Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 1 | Issue 1 Article 5

2020

An Intellectual History of Parametric Item Response Theory Models in the Twentieth Century

David Thissen

Lynne Steinberg

Chinese/English Journal of Educational Measurement and Evaluation
Vol.1, 23-39, Dec 2020

An Intellectual History of Parametric Item Response Theory Models in the Twentieth Century

David Thissen [a] and Lynne Steinberg [b]

[a] The University of North Carolina
[b] University of Houston

Abstract
The intellectual history of parametric item response theory (IRT) models is traced from ideas that originated with E.L. Thorndike, L.L. Thurstone, and Percival Symonds in the early twentieth century. Gradual formulation as a set of latent variable models occurred, culminating in publications by Paul Lazarsfeld and Federic Lord around 1950. IRT remained the province of theoreticians without practical application until the 1970s, when advances in computational technology made possible data analysis using the models. About the same time, the original normal ogive and simple logistic models were augmented with more complex models for multiple-choice and polytomous items. During the final decades of the twentieth century, and continuing into the twenty-first, IRT has become the dominant basis for large-scale educational assessment.

Keywords
Item response theory;
psychometrics;
test theory;
history

Item response theory (IRT) has a history that can be traced back nearly 100 years (Bock, 1997). The first quarter century was required for psychometrics to develop the three essential components of IRT: that items can be "located" on the same scale as the "ability" variable, that the "ability" variable is latent (or unobserved), and that the unobserved variable accounts for the observed interrelationships among the item responses. Another quarter century was needed to implement practical computer software to use the theory in practice. During the second half of the past century, IRT has provided the dominant methods for item analysis and test scoring in large-scale educational measurement, as well as other fields such as health outcomes measurement. In addition, IRT has become increasingly integrated into the larger context of models for behavioral and social data.

This presentation is organized in three large blocks: The first traces the development of models for dichotomous item responses (correct or incorrect for intelligence or achievement test items, yes-no or true-false for personality or attitude questions), because most of the ideas underlying IRT were first described for this simplest case. In the second block, we discuss models for polytomous item responses beginning with the famous Likert (1932) scale and its precursors. The third section provides a brief summary of the evolution of item parameter estimation methods that ultimately made IRT a practical tool for test assembly and scoring.

1 Models for Dichotomous Item Responses

1.1 The First Idea: The Normal Ogive Model

In work published in the middle of the 1920s, L.L. Thurstone provided the conceptual cornerstone and foundation upon which much of IRT has been built. In *A Method of Scaling Psychological and Educational Tests*, Thurstone (1925) proposed an analytic procedure to be applied "to test items that can be graded right or wrong, and for which separate norms are to be constructed for successive age- or grade-groups." Thurstone's inspiration involved data Cyril Burt (1922) collected using his translation into English of the Binet intelligence test questions; Burt's (1922) book contained a table of the percents of British children who responded correctly to each Binet item.

Thurstone (1925) graphed the percentage correct as a function of age for eleven of the questions in Burt's (1922) table. A modern plot of those data is in the lower panel of Figure 1, with the items identified by their numbers from

CONTACT: David Thissen. dthissen@email.unc.edu. L.L. Thurstone Psychometric Laboratory, 235 E. Cameron Avenue, Chapel Hill, NC 27599.

Burt's book: 6, 11, 19, ... Following Thurstone's example, the points for each age are located at the midpoint of each year (4.5, 5.5, 6.5, ...) because the data were grouped by age in years in the original tabulation. Thurstone was struck by the resemblance between those empirical curves and the cumulative normal (or "normal ogive"). In 1925 it would not have been easy to add fitted probit curves to the graphic, but those have been included as the dashed lines in the lower panel of Figure 1. The resemblance of the dashed curves to the solid lines, and the fact that the items were intended to measure "mental age", gave Thurstone the idea that the proportion of children responding correctly as a function of age can be thought to be like the area (or integral) of the normal density.

Thurstone expressed that idea in a separate graphic that is integrated into the upper panel of Figure 1. Thurstone wrote that the horizontal axis represents "achievement, or relative difficulty of test questions" (Thurstone, 1925, p. 437), and the curves (in the upper panel of Figure 1) are the distributions of mental age in the two groups. [1] He described the normal curve on the right as "the distribution of Binet test intelligence for seven-year-old children" (Thurstone, 1925, p. 434) to illustrate a sketch of an idea about why it is some children respond correctly to an item and others do not. The idea was to locate each item at the chronological age at which 50% of the children respond correctly. A line at that location divided the hypothetical Gaussian distribution of intelligence for each age group into two parts: The shaded area above the line represented children whose intelligence exceed the difficulty of the item and would respond correctly, and those whose intelligence was to the left of the item's location respond incorrectly. Dots on the x-axis of the upper panel of Figure 1 show the locations for nine of the Binet items in Burt's (1922) data; the shading indicates the area or proportion correct for 6- and 7-year-olds for item 35.

This idea has implications for data from two or more age groups. The shaded areas under the curves in the upper panel of Figure 1 correspond to two points on the increasing normal-ogive like curve for item 35 in the lower panel of Figure 1; those areas and points are connected by arrows in the graphic. The idea was that there were normal curves like those in the upper panel for each age group, each divided by vertical lines at the points on the axis in the upper panel, yielding the observed normal ogives of percent-correct in the lower panel.

Thurstone (1925) used these ideas to develop a method for placing test scores for several age groups on the same scale; Thurstone (1938) replaced his own method with a superior procedure. This process that has come to be known as developmental scaling or vertical linking [see Bock (1983), Patz and Yao (2007), Williams et al. (1998), or Yen and Burket (1997) for more extensive discussions of the topic].

Work with educational measurement was not central to Thurstone's program of research in the 1920s; he published more on psychological scaling, the assignment of numbers (scale values) to objects, be they physical or psychological. Thurstone's (1927) *Law of Comparative Judgment* used the idea that "response processes" (numerical values) associated with stimuli could be conceived of as normally distributed much like the two distributions in the upper panel of Figure 1, and that comparisons between them were like comparisons between random draws from those distributions. Thurstone did not connect these two threads of his own psychometric research in the 1920s, but the idea of a normally distributed latent response process value will reappear as IRT grows from the seeds Thurstone planted.

From a modern perspective, Thurstone's description lacks detail. It appears to be statistical, with the normal curves and all. However, there is no description of any sampling process, or statistical estimation. However, those concepts were not well defined at the time of Thurstone's writing, so this lack of sophistication is not surprising. IRT was not born whole. Rather, it has evolved; but a crucial conceptual component has been that test items can be located on the same scale as the construct they measure, and that this relationship may be used to quantify both.

Refining a suggestion by E.L. Thorndike et al. (1926), Percival Symonds (1929) contributed another idea to what was to become IRT with his analysis of Ayres' (1915) *Measuring Scale for Ability in Spelling*.[2] Ayres (1915) had obtained a list of the "1000 commonest words" in written English, and with the help of many grade-school teachers, collected spelling test data from children across a range of

[1] Thurstone used the terms "mental age", "achievement", and "intelligence" interchangeably.

[2] In *The Measurement of Intelligence*, E.L. Thorndike (1926) described sets of intelligence test items of similar difficulty he called "composites," and the ogival form of the relationship between percent correct on composites of increasing difficulty for groups and individuals. The items in Thorndike's composites were heterogeneous, measuring what Thorndike considered the four aspects of intelligence: (sentence) completions, arithmetical problems, vocabulary, and (following) directions. In contrast, Symonds' (1929) use of Ayres' spelling test as the example much more clearly foreshadowed the idea of sampling from a unidimensional domain. Thorndike was Symonds' graduate mentor at Columbia.

