Traditional and Modern Approaches to Outcomes Measurement

Ronald K. Hambleton
University of Massachusetts at Amherst

Summary

I have four goals in mind for this introductory presentation about item response theory and applications:

1. Address the strengths and weaknesses of classical test theory (CTT).

2. Introduce several of the assumptions, models, and features of item response theory (IRT).

3. Provide brief descriptions of five applications of IRT: developing tests, identifying differential item functioning (DIF), test score linking or equating, computer-adaptive testing, and score reporting

4. Offer some special concerns about applications of IRT to HRQOL, and concluding remarks.

Strengths and Weaknesses of CTT

CTT has been used in the test development field for over 80 years. The health-related research literature is full of examples of highly reliable and valid tests, and these tests have been used to produce research findings in thousands and thousands of studies. References to the CTT model with a focus on true score, observed score, and error score, the use of item difficulty (p-values) and item discrimination indices (r values), corrected split-half reliabilities and coefficient alphas, applications of the Spearman-Brown formula, corrections to correlations for range restriction, the standard error of measurement, and much more (see, for example, Gulliksen, 1950; Lord & Novick, 1968) are easily found in the research literature.

There will be no bashing from me of classical test theory and common approaches to test development and validation. Were these approaches to be used appropriately by test developers, the tests would be uniformly good, and the quality of research would be noticeably better.

At the same time, it is clear that classical test theory and related models and practices have some shortcomings and so they are not well suited for some of the demands being placed on measurement models today by two innovations: item banking and computer adaptive testing (see, for example, Mills, et al., 2002; Wainer, 2000; van der Linden & Glas, 2000). One shortcoming is that item statistics are dependent on the particular choice of respondent samples. This shortcoming makes classical item statistics (such as item difficulty levels, and biserial and point biserial correlations) problematic in

an item bank unless all of the test item statistics are coming from the same respondent sample, which of course, is highly unlikely, and to make such a requirement would lower the utility of item banks in test development. A second shortcoming is that respondent scores are highly dependent on the particular choice of items. Give an "easy" set of items to respondents and they will score high, and give a "hard" set of items to them, and they will score low. This dependence of test scores on items creates major problems when computer-adaptive testing is used. In principle, in a CAT testing environment, respondents will see items "pitched" or "matched" to their ability levels, and so, in general, respondents will be administered "non-parallel" tests and so the scores themselves will not provide an adequate basis for comparing respondents to each other or even to a set of norms.

These two shortcomings are serious drawbacks to the use of classical test theory item statistics with item banking and computer-adaptive testing but there are more. Typically, classical test models provide only a single estimate of error (i.e., the standard error of measurement) and it is applied to the scores of all respondents. But this means that the error estimate is probably too large for the bulk of "middle ability respondents" and too small for respondents scoring low or high on the ability scale. Also, classical test theory models the performance of respondents at the test score level (recall "X=T + E") and computer adaptive testing requires modeling between candidate ability and items at the item level so that optimal item selections can be made. Finally, items and respondents are reported on separate, and non-comparable scales in classical measurement. This makes it nearly impossible to implement optimal assessment where items are selected to improve the measurement properties of the test for each respondent, or a prior ability distribution.

Assumptions, Models, and Features of IRT

Item response theory is a statistical framework for linking respondent scores to the items on an test to the trait or traits that are believed to be measured by that test. A mathematical model must be specified that provides the "link" (for example, the two-parameter logistic model) between these item scores and the traits. It is common to assume there is a single trait underlying respondent performance, but models for handling multiple traits are readily available (see, van der Linden & Hambleton, 1997). With a model specified, and the respondent item scores available, respondent and item parameters can be estimated.

