Applied Psychological Measurement
2020, Vol. 44(3) 215–229
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621619843823
journals.sagepub.com/home/apm
SAGE

An Item Response Model for True–False Exams Based on Signal Detection Theory

Lawrence T. DeCarlo[1]

Abstract

A true–false exam can be viewed as being a signal detection task—the task is to detect whether or not an item is true (signal) or false (noise). In terms of signal detection theory (SDT), examinees can be viewed as performing the task by comparing the perceived plausibility of an item (a perceptual component) to a threshold that delineates true from false (a decision component). The resulting model is distinct yet is related to item response theory (IRT) models and grade of membership models, with the difference that SDT explicitly recognizes the role of examinees' perceptions in determining their response to an item. SDT also views IRT concepts such as "difficulty" and "guessing" in a different light, in that both are viewed as reflecting the same aspect—item bias. An application to a true–false algebra exam is presented and the various models are compared.

Keywords

item response models, signal detection theory, grade of membership, item difficulty, item guessing, item bias, true–false exams

A true–false exam is a selected-response exam where the examinees' task is to decide whether an item is true or false. One approach to analyzing the resulting data is to use an item response theory (IRT) model (Birnbaum, 1968; de Ayala, 2009; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). In the IRT approach, it is assumed that the probability that an examinee chooses the correct response depends on item parameters and on the examinee's ability, where ''ability'' is a latent continuous random variable. IRT models are basically *measurement models* in that examinees' response patterns are used to obtain information about their underlying abilities and about item characteristics. They are not *psychological models* because they say nothing about *how* examinees make decisions in true–false and other selected-response exams. This has earlier been noted, for example, by Hambleton, Swaminathan, and Rogers (1991), ''Much of the IRT research to date has emphasized the use of mathematical models that provide little in the way of psychological interpretations of examinee item and test performance'' (p. 154). Goldstein and Wood (1989) made a similar observation, ''As the title of Lord and Novick's (1968) book made clear, the theory is statistical, not psychological'' (p. 139). More recently, in a discussion of the three-parameter logistic (3PL) model, von Davier (2009) noted

[1] Columbia University, New York, NY, USA

Corresponding Author:
Lawrence T. DeCarlo, Department of Human Development, Teachers College, Columbia University, Box 118, 525 West 120th Street, New York, NY 10027-6696, USA.
Email: decarlo@tc.edu

". . . practical application of IRT models often picks a model out of tradition rather than out of considerations of how guessing or random response strategies are conceptualized" (p. 114).

The model developed here follows directly from a conceptualization about *how* examinees make decisions in true–false exams, using a model of decision-making based on signal detection theory (SDT; Green & Swets, 1988; Macmillan & Creelman, 2005; Wickens, 2002). Although the approach can be developed for selected-response exams in general, such as multiple choice exams, the focus here is on what is perhaps the simplest case, a true–false exam, which provides a simple and useful starting point. The psychological conceptualization underlying the model is presented, along with a comparison to ideas underlying traditional IRT models and mixed membership models. The statistical model that follows from the theoretical conceptualization is derived. Implications of the resulting item response signal detection theory (IRSDT) model for notions such as "guessing" and "item difficulty" are discussed. The model is applied to real-world data and the results are compared to those obtained with IRT models.

SDT

A novel aspect of the present approach is the application of SDT as a model of the decision process in true–false exams. SDT has been used as an account of psychological processes involved in decision-making for well over half a century (Green & Swets, 1988; Macmillan & Creelman, 2005; Wickens, 2002). Indeed, Estes (2002) noted that,

> Over ensuing decades, the SD model, with only technical modifications to accommodate particular applications, has become almost universally accepted as a theoretical account of decision making in research on perceptual detection and recognition and in numerous extensions to applied domains . . . This development may well be regarded as the most towering achievement of basic psychological research of the last half century. (p. 15)

Although the relevance of SDT to selected-response exams has previously been noted (Macmillan & Creelman, 2005, p. 249), details of the approach have not been developed.

Basic Concepts of SDT

A basic idea in SDT is that decisions are based on *perceptions* of presented events. The perceptions in turn can be represented by probability distributions, following ideas going back to Fechner (1860/1966) and later used by Thurstone (e.g., 1927). For example, the idea of representing perception as a probability distribution is consistent with the observation in psychophysics that, even if exactly the same stimulus is presented on different occasions, observers do not necessarily give the same response—clearly, the stimulus did not change, and so it must be the observer's *perception* of the stimulus that changed. As noted by Fechner,

> Even when applied in the same way, one and the same stimulus may be perceived as stronger or weaker by one subject or organ than by another, or by the same subject or organ at one time as stronger or weaker than at another. (p. 38)

Thus, a basic idea is that when an examinee encounters an item, they have a perception of the item which is used to make a decision. Furthermore, the perception is viewed as being a realization from a probability distribution. Note that even if an examinee reads the *same* item on a different occasion, then the decision is not necessarily the same, because the examinee's perception of the item can differ with each reading, which is consistent with the observation that

