Item response theory: applications of modern test theory in medical education

Steven M Downing

Context Item response theory (IRT) measurement models are discussed in the context of their potential usefulness in various medical education settings such as assessment of achievement and evaluation of clinical performance.

Purpose The purpose of this article is to compare and contrast IRT measurement with the more familiar classical measurement theory (CMT) and to explore the benefits of IRT applications in typical medical education settings.

Summary CMT, the more common measurement model used in medical education, is straightforward and intuitive. Its limitation is that it is sample-dependent, in that all statistics are confounded with the particular sample of examinees who completed the assessment. Examinee scores from IRT are independent of the particular sample of test questions or assessment stimuli. Also, item characteristics, such as item difficulty, are independent of the particular sample of examinees. The IRT characteristic of invariance permits easy equating of examination scores, which places scores on a constant measurement scale and permits the legitimate comparison of student ability change over time. Three common IRT models and their statistical assumptions are discussed. IRT applications in computer-adaptive testing and as a method useful for adjusting rater error in clinical performance assessments are overviewed.

Conclusions IRT measurement is a powerful tool used to solve a major problem of CMT, that is, the confounding of examinee ability with item characteristics. IRT measurement addresses important issues in medical education, such as eliminating rater error from performance assessments.

Keywords education, medical/*methods; *clinical competence; educational measurement/*standards; psychometrics/*education/standards; computers/*standards; England.

Medical Education 2003;37:739–745

Classical measurement theory: basic concepts and limitations

Consider the following common examination situation. A professor of anatomy gives a written test to students from a large medical school class. All of the test items were newly written for this year's test, but the test was designed to measure exactly the same anatomy content as last year's examination. The anatomy professor notes that the current year's students score considerably lower on the current test than last year's students; the current year's anatomy test appears to be far more difficult than last year's examination. The professor concludes that this year's anatomy students are less able, that is, not as proficient in anatomy, as last year's students. Would the professor's conclusion be correct?

The anatomy professor's dilemma is frequently encountered in medical education settings. When traditional classical measurement theory (CMT) is used as the measurement model, scores from all assessments suffer from a statistical confounding of student ability with the inherent difficulty (or easiness) of the test item or performance stimulus. This is so whether the assessments are written, objectively scored tests of cognitive achievement, or performance or skill examinations such as objective structured clinical examinations (OSCEs) or simulated-patient (SP) examinations. Measurement based on item response theory (IRT) will not automatically solve the anatomy professor's dilemma, but IRT has a set of easily implemented statistical procedures for placing scores on a common score scale. so that the interpretation of those scores is identical over time.

Correspondence: Steven M Downing, Associate Professor of Medical Education, University of Illinois at Chicago, College of Medicine, Department of Medical Education (MC 591), Chicago, Illinois 60612–7309, USA

Key learning points

Item response theory (IRT) measurement eliminates the confounding of test difficulty and student proficiency.

This measurement model has a procedure for effectively eliminating one major source of rater measurement error from clinical performance assessments.

IRT measurement techniques encourage the construction of tests that are closely targeted to the student's ability.

Measurement using IRT provides an estimate of the standard error of measurement for each student's score.

IRT testing techniques are essential for 'adaptive' computer-based testing.

CMT is familiar to most medical educators, although they may not label it such. Dating from the early twentieth century, CMT has been used successfully for nearly one hundred years and continues to be the most frequently used measurement model for classroom assessment, whether for testing young children's achievement in their educational fundamentals or the mastery of medical students and residents of the fund of knowledge in the domain of the medical sciences.[1–3]

All the statistics and statistical procedures that are most commonly used for evaluation and quality control of assessment in medical education, such as reliability analysis, item analysis, and the appraisal of item and test difficulty and discrimination, are derived from CMT, sometimes referred to as 'true score theory'.

Prior to discussing so-called modern measurement theory,[4,5] or item response theory, let us review the basic concepts of classical measurement. The model itself is simply stated*: Every test score is composed of two components: true score and error*. Stated as a simple formula,

$$X = T + e$$

where X = the observed score on the assessment; T = the true score, and e = random errors of measurement.

