DOCUMENT RESUME

ED 428 078 TM 029 478

AUTHOR	Henson, Robin K.
TITLE	Understanding the One-Parameter Rasch Model of Item Response Theory.
PUB DATE	1999-01-21
NOTE	51p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Ability; *Difficulty Level; *Item Response Theory; Probability; *Test Items
IDENTIFIERS	Calibration; *One Parameter Model; *Rasch Model

ABSTRACT

Basic issues in understanding Item Response Theory (IRT), or Latent Trait Theory, measurement models are discussed. These theories have gained popularity because of their promise to provide greater precision and control in measurement involving both achievement and attitude instruments. IRT models implement probabilistic techniques that yield statistics to help describe the interplay between the testing item and the respondent in terms of the unobservable latent traits that cause a given response to an item. The one-parameter IRT model, often referred to as the Rasch model, is illustrated using heuristic data. In IRT examinee ability and item difficulty estimates can be obtained that are theoretically independent of each other; therefore, that they can be used across samples of different abilities and items of varying difficulty. IRT transforms classical test theory proportions into logits, converting item difficulties and person abilities into the same linear metric across the distribution. In the Rasch model, this calibration process theoretically makes the ability statistic (theta) item-free and item difficulties ("b") examinee-free. (Contains 10 tables, 4 figures, and 9 references.) (Author/SLD)

Running Head: RASCH MODEL OF IRT

Understanding the One-parameter Rasch model of Item Response Theory

Robin K. Henson

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 21, 1999.

Abstract

The present paper will discuss basic issues in understanding Item Response Theory, or Latent Trait Theory, measurement models. These theories have gained popularity due to their promise to provide greater precision and control in measurement involving both achievement and attitude instruments. Specifically, the one-parameter Rasch model of measurement will be illustrated using heuristic data.

Understanding the One-parameter Rasch model of Item Response Theory

When they were first introduced, Item Response Theory (IRT), or Latent Trait Theory, measurement models were heralded as "one of the most important methodological advances in psychological measurement in the past half-century" (McKinley & Mills, 1989, p. 71). However the relative advantages and disadvantages of these models have since been very hotly debated (Fan, 1998; Lawson, 1991), notwithstanding their widespread use in various applications (e.g., in test equating, item selection, adaptive testing).

The popularity of IRT models lies in their promise to provide greater precision and control in measurement involving both achievement and attitude instruments. Developed largely as a response to the apparent weaknesses of classical test theory, IRT models implement probabilistic techniques that yield statistics to help describe the interplay between the testing item and the respondent (McKinley & Mills, 1989). Such statistics describe this interplay not only in terms of ratios between items as classical test theory does, but also in terms of the unobservable latent traits that cause a given response to an item. It is the modeling of these latent traits that provides the utility and power of IRT models, as against classical test theory approaches.

The purpose of this work is threefold. First, some of the weaknesses related to classical test theory will be discussed along with the basic concepts and advantages of IRT. Second, the theoretical frameworks of the one, two, and three parameter IRT models will be discussed. Finally, the one-parameter IRT model, often referred to as the Rasch model, will be performed on heuristic data to illustrate IRT's theoretical underpinnings. Simple spreadsheets are utilized to accomplish this objective. Use of such illustrative spreadsheets can serve as a powerful heuristic device.

Limitations of Classical Test Theory

Classical test theory (CTT), as applied to achievement and attitude measures, has several inherent limitations. First, the circular dependence of item characteristics cannot be separated from the characteristics of the test. To illustrate this dynamic, consider a 30-item achievement test administered to a sample of middle school students. In order to be able to reasonably compare the scores of these students with another group of students, the same 30 items must be administered to the second group of students. If the items differ, then the tests may have varying difficulty levels. Comparisons of scores across groups would then be inappropriate. This problem is called item sample dependence. That is, the examinee ability statistic (i.e., the observed score for a given set of items) is a function of the difficulty of the sample of items administered. Whenever items differ among tests, it becomes impossible to accurately report true ability levels because such statistics depend largely on how hard the test was.

Similarly, when the same items are administered to two groups of varying abilities, perhaps a regular education group as against a gifted and talent group, differences in item statistics (i.e., item difficulty and item discrimination) are a function of the abilities of the test takers. This problem is referred to as examinee sample dependence. The result is the inability to generalize item difficulty and discrimination statistics across groups. Of course, these limitations reflect a problem of circular dependence and, as such, the characteristics of items cannot be separated from the characteristics of the examinees.

A second limitation of CCT lies in the assumption that the measurement error variance is the same for scores of all persons to whom the instrument was administered. In reality, this simply is not the case since persons vary in their ability levels and thereby will vary in their accuracy in responding to items. A distribution of scores will reflect varying levels of measurement error for varying points in the distribution. Specifically, the scores for persons in the tails of a given distribution will tend to contain more measurement error if the test was designed for the average test taker and, thus, be more unreliable than scores in the middle of the distribution.

For example, if a person scores very low on the ACT, then the items may have been of such difficulty for the test taker that fatigue and disinterest occurred. Similarly, a high scorer may have also become disinterested or bored. The result is lower reliability for scores in the tails of the distribution and a subsequent inability to accurately discriminate between the true scores of persons in the tails. A test's standard error of measurement is based on the average reliability that CTT models and therefore confidence intervals are increasingly inaccurate as one examines scores in the extremes of a distribution.

A third limitation is that CTT does not allow for predicting how a person, with a given ability level, may respond to an item (Hambleton, Swaminathan, & Rogers, 1991).

This is a practical limitation of CTT. As Henard (1998) noted:

> Predicting how an individual examinee or a group of examinees will perform on a specific item is quite relevant to a number of testing applications. Consider the difficulties facing a test designer who wishes to predict test scores across multiple groups, or to design an equitable test for a particular group, or possibly to compare examinees who take either different tests or the same test at differing times. (p. 3)

Basics of IRT

IRT models were largely developed to overcome these limitations of CTT. There are several basic concepts that are central to understanding IRT.

Latent trait continuum. IRT models a person's response to a given item as a function of an unobservable latent trait that causes the response. This latent trait can be considered a continuum of ability and is continuous in scale. It is intuitive to recognize the existence of this trait. Consider, for example, the following multiple choice item intended to measure a respondent's knowledge of U.S. history (McKinley & Mills, 1989):

Which of the following states is the farthest north?

a. Iowa

b. Kentucky

c. Florida

d. Kansas

A respondent may clearly know that the answer is not Florida and probably not Kentucky. However, distinguishing between the geographic locals of Iowa and Kansas may prove more difficult. The respondent may get the item wrong, but clearly this does not indicate that the person has no knowledge of the geography examined by the question.