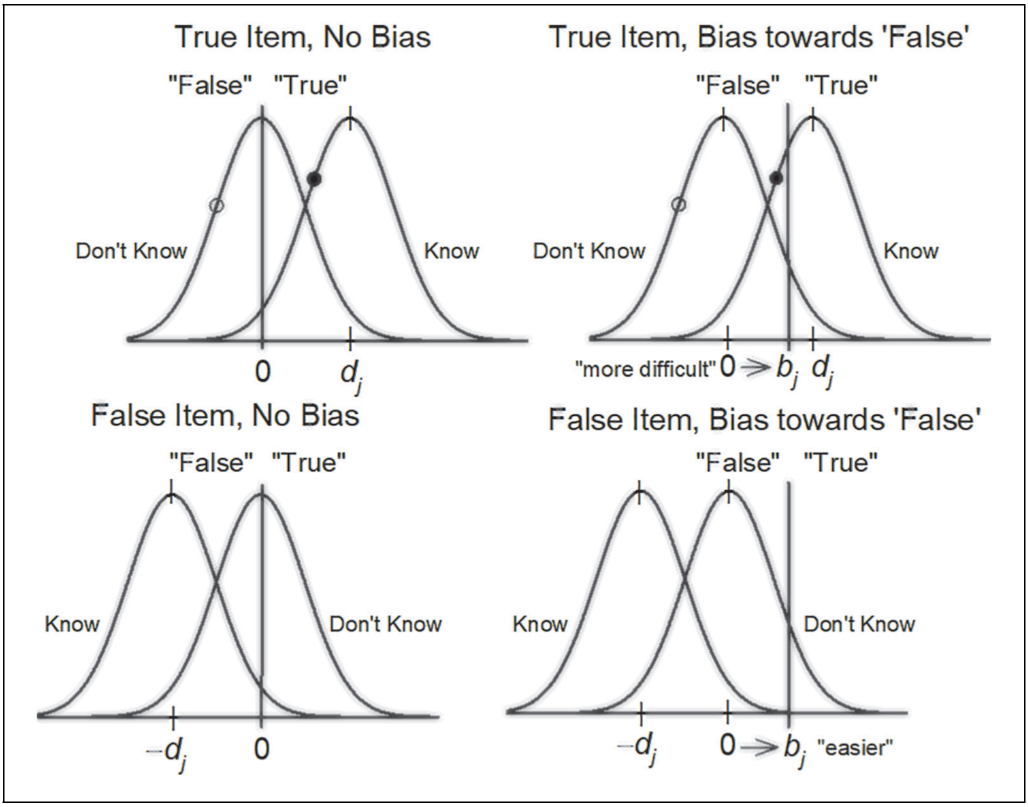

Figure 1. Signal detection theory conceptualization of a true–false test.

examinees on selected-response tests sometimes change their answers—their perception of the item differed upon a re-reading. In short, an important aspect of SDT is that it recognizes that an examinee bases his or her decision on a perception of an item (and so they can get it wrong even if they ''know'' the item).

Figure 1 presents the basic ideas of SDT as applied to true–false exams. The top two panels illustrate the theory for a true item without bias (left panel) and with bias (right panel). For the left panel, the probability distribution on the left represents a perceptual distribution of *plausibility* for examinees who do not know an item. The second distribution, to the right of the first, shows that, if an examinee knows an item, then the plausibility distribution is shifted to the right by d_j. This simply reflects that a true item will appear to be more plausible to examinees who know the item, and so they have a higher probability of deciding true. The distance parameter d_j reflects how well the item discriminates between those who know and do not know the item.

Next, examinees make a decision by comparing their perception of the item's plausibility to a *decision criterion* that delineates true from false, shown by a vertical line at zero in the top left panel of Figure 1. If an examinee perceives the plausibility of an item as being above the criterion, then the decision is that the item is ''true,'' otherwise, the decision is ''false.'' Note that locating the criterion at zero means that there is no *item bias*, in that the probability of a decision of true or false for an examinee who does not know the item is .50, which is similar to the idea of ''guessing'' in IRT. In SDT, ''guessing'' refers to the situation where an examinee does not know an item, and so guessing is not a separate process, in that the decision process is *exactly the same* irrespective of whether an examinee knows an item or not—if the perceived plausibility of the item is above the criterion, then the decision is ''true,'' otherwise, the decision is ''false.'' Thus, a basic difference between SDT and IRT is that guessing is not a separate process in SDT, but involves the same process as when ''not guessing.''‘Guessing’ can be defined in SDT as the probability of a correct response for an examinee who does not know an item, which is determined by the bias.

Also shown in Figure 1 are two possible realizations, shown as circles, from the plausibility distributions for an examinee who does not know an item (open circle) and one who does (filled circle). In the first case, an incorrect response is given because the open circle is below the criterion, and so a response of ''false'' is given to the true item, which is a ''miss.'' In the second case, a correct response of ''true'' is given, because the filled circle is above the criterion, which is a ''hit.''

The top right panel of Figure 1 shows the effect of *item bias*—the criterion is not located at zero, but is shifted, in this case to the right by b_j, as shown by the arrow. This means that examinees who either know or do not know an item both have a higher probability of deciding ''false,'' and so there is a bias toward a response of ''false.'' The right panel of Figure 2 shows that, because the item is in fact true, the probability of a correct response is lower, regardless of whether one knows the item or not. Thus, the presence of bias toward a response of false (positive bias) makes the item appear to be ''more difficult'' compared to the situation without bias (left panel). Also note that, for examinees who do not know an item, the probability of getting it correct is now considerably less than .50, and so the ''guessing'' probability is well below chance. Thus, the SDT approach can conceptually handle guessing probabilities that are well below (or above) chance, whereas this is neither predicted nor expected in the IRT view. Furthermore, the same mechanism (bias) that leads to the item being ''difficult'' also leads to low ''guessing'' rates, and vice versa.

The bottom panels of Figure 1 illustrate the theory for a false item with and without bias. In the lower left panel, the plausibility distribution for examinees who do not know the item is again located at zero, and so the probability of a decision of ''true'' or ''false'' is .50, which is the ''no bias'' situation. Also note that, for a false item, the plausibility distribution for examinees who know the item is shifted to the left, which reflects that a false item appears to be *less* plausible, and so examinees who know the item are more likely to correctly recognize that the item is false.

The lower right panel of Figure 1 shows the situation for a false item again with bias toward a response of ''false,'' shown by a rightward shift of the criterion b_j, as in the example above. The probability of a correct decision is now higher for examinees both who know and do not know the item, and so the item appears to be ''easier'' compared to the situation with no bias (left panel). For examinees who do not know the item, the probability of getting the item correct is now well above .50, and so, because of the bias, they have an above chance probability of getting the item correct.

In summary, ''difficulty'' in the SDT view depends on the decision criterion, which in turn reflects bias. As shown in Figure 1, for example, a bias toward a response of ''false'' will make a true item appear to be more difficult and a false item appear to be easier. Note that one can obtain a simple measure of item difficulty in SDT using $b_j{*}Z$, where Z is a true–false indicator coded as 1 for true items and -1 for false items, with the result that a larger value of $b_j{*}Z$ indicates a more difficult item. Second, guessing in SDT is not viewed as being a separate process but instead involves the same process that is involved when not guessing—the decision is based on whether or not the perceived plausibility of an item is above or below the decision criterion. Thus, ''item bias'' in SDT accounts for both ''item difficulty'' and ''item guessing.'' As a result, an additional parameter is not needed to account for ''guessing'' in SDT, in contrast to the 3PL model of IRT. This allows one to maintain desirable properties of the two-parameter logistic (2PL) model and also deal with guessing without the additional complications introduced by including an additional parameter (see below). Finally, SDT is consistent with "guessing" probabilities that are either well above or below chance, whereas well above or below chance guessing probabilities for true–false exams are not predicted or expected for the 3PL model; if it is random guessing, should not the probability of a correct response be around .5?