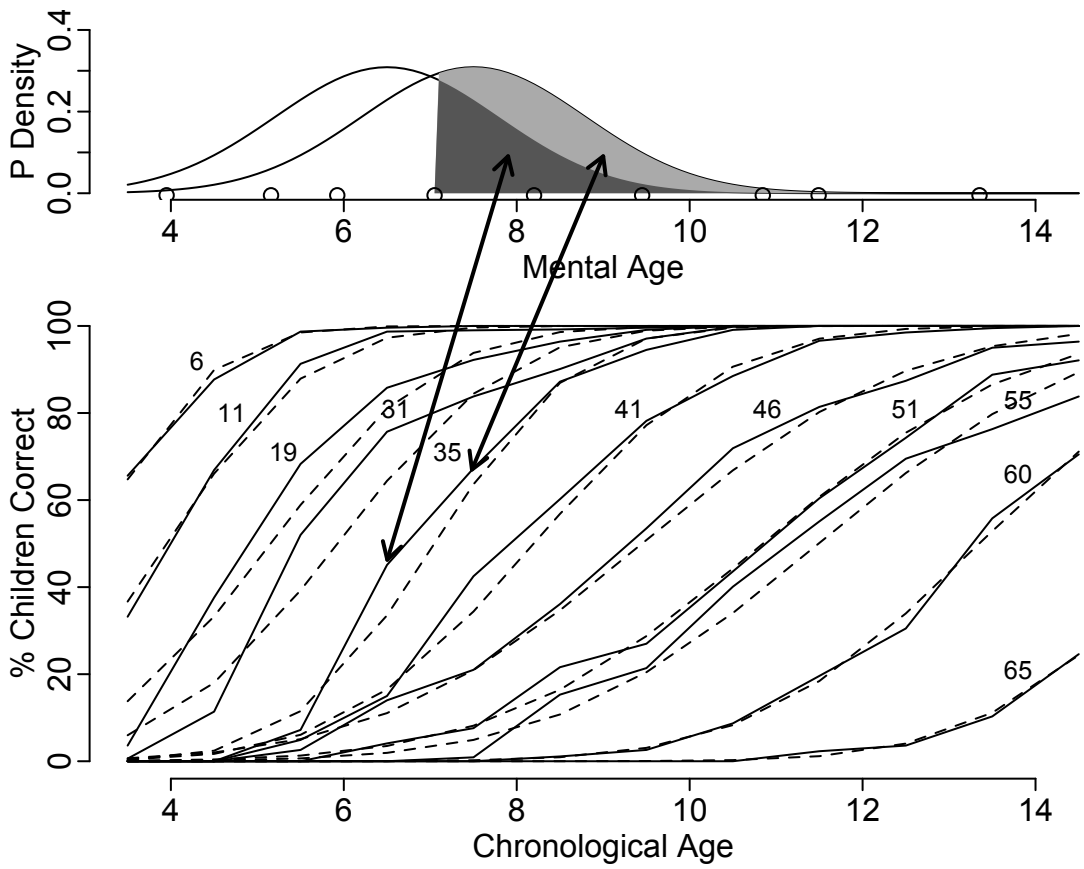

Figure 1. Upper panel: Two normal curves representing the distribution of mental age for 6- and 7-year old children [modeled after Thurstone's (1925) Figure 2], with their means at 6.5 and 7.5 years, and dots on the x-axis indicating the "location" of nine of the items. Lower panel: The observed percentages correct (solid lines) for eleven of the Binet items in Burt's (1922, pp. 132-133) data, plotted as a function of age in a graphic modeled after Thurstone's (1925) Figure 5. The arrows show the correspondence between the percentage of 6- and 7-year old children to the right of the location of item 35 and the observed percent correct. The dashed lines are cumulative normal (ogives) fitted to each of the observed solid lines.

grades in 84 cities throughout the United States. Ayres then divided the 1000 words into 26 lists, designated with the letters from *A* to *Z*. The assignment of words to lists was done by using the standard normal deviate associated the percentage of children who spelled each word correctly to put the words on lists with similar normal deviate values. List *A* was *me* and *do*, list *M* included *trust*, *extra*, *dress*, *beside*, and many other words, list *V* included *principal*, *testimony*, *discussion*, *arrangement*, and other equally difficult words, and list *Z* was *judgment*, *recommend*, and *allege*. Ayres (1915, p. 36) wrote that all the words in each list "are of approximately equal spelling difficulty," and published both the lists of words and a scoring table permitting comparison with his norming sample.

Using Ayres' spelling test as the context, Symonds (1929) described the relationship between the level of ability and the percent correct for a set of identical "tasks" or items using a graphic somewhat like that shown in Figure 2, showing (hypothetical) parallel ogives for lists *A*-*M* in Ayres' (1915) spelling test.[3]

Figure 2 is similar in some respects to the lower panel of Figure 1, but there is an important conceptual difference: Thurstone's (1925) plot (like the lower panel of Figure 1 was of the percentage of *similar children* for a constant item, whereas Symonds' (1929) plot was of the percentages of *similar items* for a child with a constant level of ability. Both conceptions recur and are sometimes confused

[3]Symonds (1929) Figure 2 was rotated 90 degrees from the modern orientation shown in Figure 2 here.

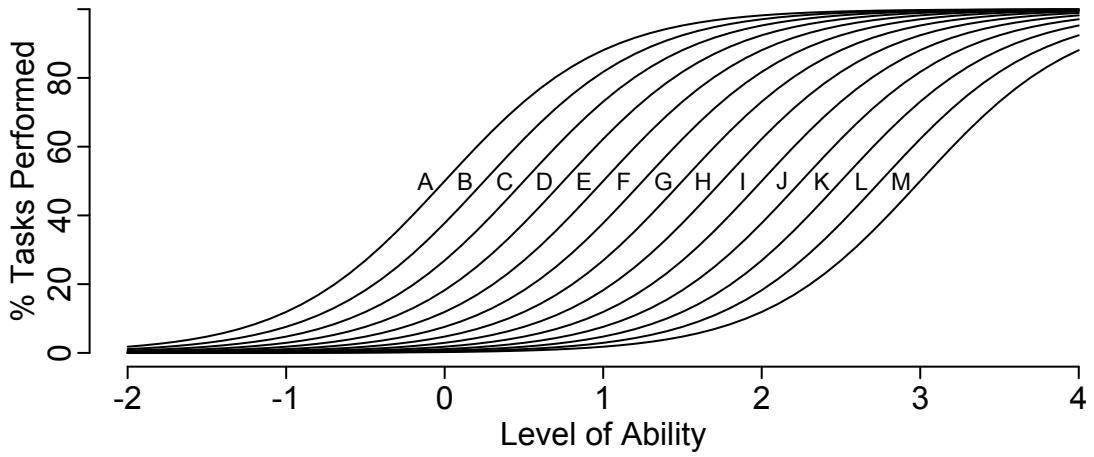

Figure 2. "Family of ogives representing items of different difficulty showing relationship between ability and correctness of performance" (Symonds, 1929, p. 483); the letters refer to sets of equally difficult items, like blocks of Ayres' (1915) spelling words.

with other conceptions, in the subsequent psychometric literature. Holland (1990) contrasts what he calls the "random sampling rationale" (Holland, 1990, p. 581) for IRT models, which harks back to Thurstone's conception of samples of children, and a "stochastic subject rationale" (Holland, 1990, p. 582), that is more closely related to Thurstone's (1927) ideas in *The Law of Comparative Judgment*. Holland (1990) expresses a lack of interest in the idea of an item sampling rationale like Symonds' because such a rationale does not apply to fixed tests. However, in the context of a spelling test, or other well-defined sets of educational objectives, reference to a domain of items certainly makes sense (Bock et al., 1997).[4]

By the time of the publication of the first edition of Guilford's (1936) *Psychometric Methods*, a standard descriptive tool for mental test items was an ogival curve illustrating the relationship between ability and "proportion of successes" (p. 427). Figure 3 is modeled after Guilford's Figure 41, which he used to discuss the concepts of difficulty and discrimination for test items: items *A* and *B* have the same difficulty; item *C* is more difficult, and *D* is most difficult. Guilford also discussed the idea that differences in the steepness, or slope, of the curves represented the "diagnostic value" of the item. Items *A* and *D* have steeper slopes, or higher diagnostic value, then items *B* or *C*. Guilford then wrote that "If one could establish a scale of difficulty in psychological units, it would be possible to identify any test item whatsoever by giving its median value and its 'precision' ... This is an ideal toward which testers have been working in recent years and already the various tools for approaching that goal are being refined" (pp. 427-428). It took more like 50 years to "refine" the methods, but by the 1980s IRT approximated Guilford's "ideal." In smaller steps, Richardson (1936), Ferguson (1943), Lawley (1943), and Tucker (1946) were among those who made further contributions to what was to become IRT.

