

The problem of lack of model or parameter identifiability arises when different combinations of parameter values are observationally equivalent, that is, imply the same distribution for the observed data. As it is generally impossible to use identically distributed data to distinguish between different parameter values, any attempt at statistical inference with respect to the parameters then breaks down.

Fortunately, the product of Bernoulli functions in Equation 2.2 is an example of a fully identified sampling model. For any shift in any combination of the components of its parameter vector $\boldsymbol{\pi} \equiv (\pi_{11}, \ldots, \pi_{1I}, \ldots, \pi_{P1}, \ldots, \pi_{PI})$, the model yields a distinct joint distribution of the responses. Conversely, for any distinct response distribution, it is thus possible to infer a unique combination of values for this vector. This last statement has only theoretical meaning though; in the practice of educational and psychological testing, due to memory or learning effects, it is mostly impossible to get more than one response per test taker to the same item without changing the ability parameters, leaving us with the extreme values of zero and one as direct estimates of each success parameter.

It is at this point, however, that second-level systems of model equations as in Equations 2.3, 2.7, and 2.8 show favor. Basically, they enable us to pool the data from multiple test takers and responses into estimates of each of the success parameters. Because of their feature of parameter separation ("different parameters for every p and i in the joint index of π_{pi}") as well as their cross-classified nature ("multiple equations with the same parameters"), these systems enable us to adjust the data on the test-taker parameters for the differences between the items, and conversely. As a result, we effectively have multiple data points per ability and item parameter, and are able to combine their estimates into estimates of each of the individual success parameters. (The problem of incidental parameters is intentionally ignored here.) As discussed below, these success parameters are key parameters in item response modeling which provide us with the empirical interpretation of the test takers' scores.

These systems of equations are only able to do their job, however, if they are identified themselves; that is, when, for the 3PL model, each distinct combination of values for the model parameters $(\theta_p, a_i, b_i, c_i)$ corresponds with a distinct value of π_{pi}. Generally, mathematical systems of equations are only identifiable if the number of equations is at least as great as the number of unknowns. Thus, at a minimum, it should hold for this model that $P \times I \geq P + 3I$—a condition which implies, for example, that for a five-item test we are unable to produce any fixed parameter estimates unless the number of test takers is at least equal to four.

However, the condition is not sufficient. For one thing, as is well known, no matter the numbers of test takers and items, it is always possible for some subset of the model parameters to compensate for certain changes in the others. More specifically, the addition of the same constant to all b_i and θ_p parameters in Equation 2.3 does not lead to a change in any of the π_{pi}. Likewise, multiplying all a_i parameters by a constant has no effect on these

success parameters when all θ_p and b_i parameters are divided by it as well. The identifiability problem is more fundamental, though. Deeper analysis has revealed cases for which the c_i parameters are not identifiable either (Maris, 2002; van der Linden and Barrett, 2016). A more comprehensive review of identifiability issues in IRT is provided by San Martín (Volume Two, Chapter 8).

For a response matrix with fixed dimensions, the only way to make the system of model equations identifiable is by adding more equations to it. A recent result by van der Linden and Barrett (2016, Volume Three, Chapter 2, Theorem 3) can be used as a check on the sufficiency of such identifiability restrictions for the logistic models in this chapter. The result is a formal characterization of the class of observationally equivalent parameters values for the 3PL model in the form of a mapping φ which, for an arbitrary test taker and item, for any given solution to the model equations gives us all other solutions. The mapping can be shown to be the vector function

$$\varphi(\theta, a, b, c) = (u\theta + v, u^{-1}a, ub + v, c) \tag{2.15}$$

with $u \equiv [\varphi_\theta(\theta) - \varphi_\beta(\beta)]/(\theta - \beta)$, $\theta \neq b$, and $v \equiv \varphi(b) - ub = \varphi(\theta) - u\theta$.

The critical quantities in this result are the (unknown) parameters u and v. Given any arbitrary combination of values for $(\theta_p, a_i, b_i, c_i)$, they index all other combinations with identical values for $(\pi_{11}, \ldots, \pi_{1I}, , \pi_{P1}, \ldots, \pi_{PI})$ in Equation 2.3.