Until about 15 years ago, most of the work with item response modeling of data was limited to models that could be applied to dichotomously-scored data—the one-, two-, and three-parameter logistic models (Hambleton, Swaminathan, & Rogers, 1991). Nationally normed achievement tests (e.g., **California Achievement Tests**, **Metropolitan Achievement Tests**), university admissions tests (e.g., **Scholastic Assessment Test**, the **Graduate Management Admissions Test**, and the **Graduate Record Exam**) and many state proficiency tests consist of multiple-choice tests that are scored 0-1. But today, more use is being made of test data arising from new item types that use polytomous-scoring (i.e., item level data that is scored in more than two

categories) such as rating scales for recording attitudes and values, and scoring writing samples and complex performances.

Numerous IRT models are available today for analyzing polytomous response data. Some of these models were available in the late 1960s (see Samejima's chapter in van der Linden & Hambleton, 1997) but were either too complex to apply, were not sufficiently well developed (for example, parameter estimation was problematic), or software for applying the models was simply not available.

These polytomous IRT response models seem especially important in the health-related area because so much of the data that is being produced from these psychological tests and scales does not fit a 0-1 item level scoring model. This seems appropriate because with many psychological questions, such as "How do you feel?" a dichotomous response such as "good" or "bad" does not capture the full range of options that might be expected from respondents. A rating scale with categories such as Excellent, Very Good, Good, Fair, and Poor would generate considerably more information from respondents and cover the range of expected responses better.

Polytomous IRT models like the graded response model can handle an unlimited number of score categories for items, but for practical reasons, current IRT software is normally limited to a maximum of about 20 score categories per item. Parameter estimation becomes problematic when the number of score categories becomes large. Five to 10 score categories is not usually a problem in practice, as long as respondent samples to the items are at least as large as 200. Larger samples are always better.

Best known of the polytomous response IRT models is Samejima's graded response model (GRM). Samejima fits a two-parameter logistic function to the probability of obtaining a particular score or a higher score on a rating scale:

$$P_{ix}^{*}(\theta) = \frac{e^{Da_i(\theta - b_{ix})}}{1 + e^{Da_i(\theta - b_{ix})}} \quad \text{where } i = 1,2,...,n \text{ and } x = 0,1,..., m_i$$

Here, $P_{ix}^{*}(\theta)$ is the probability that a respondent with a trait score of θ on item i will obtain a score of x or higher, and is often called a "cumulative score category function" or "cumulative category response curve" (CCRC). These probabilities are defined for each of n items in the test and for each possible score on the item, ranging from 0 to a maximum of m_i. Similar to dichotomous IRT models, D is the scaling constant set equal to 1.7 (sometimes this constant is even left out of the model), a_i is the item discrimination parameter while b_{ix} is called the location parameter for score x and denotes the point on the θ scale where $P_{ix}^{*}(\theta)$=.50. Trait score (the label used to describe what the test is measuring (e.g., "overall health"), and this must be determined through content, predictive, and construct validity investigations. Trait scores, or ability scores, as they are commonly called, are often scaled to a mean of zero and a standard deviation of one for convenience. If the scores are reported to respondents, it is common to apply a

linear transformation to the trait scores first to place them on a more convenient scale (for example, a scale with mean = 100, SD= 10) that does not contain negative numbers and decimals.

IRT models are based on strong assumptions about the data: In the case of the GRM, for example, they are (1) the assumption of test unidimensionality and (2) the assumption that two-parameter logistic functions will match the actual data (Hambleton, Swaminathan, & Rogers, 1991). Other models specify different assumptions about the data (see, for example, van der Linden & Hambleton, 1997). Failure to satisfy model assumptions can lead to problems—for example, expected item and ability parameter invariance may not be present, and using CCRCs to build tests, when these curves do not match the actual data, will result in tests that will function differently in practice than expected.

The assumption of test unidimensionality is that the items in the test are measuring a single dominant trait. Now, in practice, most tests are measuring more than a single trait, but good model fit requires only a reasonably good approximation to the unidimensionality assumption. One check on unidimensionality that sometimes is applied is this: From a consideration of the items in the test, would it be meaningful to report a single score for respondents? Is there a factor common to the items such as "overall health" or "attitudes about health" Multidimensionality in a dataset might result from several causes: First, the items may cluster into distinct groups of health topics that do not correlate highly with each other. Second, the use of multiple item formats (e.g., checklists, rating scales, open-ended questions) may lead to distinct "method of assessment" factors. Third, multidimensionality might result from dependencies in the data. For example, if responses to one item are conditional on responses to others, multidimensionality is introduced. (Sometimes this type of dimensionality can be eliminated by scoring the set of related items as if it were a "testlet" or "super item.")