From Theory to Model

As noted above, SDT identifies two basic aspects of the situation—examinees' perceptions of items and their use of decision criteria. Let Y_{ij} represent the response of the ith examinee to the jth item with values $y_{ij} = 1$ for a response of "true" and $y_{ij} = 0$ for "false." Let Ψ_{ij} be a random variable that represents the perceived plausibility of an item to an examinee. The *decision rule* is to respond "true" if the plausibility is above the criterion and "false" if it is below the criterion:

$$\begin{aligned} Y_{ij} &= 1, \quad \text{if } \Psi_{ij} > b_j \\ Y_{ij} &= 0, \quad \text{if } \Psi_{ij} \leq b_j, \end{aligned} \tag{1}$$

where b_j is the item bias, as illustrated in Figure 1. Note that it should be kept track of *which option* was chosen, true or false, and not simply whether the choice was correct or incorrect (the information is the same for true–false tests, but it is lost for tests with more than two alternatives).

Next, the *structural model* (DeCarlo, 2010) relates the examinee's perception of the item to its known status (i.e., true or false). Let Z_j indicate the item's actual status as true ($Z_j = 1$) or false ($Z_j = -1$); note that the coding gives the rightward (for true) or leftward (for false) distribution shift shown in Figure 1. Let δ_{ij} be a latent dichotomous variable that indicates whether or not examinee i knows item j, with a value of one indicating yes and zero indicating no. The structural model is as follows:

$$\Psi_{ij} = \delta_{ij} d_j Z_j + \varepsilon_{ij}, \tag{2}$$

where ε_{ij} is a random variable that represents variation in the perception and $\delta_{ij} \sim \text{Bernoulli}(\lambda_i)$. Equation 2 gives the two distributions of plausibility shown in Figure 1, with a shifted distribution for examinees who know the item ($\delta_{ij} = 1$) versus those who do not know ($\delta_{ij} = 0$), with d_j indicating the amount of shift; Equation 3 is a simple generalization of the structural model given in DeCarlo (2010), with the addition of δ_{ij}.

Some simplifying assumptions are made. For example, discrimination and bias are conceived of as being item characteristics (this assumption can be relaxed, as shown in the supplementary material). Second, it is assumed that the probability that examinee i has knowledge about item j is constant across the j items, $p(\delta_{ij} = 1) = \lambda_i$, which is analogous to the assumption of *parallel items* in classical test theory. The items are in essence viewed as being exchangeable, in that the probability that an examinee knows any particular item for a given test is the same across all items on the test. The examinee parameter λ_i can be viewed as being the proportion of items that an examinee knows for a given test, which is analogous to Lord's definition of the true score ς in the compound binomial model (Lord, 1965). As noted below, λ_i is also analogous to membership scores used in the grade of membership (GoM) model (Erosheva, 2002).

It follows from the decision rule and structural model that the conditional probability that an examinee responds "true" is as follows:

$$p\left(Y_{ij} = 1 | \delta_{ij}, b_j, d_j, Z_j\right) = p\left(\Psi_{ij} > b_j | \delta_{ij}, b_j, d_j, Z_j\right) = p\left(\delta_{ij} d_j Z_j + \varepsilon_{ij} > b_j\right) = p\left(\varepsilon_{ij} > b_j - \delta_{ij} d_j Z_j\right).$$

Different probability distributions can be used for ε_{ij}, such as the normal, logistic, or extreme value distributions; note that a scale indeterminancy is removed by fixing the variance of ε_{ij}. Assuming logistic distributions gives,

$$p_{ij} = p\left(Y_{ij} = 1 | \delta_{ij}, b_j, d_j, Z_j\right) = \frac{e^{-b_j + \delta_{ij} d_j Z_j}}{1 + e^{-b_j + \delta_{ij} d_j Z_j}}, \tag{3}$$

which is linearized using the logit transform,

$$\text{logit } \left(p_{ij}\right) = -b_j + \delta_{ij} d_j Z_j.$$

The above equation is an SDT model for item responses; it is referred to here as the IRSDT model and is a simple generalization of the usual logistic version of the signal detection model (DeCarlo, 1998) with the addition of a latent dichotomous variable, δ_{ij}. Equation 3 shows that if examinee i knows item j, δ_{ij} =1, then the decision depends on both the item bias b_j and item detection d_j. If the examinee does not know the item, δ_{ij} =0, then the decision only depends on the item bias. Note that the bias and discrimination parameters also have a statistical interpretation as a *log odds* and a *log odds ratio*, respectively,

$$\begin{aligned} b_j &= -\text{logit } p\left(Y_{ij} = 1 | \delta_{ij} = 0, b_j, d_j, Z_j\right) \\ d_j &= \text{logit } p\left(Y_{ij} = 1 | \delta_{ij} = 1, b_j, d_j, Z_j\right) - \text{logit } p\left(Y_{ij} = 1 | \delta_{ij} = 0, b_j, d_j, Z_j\right). \end{aligned}$$

It is also assumed that examinees are independent and that their responses are locally independent given λ_i. Written as a hierarchical Bayesian model, the IRSDT model is,

$$\begin{aligned} &Y_{ij} | \delta_{ij}, b_j, d_j, Z_j \sim \text{Bernoulli}\left(p_{ij}\right) \\ &\delta_{ij} | \lambda_i \sim \text{Bernoulli}(\lambda_i) \\ &\lambda_i \sim \text{Beta}(\upsilon, \omega). \end{aligned}$$

Note that the above equation is also related to a latent class representation of the GoM model given by Haberman (1995) and to data augmentation approaches used in Bayesian estimation (Tanner, 1996).

The remaining parameters and hyperparameters are as follows:

$$\begin{aligned} &\upsilon, \omega \sim \text{lognormal}(0, 1) \\ &d_j \sim \text{lognormal}(0, 1) \\ &b_j \sim \text{Normal}(0, 9). \end{aligned}$$

A lognormal distribution is used for d_j so that it can only take on zero or positive values, which is a *monotonicity* constraint; the constraint ensures that an examinee who knows an item has the same or higher probability of answering correctly as an examinee who does not know the item. In terms of Figure 1, monotonicity ensures that the "know" distribution is shifted to the right for true items and to the left for false items (the true–false indicator Z_j makes the negative d_j positive). Restricting d_j to positive values also helps to prevent a label switching problem that arises in latent class models, in that the switched solution has a negative d_j, which the monotonicity constraint prevents. Other options for d_j are to use a normal distribution truncated at zero, or a gamma distribution.