The three essential terms of CMT require some explanation. The observed score (X) is the observed quantity or score computed from an assessment; the raw score may be the number of correct responses to objectively scored test questions, the percent-correct score on an examination, or the ratings provided by teaching faculty of clinical performance in clerkships or residency rotations, or the checklist item scores or ratings obtained from a standardized patient assessment. This observed score is composed of two important and independent quantities, according to CMT: the true score and the error components.

The 'true score' warrants some discussion, since it is the most important component of the CMT model. The true score is typically defined as the 'long-run average or mean score'. This definition is consistent with the usual definition of μ, the population mean, in the statistical literature. The true score, which can never be known with certainty or observed directly, can only be estimated. The true score is the mean of all the scores which would be obtained on the test, if the same or an exactly equivalent test were to be repeatedly administered to the same students an infinite number of times. Generally, all of CMT is concerned in some way with the estimation of true score or errors of measurement.

The error in CMT is always random, non-systematic error, and is assumed to be independent or uncorrelated with true score. Random error consists of all the random and uncontrolled 'noise' or conditions that interfere with the precise and accurate measurement of the examinee's true ability or proficiency. Such random error can be attributed to poorly created test items or stimuli, inadequate testing conditions, internal examinee states such as illness or inattention or fatigue and so on, all contributing measurement error to the observed scores of examinees.

The strengths of CMT are many. It is fairly easily understood, at least in its basic statistics. The mathematics of CMT are no more complex than summation algebra. It is straightforward and mostly intuitive. Students take a test; they typically receive one point if they get a test item correct and 0 points if they get an item incorrect. For most locally developed assessments in medical education, CMT serves the user and the students well, and provides sufficient precision of measurement to minimize decision errors, such as passing students who really should fail.

However, in some assessment settings, the inherent confounding of the difficulty of the test item or other stimulus material with the underlying ability or proficiency of the examinee is troublesome. The anatomy professor's wish to know whether students are more or less able or proficient from year to year is a good example. CMT can not answer that question directly, since the difficulty of the questions and the ability estimates of the students are hopelessly confounded, and are reported on measuring scales that differ from each other and which change over different

administrations of the same or equivalent tests. Students' scores are test-dependent, in that the magnitude of their score depends on which test they happened to take. For example, if an anatomy student receives a test score of 60% correct on a test that is fairly difficult overall, does that 60% score have the same meaning for a student who happened to take a test that was considerably easier overall? (Both CMT and IRT offer equating methods which, when successfully applied, keep the score scale constant over time so that scores have exactly the same meaning across test administrations. IRT equating methods are typically more easily implemented than CMT equating procedures).

All the common item statistics typically used to evaluate assessments (e.g. item difficulty, item discrimination) are completely dependent on the particular sample of students who happened to take the test. This issue of confounded item difficulty and student ability is typically not a major problem if large numbers of students take the test and if the underlying ability of examinees is fairly constant over time. (CMT scores, if successfully equated statistically, can provide a legitimate comparison of student ability changes from year to year).

Item response theory (IRT) measurement models

One solution to this confounding problem is to utilize a different type of measurement model: item response theory (IRT) measurement. These psychometric models solve the sample-dependency problem and the problem of confounding of item difficulty and student ability because they have the theoretical attribute of *invariance*. This means that if a mathematical model (an IRT model) can be statistically fitted to the observed assessment data and all the assumptions are met, student ability can be estimated independently of the specific questions on the test, and the item and test characteristics, such as item difficulty, can be estimated independently of the specific group or sample of students taking the test. The IRT invariance characteristic sounds almost too good to be true and may strain credulity at times. However, it can be proven mathematically that the invariance trait is true, and empirical studies can show evidence that, in particular examination data, the assumption of invariance is a reasonable one.[6]

IRT refers to a class of psychometric measurement models used to estimate examinee ability on the trait being measured and the difficulty of the examination items on the same scale. This scale which has the property of invariance such that examinee ability is estimated independently of the particular set of items administered, and item difficulty is estimated independently of the particular sample of examinees taking those items.