In this case, it is the latent trait of U.S. geography knowledge that causes the response. While an incorrect answer may be chosen, the respondent still possesses some degree of knowledge of U.S. geography, however minimal. IRT focuses on the latent trait of interest and models item difficulties based on an estimation of the trait. This differs from CTT which models item difficulties based on correct versus incorrect responses. IRT's method of modeling item difficulties will momentarily be explained in more detail when examining the heuristic example.

Item characteristic curve. The relationship of the latent trait of interest and a given item may be represented by an item characteristic curve (ICC). The ICC is a monotonically increasing ogive in which an item difficulty is the function of the latent trait and the probability of a correct response. Figure 1 illustrates this relationship for three items. As a person's ability increases (indicated by the Greek letter theta, $\underline{\theta}$) so does his or her probability of answering the item correctly, thus making the ICC monotonically increasing across $\underline{\theta}$ levels. Note also that the ICC is asymptotic, indicating that no person has either no ability or complete ability to bring to bear on a given item.

INSERT FIGURE 1 ABOUT HERE

Three estimated item parameters. While there are many different IRT models, the one-, two-, and three-parameter models are the most commonly used. The one-parameter Rasch model estimates the item difficulties, or the $\underline{b}$ parameter. The $\underline{b}$ parameter is established by the point on $\underline{\theta}$ where there is a 50% probability of answering the item correctly. Looking at Figure 1, we see that each of the three items modeled have different $\underline{b}$ values and, as such, have different difficulty levels with item 8 ($\underline{b}$ = -1.130) being the easiest. Another way of thinking about this relationship is that the area to the right of the ICC indicates the area in which responses are correct, and, conversely, the area to the left should indicate incorrect responses given the stated probabilities.

A closer look at the ICC should reveal that the item is most discriminating among persons with $\underline{\theta}$s near the 50% probability level. Again looking at Figure 1, item 8 discriminates best for those persons around the -1.130 level of $\underline{\theta}$. The item is least discriminating among persons in the tails of the ICC because for large changes in $\underline{\theta}$, there is not a corresponding degree of change in the probability of a correct response.

The two-parameter model derives both a $\underline{b}$ value and an item discrimination statistic, called the $\underline{a}$ parameter. The $\underline{a}$ parameter depicts the slope of the ICC and thereby indicates the discriminating power of the item as described above. As the slope gets steeper, the $\underline{a}$ statistic will get larger, thereby indicating that the item discriminates better around the point at which there is a 50% probability of a correct response. Figure 2 illustrates this parameter for three items. The $\underline{a}$ statistics are not given but it should be clear that item 13 would have the largest a value. Note also that while an item may discriminate best when it has a steep slope, the range in which it discriminates well decreases. For the one-parameter Rasch model, the a parameters are all assumed to be 1.0 and are not estimated.

INSERT FIGURE 2 ABOUT HERE

Finally, the three-parameter model allows for an additional parameter, or the guessing or c parameter, to be estimated. The c parameter takes into account the probable influences of guessing on responses. For example, a four-distracter multiple-choice item would have a 25% probability of a correct response based on chance alone. Consequently, an item's ICC would be asymptotic at the lower values of θ according to the probability at which a guessing effect could occur. Figure 3 illustrates this effect with item 14 having the largest c parameter. If the optimum level of discrimination is desired among test takers of similar ability, then the researcher would want b parameters near examinee θs, a parameters to be high, and c parameters to be low. The one-parameter Rasch model assumes the c parameter to be 0 for each item.

INSERT FIGURE 3 ABOUT HERE

IRT Assumptions

Like most statistic methods, IRT has several important assumptions, including monotonicity, unidimensionality of the latent trait, and local independence of items. Monotonicity has been previously discussed and will not be revisited here.

Unidimensionality. If person abilities ($\underline{\theta}$) are used to model the item statistics, then it is important that the items measure only one latent trait, otherwise we would be unsure what latent trait was impacting the item statistics. An obvious example of the possible confounding of latent traits is found in tests that require a considerable amount of reading or use of language but that are not intended to assess verbal skills. It is possible in this case that a person's response is just as much a function of his or her reading ability rather than the latent trait of interest.

This assumption is one that can only be met in degree. For example, all paper-and-pencil type tests require the test taker to bring reading ability to bear on the items. In fact, it is difficult, if not impossible, to conceptualize a testing situation in which truly only one trait was being measured. Test performance is often a corporate function of motivation, test anxiety, experience with tests, as well as a plethora of other factors. In reality, the unidimensionality assumption demands that there be a dominant trait that is being assessed (Hambleton et al., 1991). McKinley and Mills (1989) suggested that a principal components analysis be conducted to test this assumption and that the dominant trait should account for at least 30 to 40 percent of the total variance. While a valuable starting point, this method of testing the unidimensionality assumption is an incomplete solution to a complex problem, although it is beyond the scope of the present work to discuss the issue. Suffice it to say that whether or not this assumption has been met is ultimately left to researcher judgment.

Local independence. If items are to have statistics that are attributable to the item across samples, then each of the items must be answered independently from the other items (Hambleton & Swaminathan, 1985). This means that items must not contain information that can contribute to the response of other items. If this occurs, then item statistics will be confounded with those of related items and consequently misrepresent the item's parameters. Lord and Novak (1968) indicated the this assumption does not suggest that scores on the test items are uncorrelated across examinees, but rather that the item scores are uncorrelated at a given ability level.

The Advantages of IRT Models

Hambleton and Swaminathan (1985) noted three primary advantages of using IRT models as opposed to CTT approaches.

1. Given a large pool of items measuring the same latent trait, the estimate of a person's ability is independent of a particular sample of test items that are administered to the person.
2. Given a large sample of examinees, the item statistics are independent of the particular ability levels of the persons used for calibration of the statistics.
3. IRT provides a statistic that indicates the precision with which person abilities are estimated.

IRT advantages center on the models' ability to develop item and person ability statistics that are independent of each other. As previously discussed, CTT is unable to accomplish this due to item sample dependence and examinee sample dependence. IRT, however, theoretically performs calibrations that places the item and ability statistics in the same linear metric, thereby allowing for independent comparisons across pools of items and samples of examinees.