1.2 The Introduction of the Latent Variable

Paul Lazarsfeld's (1950) chapters in *The American Soldier* series were about models for the relationship between observed item response data and lines describing the probability of a response to an item over a *latent* (unobserved) variable x. Lazarsfeld's work in mathematical sociology was only distantly related to the previously described work in psychometrics. He did not refer to the normal ogive model; he used linear trace lines. But his description of the process of testing marked the dawn of the latent variable era; Lazarsfeld wrote that "We shall now call a *pure test* of a continuum x an aggregate of items which has the following properties: *All interrelationships between the items should be accounted for by the way in which each item alone is related to the latent continuum*" (p. 367). A "pure test," in Lazarsfeld's language, is a test in which the item responses fit a model with the properties of *unidimensionality* and *local independence*.

[4]Indeed, Darrell Bock himself literally randomly sampled words from a word list to create a spelling test that yielded data used in illustrations in that article and elsewhere in the IRT literature.

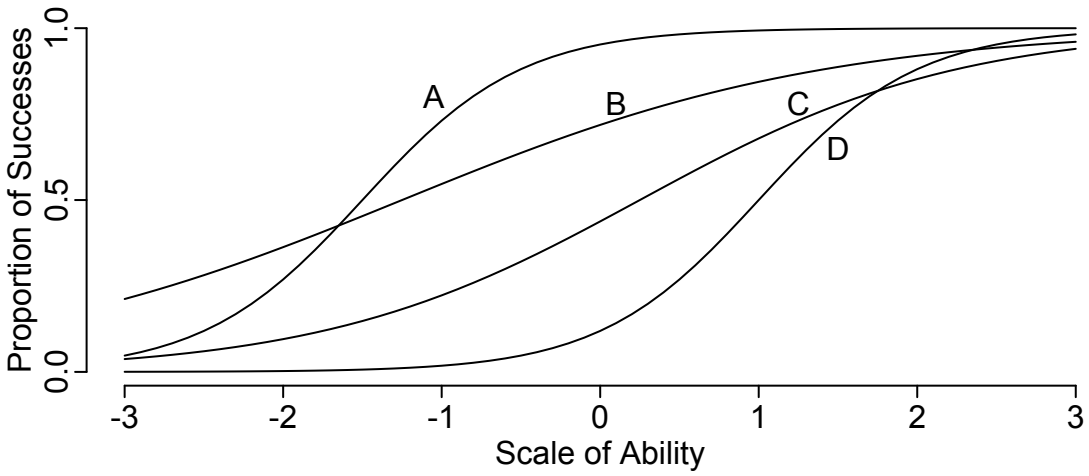

Figure 3. "Four ogives showing the increase in the probability of success in items with increase in the ability of the individual." Guilford (1936, p. 427) Ability is on the x-axis in standard units, with zero representing the population average.

Continuing to refer to the latent variable being measured as x, Lazarsfeld's (1950, p. 369) wrote that "The total sample is therefore characterized by a distribution function $\phi(x)$ which gives for each small interval dx the number of people $\phi(x)dx$ whose score lies in this interval. We can now tell what proportion of respondents in the whole sample will give a positive reply to item i with trace line $f_i(x)$ ..." That is, Lazarsfeld not only made clear that the theory was that there was a latent (unobserved) variable underlying the observed item responses, but also that there were two distinct functions: The population distribution of that latent variable he called $\phi(x)$ and the "trace line $f_i(x)$" for item i. Lazarsfeld described in equations how the joint probabilities of combinations of item responses are modeled as products of the trace lines. Lazarsfeld's work had little visible effect on quantitative psychology at the time, but in hindsight we see the importance of his conceptual contributions.

The lack of precision in the psychometric literature of the 1920s through the 1940s began to be clarified by Frederic Lord's (1952) description of *ability* as an unobserved variable defined by its relationship item responses.[5] The major point of Lord's monograph was to distinguish between the properties of the unobserved *ability* variable and observed test scores. Lord (1952, p. 1) wrote:

> A mental trait of an examinee is commonly measured in terms of a test score that is a function of the examinee's responses to a group of test items. For convenience we shall speak here of the "ability" measured by the test, although our conclusions will apply to many tests that measure mental traits other than those properly spoken of as "abilities." The ability itself is not a directly observable variable; hence its magnitude ... can only be inferred from the examinee's responses to the test items.

Lord's (1952, 1953a) early work made clear that latent ability and an observed (summed) test score are two different things.

All of the conceptual components of what was to become IRT were complete by the early 1950s. Those ideas are that items are "located" on the same scale as the "ability" variable (Thurstone, 1925), the "ability" variable is *latent* (or unobserved) (Lazarsfeld, 1950; Lord, 1952), and the unobserved variable accounts for the observed interrelationships among the item responses (Lazarsfeld, 1950).

These ideas saw some use in theoretical work concerning the structure of psychological tests, by Lord (1952, 1953a), Solomon (1956, 1961), Sitgreaves (1961a, 1961b, 1961c), and others. However, there was still no practical way to estimate the parameters (the item locations and discriminations) from observed item response data.

[5]Lord was writing his dissertation (Lord, 1952) in New York City at about the same time as Lazarsfeld's chapters were published in *The American Soldier*. However, it is not clear how much direct influence Lazarsfeld's work might have had on Lord. Lord (1952) did not cite Lazarsfeld; Lord (1953a, 1953b) later mentions Lazarsfeld (1950) only in passing.

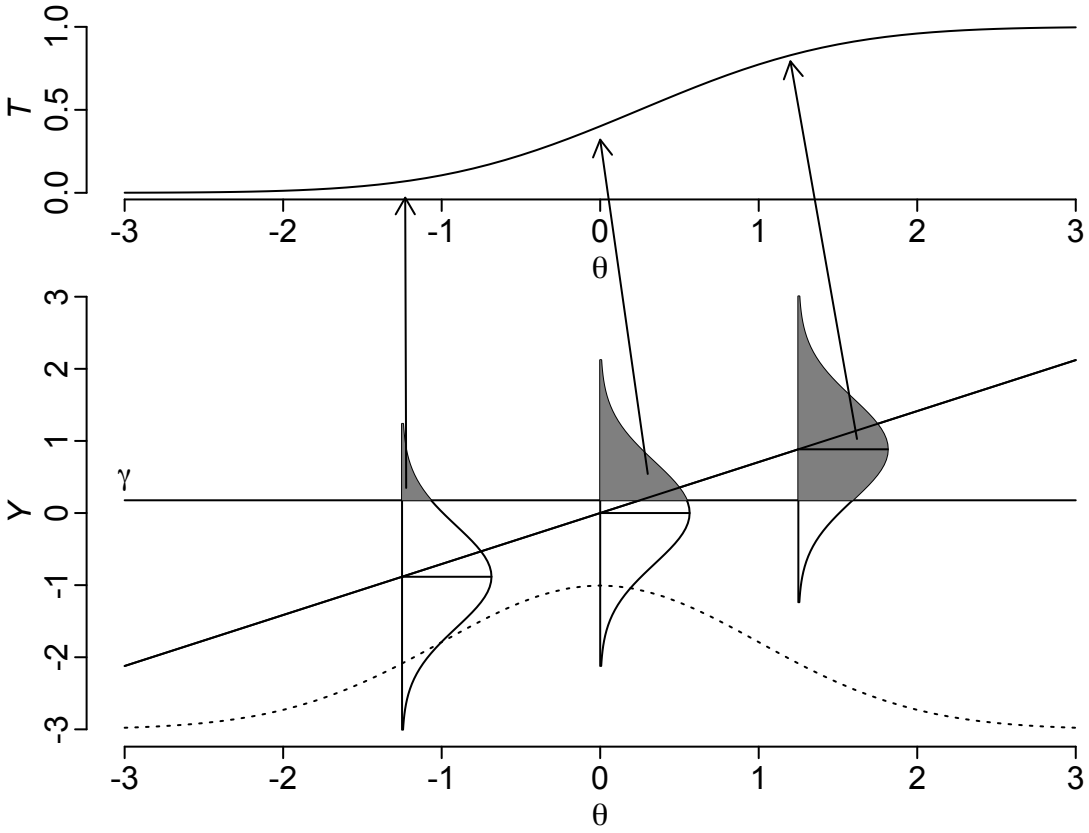

Figure 4. The hypothetical relationships involved in the normal ogive model, elaborating on Figure 16.6.1 of Lord and Novick (1968). The ogive in the upper panel is T, the trace line or probability of a correct or positive response, which in turn is a graph of the areas above γ of normal response process densities with means that are a linear function of θ. Three illustrative response process densities are shown in the lower panel. The dotted curve at the bottom of the lower panel is the population density for θ.