These psychometric models, with roots in the 1920s, have been actively researched since at least the 1960s and have been applied widely since the 1980s.[7–9] The advent of computer-based testing, and especially computer-adaptive testing, has moved the IRT discipline forward in major ways. The availability of IRT software for personal computers has further extended the use and the usefulness of IRT measurement.[6]

What is the catch? If these IRT measurement models have this almost magical method of estimating student ability in a manner that is independent of the particular set of questions asked, why are not all assessments currently using IRT measurement?

Dimensionality

In IRT measurement, as with all statistics, there are a set of assumptions that must be met. The fundamental assumption for all of the commonly used IRT models is *unidimensionality*. The test or assessment must measure a single, unified underlying 'trait' or construct. For the most frequently used IRT models, if the test measures more than a single construct (is multidimensional), the IRT model will not fit the data and thus can not be used to estimate examinee ability and item characteristics. There are many ways to empirically evaluate the dimensionality of tests, the most common of which are factor-analytic methods using item-level intercorrelations as the input to the factor analysis.[10]

Local independence

Test item characteristics and population parameters, such as difficulty and discrimination, are estimated in a process called 'item calibration'. For test questions to be successfully calibrated by an IRT measurement model, the items in the test must be 'locally independent', meaning that, for example, the answer to one test question cannot depend on the answer to any other test question on the examination. This prerequisite also pertains to CMT, but the IRT models may not be as robust to violations of this assumption as CMT.

Sample size

To work properly, IRT models require fairly sizable samples of examinees. The minimum number of

examinees required to properly fit the simplest IRT model (the one-parameter model) is approximately 200. As model complexity increases (i.e. as more parameters or population item characteristics are estimated), the greater the number of subjects required. The relatively large numbers of examinees required for IRT measurement may be the most limiting factor for its use in typical medical education settings. But sample size is also an issue with CMT item analysis data, such as item difficulty and item discrimination indices; for these indices to be completely reliable and stable, sample sizes must also be near 200.

The large number of students required for IRT is much more critical for item characteristic estimation than for the estimation of student ability. If estimation of examinee ability is of most interest, as is likely in typical classroom settings, sample sizes as low as 50 could be sufficient, in some cases, for fairly accurate estimation of student ability using the one-parameter IRT model.[6] However, in order to accurately estimate student ability, the IRT statistical model must be shown to 'fit the data' by applying appropriate model fit statistics. These fit statistics typically require a minimum of about 200 examinees to work properly. So, in actual practice, IRT measurement models require large sample sizes, with approximately 200 students needed, as a minimum, for the most straightforward model, the one-parameter IRT model.

How IRT works

The core assumption of IRT is that the probability of an examinee's answering a test question correctly depends on the examinee's underlying ability with regard to the trait being measured and on the statistical characteristics of the test item. Further, this relationship between the probability of answering the question correctly and the examinee's ability can be described by a mathematical function called an 'item characteristic curve' (ICC). In order to carry out this IRT scaling analysis, an appropriate mathematical model must be selected: a model that can be empirically demonstrated to fit the data and meets all of the required assumptions. (Most of the common IRT measurement models use a logistic function to mathematically model this relationship between examinee ability and test item characteristics.)

Much of the IRT literature is concerned with the advantages and disadvantages of various IRT models, in various testing settings, and used for varying purposes. Research on the appropriateness of fit statistics and item characteristic estimation methods comprises much of the contemporary IRT measurement literature.[5]

Common IRT models

There are three IRT models commonly used for tests that are scored as 'right' or 'wrong' (i.e. as 0 or 1, or dichotomously). These models are named for the number of parameters they use to estimate examinee ability. (A parameter is a population statistic, such as the population mean, μ, which is usually estimated by a sample statistic, such as the sample mean).

The one-parameter IRT model is also known as the Rasch model, after its originator.[9,11] The Rasch model uses only a single parameter, item difficulty, to estimate item and student characteristics. If the assessment data fit the Rasch model, examinee ability on the trait measured can be accurately estimated. The one-parameter IRT model is widely used throughout the world, in many different medical education settings. For example, the National Board of Medical Examiners (NBME), the agency providing the national medical licensure examinations in the United States, uses the Rasch model to calibrate its tests and to provide basic psychometric information to test developers.