Illustrating the One-Parameter Rasch Model

While IRT models can be conceptually quite complex and powerful in their applications, the actual calculations in attaining the statistics of interest are relatively straightforward (Cantrell, 1997; Henard, 1998). Understanding the process by which the various IRT statistics are estimated is important to fully grasp the advantages that IRT purportedly possesses. Here the calculations for the Rasch model are illustrated using commonly available spreadsheets. Such a method is useful in providing concrete examples of IRT calibrations.

Deriving the Calibrations

The Rasch model estimates the $\underline{b}$ parameters for each item and person abilities ($\underline{\theta}$) for the sample. The $\underline{a}$ and $\underline{c}$ parameters are considered negligible and assumed to be 1.0 and 0, respectively. In the present example, suppose that 25 persons were tested on 20 items. These data are presented in Table 1 where a '1' indicates that the person answered the item correctly and a '0' marks an incorrect response. The final column lists the proportion of correct responses by each person (CTT person ability statistic) and the final row gives the proportion of correct responses to each item (CTT item difficulty). These data, as shown in Table 1, are then sorted in ascending order by person ability and descending order by item difficulty.

INSERT TABLE 1 ABOUT HERE

Persons either answering all items correctly or incorrectly will be removed from further analysis. In addition, all items that are either answered all correctly or incorrectly will be discarded. This is necessary because such persons and items provide indeterminant information regarding person ability or item difficuly. For example, item 20 was missed by all examinees. Responses to item 20 therefore contain no information to contribute to the calibration process. We simply do not know if the item was exactly hard enough for everyone to miss it or if its difficulty level far exceeded the abilities of the examinees. Likewise, persons answering all items correctly contribute no information because we do not know if they were just smart enough to get all items right or if their abilities go well beyond the difficulty of the test.

Using these criteria, we find that items 1 and 20 must be eliminated. After the removal of these items, persons 3, 9, and 22 are discarded because they have missed all remaining items or answered them all correctly. This process is repeated until no items or persons meet the noted criteria. Table 2 provides the sorted and edited data after this process.

INSERT TABLE 2 ABOUT HERE

The next step in the Rasch model is to perform conversion calibrations on the item difficulties and person abilities. This conversion is done to place both statistics in the *same metric* and in *linear* form. Such a conversion is necessary if we are going to be able to make item-free and examinee-free testing predictions with the statistics, such as the probability of answering an item correctly for a person with a given ability.

The conversion is partially accomplished by transforming the difficulties and abilities into *logits*. Logits for item difficulties are computed by taking the natural log of the proportion of items answered incorrectly divided by the proportion of items answered correctly, {$\underline{ln}[(\underline{p}-1)/\underline{p}]$}. Similarly, the calibration of person abilities is the natural log of the proportion of items a person answered correctly divided by the proportion he or she answered incorrectly, $\{\underline{\ln}[\underline{p}/(\underline{p}-1)]\}$.

Two important transformations occur at this stage. First, the manipulation of the proportions in this manner is necessary to begin to account for the previously discussed measurement error for those scores that lie in the tails of the distribution. Specifically, the extreme scores and abilities are spread out to allow for more discrimination among them. Second, the natural log conversion places the scores and abilities in a linear metric. Tables 3 and 4 illustrate the initial calibration process for each difficulty and ability, respectively, along with the calculations of logit variances which will be used in later computations.

INSERT TABLES 3 AND 4 ABOUT HERE

The relationship between proportions for item difficulties and subsequent logit conversions is illustrated in Figure 4. Note that the non-linear proportions are converted to a linear metric. Additionally, Table 5 expresses the relationship between person ability proportions and corresponding logit conversions. Note that the logits representing extreme proportions possess a greater spread between values. Looking at Table 5, we see that for proportions 0.98 and 0.99 the logits are 3.89 and 4.60, respectively, yielding a difference between logits of .71. However, for the proportions 0.50 and 0.51 the logits are 0.00 and 0.04, respectively, yielding a difference of .04 between logit values. Here again we see that the logit conversion allows for greater discrimination in the tails of a distribution.

INSERT FIGURE 4 AND TABLE 5 ABOUT HERE

The conversion to logits also overcomes yet another limitation of CTT. Henard (1998) correctly pointed out that "while item difficulty and person ability levels realistically range from negative infinity to positive infinity, the proportion correct/incorrect are bound by the values of zero and one" (p. 9). Logits, on the other hand, are not bound by zero and one, and thereby reflect the reality of an infinite range of difficulty and ability. Another step in the calibrations includes finding the logit mean and centering the logits to a mean of zero (see Tables 3 and 4). While logits theoretically may range from positive to negative infinity, as a practical matter the majority of difficulty and ability logits will range from +3 to -3.

Final estimates of item difficulties (b) and person abilities (θ) are computed by applying expansion factors to the initial logits. The item difficulties are corrected by the expansion factor allowing for sample spread. Such a correction yields final estimates of item difficulties that are independent of examinee characteristics. Additionally, person abilities are corrected by the expansion factor for test width, yielding estimates of person abilities that are independent of item characteristics. This, remember, is the goal and primary advantage of IRT. The formulas for the expansion factors and final estimates of difficulties and abilities are presented in Tables 6 and 7, respectively.

INSERT TABLES 6 AND 7 ABOUT HERE

Evaluating Model Fit

One step remains in the Rasch calculations. It is not appropriate to simply assume that the modeled difficulty and ability estimates are accurate. Rather, the model must be evaluated for goodness of fit with the data that were used to derive the calibrations (Hambleton & Cook, 1977). Table 8 contains the sorted and edited data as found in Table 2 but lists the calibrated estimates of person ability ($\underline{\theta}$) in the final column and item difficulty ($\underline{b}$) in the final row. Since the persons and items were initially sorted in ascending and descending order, respectively and by CTT proportions, we can see a pattern in which most of the correct responses are located toward the bottom and left of the table and incorrect responses are found in the top and right of the table.