The model as specified above can be fit using Bayesian estimation. For the data analyzed below, first examined were situations where a distribution for λ_i was specified, such as uniform, λ_i ~ Beta (1, 1); also examined were situations where the model was specified with arbitrary shape parameters for λ_i, that is λ_i ~ Beta (v, w); some simulations that examine parameter recovery in several situations are presented in the supplementary material.

The IRSDT model can also be written as follows:

$$p\left(Y_{ij} = 1 | \lambda_i, b_j, d_j, Z_j\right) = \lambda_i \frac{e^{-b_j + d_j Z_j}}{1 + e^{-b_j + d_j Z_j}} + (1 - \lambda_i) \frac{e^{-b_j}}{1 + e^{-b_j}}. \tag{4}$$

That is, the probability of a response of ''true'' is a mixture of the probability of knowing and not knowing an item, along with hit and false alarm probabilities. Equation 4 shows that the IRSDT model is a *probabilistic mixture model* with a random mixing parameter λ_i. It can be viewed as a type of generalized latent class model—note that if λ_i is restricted to take on values of only zero or one, and so examinees are assumed to either know or not know all of the items, then the IRSDT model reduces to a simple latent class model. Given this relation, it is informative to compare results for the simple latent class model to those obtained for the IRSDT model.

GoM Model

The IRSDT model is closely related to the GoM model (Erosheva, 2002, 2005). The GoM is often written as follows:

$$p\left(Y_{ij}=1|\lambda_{jk},g_{ik}\right)=\sum_{k=1}^{K}\lambda_{jk}g_{ik},$$

with $\sum_k g_{ik}=1$ and $0\leq g_{ik}\leq 1$, where K is the number of latent classes, λ_{jk} are structural coefficients (item parameters), and g_{ik} are GoM scores. Note that the above is equivalent to Equation 4 with K = 2, g_{i1} = $(1-\lambda_i)$, g_{i2} = λ_i, λ_{j1} = expit $(-b_j)$, and λ_{j2} = expit $(-b_j+d_j\ Z_j)$, where the *expit function* is $e^{\mathrm{a}}/(1+e^{\mathrm{a}})$. Thus, for K = 2, the IRSDT model can be viewed as a reparameterized version of the GoM model with λ_i of IRSDT being analogous to the grade of membership in classes of knowing or not knowing all of the items in the GoM. An interesting consequence of this view is that it shows that the GoM item parameters can be transformed and interpreted as signal detection parameters:

$$b_j=-\ln\frac{\lambda_{j1}}{1-\lambda_{j1}},\quad d_j=\left(\ln\frac{\lambda_{j2}}{1-\lambda_{j2}}-\ln\frac{\lambda_{j1}}{1-\lambda_{j1}}\right)Z_j.$$

It also follows that GoM software can be used to fit a version of the IRSDT model (with maximum likelihood), with parameters transformed as shown above. Some small simulations using the R package SIRT (Robitzsch, 2018) to fit the nonparametric GoM model (with 16 points for g_{ik}, the maximum) showed good recovery of the IRSDT parameters.

Note that a difference between the models, however, is that Equation 4 is a probabilistic mixture model, in that λ_i is the probability that an examinee knows an item, whereas membership scores in the GoM model are not probabilities, but are measures of the distance from the extremal categories (in this case whether an examinee knows all of the items or not).

IRT

The left side of Figure 2 shows another representation of the IRSDT model. In the first branch, it is assumed that an examinee either knows or does not know an item, with probabilities λ_i or $1-\lambda_i$, respectively. In the second branch, the decision depends on an examinee's perception of an item, along with the item parameters. Adding the two branches shown in Figure 2, multiplied by their weights, gives the probability of a choice of ''true'' (Y_{ij} = 1) as given by Equation 4.

Although IRT models are commonly derived directly as measurement models, a ''psychological motivation'' for the 3PL model has been noted by several authors (Birnbaum, 1968, p. 404; San Martín, del Pino, & De Boeck, 2006). The motivation offers a useful comparison to the SDT approach and shows similarities and differences between the approaches.

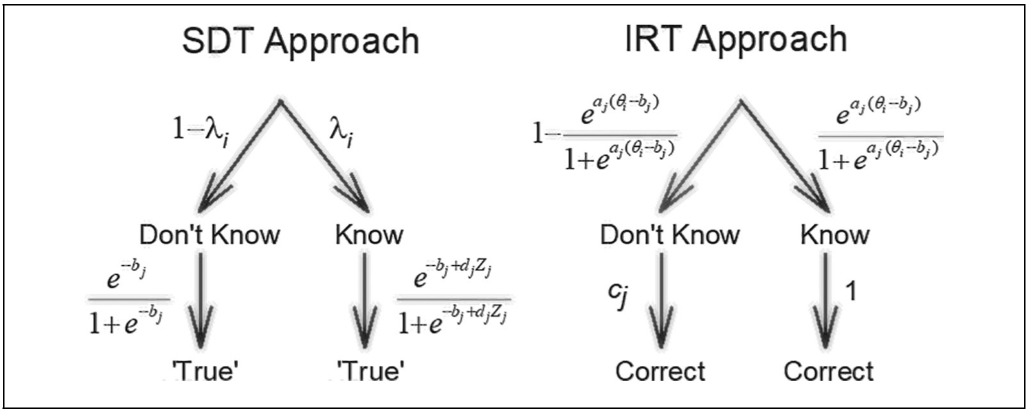
SDT Approach
IRT Approach
$1-\lambda_i$
λ_i
Don't Know
Know
$\frac{e^{-b_j}}{1+e^{-b_j}}$
$\frac{e^{-b_j+d_jZ_j}}{1+e^{-b_j+d_jZ_j}}$
'True'
'True'
$1-\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}$
$\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}$
Don't Know
Know
c_j
1
Correct
Correct

Figure 2. Decision structures for SDT and IRT approaches.
Note. SDT = signal detection theory; IRT = item response theory.