Because the one-parameter IRT model requires the fewest number of examinees, this IRT model is potentially the most useful for medical educators who have reasonably large class sizes.

Two other IRT models, the two-parameter and the three-parameter models, are also widely used, especially for large-scale assessments. The two-parameter model adds an item discrimination parameter (in addition to item difficulty) and the three-parameter model adds a 'guessing' parameter (pseudo-chance) to item difficulty and item discrimination. (The 'guessing' parameter accounts for the probability of arriving at the correct answer, in a selected-response question, by chance alone.) The two- and three-parameter models typically fit large-scale unidimensional achievement data well. These models are used, at least experimentally or in conjunction with CMT statistics, for many large-scale assessments in North America, including some medical specialty board certifying examinations and professional licensure examinations, such as nursing.

IRT statistics

Testing programmes using IRT models usually continue to use and report CMT statistics, so that many users of these examinations are not aware that IRT models are used to construct, score and analyse test data, and may even be used in the determination of who passes or fails these tests.

In the jargon of IRT measurement, examinee ability is estimated on what is usually called a theta scale, which typically is constrained to range from −4·0 to +4·0, with 0 as the mean. (The theta scale is rarely used for reporting scores, since many examinees would be upset to receive negative scores or be told that their score of '0' is an 'average' score.)

One of the major advantages of IRT measurement models is that they provide a number of very useful statistics which are not available in CMT. For example, the standard error of measurement (SEM) in IRT is computed for every observed score, rather than only for the test as a whole (as SEM is computed in CMT). Thus, test developers have a good estimate of the exact precision of measurement for each obtained score on the test, and thus, for every individual examinee's score. This is especially important for tests that have a pass–fail point. Ideally, achievement tests are constructed to have their greatest precision of measurement near the passing score. IRT measurement, with its emphasis on SEMs for each score level, facilitates the selection of test questions that maximize the precision of measurement at the pass–fail point.

While the concept of 'test reliability' certainly pertains to all assessments, including those which are IRT-based, the more general concept of test reliability is augmented and expanded in IRT measurement by statistics called 'item information functions' or 'item information curves'.

Item information functions graphically display the contribution of test items to the assessment of examinee ability at various ability levels. In general, the higher the item discrimination, the more 'information' is contributed by that item to the estimation of ability. The item information function is maximized by highly discriminating items closely targeted to the students' ability.

IRT and computer-based testing

Computer-based testing (CBT) is growing in popularity throughout the world. Many large-scale, high-stakes examinations in North America and Europe are currently administered in CBT formats, as either linear or adaptive examinations. Linear CBTs use a computer to present test questions and generally score the examination as soon as the examinee completes the test. Linear CBTs typically present fixed forms of a test, which means that a fixed number of test questions are pre-selected by human test developers as a test form. Several different forms may be prepared for each test administration and a specific test form is then randomly selected to be administered to an individual test-taker.

Computer-adaptive tests, on the other hand, are a very specific type of CBT in which each individual examination item is selected by the computer to be presented to an examinee, based on the (estimated) ability of the examinee and the known difficulty of the question. Thus, the test 'adapts' to the ability of the examinee as the test proceeds. One great advantage of adaptive CBTs is that, using IRT measurement models, highly precise estimates of examinee ability can be achieved using tests that are shorter than typical paper-and-pencil achievement tests. However, in practice, the actual length of the test will also be determined by the content requirements of the assessment, since the adaptive test must also meet specific content specifications in order to adequately sample the domain of knowledge being assessed.

Adaptive testing, although very efficient in terms of examinee time, requires large numbers of IRT pre-calibrated test items, since the 'exposure' of each unique item must be carefully controlled to ensure the continuing security of test items and the integrity of the examination scores. A completely adaptive test is administered as follows. An examinee is first presented with a test question of medium difficulty. If the examinee gets that item correct, the computer selects a slightly harder test question as the second question, and so on, until the examinee answers one or two questions incorrectly. Then, the computer is programmed to administer a slightly easier question, and so on, until the IRT-estimated difficulty of the items matches the estimated ability of the candidate. (Typical 'stopping rules' for adaptive CBTs specify the exact precision of ability estimation required for the test to stop, which means that the test administrator knows the IRT-derived standard error of measurement for the examinee's score just as the testing ends).