The diagonal-type line running through the table represents the point at which the $\underline{b}$ and $\underline{\theta}$ values are equal. Remember that this is the point where there is a 50% probability of a correct response. As such, all responses above the line should be incorrect and marked by a '0' because the item's $\underline{b}$ exceeds the person's $\underline{\theta}$. In like fashion, all responses below the line should be correct and marked by a '1' because the person's ability exceeds the item difficulty. Those responses that do not fit this expectation are considered aberrant and must be evaluated. For example, responses to items 8, 12, and 19 by person 23 are aberrations. Similarly, item 8 has six aberrant responses by persons 12, 23, 1, 24, 7, and 18. Each of these items are circled in Table 8 for easy identification.

Both person 23 and item 8 must be evaluated for fit with the model.

INSERT TABLE 8 ABOUT HERE

A simulated fit analysis for person 23 is shown in Table 9. Here each of the person's responses are evaluated for fit by calculating the difference between their θ level and the item's difficulty depending upon whether they answered correctly or not. When the response is incorrect (x=0), we would expect that b would exceed θ. Subtracting b from θ should yield a negative value. Likewise, when the item is answered correctly (x=1) the expectation is that θ exceeds b and that a negative value would be attained from subtracting θ from b. When we do not attain the expected negative values, the person's responses are considered aberrant and the magnitude of the positive values attained indicate the degree of aberration.

For example, person 23's correct response to item 19 is very unexpected given the item's high level of difficulty (b=4.804) and person 23's ability (θ=-1.576). The 6.380 value (b-θ) illustrates the "misfit." For each "misfit," a z^2 value is calculated using the formula $z^2 = \exp|\ \theta - b\ |$, which yields the squared distance of an actual response from a given $\underline{\theta}$. The variance (V) is determined and used to calculate a $\underline{t}$-statistic using the formula indicated in Table 9. For person 23, $\underline{t}(16)=57.161$, which clearly exceeds the $\underline{t}$ critical value of 2.120 at alpha = .05. This indicates that person 23 is not a good fit with the model and should be removed.

INSERT TABLE 9 ABOUT HERE

In like manner, specific items are evaluated for fit using a similar process. Table 10 illustrates a fit analysis for item 8. Here we find that item 8 is a relative good fit for the model with t(20)=0.586 and should be retained.

INSERT TABLE 10 ABOUT HERE

Looking again at Table 8, we can see that these findings are not all that surprising. Aberrant but correct responses to item 8 (by person's 12 and 19) cause the fit line to be drawn between person's 23 and 1. However, we have discovered that person 23 is a poor fit with the model which can also be inferred from Table 8 by noting his or her seemingly random pattern of responses, possibly indicating a guessing effect. This is especially highlighted by person 23's correct response to the most difficult item on the test (19). In light of these findings, the aberrations for item 8 are probably due, at least in large part, to person 23's response. Therefore person 23 should be eliminated and item 8 retained.

After all items and persons are evaluated for fit with the Rasch model, the entire process iterates. New logits are calibrated and all items and persons are again evaluated. This iterative process continues until all items and persons fit well with the model developed. The final result yields item difficuties and person abilities that are theoretically independent of each other and can be used across samples and tests.

To test if the final calibrations are actually independent, Cantrell (1997) suggested that researchers perform a cross validation. Typically, this is accomplished by dividing a large sample with a large spread of ability into six ability groupings. The intent is to mirror the six groupings that are +/- 3 standard deviations in the normal curve. If the Rasch model worked well, the item difficulties for each of the six groups should resemble those found with the entire sample, thus indicating that the _b_ values are independent of the ability of the sample used to calculate them!

Of course, computer software packages perform all of these calculations. The present use of spreadsheets is an heuristic illustration only and certainly not necessary to develop Rasch models of measurement! However, this approach does help to conceptualize what is actually going on in Rasch modeling, which is critical in appreciating the utility of the method and using the statistics it provides. McKinley and Mills (1989) provide a listing of various software packages that are available to derive IRT models.

IRT versus CTT Revisited

It is important to note that not all researchers agree that the theoretical advantages of the Rasch model are real. Specifically, when comparing three data sets, Lawson (1991) found that both CTT and Rasch models yielded similar results with little variation between the obtained item difficulties and person abilities. Lawson questioned the need for a Rasch model given the mathematical procedures that it demands. In fact, the ICCs in Figure 1 and the proportion to logit plots in Figure 4 illustrate this point. In the middle of the distribution, the proportions actually approach a linear form. When sampling in the middle of a populations distribution there will be fewer differences between IRT and CTT results, because CTT targets the average of the distribution anyway. However, when sampling in the tails of the population distribution, IRT becomes more valuable.

Consider, for example, a graduate psychology training program that wishes to use GRE scores to determine entry cutoffs for new applicants. Successful applicants will most likely be high-end scorers on the GRE which would place their scores in the positive tail of the population distributions of all GRE scores. As previously emphasized, CTT is less accurate in discriminating between persons at these points in the distribution due to increased measurement error in extreme scores. Again, CTT reliability of scores is an average and is most accurate in the middle of the distribution. In our example of high-end GRE scores, items can be developed to target the higher ability levels of potential applicants and thereby be more accurate in discriminating between scores.

Adaptive Testing

The rise of computerized, adaptive testing testifies to this advantage. All test takers are administered items that have been previously calibrated through an IRT model. Each examinee is given those test items that are centered on examinee ability. As such, tests can obtain a more reliable estimate of person ability in a far shorter amount of time, also reducing fatigue related effects that would compound the inherent limitations of CTT.

Summary

IRT models of measurement, including the Rasch model, were largely developed to overcome some of the weaknesses in CTT. In IRT, examinee ability and item difficulty estimates can be obtained that are theoretically independent of each other and, therefore, can be used across samples of different abilities and items of varying difficulty. Because classical models attempt to maximize the average reliability across scores, scores in the tails of a distribution contain more measurement error than scores toward the middle of the distribution. IRT transforms CTT proportions into logits, thereby converting item difficulties and person abilities into the same linear metric across the distribution. In the Rasch model, this calibration process theoretically makes the ability statistic ($\underline{\theta}$) item-free and item difficulties ($\underline{b}$) examinee-free.

References

Cantrell, C. E. (1997, January). Item Response Theory: Understanding the one-parameter Rasch model. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED 415 281)

Fan, X. (1998). Item Response Theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement, 58, 357-381.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and applications. Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage.

Henard, D. H. (1998, April). Using spreadsheets for implement the one-parameter Item Response Theory (IRT) model. Paper presented at the annual meeting of the Southwestern Psychological Association, New Orleans. (ERIC Document Reproduction Service No. forthcoming)

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), Advances in education research: Substantive findings, methodological developments (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

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

McKinley, R. L., & Mills, C. N. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 71-135). Greenwich, CT: JAI Press.