Observe that for the choice of $(u,v) = (1,0)$ the function just returns its input. It follows that adding equations to the system that effectively restrict u and v to this pair of values does restrict it to have only one solution as well—in other words, makes the model identifiable. As an example, suppose we set $\theta_p = \kappa_1$ and $\theta_{p'} = \kappa_2$ for two arbitrary test takers p and p' and constants κ_1 and κ_2. The first component of Equation 2.15 can then only take the form

$$\theta_p = u\theta_p + v$$

$$\theta_{p'} = u\theta_{p'} + v$$

for these two test takers, leaving $(u,v) = (1,0)$ as the only admissible values.

These two restrictions are definitely not the only option. More generally, it can be shown that in order to make the fixed-effect 3PL model identifiable, it is sufficient to add (i) one linear restriction on one of the a_i parameters in combination with another on one of the b_i or θ_p parameters or (ii) two independent linear restrictions on the θ_p and/or b_i parameters (van der Linden, 2016, Theorem 1).

The sets of identifiability restrictions for the 3PL and 2PL model are identical. The restrictions for the 1PL model follow if we ignore the a_i parameters. As for the c_i parameters, the fact that they do not need to be restricted does not imply that they are automatically identifiable though. The mapping in Equation 2.15 is a vector function with components that hold only simultaneously; we should not isolate one of the components and declare it to hold independently of the others. The correct conclusion from Equation 2.15 is that the c_i parameters are identifiable once all other model parameters are. In fact, our earlier references to specific cases for which the c_i parameters have shown to lack identifiability already implied the necessity of additional conditions.

A similar analysis for the random-effects versions of the logistic models has not resulted in any final conclusions yet. Remember that their sampling distribution is multinomial with the vector of marginal probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{2^I})$ in Equation 2.12 as parameters.

The good news is that the multinomial model is identifiable as well; its family of distributions is known to have distinct members for distinct vectors of success probabilities. The problematic part, however, is the more complicated system of equations for these success probabilities as a function of the model parameters in Equation 2.12, which now includes the unknown parameters of the density $f(\theta)$ as well.

A common practice for the 3PL and 2PL models is to set the mean and variance of the ability parameters equal to zero and one, respectively:

$$\mu_\theta = 0 \quad \text{and} \quad \sigma_\theta^2 = 1. \tag{2.16}$$

For the fixed-effects versions of these models, these restrictions were typically implemented by renorming all ability parameters estimates upon each iteration step in the JML estimation procedure. It is easy to show that these restrictions pass our earlier check on the validity of the identifiability restrictions for these models (van der Linden, 2016). For the random-effects versions, the restrictions are typically implemented through the choice of the standard normal density for $f(\theta)$ in Equation 2.14. Although, a formal proof is still missing, numerous parameter recovery studies have invariably supported the validity of this practice.

2.2.4 Parameter Linking

One of the consequences of the fundamental lack of identifiability of the logistic response models is the necessity to link parameter values obtained in different calibration studies through common items and/or test takers. This is necessary even when formally identical identifiability restrictions are imposed.

A simple case illustrates the necessity. Suppose the same test taker responds to different items in two calibration studies, both with Equation 2.16 as identifiability restrictions. Owing to the presence of other test takers, the restrictions have a differential effect on the ability parameter for this common test taker. For example, if the test taker would be among the most able in the first study but the least able in the second, the use of $\mu_\theta = 0$ would force his parameter value to be positive in the former but negative in the latter. A similar analysis is possible for common items in different calibration studies.

Observe that the issue has nothing to do with the impact of estimation error; all changes are in what commonly is referred to as true parameter values. "True" should not be taken to mean unique, though. Lack of identifiability means that each parameter has an entire class of observationally equivalent values. Identifiability restrictions are necessary to reduce each of these classes to a unique value. These restrictions do not operate on each parameter in isolation but, as an integral part of the system of model equations, have a joint effect on all of them.

The close relationship between the lack of model identifiability and the necessity of parameter linking suggests an equally important role of the characterization of the equivalent parameter values in Equation 2.15 with respect to the latter. Indeed, the linking functions required to map the unique parameters values in one calibration onto those in another have to be derived from Equation 2.15 as well; details are provided in van der Linden and Barrett (Volume Three, Chapter 2).