There are several types of computer-adaptive test delivery systems in use.[12] The completely adaptive test delivery method, discussed above, is the most complex to manage, since many variables (for example, item difficulty, current examinee ability estimate, item exposure rate and content considerations) are in play simultaneously and must be managed by the computer system in real time. One variation that simplifies CBT administration and content control is an adaptive test that is fixed in length and has predetermined content coverage, so that examinees are routed through a series of smaller tests (testlets) of varying difficulty (easy, medium, hard), depending on their IRT-estimated ability at the branching points.

One great advantage of computer-adaptive testing is that very precise estimates of examinee ability can be achieved with the administration of fewer test

questions, within the constraints of testing the appropriate content. Computer-adaptive testing requires IRT measurement, since every examinee may receive a unique set of test questions. Thus, test items must have item difficulty estimates which have been precalibrated using an IRT measurement model, so that the computer can select test questions that match, as closely as possible, the actual ability of the examinee. The exact targeting of item difficulties to examinee ability usually greatly increases the precision of measurement and the efficiency of test administration. The IRT property of invariance is essential for computer-adaptive testing.

The adaptive testing model does not, however, alleviate the requirement for the test developer to pay careful attention to the content tested and to balance the content with exact specifications for the examination. In situations where the content is complex and covers many unique disciplines (e.g. all the basic sciences), the requirement for adequate content sampling may considerably increase the numbers of items needed for adequate assessment, even in an adaptive testing environment. IRT measurement can not override the need for content-related validity evidence.

Other applications of IRT in medical education

In typical medical education settings, many assessments use rating scale data. For example, most clinical performance assessments consist of ratings by clinical teachers, preceptors or other teaching faculty. Often, multiple raters complete clinical performance ratings for medical students and residents over the course of a clinical learning experience such as a clerkship, preceptorship or residency rotation. Most measurement error associated with clinical performance ratings is attributed to the raters, as opposed to the rating scale, the items rated or the students.[13]

While several methodological techniques, such as generalizability theory[14] are available for evaluating the measurement error contributed by raters of clinical performance, IRT measurement has a tool that not only estimates the measurement error contributed by raters but allows for the adjustment of ratings to statistically eliminate this type of rater error.

A variant of the one-parameter model calibrates rating scale data (estimates the ability of students) and the rating score differences due to rater measurement error, and adjusts the final rating data to reduce or eliminate rater error. Thus, one major source of unreliable rating data is removed and the validity evidence for the clinical performance ratings is improved. (Other sources of measurement error may remain in the ratings, however. For example, this IRT procedure will not reduce or eliminate measurement error resulting from the rater-by-student interaction.)

Software to accomplish these rater adjustments, using an extension of the Rasch model, is available.

Software

IRT measurement is no longer just experimental or primarily used for methodological research. Many different PC software applications for calibrating tests, using various IRT models, are commercially available and fairly inexpensive. At least one open-source IRT software application is currently available as freeware. While many of these software applications are less than 'user-friendly', training workshops are frequently offered by software authors and professional organizations.

Issues in IRT measurement

IRT measurement can be a useful tool. But, like all tools, IRT must be used properly or more harm than good may result. The assumptions for IRT measurement, especially those of unidimensionality and local independence, must be met for successful application of IRT models to real test data. Both assumptions are empirically testable using various correlational techniques, but to carry out these analyses successfully, sufficiently large representative samples of examinees must be available.

Sample sizes are also a practical issue and a limitation for successful application of IRT methods. As discussed earlier, at least 200 examinees are typically needed for the Rasch model in order to test the fit of the model to the actual test data. Up to 1000 or more examinees may be needed for the three-parameter IRT model. Fewer examinees could be used for estimation of student ability levels, but the standard errors would be greater than for estimates based on larger samples of students.