Table 1
Sorted Responses for 25 Persons on 20 Items

	Items																				
	1	2	3	4	5	7	6	8	9	10	13	11	12	14	15	16	17	18	19	20	Score	p
Person																						
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.05
5	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0.10
12	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3	0.15
16	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0.15
6	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	4	0.20
4	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0.25
10	1	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0.30
23	1	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	6	0.30
1	1	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	7	0.35
24	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	7	0.35
7	1	1	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	8	0.40
18	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	8	0.40
21	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	10	0.50
14	1	1	1	0	1	1	0	1	1	1	1	1	0	1	0	0	0	0	0	0	11	0.55
2	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0	12	0.60
11	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	12	0.60
20	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	13	0.65
8	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0	0	0	14	0.70
13	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	15	0.75
19	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	15	0.75
25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	16	0.80
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	18	0.90
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	18	0.90
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	19	0.95
22	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	19	0.95
Item Score	25	24	20	17	17	16	15	15	13	13	12	11	11	10	9	9	7	5	3	0		
p	1.00	0.96	0.80	0.68	0.68	0.64	0.60	0.60	0.52	0.52	0.48	0.44	0.44	0.40	0.36	0.36	0.28	0.20	0.12	0.00		

Table 2
Edited and Sorted Responses for 21 Persons on 17 Items

	Items																	
	3	4	5	7	6	8	9	10	13	11	12	14	15	16	17	18	19	Score	p
Person																			
12	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	.059
16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	.059
6	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	.118
4	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	.176
10	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	4	.235
23	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	4	.235
1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	5	.294
24	1	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	5	.294
7	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	6	.353
18	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	6	.353
21	1	1	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	8	.471
14	1	0	1	1	0	1	1	1	1	1	0	1	0	0	0	0	0	9	.529
2	1	0	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	10	.588
11	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	10	.588
20	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	11	.647
8	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0	0	12	.706
13	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	13	.765
19	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	13	.765
25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	14	.824
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	16	.941
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	16	.941
Item Score	18	15	15	14	13	13	11	11	10	9	9	8	7	7	5	3	1		
p	.857	.714	.714	.667	.619	.619	.524	.524	.476	.429	.429	.381	.333	.333	.238	.143	.048		

Table 3
Initial Calibration of Item Difficulties: Converting Proportions into Logits

Item	Item Score	Frequency (F)	p	1-p	1-p/p	Logit Ratio=(x)	F(x)	x^2	$F(x^2)$	x-M=(b)	$F(M^2)$
3	18	1	0.857	0.143	0.167	-1.792	-1.792	3.210	3.210	-1.969	0.031
4,5	15	2	0.714	0.286	0.400	-0.916	-1.833	0.840	1.679	-1.093	0.063
7	14	1	0.667	0.333	0.500	-0.693	-0.693	0.480	0.480	-0.870	0.031
6,8	13	2	0.619	0.381	0.615	-0.486	-0.972	0.236	0.471	-0.663	0.063
9,10	11	2	0.524	0.476	0.909	-0.095	-0.191	0.009	0.018	-0.272	0.063
13	10	1	0.476	0.524	1.100	0.095	0.095	0.009	0.009	-0.082	0.031
11,12	9	2	0.429	0.571	1.333	0.288	0.575	0.083	0.166	0.111	0.063
14	8	1	0.318	0.619	1.625	0.486	0.186	0.236	0.236	0.309	0.031
15,16	7	2	0.333	0.667	2.000	0.693	1.386	0.480	0.961	0.516	0.063
17	5	1	0.238	0.762	3.200	1.163	1.163	1.353	1.353	0.986	0.031
18	3	1	0.143	0.857	6.000	1.792	1.792	3.210	3.210	1.615	0.031
19	1	1	0.048	0.952	20.000	2.996	2.996	8.974	8.974	2.819	0.031
		v = 17				Sums =	3.014		20.769		0.533
						M(logit mean) =		3.014 / 17			
						M =		0.177			
						M^2 =		0.031			
						V item =		20.769 - 0.533 / (17 - 1)			
						V item =		20.236 / 16			
						V item =		1.265			

Table 4
Initial Calibrations of Person Abilities: Changing Proportions into Logits

Possible Correct	Person Frequency (F)	p	1-p	1-p/p	Logit Ratio=(y)	F(y)	y^2	$F(y^2)$	y-M=($\underline{\theta}$)	$F(M^2)$
1	2	0.059	0.941	0.063	-2.773	-5.545	7.687	15.374	-2.640	0.036
2	1	0.118	0.882	0.133	-2.015	-2.015	4.060	4.060	-1.882	0.018
3	1	0.176	0.824	0.214	-1.540	-1.540	2.373	2.373	-1.407	0.018
4	2	0.235	0.765	0.308	-1.179	-2.357	1.389	2.778	-1.046	0.036
5	2	0.294	0.706	0.417	-0.875	-1.751	0.766	1.533	-0.742	0.036
6	2	0.353	0.647	0.545	-0.606	-1.212	0.367	0.735	-0.473	0.036
8	1	0.471	0.529	0.889	-0.118	-0.118	0.014	0.014	0.015	0.018
9	1	0.529	0.471	1.125	0.118	0.118	0.014	0.014	0.251	0.018
10	2	0.588	0.412	1.429	0.357	0.713	0.127	0.254	0.490	0.036
11	1	0.647	0.353	1.833	0.606	0.606	0.367	0.367	0.739	0.018
12	1	0.706	0.294	2.400	0.875	0.875	0.766	0.766	1.008	0.018
13	2	0.765	0.235	3.250	1.179	2.357	1.389	2.778	1.312	0.036
14	1	0.824	0.176	4.667	1.154	1.540	2.373	2.373	1.673	0.018
16	2	0.941	0.059	16.000	2.773	5.545	7.687	15.374	2.906	0.036
	n = 21				Sums =	-2.783		48.795		0.378

M(logit mean) = -2.783 / 21
M = -0.133
M^2 = 0.018

V ability = 48.795 - 0.378 / (21-1)
V ability = 48.417 / 20
V ability = 2.421

Table 5
Logit Converstion Chart: Proportion Correct to Personal Ability Logits (Uncorrected Theta)