2.2.5 Model Interpretation

Several times so far, we have emphasized the formal nature of the response models as a system of equations for the test takers' success probabilities on the items. These success

probabilities are the "empirical reality" explained by the models. Conversely, in order to empirically interpret the features of the model, we have to restrict ourselves to the given set of probabilities. Any appeal to a reality outside of it is speculative at best.

It follows that formal features of the response models that can be changed without changing the success probabilities for the test takers are meaningless. Only features that remain invariant under such changes can have a valid empirical interpretation. A prime example of features of the logistic models that are not invariant are the values of their parameters. As just noted, each of them can be made to vary as a result of the infinitely many choices for the identifiability restrictions that have to be imposed on the models. Hence, statements as "John's ability score on this test is equal to 50" or "the difficulty of this item is equal to –2" should not be taken to have any absolute meaning.

Actually, the same lack of invariance holds for the entire parameter structure of the model. As already demonstrated, it is possible to use Equation 2.9 to reparameterize the 1PL model into the Rasch model. Both representations have a completely different structure. But empirically, it is impossible to tell one from the other; for the same test takers and items, they imply the same probabilities of success and thus identical response distributions. This observation forces us to extend our earlier notion of observational equivalence to include equivalence across reparameterizations of the model as well.

Note that it is possible to go back and forth between the 1PL and Rasch models because the logarithmic transformation in Equation 2.9 has the exponential as its inverse. More generally, it is always possible to reparameterize probabilistic models provided the vectors of old new and parameters are of the same dimension and have a reversible (bijective) relationship. The logistic models are monotone, continuous functions in their parameters (the only exception to the monotonicity is when $\theta \neq b$, which case we exclude). If we want to keep these features, the required relationship reduces to a similar function between the old and new parameters (van der Linden and Barrett, 2016, theorem 3). Hence, the following conclusion: For the logistic models, it is always possible to replace their parameters by new parameters provided the old and new parameters are monotone continuous functions of each other.

For each of these infinitely many possible reparameterizations, the response probabilities for the test takers remain the same but the response functions look entirely different. Figure 2.2 illustrates the impact of a few reparameterizations of the Rasch model. Its first plot shows the response functions for 20 items for the original version of the model in Equation 2.10. Following S. S. Stevens' (1946) classification of different levels of measurement, the choice is sometimes claimed to yield an absolute zero for the ability scale, and thus measurement on a ratio scale. The second plot shows the shapes of the response functions for the same items for the standard parameterization of the 1PL model. The fact that the curves now run "parallel" seems to suggest equal measurement units along the ability scale and this feature has led to claims of measurement on an interval rather than a ratio scale. These two claims are already inconsistent. But actually, almost every claim of special features for the scale of ability parameters can be "reparameterized away." The third plot shows the shift in response functions obtained for an alternative parameterization of the Rasch model, using

$$\begin{aligned} &\tau_p = \lambda_p - 1, \lambda_p > 1, \\ &\xi_i = \beta_i, \beta_i > 0 \end{aligned} \tag{2.17}$$