Many different methods are available to explore the dimensionality of item response data (see, Hambleton, Swaminathan, & Rogers, 1991). Various reviews of several older methods have found them all in one way or another to have shortcomings. While this point may be discouraging, methods are available that will allow the researcher to draw defensible conclusions regarding unidimensionality. For example, linear factor analysis (e.g., principal components analysis), nonlinear factor analysis, and multidimensional scaling may be used for this purpose, though not without some problems at the interpretation stage. Further, it is recognized that few constructs are reducible to a strictly unidimensional form and that demonstration of a dominant single trait may be all that is reasonable. For example, using principal components analysis we would expect a dominant first factor to account for roughly 20 percent or more of the variance in addition to being several times larger than the second factor. Were these conditions met, the assumption of unidimensionality holds to a reasonable degree.

Descriptions of Five IRT Applications

Very brief introductions to five popular applications of IRT follow. Books by Hambleton, Swaminathan, and Rogers (1991), Mills, et al. (2002), van der Linden and Glas (2000), and Wainer (2000) provide many more details on the applications.

Developing Tests. Two of the special features of IRT modeling are item and test information functions. For each item, a function is available indicating where on the reporting scale an item is helpful in estimating ability and how much it contributes to an increase in measurement precision. Basically, items provide the most measurement around their "b-value" or level of difficulty and the amount of information depends on its discriminating power. The test information function (which is a simple sum of the information functions for items in a test) provides an overall impression of how much information a test is providing across the reporting scale. The more information a test provides at a point on the reporting scale, the smaller the measurement error will be. In fact, the standard error of measurement at a point on the reporting scale (called the "conditional standard error of measurement") is inversely related to the square root of the test information at that point.

A test information function is the result of putting a particular set of items into a test. Therefore, sometimes, it is specified in advance as the "target" and then items can be selected to produce the test of statistical interest. Item selection often becomes a task of selecting items to meet content specifications, and statistical specifications (as reflected in a "target information function"). One of the newest IRT topics (called "automated test assembly" is the development of procedures for allowing test developers to define the test of interest in considerable detail, translate those specifications into mathematical equations, and then with the appropriate software, actually select test items from a bank of calibrated test items to meet the requirements for the test (see, van der Linden, in press; van der Linden & Glas, 2000).

Identifying DIF. The property of item parameter invariance is immensely useful in test development work, but not something that can be assumed with IRT models. The property must be demonstrated across sub-populations of the population for whom the test is intended. This might be male and females; Blacks, Whites, and Hispanics; well-educated and less well-educated; older, middle age, and younger respondents; etc. Basically, IRT DIF analyses involve comparing the item characteristic curves (for 0-1 data) or, say, the CCRCs (for polytomous response data fitting the GRM) obtained in these subpopulations. Much of the research has investigated different ways to summarize the differences between these ICCs/CCRCs (see, Hambleton, Swaminathan, & Rogers, 1991). DIF via IRT modeling is not especially easy to implement (because of the number of steps involved), but the easy graphing capabilities of ICCs and CCRCs makes DIF interpretation more understandable to many practitioners.

Test Score Linking or Equating. In many practical testing situations, such as achievement testing, it is desirable to have multiple versions or forms of a test. For example, a test like the **Scholastic Assessment Test** would quickly become of limited value if the same test items were used over and over again. Every high school senior would be going to Kaplan to get an advanced look at the questions. Items would become

known to test takers and passed on to others about to take the test. Test validity would drop to zero quickly. So, while multiple forms of a test may be a necessity, it is also important that these tests be statistically equivalent so that respondents do not benefit or placed at a disadvantage because of the form of the test they were administered. Proper test development is invaluable in producing near equivalent tests, but it is no guarantee, and so "statistical equating" is carried out to link comparable scores on pairs of tests. Statistical equating can be carried out with classical or IRT modeling, but it tends to be easier to do with IRT models and with a bit more flexibility. There is some evidence too that IRT equating may produce a better matching of scores at the low and high end of the ability scale (see, for example, Hambleton, Swaminathan, & Rogers, 1991).