The statistics used to test the goodness of fit of the IRT model to the data are also controversial and problematic, especially for the two- and three-parameter models.[15] There is little consensus on which statistics to use, or how to definitively evaluate those that are used or on what actions to take if the statistics indicate misfitting data.

The most active areas of current theoretical research in IRT concern new and complex IRT models (for example, models useful for nominal data; for polytomous data, such as those obtained from rating scales; for multidimensional data; for locally dependent data, and

for performance examination data, such as from complex simulations). Much of the current applied research in IRT deals with the application of IRT models to computer-adaptive testing. Major themes include the following: types of adaptive delivery (for example, real-time, completely adaptive testing, various mastery methods, and testlet or panel delivery of preconstructed mini-tests); control of item exposure methods; adaptive test security issues, and item development issues, focusing on how to develop and pre-test the large quantities of test items which are required for adaptive testing, especially test items of medium difficulty and maximum IRT information.

Conclusion

Item response theory measurement is a powerful testing tool which estimates examinee proficiency and item and test difficulty on the same scale. If all the statistical assumptions are met, and the test data fit the IRT model, the vexing problem of the confounding of item difficulty and student ability estimates is eliminated. IRT measurement provides powerful new tools to medical educators for more precisely estimating true student ability in cognitive disciplines, and provides a method of adjusting and effectively eliminating one major source of measurement error (rater error) in contexts such as the assessment of clinical performance.

The application of IRT measurement is not beyond the reach of many medical educators, in appropriate and useful settings. Given the general availability of IRT software, and with some collaborative assistance from personnel with training and experience in the application of IRT to typical medical education measurement problems, these measurement techniques can be successfully applied in many different medical education settings.

Acknowledgements

The author wishes to thank Georges Bordage MD PhD, Thomas M. Haladyna PhD, Rachel Yudkowsky MD MHPE, and an anonymous reviewer for their critical reviews of this manuscript.

Funding

There was no external funding for this project.

References

1 Spearman C. The proof and measurement of association between two things. *Am J Psychol* 1904;**15**:72–101.
2 Novick MR. The axioms and principal results of classical test theory. *J Mathemat Psychol* 1966;**3**:1–18.
3 Lord FM. *A Theory of Test Scores. Psychometric Monographs no. 7.* New York: Psychometric Society; 1952.
4 Lord, FM. *Application of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Erlbaum; 1980.
5 Van der Linden W, Hambleton R, eds. *Handbook of Modern Item Response Theory*. New York: Springer; 1997.
6 Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of Item Response Theory*. Newbury Park, California: Sage Publications; 1991.
7 Fisher RA. On the mathematical foundations of theoretical statistics. *Philosoph Trans* 1921;**222**:309–68.
8 Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In: FM Lord, MR Novick, eds. *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley; 1968: 395–479.
9 Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press; 1960.
10 McLeod L, Swygert KA, Thissen D. Factor analysis for items scored in two categories. In: D Thissen, H Wainer, eds. *Test Scoring.* Mahwah, New Jersey: Erlbaum; 2001; 189–216.
11 Wright BD, Stone MH. *Best Test Design*. Chicago: MESA Press; 1979.
12 Folk VG, Smith RL. Models for delivery of CBTs. In: CN Mills, MT Potenza, JJ Fremer, WC Ward, eds. *Computer-Based Testing: Building the Foundation for Future Assessments*. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2002; 41–66.
13 Kreiter CD, Ferguson KJ. Examining the generalizability of ratings across clerkships using a clinical evaluation form. *Eval Health Professions* 2001;**24**:36–46.
14 Brennan RL. *Generalizability Theory*. New York: Springer-Verlag; 2001.
15 Van der Linden WJ, Hambleton RK. Item response theory. Brief history, common models, and extensions. In: WJ van der Linden, RK Hambleton, eds. *Handbook of Modern Item Response Theory*. New York: Springer-Verlag; 1997: 1–28.

Received 30 *October* 2002; *editorial comments to authors* 27 *January* 2003; *accepted for publication* 26 *March* 2003