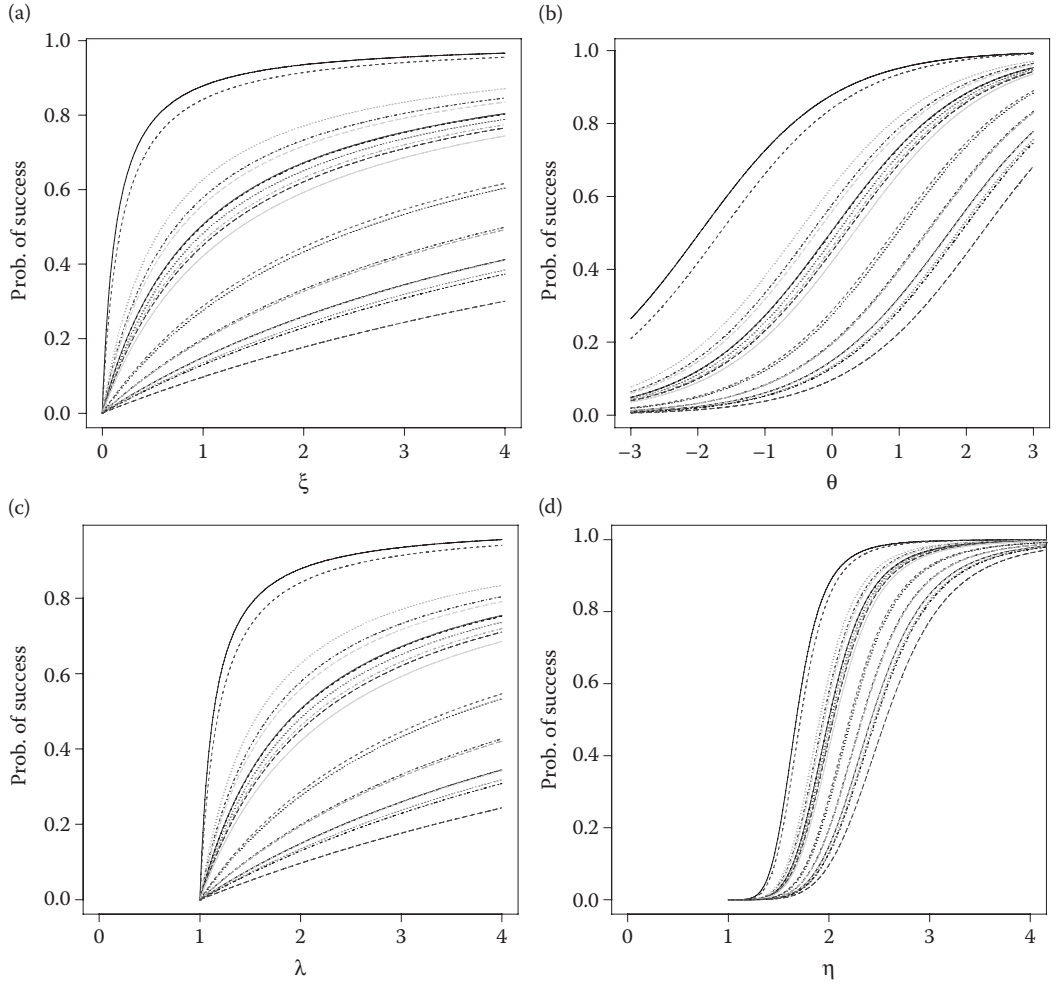

FIGURE 2.2
Four different parameterizations of the Rasch model, each with the same goodness of fit to the same response data. (a) Equation 2.10; (b) Equation 2.8; (c) Equation 2.18; (d) Equation 2.20.

for all p and i. The result is

$$\pi_{pi} = \frac{\lambda_p - 1}{\lambda_n + \beta_i - 1}, \tag{2.18}$$

which now has a scale for the ability parameters with the earlier "zero" at the point $\lambda = 1$. Our last example is

$$\begin{aligned} \tau_p &= \eta_p^5 - 1,\ \eta_p > 1, \\ \xi_i &= \varepsilon_i,\ \varepsilon_i > 0 \end{aligned} \tag{2.19}$$

for all p and i, which gives us

$$\pi_{pi} = \frac{\eta_p^5 - 1}{\eta_p^5 + \varepsilon_i - 1} \tag{2.20}$$

as another representation of the Rasch model. The result has the same shift of "zero" but now gives us ogive-shaped response functions squeezed together at the lower end of range of possible values for the ability parameters. The squeeze clearly illustrates how the same interval of values for the ability parameters at different positions along the scale is associated with different changes in success probability. (Actually, the same observation could already have been made directly for the standard 1P model. The "parallelness" of its response functions in the second plot is just an optical illusion. How could monotone curves each mapping the real line onto [0,1] ever be made to run completely parallel?)

Figure 2.2 illustrates how dangerous it is to ascribe absolute meaning to apparent features of the logistic response models. The possibilities of reparameterization are virtually endless. Other choices could have been made with more surprising changes in the shapes of the response functions, especially if we would include other transformations of the item parameters than with the identify functions above. The same conclusions hold for the 2PL and 3PL models, with even a larger variety of changes due to the possibility to play with their a_i and c_i parameters as well.

What, if any, features of the logistic models are invariant across all possible reparameterizations and thus allow for absolute interpretation? There are three. The first is the order of the ability parameters. Although their scale does have an arbitrary zero and unit, the models do order all ability parameters identically across all possible parameterizations—a fundamental prerequisite for measurement of a unidimensional variable.