Prop.	Logit	Prop.	Logit	Prop.	Logit	Prop.	Logit
0.01	-4.60	0.26	-1.05	0.51	0.04	0.76	1.15
0.02	-3.89	0.24	-0.99	0.52	0.08	0.77	1.21
0.03	-3.48	0.28	-0.94	0.53	0.12	0.78	1.27
0.04	-3.18	0.29	-0.90	0.54	0.16	0.79	1.32
0.05	-2.94	0.30	-0.85	0.55	0.20	0.80	1.39
0.06	-2.75	0.31	-0.80	0.56	0.24	0.81	1.45
0.07	-2.59	0.32	-0.75	0.57	0.28	0.82	1.52
0.08	-2.44	0.33	-0.71	0.58	0.32	0.83	1.59
0.09	-2.31	0.34	-0.66	0.59	0.36	0.84	1.66
0.10	-2.20	0.35	-0.62	0.60	0.41	0.85	1.73
0.11	-2.09	0.36	-0.58	0.61	0.45	0.86	1.82
0.12	-1.99	0.37	-0.53	0.62	0.49	0.87	1.90
0.13	-1.90	0.38	-0.49	0.63	0.53	0.88	1.99
0.14	-1.82	0.39	-0.45	0.64	0.58	0.89	2.09
0.15	-1.73	0.40	-0.41	0.65	0.62	0.90	2.20
0.16	-1.66	0.41	-0.36	0.66	0.66	0.91	2.31
0.17	-1.59	0.42	-0.32	0.67	0.71	0.92	2.44
0.18	-1.52	0.43	-0.28	0.68	0.75	0.93	2.59
0.19	-1.45	0.44	-0.24	0.69	0.80	0.94	2.75
0.20	-1.39	0.45	-0.20	0.70	0.85	0.95	2.94
0.21	-1.32	0.46	-0.16	0.71	0.90	0.96	3.18
0.22	-1.27	0.47	-0.12	0.72	0.94	0.97	3.48
0.23	-1.21	0.48	-0.08	0.73	0.99	0.98	3.89
0.24	-1.15	0.49	-0.04	0.74	1.05	0.99	4.60
0.25	-1.10	0.50	0.00	0.75	1.10		

Table 6
Final Estimate of Item Difficuties: Applying the Expansion Factor

Item	Initial Item Calibration	Sample Spread Expansion Factor	Corrected Item Calibration (b)
3	-1.969	1.704	-3.355
4, 5	-1.093	1.704	-1.862
7	-0.870	1.704	-1.482
6, 8	-0.663	1.704	-1.130
9, 10	-0.272	1.704	-0.463
13	-0.082	1.704	-0.140
11, 12	0.111	1.704	0.189
14	0.309	1.704	0.527
15, 16	0.516	1.704	0.879
17	0.986	1.704	1.680
18	1.615	1.704	2.752
19	2.819	1.704	4.804

Calculation of Expansion Factor for Sample Spread
EF = SQRT ([1+(V ability/2.89)]/{1-[(V ability*V item)/8.35]})
SQRT ([1+(2.421/2.89)]/{1-[2.421*1.265)/8.35]})
SQRT (1.8377/0.6332)
SQRT (2.9022)
EF = 1.704

Table 7
Final Estimate of Person Abilities: Applying the Expansion Factor

Possible Correct	Initial Ability Calibration	Test Width Expansion Factor	Corrected Ability Calibration ($\underline{\theta}$)
1	-2.640	1.507	-3.978
2	-1.882	1.507	-2.836
3	-1.407	1.507	-2.120
4	-1.046	1.507	-1.576
5	-0.742	1.507	-1.118
6	-0.473	1.507	-0.713
8	0.015	1.507	0.023
9	0.251	1.507	0.378
10	0.490	1.507	0.738
11	0.739	1.507	1.114
12	1.008	1.507	1.519
13	1.312	1.507	1.977
14	1.673	1.507	2.521
16	2.906	1.507	4.379

Calculation of Expansion Factor for Test Width

```
EF = SQRT ([1+(V item/2.89)]/{1-[(V item*V ability)/8.35]})
     SQRT ([1+(1.265/2.89)]/{1-[(1.265*2.421)/8.35]})
     SQRT (1.4377/0.6332)
     SQRT (2.2705)
EF = 1.507
```

Table 8
Model Fit Analysis: Identification of Aberrations for 21 Persons on 17 Items

	Items																	
	3	4	5	7	6	8	9	10	13	11	12	14	15	16	17	18	19	θ
Person																		
12	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	-3.978
16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-3.978
6	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	-2.836
4	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-2.120
10	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	-1.576
23	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	-1.576
1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	-1.118
24	1	1	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	-1.118
7	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	-0.713
18	1	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	-0.713
21	1	1	0	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0.023
14	1	0	1	1	0	1	1	1	1	1	0	1	0	0	0	0	0	0.378
2	1	0	1	1	1	1	0	1	1	1	0	1	1	0	0	0	0	0.738
11	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0.738
20	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1.114
8	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0	0	1.519
13	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	1.977
19	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1.977
25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	2.521
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	4.379
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	4.379
b	-3.355	-1.862	-1.862	-1.482	-1.130	-1.130	-0.463	-0.463	-0.140	0.189	0.189	0.527	0.879	0.879	1.680	2.752	4.840	

Note. For visual clarity, vertical lines may be drawn connecting the ends of the short horizontal lines.