1.3 The Synthesis by Lord and Novick (1968)

IRT remained primarily a conceptual model for test theorists (as opposed to testing practitioners) until the 1970s, after the publication of Lord and Novick's (1968) *Statistical Theories of Mental Test Scores*. Lord and Novick integrated much of the preceding work, and their volume, combined with newly available electronic computers, signaled a new era in test theory.

Lord and Novick (1968, p. 366) codified the theoretical development of IRT up to that time; they described a kind of psychological theory underlying the normal ogive model that had (almost) all of the necessary elements. Figure 4, the lower panel of which is inspired by Figure 16.6.1 of Lord and Novick (1968, p. 371), illustrates the relationships among the latent variable θ (most often "ability"), an unobserved response process variable Y, a threshold parameter γ, and the probability of a correct response T, all for a single item.

The ideas were that there is a latent response process variable Y that is linearly related to the latent variable θ; the item parameters are the slope and intercept of that linear relationship, shown by the regression line in Figure 4. At any value of θ, there is a distribution of values of Y, shown by the vertical normal densities.[6] Those densities are divided at

[6]Holland (1990) pointed out that the vertical densities in Figure 4 can have several interpretations. One of those is frequentist, in which one imagines a subpopulation of examinees, all with the same value of θ, some of whom know the answer $(Y > \gamma)$ and others who do not. Or one can take what Holland called the "stochastic subject" view that the vertical densities represent some psychological process that varies within a single examinee; this interpretation is closely related to Thurstone's (1927) *Law of Comparative Judgment* for the comparison of objects. A third story, related to Symond's (1929) analysis of the spelling test, is that the vertical densities may represent a population of interchangeable

some constant γ, with the shaded area above γ corresponding to the conditional probability of a correct response plotted in the upper panel as the trace line (Lazarsfeld's phrase) *T*.

In a subsequent figure (Lord & Novick, 1968, Figure 16.11.1, p. 380) they plot a *second* kind of normal density that is the distribution of θ in the population (that Lazarsfeld had referred to as $\phi(x)$); that distribution is shown as the dashed curve in the lower panel of Figure 1. This representation transforms Thurstone's (1925) story into a fully-fledged statistical model, distinguishing between the population distribution and the response process variable, both of which, confusingly, are Gaussian. Thus, by the time of the publication of Lord and Novick's (1968) text, the normal ogive model had changed from an attempt to describe observed empirical data into a theory about an underlying, unobservable response process that might have produced the observable data.

1.4 Logistic IRT Models

In chapters contributed to Lord and Novick's (1968) volume, Allan Birnbaum (1968) pointed out that the logistic function had already been used in bioassay (Berkson, 1953, 1957) and other applications as a computationally convenient replacement for the normal ogive. Haley (1952, p. 7; see Camilli, 1994) had shown that if the logistic is rescaled by multiplication of the argument by 1.7, giving

$$\Psi(x) = e^{1.7x}/\left(1+e^{1.7x}\right) = 1/\left(1+e^{-1.7x}\right)$$

the resulting curve differs by less than 0.01 from the normal ogive $\Phi(x)$ for any value of x.[7]

Birnbaum (1968) also provided several mathematical statistical results for the now-ubiquitous three-parameter logistic (3PL) model for multiple choice items. The seed of the idea came from Lord (1953b, p. 67), who wrote "Suppose that any examinee who does not know the answer to a multiple-choice item guesses at the answer with 1 chance in k of guessing correctly. If we denote the item characteristic function for this item by P_i', we have

$$P_i' = P_i + Q_i/k."^{8}$$

Lord (1953b) did not pursue the idea, but Birnbaum elaborated on it as follows:

> Even subjects of very low ability will sometimes give correct responses to multiple choice items just by chance. One model for such items has been suggested by a highly schematized psychological hypothesis. This model assumes that if an examinee has ability θ, then the probability that he will *know* the correct answer is given by a normal ogive function $\Phi[a_g(\theta - b_g)] \ldots$ [I]t further assumes that if he does not know it he will guess, and, with probability c_g, will guess correctly. It follows from these assumptions that the probability of an incorrect response is
>
> $$Q_g(\theta) = \{1 - \Phi[a_g(\theta - b_g)]\}(1 - c_g)$$
>
> and the probability of a correct response is the item characteristic curve
>
> $$P_g(\theta) = c_g + (1 - c_g)\Phi[a_g(\theta - b_g)].$$
>
> ... Similarly, with the logistic model, ...
>
> $$P_g(\theta) = c_g + (1 - c_g)\Psi[a_g(\theta - b_g)].$$

Because the model had three item parameters (a_g, b_g, and c_g), it came to be called the "three-parameter logistic" (3PL) model, and by extension the logistic replacement for the original normal ogive model, became the "two-parameter logistic" (2PL) model.

1.5 The Rasch Model and the One-Parameter Logistic (1PL) Model

Georg Rasch (1960; Fischer, 2007) developed an item response model based on the mathematical requirement that one could meaningfully say one person has twice the ability (ξ) of another ($\xi_1 = 2\xi_2$), or that one problem is twice as difficult (δ) as another ($\delta_1 = 2\delta_2$).[9] Rasch (1960, pp. 74ff)

items, like spelling words of equal difficulty that a particular examinee may know or not. Which of these three stories make sense depends on the items and the construct being measured.

[7] The scaling constant 1.7 makes the numerical value of a (the slope) approximately the same for the logistic or normal ogive. However, for decades now since logistic IRT models have become dominant, the 1.7 is frequently omitted and absorbed in the value of a.

[8] In Lord's equation, $Q_i = 1 - P_i$.

[9] The Rasch model appears to have been developed nearly independently from the previous pre-history of IRT. Rasch (1960, p. 116) mentioned the normal ogive model (attributing it to Lord (1953a)), but only to say it was equally arbitrary (in Rasch's view) with the logistic, and that "with its extra set of parameters it falls outside the scope of the present work."

wrote that it would follow that:

$$\frac{\xi_1}{\delta_1} = \frac{\xi_2}{\delta_2}$$

> the probability that person no.1 solves problem no.1 should equal the probability that person no.2 solves problem no.2. This means, however, that *the probability is a function of the ratio, $\frac{\xi}{\delta}$, between the degree of ability of the person and the degree of difficulty of the problem, while it does not depend on the values of the two parameters ξ and δ separately*....
>
> If we put $\frac{\xi}{\delta} = \zeta$,... the simplest function I know of, which increases from 0 to 1 as ζ goes from 0 to ∞, is $\frac{\zeta}{(1+\zeta)}$.

Written as it was by Rasch (1960), the model appears different from those previously discussed. However, if it is reparameterized by changing ξ to e^{θ} and δ to e^{b}; then the model becomes a logistic function with no explicit slope or discrimination parameter. Birnbaum (1968, p. 402) noted that Rasch's (1960) model was logistic with the restriction of a common (equal) discrimination parameter for all items, and observed that might be plausible for some tests.

While Rasch originally wrote that he based his choice of the logistic function on simplicity, in subsequent writings, Rasch and others have stated that the assumptions of the Rasch model must be met to obtain valid measurement. (Rasch 1966, pp. 104-105) wrote:

> In fact, *the comparison of any two subjects* can be carried out in such a way that *no other parameters are involved than those of the two subjects* — neither the parameter of any other subject nor any of the stimulus parameters. Similarly, *any two stimuli can be compared independently of all other parameters than those of the two stimuli*...
>
> It is suggested that comparisons carried out under such circumstances be designated as "specifically objective."