The second is the quantile rank of the test takers in *any* norm group for which the test is in use. The feature follows from a simple statistical fact. Let $q_p \in [0,1]$ be the rank of test taker p with score θ_p in a norm group with any distribution $F_\Theta(\cdot)$ defined as the proportion of test takers below this person; that is, $F_\Theta(\theta_p) = q_p$. For any monotone reparameterization $\vartheta(\cdot)$, it holds that

$$F_{\vartheta(\Theta)}(\vartheta(\theta_p)) = F_\Theta(\theta_p). \tag{2.21}$$

Thus, $F_{\vartheta(\Theta)}(\vartheta(\theta_p)) = q_p$ as well, and p has the same quantile rank in the new as in the old distribution.

The final feature, not surprisingly, is the invariant link between each of the test takers and their probabilities of success on the items. No matter how we reparameterize the model, for the same items, it always assigns the same probability of success to each of the test takers. The feature can be used to map the content of the items along the current ability scale, providing it with a point-by-point empirical interpretation. Graphically, each reparameterization of the model amounts to local stretching or shrinking of the scale, but the order of all points remains the same and so does their empirical interpretation. Item maps have been shown to be key to IRT-based test score interpretation (Hambleton and Zenisky, Volume Three, Chapter 7), standard setting (Lewis and Lord-Bessen, Volume Three, Chapter 11), and more generally item writing and test construction (Wilson, 2005). In our example below, we use a logistic response model to construct an item map that illustrates the behavioral meaning of the variable of body height.

In educational testing, it is common to distinguish between norm-referenced and criterion-referenced interpretations of test scores. The former refers to the quantile ranks of test scores in the distributions of the norm groups for which the test is in use; the latter to empirical information on what test takers with a given score can and cannot perform. From the preceding discussion, we are able to conclude that the logistic response models in this chapter provide us with measurement of a unidimensional variable on scale that, except for nonmonotone transformation, is entirely arbitrary but nevertheless has absolute norm- and criterion-referenced interpretations.

The fact that the parameter structure of the logistic response models is arbitrary should not lead to relativism or carelessness. As highlighted by Hand (2004), measurement in any science has both a representational and pragmatic aspect. The pragmatic aspect is the result of the necessity to make arbitrary choices with respect to such steps as the design of the measurement instrument, the actual measurement operations, definition of the data, and auxiliary assumptions required to infer the measurements from them. These choices are entirely separate from the empirical features the measurements present, but are necessary to map them on numbers.

As for the item parameters, the standard parameterization of the 3PL model in Equation 2.3 involves parameters (a_i, b_i, c_i) that, strictly speaking, do not have any meaning beyond their formal definitions in Equations 2.4 through 2.6. Our references to them as parameters for the discriminating power, difficulty, and probabilities of successful guessing on an item are just empirical metaphors. Although fundamentally arbitrary, these metaphors are well established and have helped us to carefully communicate differences between test items. Any other choice of parameterization would lead to loss of these metaphors and seriously disrupt our communications.

2.3 Parameter Estimation

The most frequently used methods for estimating the parameters in the logistic models are Bayesian and MML methods, where the former have taken the lead due to the current popularity of its Markov chain Monte Carlo (MCMC) methods. This handbook has several other chapters exclusively devoted to these methods, which use one or more of our logistic models to illustrate their ideas, equations, and algorithms, and highlight important details of their implementation. It is pointless to duplicate their materials in this chapter.

A general introduction to Bayesian statistical methods, explaining their logic of inference with its focus on the joint posterior distribution of all model parameters, is provided by Johnson and Sinharay (Volume Two, Chapter 13). As a continued example, these authors use the one-parameter normal-ogive model. The treatment of their example easily transfers to the 1PL model though. Junker, Patz and VanHoudnos (Volume Two, Chapter 15) should be consulted for a comprehensive introduction to the theory and practice of MCMC methods for sampling posterior distributions. Specifically, they introduce the history of general ideas underlying these methods, address all choices that have to be made when implementing them, and use the 2PL model extensively to illustrate how to use a Metropolis–Hastings algorithm to sample the parameters in the 2PL model. Readers interested in *R* code to run the algorithm find it in their example.