With attitude and personality tests, equating of multiple forms of a test may not be important because often only a single form of each test exists. But what may be important are efforts to link or equate scores on tests designed by different researchers but, in the main, that purport to assess the same construct. With linking procedures, aggregation of results obtained across multiple tests measuring the same construct may be possible.

Computer-Adaptive Testing (CAT). "Adapting a test" to the performance of the respondent as he/she is working through the test has always been viewed as a very good idea because of the potential for shortening the length of tests. The basic idea is to focus on administering items where the particular answer of the respondent is the most uncertain—that is, items of "medium difficulty" for the respondent. When testing is done at a computer, ability can be estimated after the administration of each item, and then the ability estimate provides a basis for the selection of the next test items. Testing can be discontinued when the level of precision that is desired for ability estimates is achieved. It is not uncommon for the length of testing to be cut in half with "computer adaptive testing." The computer provides the mechanism for ability estimation and item selection. Item response theory provides a measurement framework for estimating abilities and choosing items. The property of "ability parameter invariance" makes it possible to compare respondents to each other or standards that may have been set despite the fact that they almost certainly took collections of items that differed substantially in their "difficulty." Without IRT, computer-adaptive testing would lose many of its advantages. Computer-adaptive testing remains today as one of the best applications of IRT (see, Wainer, 2000).

Score Reporting. One of the special features of IRT modeling of data is that the item statistics and respondent abilities are reported on the same scale. This feature creates the possibility to make a score reporting scale more meaningful by defining points along the scale in terms of the test items and how they are functioning at points along the reporting scale. For example, a point on the scale might be defined by the statements to which a candidate might agree with, with a probability of, say, 80% probability or higher. By defining a number of points in the same way, it becomes possible to make meaningful distinctions among respondents scoring at different points along the ability continuum.

Special Concerns and Concluding Remarks

The health outcomes field provides some unique challenges for persons interested in applying IRT models to their data. One challenge arises because of the potential for multidimensional data. All of the popular IRT models are based on the assumption of a single dominant factor underlying performance on the test. It remains to be seen to what extent health outcome measures are multidimensional, how that multidimensionality can be detected, and how it might be handled or modeled when it is present. A second challenge is associated with model fit. IRT models are based on strong assumptions about the data, and when they are not met, advantages of IRT modeling are diminished or lost. At the same time, approaches to addressing model fit, remain to be worked out, especially the extent to which model misfit can be present without destroying the validity of the IRT model application.

Finally, and as I have said in other papers, IRT is not a magic wand that can be used to fix all of the mistakes in test development such as (1) the failure to properly define the construct of interest, (2) ambiguous items, and (3) flawed test administrations. At the same time, it has been demonstrated many times over that IRT models, when they fit the data, and when other important features of sound measurement are present, IRT models provide an excellent basis for developing tests, and providing valid scores for making decisions about individuals and groups.

References

Gulliksen, H. (1950). **Theory of mental tests**. New York: Wiley.

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

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

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (Eds.). (2002). **Computer based testing: Building the foundation for future assessments**. Hillsdale, NJ: Lawrence Erlbaum Associates.

Van der Linden, W. (in press). **Linear models for optimal test design**. New York: Springer.

Van der Linden, W., & Glas, C. (Eds.). (2000). **Computer adaptive testing: Theory and practice**. Boston, MA: Kluwer Academic Publishers.

Van der Linden, W., & Hambleton, R. K. (Eds.). (1997). **Handbook of modern item response theory**. New York: Springer-Verlag.

Wainer, H. (Ed.). (2000). **Computerized adaptive testing: A primer** (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Publishers.

Version: July 1, 2004