Table 9
Fit Analysis for Person 23

θ = -1.576 (Calibrated Ability for Person 23)

Items	3	5	8	12	19
Aberrant Responses by Person 23	0	0	1	1	1
b	-3.355	-1.862	-1.130	0.189	4.804
"Misfits" If x=0, (2-b)	1.779	0.286			
If x=1, (b-2)			0.446	1.765	6.380
z^2	5.92	1.33	1.56	5.84	589.93

Sum of z^2 = 604.59

V = SOS/(v-1) where v=number of items
V = 604.59/(17-1)
V = 37.787

t(df=v-1) = {[ln(V)]+(V-1)}*{[(v-1)/8]**.5}
t(16) = [ln(37.787)]+(37.787-1)]*[(16/8)**.5]
t(16) = 57.161
t crit. = 2.120 at alpha = .05

Table 10
Fit Analysis for Item 8

b = -1.130 (Calibrated difficulty for Item 8)

			"Misfits"		
Person	Theta (θ)	Aberrant Responses	If x=0, (θ-b)	If x=1, (b-θ)	z^2
12	-3.978	1		2.848	17.253
23	-1.576	1		0.446	1.562
1	-1.118	0	0.012		1.012
24	-1.118	0	0.012		1.012
7	-0.713	0	0.417		1.517
18	-0.713	0	0.417		1.517

Sum of z^2 = 23.874

V = SOS/(n-1) where n=number of persons
V = 23.874/(21-1)
V = 1.1937

t(df=(n-1) = {[ln(V)]+(V-1)}*{[(n-1)/8]**.5}
t(20) = [ln(1.1937)]+(1.1937-1)]*[(20/8)**.5]
t(20) = 0.586
t crit. = 2.086 at alpha = .05

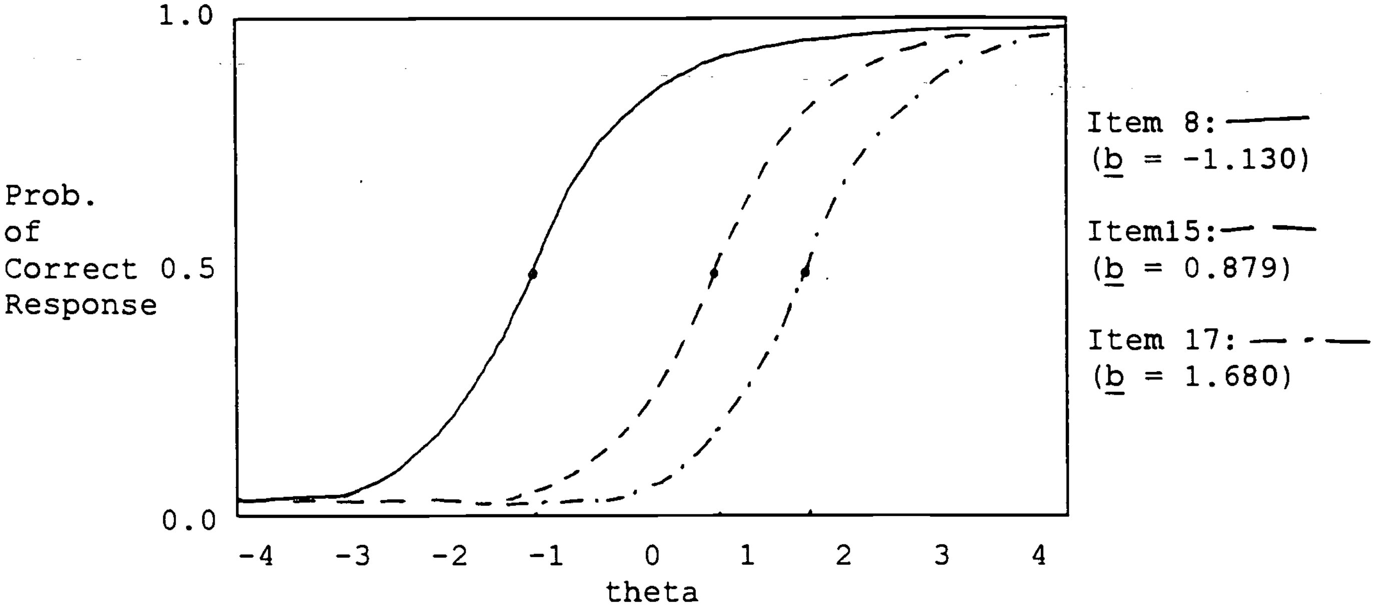

Figure 1. One-parameter item characteristic curves.

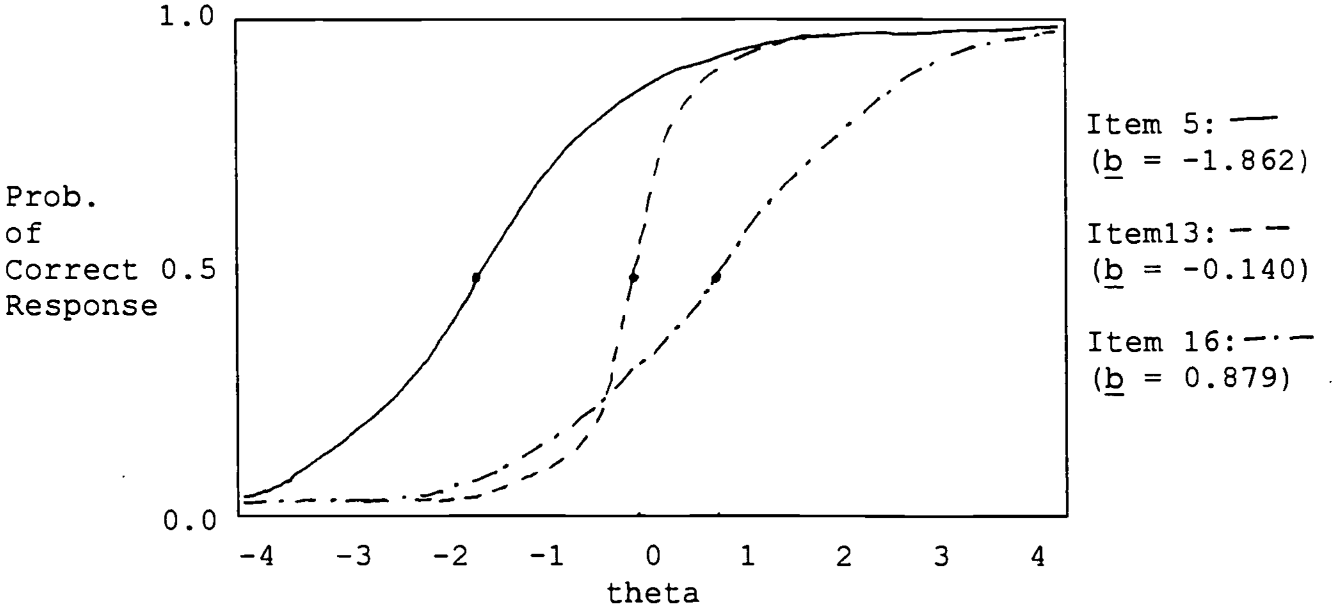

Figure 2. Two-parameter item characteristic curves.