Glas (Volume Two, Chapter 11) offers an extensive introduction to MML estimation of item parameters in IRT models, including the logistic models in this chapter, based on Fisher's identity as organizing principle. The identity equates the derivatives of the log of

the marginal likelihood in Equation 2.14 with respect to the parameters to their posterior predictive expectation across the missing ability parameters. The resulting systems of estimation equations can be solved numerically or using an EM algorithm. The use of the EM algorithm in this context is carefully explained in Aitkin (Volume Two, Chapter 12). Unlike the Bayesian methods, MML estimation only provides us with estimates of the item parameters. Estimates of the ability parameters are typically obtained either through a second maximum-likelihood step treating the item parameters as known constants or in a Bayesian fashion that accounts for their remaining uncertainty.

Almost every computer program in an extensive section with reviews of programs in *Handbook of Item Response Theory, Volume 3: Applications* can be used to estimate the current logistic models, including programs that embed them in larger modeling frameworks, such as *Mplus*, generalized linear latent modeling (GLLAM), and *Latent GOLD*.

2.4 Model Fit

Bayesian model fit methodology for item response models is reviewed in Sinharay (Volume Two, Chapter 19). The methodology enables us to check on such features of the models as monotonicity of their response functions, local independence between the responses, differential item function across subgroups of test takers, or an assumed shape for the ability distribution (given the current parameterization). The methods include Bayesian residual analysis (i.e., evaluation of the size of the observed u_{pi} under the posterior distribution of π_{pi}) and predictive checks on a variety of test quantities given the prior or posterior distributions of the parameters. The chapter also shows how easy it is to implement these methods for MCMC sampling of the posterior distribution of the parameters, and extensively demonstrates an implementation for the 1PL model.

Glas (Volume Two, Chapter 17) shows how to test the validity of each logistic models against a variety of alternative hypotheses using a coherent framework of Lagrange multiplier (LM) tests. His chapter also reviews alternatives based on likelihood-ratio and Wald tests.

For a more descriptive analysis of model fit based on the classical residuals $U_{pi} - \pi_{pi}$ aggregated across the test takers and/or items, the reader should consult Wells and Hambleton (Volume Two, Chapter 20). This review is built entirely around a demonstration of the application of their methods to an empirical dataset for the 3PL model.

The logistic models in this chapter are nested in the sense that a version with fewer parameters is obtained constraining a version with more. For comparison with alternative models outside this nested family, the information criteria in Cohen and Cho (Volume Two, Chapter 18) or Bayes factors in Sinharay (Volume Two, Chapter 19) can be used. The chapter by Cohen and Cho has an entire section reviewing the applications of information criteria for model comparison throughout the most recent IRT literature.

2.5 Empirical Example

The purpose of the example is to illustrate the use of a logistic response model for the measurement of body height. The application is somewhat unusual in that body height seems a

I bump my head quite often
For most people, my shoes would be too large
When a school picture was taken, I was always asked to stand in the first row
In bed, I often suffer from cold feet
When walking down the stairs, I usually take two steps at a time
I think I would do reasonably well in a basketball team
As a police officer, I would not make much of an impression
In most cars, I sit uncomfortably
I literally look up to most of my friends
I often have to stand on my toes to look in the mirror
Seats in theaters are usually too small for me

FIGURE 2.3
Sample items from 32-item test of body height.

physical variable only to be measured by a yardstick. However, differences in body height do have behavioral consequences. Conversely, it is thus perfectly possible to treat these consequences as indicators of body height and infer measurements from them.

The measurement instrument used to collect the current dataset was a test of 32 dichotomous items, each formulating a different behavioral consequence. Examples of the items are given in Figure 2.3. The subjects were 214 students in a class on test theory at Free University, Amsterdam, who were asked to respond to each of the items indicating whether or not they endorsed the experience formulated in them. As nothing was at stake for the test takers, guessing was not expected to play any role, and the responses were analyzed using the 2PL model. The method of estimation was MML estimation with the EM algorithm as implemented in *BILOG-MG*, version 3 (Volume Three, Chapter 23; Zimowski et al., 2003), using its default option of a standard normal distribution for the θ_p parameters.