Rasch (1966, p. 107) concluded: "I must point out that the problem of the relation of data to models is not only one of trying to fit data to an adequately chosen model from our inventory to see whether it works; it is also *how to make observations in such a way that specific objectivity obtains*." Subsequently, Rasch (1977) and others (Fischer, 1974, 1985; Wright & Douglas, 1977; Wright & Panchapakesan, 1969) emphasized the idea that "specific objectivity" was a *requirement* of psychological measurement.

There is no universal agreement that specific objectivity is necessary, even among scholars in the Rasch tradition; de Leeuw and Verhelst (1986, p. 187) wrote that although "the factorization that causes ... *specific objectivity* ... is certainly convenient, its importance has been greatly exaggerated by some authors." Because both the Rasch and the Thurstone-Lazarsfeld-Lord-Birnbaum traditions lead to logistic item response functions with (potentially, in the latter case) equal discrimination parameters, but arise from different conceptual frameworks, it is useful for that model to have two different names. Wainer et al. (2007) suggested reference to models from the Rasch tradition as "Rasch models," and the term "one-parameter logistic" (1PL) for logistic item response functions with equal discrimination parameters arising from the Birnbaum tradition.

2 Models for Polytomous Item Responses

2.1 The Likert Scale

Rensis Likert[10] (1932) introduced the now-ubiquitous "Likert-type" response scale in his monograph (and dissertation), *A Technique for the Measurement of Attitudes*. Before Likert's suggestion, polytomous item response data were collected in clumsier ways: At Thurstone's Psychometric Laboratory in Chicago, research participants were given the item-stems typed on individual cards, and sorted them into eleven piles as a method of responding from most positive through neutral to most negative (Thurstone & Chave, 1929). At the University of Iowa, in a study described by Hart (1923), participants made a first pass through the items to indicate a positive response, neutrality, or a negative response, followed by a second pass to underline and double underline some responses for emphasis, yielding a seven-point scale. Such ratings served as item-scores representing "difficulty" of endorsement. Total scores were the sum of the item-scores for statements that respondents endorsed. At Teachers College, before Likert's dissertation, Neumann (1926) followed a similar procedure, but in a second study began to use a 5-point scale of the general form of a Likert-type scale as a time-saving measure.

However, the intended point of Likert's monograph was not the response scale which has ultimately been his most

[10] There has often been confusion over the pronunciation of Likert's name. According to people who knew him, it was pronounced lick-ert, not like-ert (Wimmer, 2012; Likert scale, 2020).

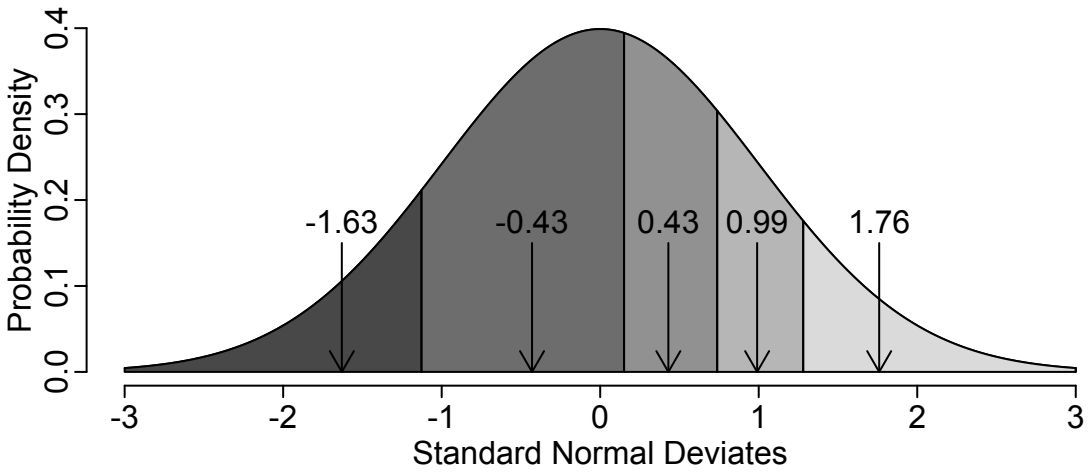

Figure 5. Graphical expression of Likert's (1932) idea that the means of five ordered segments of the normal density could be used as scoring values for five graded response alternatives (like *Strongly Approve*, *Approve*, *Undecided*, *Disapprove*, and *Strongly Disapprove*)

widely-known contribution; it was to propose "sigma scoring" that was a variant on Thurstone's scaling ideas based on the normal-distribution. For attitude questions with a five-point response scale labeled *Strongly Approve*, *Approve*, *Undecided*, *Disapprove*, and *Strongly Disapprove*, Likert proposed using the average standard normal deviate for the corresponding percentile range of the normal distribution as shown in Figure 5 as the numerical value of each choice.[11] Then Likert proposed scores computed as the sum or average of the "sigma" values so-computed. The motivation was to obtain an easier method of scoring than the elaborate judgment systems used in Thurstone's Psychometric Laboratory for similar purposes (Thurstone, 1928; Thurstone & Chave, 1929). Likert compared the performance of "sigma scoring" with the "simple" method of summing the numeric values 1-5 for the five responses, using correlations of the scores with other variables as the criterion; he found little difference. Sigma scoring faded into oblivion, but that scoring method anticipated polytomous IRT.

2.2 Samejima's Graded Models

While visiting Fred Lord's group at ETS in the late 1960s, Samejima (1969, 2016) developed graded item response models for items with more than two ordered response alternatives. The original impetus for the model was fitting data for all response alternatives to educational multiple choice items. Although better models have been developed for that purpose (see Thissen and Steinberg, 1984, 1997), Samejima's graded models have seen widespread use for items with categorical response scales in the Likert-style format. The basic idea was simple (once pointed out): Use the existing normal ogive (or logistic) model for successive dichotomies formed by comparing responses 2 or higher vs. lower (1), and then 3 or higher vs. lower (1 or 2), and then 4 or higher vs. lower (1, 2, or 3), and so on. Then differences between those "response or higher" curves are the trace lines for the response categories themselves. Samejima's (1969) monograph included the core mathematical development for both the normal ogive and logistic versions of the model. The left panel of Figure 6 shows the trace lines for a prototypical item with five graded response alternatives.

2.3 Bock's Nominal Model

The nominal categories item response model (Bock, 1972; Thissen & Cai, 2016) was inspired by Samejima's (1969, 2016) graded response model, and was also originally proposed as a model for trace lines for all of the response alternatives on multiple choice items. Like Samejima's (1969) model it has been superseded for that purpose by the multiple-choice model (Thissen & Steinberg, 1984, 1997). However, the nominal model continues to have three uses (Thissen et al., 2010): (1) item analysis and scoring for items that elicit purely nominal responses; (2) to provide an empirical check that items intended to yield ordered re-

[11] Likert's (1932) monograph includes no graphics. Likert made use of a table provided by Thorndike (1913) to compute the averages for any percentile range of the normal distribution.