The right side of Figure 2 shows basic ideas underlying the 3PL IRT model. As before, it is assumed that an examinee either knows or does not know an item, as indicated by δ_{ij}. The first branch shows the probability that an examinee knows an item is a function of both examinee characteristics and item characteristics:

$$p(\delta_{ij}=1)=\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}},$$

where a_j and b_j are item discrimination and difficulty parameters, respectively, and θ_i is the ability of examinee i. The second branch shows that it is assumed that if an examinee knows an item, then they answer it correctly, and so the probability of a correct response is unity. However, if an examinee does not know an item, then the probability of a correct response is given by c_j, which is an item parameter that reflects ''guessing.'' For example, in a true–false exam, the probability of a correct response should be around .5 if an examinee simply ''guesses.'' The parameter c_j is often referred to as the ''pseudo-guessing'' parameter because values that are less than chance are often found in practice (Lord, 1980).

Multiplying by the weights and adding, it follows from the two branches shown in Figure 2 that the probability of a correct response for a given examinee is,

$$p(Y_{ij}=\text{correct}|\theta_i,a_j,b_j)=1\times\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}+c_j\times\left(1-\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}\right),$$

which can be rearranged to give the usual 3PL model,

$$p(Y_{ij}=\text{correct}|\theta_i,a_j,b_j)=c_j+(1-c_j)\frac{e^{a_j(\theta_i-b_j)}}{1+e^{a_j(\theta_i-b_j)}}. \tag{5}$$

Note that the response Y_{ij} is usually coded as 1 or 0 to indicate whether a choice was correct or incorrect, but ''correct'' is used explicitly in the above (instead of 1) to distinguish Y_{ij} from that used in the IRSDT model, where 1 indicates a response of ''true.'' There are well-known estimation problems associated with the 3PL (see Baker, 1992) as well as identification issues for various versions of the model (Maris & Bechger, 2009; San Martin, Gonzalez, & Tuerlinckx, 2015; Wu, 2016). A common approach to estimation problems is to set c_j to a fixed value or to restrict c_j to be equal across the J items. There are also other ways to derive the model (San Martín et al., 2006), but the current derivation is useful for a comparison to the IRSDT approach.

Setting $c_j = 0$ in Equation 4 gives the 2PL model,

$$p\left(Y_{ij} = \text{correct}|\theta_i, a_j, b_j\right) = \frac{e^{a_j\left(\theta_i - b_j\right)}}{1 + e^{a_j\left(\theta_i - b_j\right)}}.$$

In terms of Figure 2, the model implies that if an examinee does not know an item, then they get the item incorrect. Thus, both branches of the second stage of the 2PL are deterministic: if an examinee knows an item, then he or she gets it correct; if the examinee does not know an item, then he or she gets it incorrect. The use of the 2PL for true–false exams is conceptually questionable, because ''guessing'' clearly gives a chance of being correct that is greater than zero, and so the assumption that $c_j = 0$ is not appropriate, although one can of course still fit the model (or fix c_j to 0.5 in the 3PL). However, the IRSDT model considers guessing to be a consequence of response bias and so it has no difficulties dealing with guessing with only two item parameters and can also deal with above and below ''chance'' levels of guessing.

Also note that IRT does not consider the role of perception in decision-making, which is why at least one branch in the second stage on the right side of Figure 2 is deterministic—if an examinee knows an item, then he or she gives the correct answer. In contrast, even if an examinee knows an item, IRSDT recognizes that the examinee can still give an incorrect answer, due to the role of perception in decision-making (e.g., the examinee can misperceive the item in some way), along with effects of item bias and discrimination.

Application to a True–False Algebra Exam

Given that the IRSDT model is new, a comparison of results across IRSDT and IRT models for real-world data is of primary interest. The example is an algebra exam that was developed by the author years ago, with items based on observed algebraic mistakes made by students in courses on measurement and statistics. The exam is used as a screening instrument in these courses to help identify students who might lack algebraic skills needed for the course; use of the exam has also been approved by the institution's institutional review board (IRB). The exam consists of 16 true–false items and is shown in the supplementary material; half of the items are true and the other half are false. Students were instructed to not leave any items blank and were given about 20 min to complete the exam. Results for 829 students who took the exam are analyzed.

The IRSDT and IRT models were fit using Bayesian estimation, implemented via PROC MCMC of SAS (using the Metropolis–Hastings algorithm). About 20,000 iterations seemed adequate for convergence; however, because of the low computational time, each model was fit with 20,000 burn-ins followed by 100,000 iterations. For the IRSDT, the priors and hyperpriors were d_j ~ lognormal (0, 1), b ~ N (0, 9), and λ_i ~ Beta (v, w) with v and w ~ lognormal (0, 1); details about a random parameter version of the model, the IRSDTr model, are given in the supplementary material. For the 2PL and 3PL models, the priors used were a_j ~ lognormal (0, 1), b_j ~ N (0, 9), θ ~ N (0, 1), and c_j ~ Beta (1, 1).

Model Fit

Posterior predictive checks (PPCs; Gelman et al., 2014; Sinharay, Johnson, & Stern, 2006) are useful for comparisons of the models. For each model, 1,000 samples from the posterior predictive distributions were generated and checks at both the test level and item level were conducted. For the IRSDT models, the checks reported here are for the model with estimated shape parameters, that is, $\lambda_i \sim \text{Beta}\ (\nu, \omega)$; models with fixed shape parameters were also examined (see the supplementary material).

At the test level, a PPC of the total test score was performed, that is, the score frequencies predicted by the models over the samples were compared to the score frequencies for the observed data. Figure A1 in the supplementary material shows results for four models. The solid circles indicate the observed data, whereas the solid lines show, for the replicated samples, the 5th, 50th, and 95th percentiles (across 1,000 samples). The figure shows that the IRSDT model performs well throughout the range except for the top two scores; it predicts too many scores of 15 and too few scores of 16. The IRSDTr model corrects for this somewhat and better predicts scores of 15, but it still has too few scores of 16. The 2PL model performs poorly, with many points outside of the 90% range, whereas the 3PL is consistent with the observed score distribution. Overall, the two IRSDT models and the 3PL model appear to be adequate with respect to accounting for the observed total scores.

At the item level, a PPC of the item-total score curves was performed. Figure A2 in the supplementary material shows the observed data (solid circles) along with the 5th, 50th, and 95th percentiles for the predicted data for the IRSDT model. The figure shows the proportion correct for each item plotted against the observed total score (item fit plots; see Sinharay, 2006). Overall, the predicted curves generally include the observed data; plots for the 3PL are similar to those obtained for the IRSDT. The item fit plots show that the IRSDT model adequately accounts for the observed item-total score relationships.