Three of the items yielded values for the residual-based chi-square type statistic in *BILOG-MG* with probabilities of exceedance lower than 0.10. The items were "I often sit uncomfortably in the back seat of a car" ($p = 0.08$), "When people were chosen for a school basketball team, I was usually chosen last" ($p = 0.04$), and "When I sit at a table, I usually have trouble with legroom" ($p = 0.05$). It is tempting to speculate about the reasons of misfit—Dutch students generally bike? Not too many Dutch schools with a basketball team? Other people at the same table creating the problem?—but we just accepted the lack of fit of these items as a statistical fact. In addition, we removed two items with rather skewed response distributions from the analysis. One had only seven zeroes among its 214 responses, which resulted in an estimate of its b_i parameters of −4.62 with a standard error of estimation equal to 1.34. The other had only 15 ones, with a b_i estimate of 6.18 and a standard error equal to 2.19. These extreme values were taken to point at poor identifiability of the parameters from the given data.

The estimates of the b_i parameters of the remaining 27 items varied across [–2.23, 3.55] with standard errors in the range of [0.09, 0.90]. Likewise, the estimates of the a_i parameters ran between [0.36, 2.94] with standard errors of [0.10, 0.76].

Figure 2.4 shows the item map for a selection from the remaining items with brief content descriptions at their points of 0.50 probability of endorsement. Our selection is largely arbitrary; in an electronic version of the map, we would have put all items in it, with the options to zoom in on certain portions of the map, click on them to see full items or an interval of uncertainty about their location, etc. But the current version already illustrates how much richer a plain variable as body height becomes when it is presented lined with

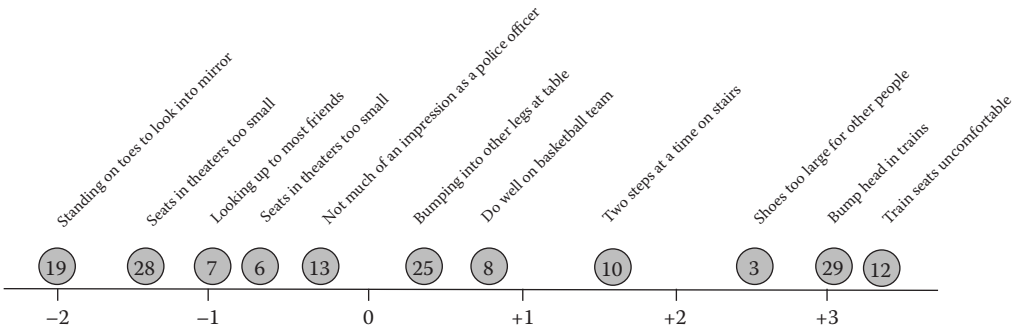

FIGURE 2.4
Example of an item map for a test of body height (locations of the items at 0.50 probability of their endorsement).

behavioral consequences rather than a single number for each subject. Unfortunately, our dataset did not include physical measurements of the body heights of its subjects. Otherwise we could have used the arbitrariness of the scale of the 2PL model to introduce a transformation of it that would have mapped the item content along the physical scale used to obtain the measurements.

2.6 Discussion

The logistic models in this chapter belong to the best-researched and most frequently used models in the field of IRT. At times, however, their use has led to debates between supporters of the 1PL and 3PL model pleading an *a priori* preference for their model. The former have typically supported their plea by pointing at such unique properties of the 1PL model as the presence of sum scores as sufficient statistics for its model parameter, measurement on interval or ratio scales, and similar ordering of the response functions across the ability scale. The latter have pointed at a higher likelihood of fit of their model to empirical response data and its better accounting for the presence of guessing on test items, especially in high-stakes testing.

Either position has both something against and in favor of it. The 1PL model belongs to the exponential family and consequently does have sufficient statistics for its parameters, a feature that gives us the option of conditional inference about them. Except for the elegance of its estimation equations and the expressions of its goodness-of-fit statistics, conditional reference does not have any serious advantages though. It certainly does not have any computational advantages, a criterion that, given the overwhelming computational power currently at our disposal, has lost its significance anyhow. More importantly, the 2LP and 3PL models have sufficient statistics too—the complete response vectors—that contain all statistical information available about their ability and item parameters. In fact, if we condition on sum scores for the 1PL model, some of the information in the response vectors (not much) is lost because their distribution still depends on the parameters we try to eliminate (Eggen, 2000).

As for the unique scale properties claimed to hold for the 1PL model, Figure 2.2 illustrates that none of the logistic models has any of them. The common thing they provide are measurements of a unidimensional variable on a scale with an invariant order of the test takers.