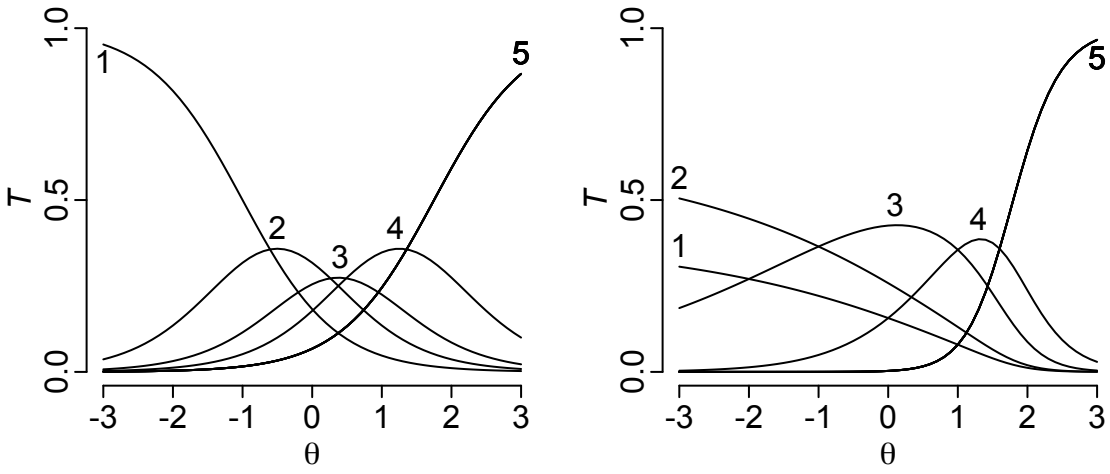

Figure 6. Trace lines for the probability of responding in each of the five response categories as a function of the value of the underlying construct. The left panel shows the trace lines for a prototypical item fitted with Samejima's (1969) graded model. The right panel shows trace lines fitted with Bock's (1972) nominal model: The leftmost two trace lines are for two responses that equally indicate low levels of the trait being measured, one more likely than the other; then there are two ordered responses "below" a very discriminating highest (fifth) alternative.

sponses actually do so (Thissen et al., 2007); and (3) to provide a model for testlet responses (Wainer et al., 2007). The right panel of Figure 6 shows trace lines for an item with five alternatives: The leftmost two trace lines are for two responses that equally indicate low levels of the trait being measured, one more likely than the other; then there are two ordered responses "below" a very discriminating highest (fifth) alternative. The graded model could not fit data with a generating process like that illustrated in the right panel of Figure 6.

2.4 "Rasch Family" Models for Polytomous Responses

Rasch (1961) suggested multidimensional and unidimensional generalizations of his original dichotomous logistic model to polytomous responses. However, little was made of that until Andersen (1977) proved that the so-called "scoring functions" of the polytomous Rasch model had to be proportional to successive integers for the model to have the original Rasch model property that the simple summed score is a sufficient statistic for scoring.

Andrich (1978, 2016) proposed the *rating scale* (RS) model for Likert-type ordered responses; that model used successive integers as the scoring function values, and divided the "threshold" or "location" parameter set for an item into a single overall "difficulty" parameter and a set of thresholds that reflect the relative widths of the categories on the graded scale. The idea was that the thresholds could be properties of the response scale, and the same for all items, which then differ only in overall degree of endorsement. Masters (1982, 2016) developed the Rasch-family *partial credit* (PC) model for, as the name suggests, use with graded judged multi-point ratings of constructed responses to open-ended questions in educational testing.

The RS and PC models were derived from very different mathematical principals than Bock's (1972) nominal categories model; as a result, even though the trace line equations appeared to be similar in many respects, it took some time before it was recognized that the RS and PC models are restricted parameterizations of the nominal model (Thissen & Steinberg, 1986). Indeed, the nominal model may be viewed as a kind of template from which models with specific properties can be obtained with constraints. Examples include Muraki's (1992; Muraki and Muraki, 2016) *generalized partial credit* (GPC) model, or (equivalently) Yen's (1993) *two-parameter partial credit* model, both of which were developed (separately) by analogy with the relation between the 2PL model and the Rasch model, extending the PC model to provide items with potentially unequal discrimination parameters.

3 Parameter Estimation

IRT was not used for operational item analysis or test scoring before the 1970s, because there were no computationally feasible ways to estimate the parameters of the trace

line models. Sitgreaves (1961c) worked out the required equations to do normal ogive model parameter estimation by minimizing the expected squared error; but her results were extremely complex, and she concluded, "In general, these results are not very useful" (Sitgreaves, 1961c, p. 59).

The first fully maximum likelihood (ML) procedure for estimating the parameters of the normal ogive model was published by Bock and Lieberman (1970). They fitted the model to sets of five dichotomous items, using the now famous (or infamous) "LSAT sections 6 and 7" datasets provided to them by Fred Lord, at a time when ETS did the data analysis for the LSAT. A problem with the Bock and Lieberman (1970) estimation procedure was that it was barely manageable by the computers of the time. In their conclusion, Bock and Lieberman (1970, p. 180) wrote that "the maximum likelihood method presented here cannot be recommended for routine use in item analysis. The problem is that computational difficulties limit the solution to not more than 10 or 12 items in any one analysis — a number too small for typical psychological test applications. The importance of the present solution lies rather in its theoretical interest and in providing a standard to which other solutions . . . can be compared."

3.1 Heuristics and "Joint Maximum Likelihood" Estimation

Lord and Novick's (1968) chapters on the normal ogive model included the equations for the relationships between the parameters of the IRT model and the proportion correct on the one hand, and factor loadings for a one-factor model on the other. They suggested that factor analysis of the matrix of inter-item tetrachoric correlations, along with the proportion correct for each item, could be transformed to yield heuristic estimates of the slope and threshold parameters of the normal ogive model. That suggestion did not come into widespread use, probably because factor analysis based on tetrachoric correlations was itself nearly as difficult as the IRT parameter estimation problem.

A solution offered by Fred Lord's group at ETS was called "joint maximum likelihood" (JML) estimation, because it computed maximum likelihood estimates for the item parameters and maximum likelihood estimates of the latent variable (θ) values for the examinees "jointly." This followed a suggestion Lord (1951) made long before it was computationally feasible. But by the 1970s it could be done using the mainframe computers, with an alternating algorithm that used provisional estimates of θ to estimate logistic model item parameters in what amounted to logistic regression for the item responses, and then in the alternating stage replaced the provisional estimates of θ with ML estimates computed essentially as per the procedure provided by Lawley (1943). The computer program LOGIST (Wingersky et al., 1982) implemented this algorithm and became widely used, first inside ETS and then elsewhere. Other less widely known or distributed software also used variations on this algorithm in the 1970s.

A downside to JML is that Neyman and Scott (1948) had shown before the IRT programs were written that such procedures could not work, with the number of parameters estimated increasing with the number of observations. Indeed, the JML IRT software did not work very well; it was made to appear to function with a variety of *ad hoc* fixes. Haberman (in press) expresses dismay that computer programs implementing joint estimation algorithms are still in use, given their well known statistical failings and the fact that superior algorithms have long been available.

3.2 Rasch Family Models: Conditional and Loglinear Estimation

In the first decade of development of the Rasch model, Wright and Panchapakesan (1969) published a JML algorithm and associated computer program to estimate the item parameters (for the Rasch model, those are the item difficulty values). It was not long before Andersen (1973) showed that the JML estimates were, as expected, not consistent, and for the two-item example Andersen considered, not very good.

But Andersen (1970, 1972) had already worked out the mathematical statistics for conditional ML (CML) estimation, and shown that it produces consistent estimates of Rasch model item parameters. The Rasch model is unique among latent variable models for dichotomous item responses in that the simple summed score is a sufficient statistic to characterize the latent variable for a respondent; that characterization is the same regardless of the pattern of responses across items, if the total score is the same. A likelihood can be written for the IRT model within (conditional on) each summed score group, and then those conditional likelihoods can be combined into an overall likelihood that is maximized to yield item parameter estimates. The algorithm does require computation of values that (at least appear to) involve all response patterns; the Rasch model literature of the 1970s is filled with solutions to that computational challenge, making CML practical.

In the early 1980s, researchers from several perspectives showed that the Rasch model is also a loglinear model for

the 2^n table of frequencies for each response pattern to n dichotomous items (Tjur, 1982; Cressie & Holland, 1983; Duncan, 1984; Kelderman, 1984). This meant that algorithms already developed and implemented in software could be used to compute ML estimates of the parameters of the Rasch model. de Leeuw and Verhelst (1986) showed that the loglinear model estimates and CML estimates are identical for the Rasch model.[12]

3.3 The Bock-Aitkin EM Algorithm

Bock and Aitkin (1981) used elements of the *EM algorithm* (Dempster et al., 1977) to re-order the computations implicit in the Bock-Lieberman maximum likelihood estimation procedure in such a way as to make item parameter estimation possible for truly large numbers of items. They called the procedure "marginal maximum likelihood" (MML) to indicate that it was "marginal" with respect to (or involved integrating over) the population distribution of θ, and to distinguish the procedure from JML and CML. Subsequently, the words were often rearranged to become the more semantically correct "maximum marginal likelihood" (which is still MML). Statisticians just call it maximum likelihood, because it is standard statistical practice to "integrate out" latent or nuisance variables.