Parameter Estimates

Table 1 shows parameter estimates (posterior means and standard deviations) for a fit of the IRSDT and IRT models. For the IRSDT model, a few items (e.g., 1, 5, and 7) show large negative bias, which is a bias toward a response of ''true,'' and a few items show large positive bias (e.g., 3, 4, 14, and 16), which is a bias toward a response of ''false.'' The discrimination parameters are generally in the range of 2 to 6 (with one large value, with large posterior standard deviation), which indicates good to excellent discrimination. The posterior standard deviations for the bias parameter are fairly small, whereas those for discrimination are larger. The estimated shape parameters for the beta distribution of lambda are 0.33 and 0.31, which indicates a bimodal distribution. For the IRSDTr model, the right side of the table shows that the parameters are generally consistent with those found for the IRSDT model, with similar bias but with smaller discrimination estimates. The variance of the bias parameter is small (0.10), whereas that for the discrimination parameter is considerably larger (2.33).

For a fit of the 3PL model, shown in the lower half of Table 1, five items (4, 5, 8, 9, and 11) show very low probabilities of guessing correctly, less than .26, whereas three items (1, 6, and 14) have high guessing probabilities, larger than .63. Given that guessing should give a probability of a correct response of around .50 in a true–false exam, it is not clear why some items have guessing probabilities that are considerably higher or lower.

It is informative to examine the relation of the parameter estimates across the different models; Table 2 shows Spearman's correlations and p-values. The top part of the table shows that the difficulty measure of IRSDT (b_jZ) is highly correlated (.89) with the difficulty parameter b_j of the 2PL, and so both models rank order the item difficulties similarly. The IRSDT discrimination parameter d_j is also highly correlated (.81) with the 2PL discrimination parameter a_j. Thus, the IRSDT and 2PL models lead to similar conclusions about item difficulty and item discrimination. The top right side of Table 2 shows that the b_j and a_j across the 2PL and 3PL models are not significantly correlated, and so the 2PL and 3PL models lead to different conclusions about

Table 1. Posterior Means and Standard Deviations for Algebra Data, IRSDT, and IRT Models.

IRSDT						IRSDTr					
	PM	PSD		PM	PSD		PM	PSD		PM	PSD
b_1	−1.10	0.13	d_1	2.69	0.50	b_1	−1.19	0.14	d_1	2.04	0.31
b_2	−0.18	0.12	d_2	4.60	0.87	b_2	−0.35	0.16	d_2	2.11	0.31
b_3	1.44	0.13	d_3	6.10	0.75	b_3	0.98	0.15	d_3	4.85	0.53
b_4	1.78	0.22	d_4	3.11	0.33	b_4	1.71	0.18	d_4	2.32	0.26
b_5	−1.19	0.15	d_5	3.20	0.30	b_5	−1.30	0.19	d_5	2.60	0.24
b_6	0.45	0.11	d_6	2.52	0.37	b_6	0.57	0.18	d_6	1.88	0.34
b_7	−1.12	0.12	d_7	10.56	2.62	b_7	−0.58	0.16	d_7	5.21	0.28
b_8	0.33	0.12	d_8	6.00	0.83	b_8	−0.42	0.17	d_8	5.91	0.35
b_9	0.82	0.12	d_9	5.16	0.92	b_9	0.44	0.17	d_9	3.75	0.37
b_{10}	−0.43	0.11	d_{10}	4.48	0.90	b_{10}	0.16	0.17	d_{10}	4.69	0.38
b_{11}	0.18	0.12	d_{11}	4.89	0.74	b_{11}	−0.34	0.21	d_{11}	3.99	0.36
b_{12}	−0.15	0.11	d_{12}	2.43	0.36	b_{12}	−0.42	0.17	d_{12}	1.43	0.28
b_{13}	0.35	0.12	d_{13}	3.30	0.46	b_{13}	0.94	0.35	d_{13}	3.59	0.36
b_{14}	0.81	0.12	d_{14}	2.11	0.40	b_{14}	1.20	0.20	d_{14}	1.17	0.30
b_{15}	−0.71	0.12	d_{15}	2.81	0.29	b_{15}	−0.29	0.17	d_{15}	1.44	0.27
b_{16}	1.04	0.19	d_{16}	5.36	0.71	b_{16}	0.95	0.18	d_{16}	2.86	0.30
ν	0.33	0.03	ω	0.31	0.02	ν	0.50	0.06	ω	0.25	0.02
						σ_b^2	0.10	0.03	σ_d^2	2.33	0.32

	2PL					3PL								
	PM	PSD		PM	PSD		PM	PSD		PM	PSD		PM	PSD
b_1	−2.51	0.32	a_1	0.84	0.13	b_1	0.14	0.22	a_1	2.44	0.57	c_1	0.75	0.04
b_2	−1.24	0.11	a_2	1.31	0.15	b_2	−0.03	0.10	a_2	3.34	0.40	c_2	0.53	0.04
b_3	−1.67	0.15	a_3	2.16	0.33	b_3	−1.35	0.20	a_3	2.14	0.27	c_3	0.37	0.13
b_4	0.12	0.07	a_4	1.29	0.12	b_4	0.57	0.07	a_4	3.40	0.45	c_4	0.22	0.03
b_5	−0.27	0.08	a_5	1.33	0.13	b_5	0.30	0.11	a_5	2.43	0.31	c_5	0.26	0.04
b_6	−1.85	0.26	a_6	0.80	0.12	b_6	0.28	0.17	a_6	2.08	0.38	c_6	0.63	0.04
b_7	−2.03	0.24	a_7	1.23	0.19	b_7	−1.11	0.33	a_7	1.55	0.30	c_7	0.48	0.13
b_8	−1.12	0.08	a_8	1.98	0.22	b_8	−0.95	0.12	a_8	2.25	0.27	c_8	0.12	0.08
b_9	−1.43	0.11	a_9	1.83	0.23	b_9	−1.15	0.22	a_9	1.93	0.26	c_9	0.25	0.12
b_{10}	−1.37	0.12	a_{10}	1.36	0.15	b_{10}	−0.81	0.31	a_{10}	1.73	0.30	c_{10}	0.30	0.15
b_{11}	−1.00	0.08	a_{11}	2.34	0.27	b_{11}	−0.76	0.10	a_{11}	2.89	0.29	c_{11}	0.16	0.07
b_{12}	−1.49	0.23	a_{12}	0.79	0.12	b_{12}	0.39	0.13	a_{12}	2.53	0.36	c_{12}	0.57	0.03
b_{13}	−0.88	0.10	a_{13}	1.11	0.12	b_{13}	0.33	0.08	a_{13}	4.78	0.96	c_{13}	0.49	0.03
b_{14}	−2.31	0.31	a_{14}	0.74	0.11	b_{14}	0.54	0.15	a_{14}	3.20	0.39	c_{14}	0.73	0.02
b_{15}	−0.56	0.07	a_{15}	1.05	0.12	b_{15}	0.39	0.09	a_{15}	2.74	0.25	c_{15}	0.38	0.03
b_{16}	−0.52	0.07	a_{16}	1.64	0.16	b_{16}	0.14	0.07	a_{16}	4.51	0.45	c_{16}	0.33	0.03