The 1PL model does have one invariant property in addition to those shared with the other logistic models though—similar order of its response functions across all possible reparameterizations. This feature of noncrossing functions, which is not a prerequisite for the identical ordering of the test takers, definitely allows for an easier choice of the response probability necessary to create an item map, in the sense that each choice leads to the same ordering of the item content in the map. For the other two logistic models, the order may change locally as a result of crossing response functions, a fact sometimes difficult to understand by consumers with lack of statistical training (for more on this issue, see Lewis and Bessen-Lord, Volume Three, Chapter 11).

The 3PL does offer free parameters to account for the possibility of guessing on test items. However, though empirical estimates of these parameters do a better job of catching the effects of guessing than versions of the model with these parameters set to zero, the assumption of knowledge-or-random guessing on which they are based is definitely too simple to represent all the subtle processes that may go on in test takers who are uncertain about their answers (von Davier, 2009).

As the 3PL is more flexible than the two other logistic models, it does have a higher likelihood of fitting a given response matrix. But it does not automatically follow that its ability parameter estimates are therefore always better. For smaller datasets (short test lengths and/or few test takers), its greater flexibility actually make it adjust to random peculiarities in the data that better be avoided. The statistical issue at play here is the bias-accuracy trade-off fundamental to all statistical modeling: For a given dataset, if we increase the flexibility of a model by adding more parameters, the likelihood of bias in their estimates as a result of misfit decreases, but always at the price of a greater inaccuracy due to a smaller number of observations per parameter. For small datasets, it may therefore pay off to choose a more constrained model as the 1PL model. Lord caught this point exactly when he gave his 1983 article the title "Small *N* Justifies Rasch Model."

Acknowledgments

The idea of a test of body height has been circulating for quite some time in the Netherlands. As far as I have been able to trace back, the original idea is due to Frits Zegers, who introduced an early version of the test as a teaching tool in his test theory class in the 1970. Klaas Sijtsma adjusted and expanded the set of items in the 1980s. Henk Kelderman was so generous to give me a copy of the dataset, which he had collected more recently using students in one of his test theory classes as subjects. Although classical item analyses for the test have been distributed, I am not aware of any IRT analysis of it. I thank all my colleagues for the opportunity they have given me to use the items and dataset for the example in this chapter.

References

Barton, M. A. and Lord, F. M. 1981. An upper asymptote for the three-parameter logistic item-response model. Research Bulletin, 81–20. Princeton, NJ: Educational Testing Service.

Birnbaum, A. 1968. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical Theories of Mental Test Scores* (pp. 392–479). Reading, MA: Addison-Wesley.

Bock, R. D. and Aitkin, M. 1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.

Bock, R. D. and Lieberman, M. 1970. Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179–197.

De Boeck, P. 2008. Random item IRT models. *Psychometrika*, 73, 533–599.

Eggen, T. J. H. M. 2000. On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika*, 65, 337–362.

Hand, D. J. 2004. *Measurement Theory and Practice: The World through Quantification*. London: Oxford University Press.

Lord, F. M. 1952. A theory of test scores. *Psychometric Monograph No. 7*. Richmond, VA: Psychometric Corporation. Retrieved from: http://www.psychometrika.org/journal/online/MN07.pdf

Lord, F. M. 1983. Small N justifies Rasch model. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing* (pp. 51–61). New York: Academic Press.

Maris, G. 2002. Concerning the identification of the 3PL model (Measurement and Research Department Reports 2002-3). Arnhem, The Netherlands: Cito.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Rasch, G. 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press.

San Martín, E., Del Pino, G., and De Boeck, P. 2006. IRT models for ability-based guessing. *Applied Psychological Measurement*, 30, 183–203.

Stevens, S. S. 1946. On the theory of scales of measurement. *Science*, 103, 677–680.

van der Linden, W. J. 2016. Identifiability restrictions for the fixed-effects 3PL model. *Psychometrika*, 81, in press.

van der Linden, W. J. and Barrett, M. D. 2016. Linking item response model parameters. *Psychometrika*, 81, in press (doi: 20.1007/211336-015-9496-6).

von Davier, M. 2009. Is there a need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspectives*, 7, 110–114.

Wilson, M. 2005. *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R., and Bock, R. D. 2003. *BILOG-MG: Multiple Group IRT Analysis and Test Maintenance for Binary Items* (3rd ed.) [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.