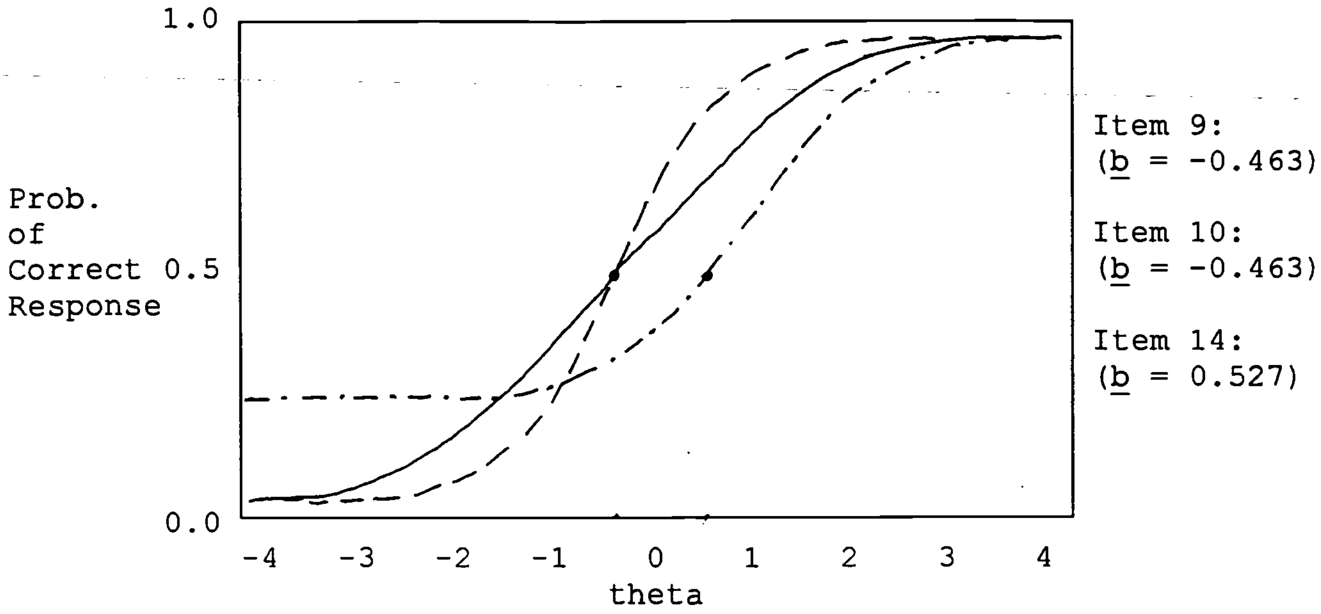

Figure 3. Three-parameter item characteristic curves.

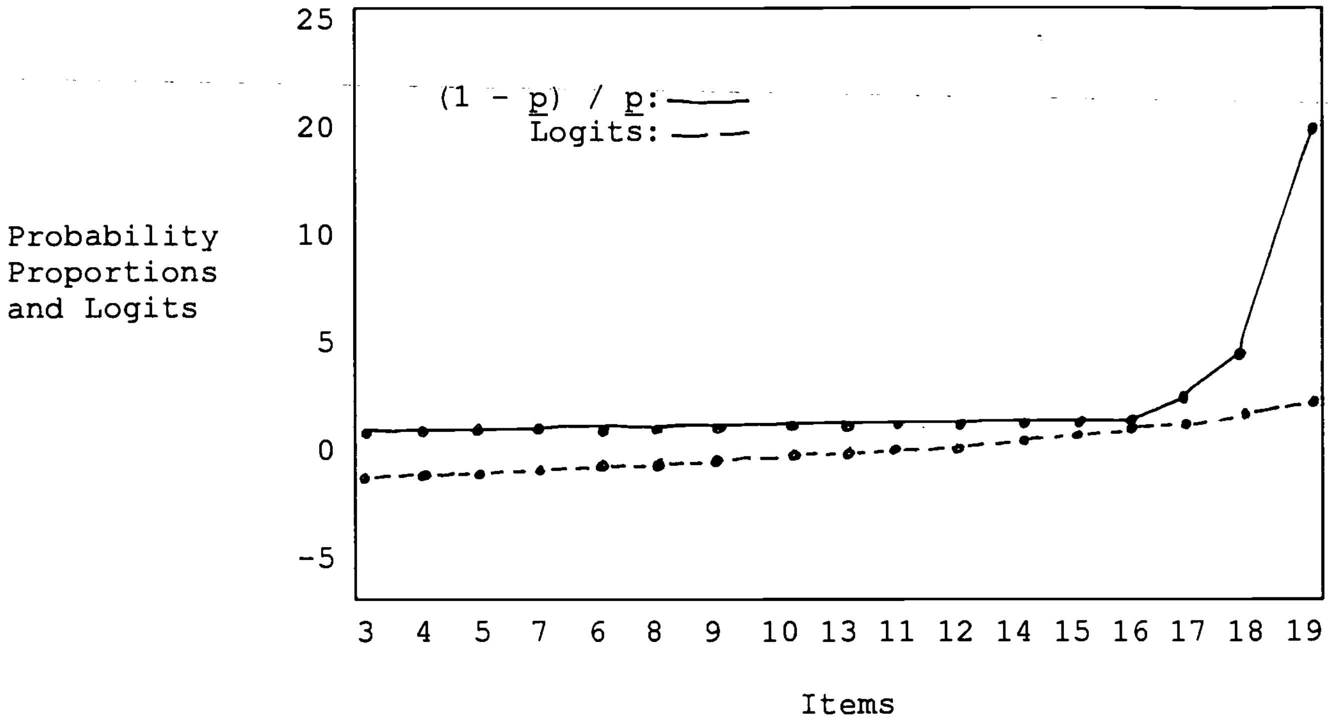

Figure 4. Graph of probability proportions and logits for all items. Items listed in descending order according to total item score.

TM029478

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: UNDERSTANDING THE ONE-PARAMETER RASCH MODEL OF ITEM RESPONSE THEORY	
Authors): ROBIN K. HENSON	
Corporate Source:	Publication Date: 1/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ ← **Sample sticker to be affixed to document**

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROBIN K. HENSON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document → □

or here
Permitting reproduction in other than paper copy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: X Robin K. Henson	Position: RES ASSOCIATE
Printed Name: Robin K. Henson	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1831
	Date: 1/13/99

OVER