Note. For IRSDT, a positive value of b_j indicates a bias toward a response of "false." IRSDT = item response signal detection theory; IRT = item response theory; IRSDTr = random parameter version of IRSDT; PM = posterior mean; PSD = posterior standard deviation; 2PL = two-parameter logistic; 3PL = three-parameter logistic.

item difficulty and discrimination. For example, according to the 2PL, Items 1 and 14 are the "easiest" items (see Table 1), which is also consistent with the IRSDT estimates of b_j*Z, and so conclusions about relative difficulty are consistent across IRSDT and the 2PL (and the GoM, though not shown), whereas estimates for the 3PL model suggest that Items 1 and 14 are relatively difficult. Furthermore, both items show high guessing rates (.75 and .73). If the items are relatively difficult, as indicated by the 3PL estimates, then why are they easy to guess? This differs from the IRSDT model, where difficult items are also more difficult to guess, because of the common mechanism (bias).

Table 2. Spearman's Correlations of Parameter Estimates for SDT and IRT Models.

	IRSDT		3PL		
Model	$\hat{b}_j * Z$	$\hat{d}_j$	$\hat{b}_j$	$\hat{a}_j$	$\hat{c}_j$
2PL					
$\hat{b}_j$	.89 (<.01)	.15 (.60)	.26 (.32)	.50 (.05)	−.65 (<.01)
$\hat{a}_j$	−.02 (.96)	.81 (<.01)	−.72 (<.01)	−.16 (.58)	−.82 (<.01)
IRSDT					
$\hat{b}_j * Z$			.62 (.01)	.65 (<.01)	−.34 (.20)
$\hat{d}_j$			−.84 (<.01)	−.28 (.29)	−.57 (.02)

Note. Two-tailed probabilities are in parentheses. GoM was fit with 16 points. SDT = signal detection theory; IRT = item response theory; IRSDT = item response signal detection theory; 3PL = three-parameter logistic; 2PL = two-parameter logistic.

Examinee Variables

Figure 3 shows the distribution of the estimates (posteriors) of the examinee variables lambda and theta across the IRSDT and IRT models. As noted above, the shape parameter estimates for the beta-distributed lambda of IRSDT indicate bimodality, and this can be seen in the top left panel of Figure 3. The figure also shows that many examinees obtained perfect scores of 16. The top right panel for the 2PL shows that theta appears normally distributed except for a cluster of high values, which reflects the influence on the posterior of the large number of perfect scores. The results suggest possibly exploring using a mixture extension for either the IRT (or IRSDT) models.

The lower panel of Figure 3 shows that estimates of lambda for the IRSDT model and theta for the 2PL model are highly correlated (Spearman's correlation of .98) and so the models tend to rank order examinees in the same way; the right panel also shows a high correlation with theta from the 3PL, with more scatter however at the lower end. Thus, although the IRSDT and IRT models follow from somewhat different conceptual frameworks, they lead to similar conclusions about examinees.

For a subset of the examinees (493 students), the numerical grade they received in the course (weighted average of exams and homework) was also available. Although algebra was only a small part of the course content, it is of interest to examine the relation of the various latent variables in IRSDT and IRT to the numerical grade obtained in the course. The Spearman's correlations between predicted values of the latent variables and observed measures were .30 for λ_i of IRSDT, .31 for θ_i of 2PL, .31 for θ_i of 3PL, and .31 for the proportion correct. The correlations are all significant ($p < .01$) and positive and are about the same size in magnitude. Thus, the correlations provide some evidence as to the validity of the examinee measure; however, they also do not support one model over another or any of the models over the simple proportion correct.

Discussion

A psychological model of perceptual and decision processes involved when examinees answer true–false items is presented in this article. SDT recognizes the role of an examinee's perception of an item as a source of variability, as shown in Figure 1, whereas perception does not have an explicit role in IRT. SDT recognizes that, even if an examinee knows an item, they might still get it wrong because they misperceive (misinterpret, misread, etc.) the item in some way. The role of perception also provides an account as to why examinees sometimes change responses (as evidenced by cross-outs and erasures on exams)—the item was perceived differently upon a