The Bock-Aiktin algorithm was implemented in specialized software such as Bilog-MG, Parscale, and Multilog (du Toit, 2003) and could be used to estimate the parameters of IRT models for data involving realistic numbers of items and respondents. With the exception of Bilog-MG, those software packages are retired, and a second generation of software that includes IRTPRO (Cai et al., 2011), flexMIRT (Cai, 2017), mirt in R (Chalmers, 2012), the IRT procedure in Stata (StataCorp, 2019), and others, implement the Bock-Aitkin algorithm for most of the models described in previous sections. These software packages make IRT the basis of most large-scale testing programs.

3.4 MCMC Estimation for IRT

While there had been some previous Bayesian work on aspects of estimation for IRT models, Albert's (1992) Markov chain Monte Carlo (MCMC) algorithm for estimation of the parameters of the normal ogive model is of historical interest for two reasons. The first reason is that it marks the beginning of the recent era in which many new IRT models are first "tried out" using MCMC estimation, which can be quicker and easier to implement than ML. The second is that Albert (1992) used Tanner and Wong's (1987) idea of "data augmentation" to produce a Gibbs sampling algorithm in which all of the sampling steps are in closed form. While that was done for entirely statistical reasons, the interesting thing is that the augmenting data are both the values of the latent variable θ and the values of the response process variables Y from Figure 4, or from Lord and Novick (1968)! So the statistical and psychological theories merged.

Albert's (1992) data augmentation strategy only works well for the normal ogive model. But once the door was opened, others followed with other Gibbs sampling algorithms for many IRT models; examples from the twentieth century (if barely) include MCMC algorithms by Patz and Junker (1999a, 1999b) and Bradlow et al. (1999). Fully Bayesian estimation involves computing the *mean* of the posterior distribution of the parameters, as opposed to the mode of the likelihood, which is located using an ML algorithm. MCMC estimation is computationally intensive, but for the past couple of decades, and looking forward, computational power has been and will be inexpensive and plentiful, which has made MCMC estimation the tool of choice for trying out novel or custom IRT models.

4 Conclusion

We have traced the early development of parametric IRT models from their origins in the work of Thurstone, Lazarsfeld, Lord, Birnbaum, and Rasch. Current uses of the "standard" IRT models we have described include item analysis, scale development, detecting group differences in item responses, estimating item parameters for computerized adaptive testing, accounting for violations of local dependence with the use of testlets, as well as developing an understanding of the psychological processes underlying responses to academic, social, and personality questions.

In the past three or four decades there has been a veritable explosion of development of IRT models for increasingly specialized uses. The recently published *Handbook of Item Response Theory, Volume One: Models* (van der Linden, 2016b)[13]comprises 33 chapters and nearly 600 pages; this article has mentioned only a fraction of the models

[12]Cressie and Holland (1983) showed that there is a "catch" to either loglinear or CML Rasch model estimation: While no population distribution for θ appears in the equations, there must *be* one, and it has to satisfy the moment inequalities for any proper density. There is no explicit check that those inequalities are satisfied in either CML or loglinear estimation. Checking is required; de Leeuw and Verhelst (1986) expand on Cressie and Holland's (1983) specifications for checking.

[13]Space does not permit citation of more than token references for these topics;

that volume covers, most of which have appeared in the past few decades. Large general-purpose classes of models include extensions of all IRT models to accommodate multidimensional latent variables (multidimensional IRT, or MIRT; Reckase, 2009), and hierarchical or multilevel item response models (e.g. Fox and Glas, 2001). A modern synthesis of disparate traditions merges IRT with the factor analytic framework, within the scope of generalized latent variable models (Skrondal & Rabe-Hesketh, 2004; Rabe-Hesketh, Skrondal, & Pickles, 2004; Bock & Moustaki, 2007). Cognitive diagnostic models are used with structured educational assessments to support inference about mastery or non-mastery of specific skills (von Davier & Lee, 2019). More specialized models include non-compensatory multidimensional models for achievement or ability test items believed to measure multiple components of processing (e.g., Embretson and Yang, 2013), or to measure response sets in personality or attitude measurement (e.g. Thissen-Roe and Thissen, 2013). There are also models for less commonly used response formats and processes (e.g., Mellenbergh, 1994; Roberts, Donoghue, and Laughlin, 2000), and for the response times now routinely collected in the process of computerized testing (e.g. van der Linden, 2016). *Explanatory item response models* are specially customized models built to express and test psychological hypotheses about processing (De Boeck & Wilson, 2004). And there are several traditions of nonparametric analyses intended to provide similar, or complementary, data analysis to that obtained with parametric IRT models (e.g. Sijtsma and Molenaar, 2002; Ramsay, 2016).

But that brings us to contemporary developments rather than history. We conclude that IRT is an active field that continues to grow and develop.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269. https://doi.org/10.2307/1165149

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society: Series B (Methodological)*, *32*(2), 283–301. https://doi.org/10.1111/j.2517-6161.1970.tb00842.x

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*, 42–54. https://doi.org/10.1111/j.2517-6161.1972.tb00887.x

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk forlag.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69–81. https://doi.org/10.1007/BF02293746

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. https://doi.org/10.1007/BF02293814

Andrich, D. (2016). Rasch rating-scale model. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 75–94). Boca Raton, FL: Chapman & Hall/CRC.

Ayres, L. P. (1915). *A measuring scale for ability in spelling*. N.Y.: Russell Sage Foundation.

Berkson, J. (1953). A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association*, *48*, 565–599. https://doi.org/10.1080/01621459.1953.10483494

Berkson, J. (1957). Tables for the maximum likelihood estimate of the logistic function. *Biometrics*, *13*, 28–34. https://doi.org/10.2307/3001900

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51. https://doi.org/10.1007/BF02291411

Bock, R. D. (1983). The mental growth curve reexamined. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 205–219). N.Y.: Academic Press.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, *16*, 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02291262

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*, 179–197. https://doi.org/10.1007/BF02291262

Bock, R. D., & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics Volume 26: Psychometrics* (pp. 469–513). Amsterdam: North-Holland.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, *34*, 197–211. https://doi.org/10.1111/j.1745-3984.1997.tb00515.x

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168. https://doi.org/10.1007/BF02294533

Burt, C. (1922). *Mental and scholastic tests*. London, P.S.King.

Cai, L. (2017). *flexMIRT® version 3.51: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.

Camilli, G. (1994). Origin of the scaling constant d=1.7 in item response theory. *Journal of Educational and Behavioral Statistics*, *19*, 293–295. https://doi.org/10.2307/1165298

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*, 1–29. https://doi.org/10.18637/jss.v048.i06

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika*, *48*, 129–141. https://doi.org/10.1007/BF02314681

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, *11*, 183–196. https://doi.org/10.3102\%2F10769986011003183

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*, 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Duncan, O. D. (1984). Rasch measurement: Further examples and discussion. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena, volume 2* (pp. 367–403). New-York, NY: Russell Sage Foundation.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG MULTILOG PARSCALE TESTFACT*. Lincolnwood, IL: Scientific Software International.

Embretson, S. E., & Yang, X. (2013). A Multicomponent Latent Trait Model for Diagnosis. *Psychometrika*, *78*, 14–36. https://doi.org/10.1007/s11336-012-9296-y

Ferguson, G. A. (1943). Item selection by the constant process. *Psychometrika*, *7*, 19–29. https://doi.org/10.1007/BF02288601

Fischer, G. H. (1974). *Einführung in die theorie psychologischer tests*. Bern: Huber.

Fischer, G. H. (1985). Some consequences of specific objectivity for the measurement of change. In E. E. Roskam (Ed.), *Measurement and personality assessment* (pp. 39–55). Amsterdam: North-Holland.

Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics volume 26: Psychometrics* (pp. 515–585). Amsterdam: North-Holland.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 269–286. https://doi.org/10.1007/BF02294839

Guilford, J. P. (1936). *Psychometric methods*. N.Y.: McGraw-Hill. https://doi.org/10.1007/BF02287877

Haberman, S. (in press). Statistical theory and assessment practice. *Journal of Educational Measurement*.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error.* Stanford: Applied Mathematics and Statistics Laboratory, Stanford University, Technical Report 15.

Hart, H. N. (1923). Progress report on a test of social attitudes and interests. In B. T. Baldwin (Ed.), *University of Iowa Studies in Child Welfare (Vol.2)* (pp. 1–40). Iowa City: The University.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601. https://doi.org/10.1007/BF02294609

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, *49*, 223–245. https://doi.org/10.1007/BF02294174

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 62-A*, Part I, 74–82. https://doi.org/10.1017/S0080454100006282

Lazarsfeld, P. F. (1950). The logical and mathematical

foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and Prediction* (pp. 362–412). New York: Wiley.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 4-55.

Likert scale. (2020, June 11). Retrieved June 16, 2020, from https://en.wikipedia.org/wiki/Likert_scale#Pronunciation

Lord, F. M. (1951). *A maximum likelihood approach to test scores* (ETS Research Bulletin Series No. RB-51-19). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1951.tb00219.x

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, Whole No.7.

Lord, F. M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, *18*, 57–76. https://doi.org/10.1007/BF02289028

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, *13*, 517–548. https://doi.org/10.1177/001316445301300401

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/BF02296272

Masters, G. N. (2016). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 109–126). Boca Raton, FL: Chapman & Hall/CRC.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*, 223–236. 10.1207/s15327906mbr2903_2

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *29*, 159–176. https://doi.org/10.1177/014662169201600206

Muraki, E., & Muraki, M. (2016). Partial credit model. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 127–137). Boca Raton, FL: Chapman & Hall/CRC.

Neumann, G. B. (1926). *A study of international attitudes of high school students*. New York,NY: Teachers College, Columbia University, Bureau of Publications.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32. https://doi.org/10.2307/1914288

Patz, R. J., & Junker, B. W. (1999a). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366. https://doi.org/10.3102/10769986024004342

Patz, R. J., & Junker, B. W. (1999b). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178. https://doi.org/10.3102/10769986024002146

Patz, R. J., & Yao, L. (2007). Vertical scaling: Statistical models for measuring growth and achievement. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics volume 26: Psychometrics* (pp. 955–975). Amsterdam: North-Holland. https://doi.org/10.1016/S0169-7161(06)26030-9

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM Manual (Second Edition).* Berkeley, CA: U.C. Berkeley Division of Biostatistics Working Paper Series University of California Working Paper 160.

Ramsay, J. O. (2016). Functional approaches to modeling response data. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 337–350). Boca Raton, FL: Chapman & Hall/CRC.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedagogiske Institut.

Rasch, G. (1961). On General Laws and the Meaning of Measurement in Psychology. *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, *4*, 321–333.

Rasch, G. (1966). An individualistic approach to item analysis. In P. Lazarsfeld & N. V. Henry (Eds.), *Readings in mathematical social science* (pp. 89–108). Chicago: Science Research Associates.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Blegvad (Ed.), *The Danish yearbook of philosophy.* Copenhagen: Munksgaard.

Reckase, M. D. (2009). *Multidimensional item response theory models*. N.Y.: Springer. https://doi.org/10.1007/978-0-387-89976-3

Richardson, M. W. (1936). The relationship between

the difficulty and the differential validity of a test. *Psychometrika*, *1*, 33–49. https://doi.org/10.1007/BF02288003

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A General Item Response Theory Model for Unfolding Unidimensional Polytomous Responses. *Applied Psychological Measurement*, *24*, 3–32. https://doi.org/10.1177/01466216000241001

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17*, *34*, Part 2. https://doi.org/10.1007/BF03372160

Samejima, F. (2016). Graded response models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 95–107). Boca Raton, FL: Chapman & Hall/CRC.

Sijtsma, K., & Molenaar, I. W. (2002). *Measurement Methods for the Social Science: Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage Publications, Inc. https://doi.org/10.4135/9781412984676

Sitgreaves, R. (1961a). Further contributions to the theory of test design. In H. Solomon (Ed.), *Studies in item analysis and prediction* (pp. 46–63). Stanford, CA: Stanford University Press.

Sitgreaves, R. (1961b). Optimal test design in a special testing situation. In H. Solomon (Ed.), *Studies in item analysis and prediction* (pp. 29–45). Stanford, CA: Stanford University Press.

Sitgreaves, R. (1961c). A statistical formulation of the attenuation paradox in test theory. In H. Solomon (Ed.), *Studies in item analysis and prediction* (pp. 17–28). Stanford, CA: Stanford University Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9780203489437

Solomon, H. (1956). Probability and statistics in psychometric research: item analysis and classification techniques. In J. Neyman (Ed.), *Proceedings of the third berkeley symposium on mathematical statistics and probability* (Vol. 5, pp. 169–184). Berkeley, CA: University of California Press.

Solomon, H. (1961). Classification procedures based on dichotomous response vectors. In H. Solomon (Ed.), *Studies in item analysis and prediction* (pp. 177–186). Stanford, CA: Stanford University Press.

StataCorp. (2019). *Stata: Release 16* [Statistical Software]. College Station, TX: StataCorp LLC.

Symonds, P. M. (1929). Choice of items for a test on the basis of difficulty. *Journal of Educational Psychology*, *20*, 481–493. https://doi.org/10.1037/h0075650

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American statistical Association*, *82*, 528–540. https://doi.org/10.1080/01621459.1987.10478458

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 51–73). Boca Raton, FL: Chapman & Hall/CRC.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 43–75). New York, NY: Routledge.

Thissen, D., Reeve, B. B., Bjorner, J. B., & Chang, C.-H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research*, *16*, 109–116. https://doi.org/10.1007/s11136-007-9169-5

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501–519. https://doi.org/10.1007/BF02302588

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577. https://doi.org/10.1007/BF02295596

Thissen, D., & Steinberg, L. (1997). A response model for multiple choice items. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–65). New York: Springer-Verlag. https://doi.org/10.1007/978-1-4757-2691-6_3

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*, 522–547. https://doi.org/10.3102/1076998613481500

Thorndike, E. L. (1913). *An introduction to the theory of mental and social measurements* (Second Edition). New York, NY: Teachers College, Columbia University. https://doi.org/10.1037/10866-000

Thorndike, E. L., Bregman, E. O., Cobb, M. V., Woodyard, E., & Institute of Educational Research, Division of Psychology, Teachers College, Columbia University. (1926). *The measurement of intelligence*. Teach-

ers College Bureau of Publications. https://doi.org/10.1037/11240-000

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433–449. https://doi.org/10.1037/h0073357

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273—286. https://doi.org/10.1037/h0070288

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554. https://doi.org/10.1086/214483

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Thurstone, L. L., & Chave, E. J. (1929). *The Measurement of Attitude*. Chicago, IL: University of Chicago Press.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative poisson model. *Scandinavian Journal of Statistics*, *9*, 23–30.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, *11*, 1–13. https://doi.org/10.1007/BF02288894

van der Linden, W. J. (2016a). *Handbook of item response theory, volume one: Models*. Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9781315374512

van der Linden, W. J. (2016b). Lognormal response time model. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 261–282). Boca Raton, FL: Chapman & Hall/CRC. https://doi.org/10.1201/9781315374512

von Davier, M., & Lee, Y.-S. (Eds.). (2019). *Handbook of Diagnostic Classification Models*. New York, NY: Springer. https://doi.org/10.1007/978-3-030-05584-4

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement*, *35*, 93–107. https://doi.org/10.1111/j.1745-3984.1998.tb00529.x

Wimmer, R. (2012). Likert Scale-Dr. Rensis Likert Pronunciation-Net Talk. Retrieved June 16, 2020, from https://www.allaccess.com/forum/viewtopic.php?t=24251

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample free item analysis. *Applied Psychological Measurement*, *1*, 281–295.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23–48. https://doi.org/10.1177/001316446902900102

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement*, *34*, 293–313. https://doi.org/10.1111/j.1745-3984.1997.tb00520.x