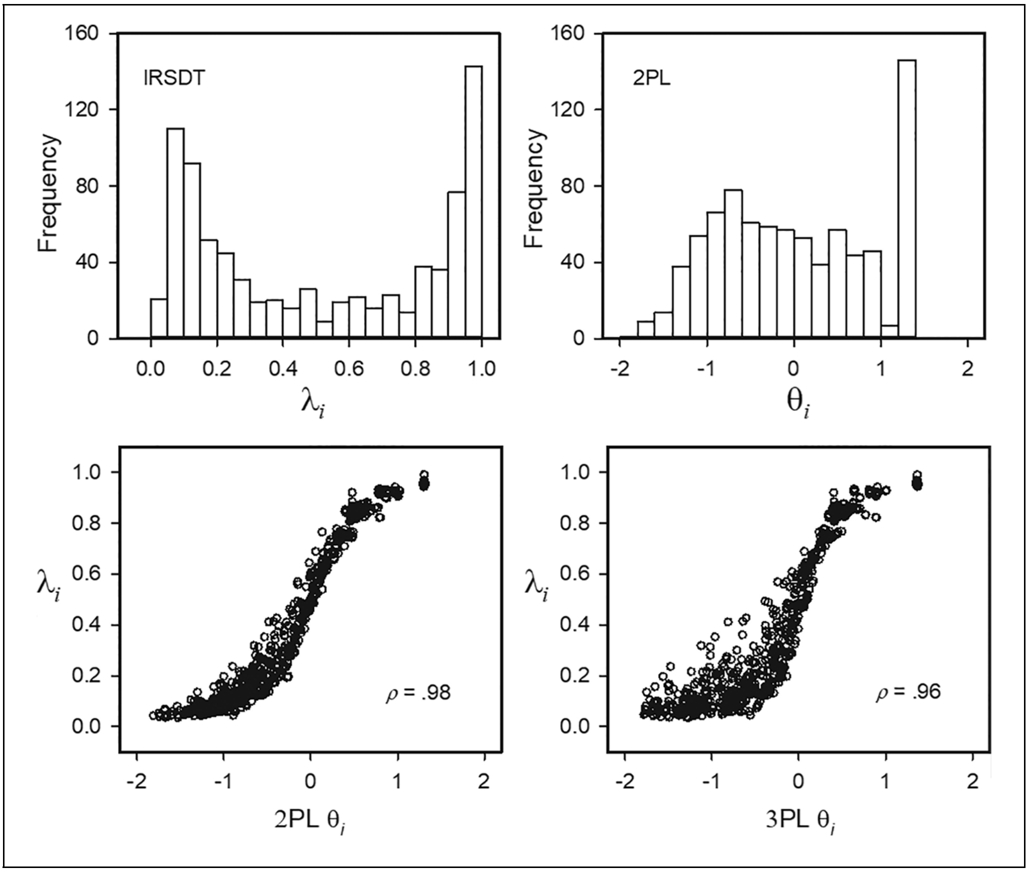

Figure 3. Distributions and relation of IRSDT and IRT examinee variable estimates.
Note. IRSDT = item response signal detection theory; IRT = item response theory.

re-reading. This variation due to perception is exactly what was recognized by Fechner (1860/1966) and others (Thurstone, 1927) in early psychological studies, and here it is noted that perception continues to play an important role when examinees are presented with items in exams. This role is explicitly recognized in the signal detection approach.

SDT also places traditional IRT concepts such as item difficulty and item guessing in a somewhat different light, in that they are not seen as separate item aspects, but rather as reflecting one aspect—item bias. As a result, the IRSDT model does not require an additional set of J parameters to account for guessing, and so the model has only two item parameters, as in the 2PL model, rather than three, as in the 3PL model. The IRSDT model also defines ‘‘guessing’’ in a simple way, namely, as the probability of getting an item correct for an examinee who does not know the item, instead of as the probability of a correct response for an examinee with (infinitely) low ability, as in the 3PL model. The model is also consistent with high or low guessing rates, given that this simply reflects item bias. In the SDT view, bias also affects examinees who know the item, thus making the item overall more ‘‘easy’’ or ‘‘difficult.’’ The IRSDT model also avoids issues with respect to inconsistencies of item difficulty and guessing that arise across the 2PL and 3PL models for true–false data, as shown above.

The IRSDT approach also offers the advantage of potentially providing insights into items. For example, for the Algebra data, Item 1 has the highest guessing rate of all items (.75), according to the 3PL estimate. What this means and what to do about it are not clear: Does it mean that the item is poor and should be eliminated or revised? From the IRSDT point of view, the bias estimate simply indicates that there was a large bias toward a response of ''True'' for Item 1, and given that the item is true, the bias leads to a high correct ''guessing'' rate for examinees who do not know the item. Thus, the high guessing rate found for the 3PL is consistent with the bias revealed by the IRSDT model. The bias in this case also implies that the item overall is ''easy,'' as found for a fit of the 2PL model, but not for a fit of the 3PL model. Thus, the IRSDT model gives a simple account of the results and shows that the item is not necessarily poor because of what appears to be high guessing in the 3PL model; in fact, discrimination is quite high for this item. Furthermore, SDT suggests *things to do*. For example, Item 1 could be re-written in a way to make it look ''falser,'' which might reduce the bias toward ''true,'' and so the guessing rate would be lower, both for the IRSDT and 3PL models. This could be tested by revising the item and seeing if and how the bias is affected, and fitting the various models. IRT does not directly suggest these types of experiments.

The model can be also extended in a straightforward way to multiple choice items using an m-alternative forced choice version of SDT (DeCarlo, 2012). In that case, the IRSDT models differ somewhat from standard IRT models, in that, for the IRSDT approach, one must keep track of which alternative was chosen, whereas simply using correct/incorrect leads to a loss of information about bias (when there are more than two alternatives). The IRSDT approach in that case is more similar to IRT models with an analysis of distractors (Penfield & de la Torre, 2008).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Lawrence T. DeCarlo https://orcid.org/0000-0001-9510-0212

Supplemental Material

Supplemental material is available for this article online.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186-205.

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*, 304-313.

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology*, *56*, 196-207.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Erosheva, E. A. (2002). *Grade of membership and latent structure models with application to disability survey data* (Doctoral dissertation), Carnegie Mellon University, Pittsburgh, PA, USA.

Erosheva, E. A. (2005). Comparing latent structures of the grade of membership, Rasch, and latent class models. *Psychometrika*, *70*, 619-628.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*, 3-25.

Fechner, G. (1860/1966). *Elements of psychophysics*. New York, NY: Holt, Rinehart and Winston.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). New York, NY: Chapman & Hall.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, *42*, 139-167.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. Los Altos, CA: Peninsula.

Haberman, S. J. (1995). Review of Statistical applications using fuzzy sets. *Journal of the American Statistical Association*, *90*, 1131-1133.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. R., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, *30*, 239-270.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement*, *7*, 75-88.

Penfield, R. D., & de la Torre, J. (2008, April). A new *response model* for *multiple*-choice *items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Robitzsch, A. (2018). Package 'sirt': Supplementary item response theory models (Version 2.6-9). Retrieved from https://cran.r-project.org/web/packages/sirt/

San Martín, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*, 183-203.

San Martin, E., Gonzalez, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*, 450-467.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*, 429-449.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298-321.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distributions and likelihood functions* (3rd ed.). New York, NY: Springer.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286.

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement*, *7*, 110-114.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford University Press.

Wu, H. (2016). A note on the identifiability of fixed-effect 3PL models. *Psychometrika*, *81*, 1093-1097.