Item response theory

In psychometrics, **item response theory** (**IRT**) also known as **latent trait theory**, **strong true score theory**, or **modern mental test theory**, is a paradigm for the design, analysis, and scoring of tests, questionnaires, and similar instruments measuring abilities, attitudes, or other variables. Unlike simpler alternatives for creating scales as the simple sum questionnaire responses it does not assume that each item is equally difficult. This distinguishes IRT from, for instance, the assumption in Likert scaling that "*All items are assumed to be replications of each other or in other words items are considered to be parallel instruments*" [1] (p. 197). By contrast, item response theory treats the difficulty of each item (the ICCs) as information to be incorporated in scaling items.

It is based on the application of related mathematical models to testing data. Because it is generally regarded as superior to classical test theory, it is the preferred method for developing scales, especially when optimal decisions are demanded, as in so-called high-stakes tests e.g. the Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT).

The name *item response theory* is due to the focus of the theory on the item, as opposed to the test-level focus of classical test theory. Thus IRT models the response of each examinee of a given ability to each item in the test. The term *item* is generic: covering all kinds of informative item. They might be multiple choice questions that have incorrect and correct responses, but are also commonly statements on questionnaires that allow respondents to indicate level of agreement (a rating or Likert scale), or patient symptoms scored as present/absent, or diagnostic information in complex systems.

IRT is based on the idea that the probability of a correct/keyed response to an item is a mathematical function of person and item parameters. The person parameter is construed as (usually) a single latent trait or dimension. Examples include general intelligence or the strength of an attitude. Parameters on which items are characterized include their difficulty (known as "location" for their location on the difficulty range), discrimination (slope or correlation) representing how steeply the rate of success of individuals varies with their ability, and a pseudoguessing parameter, characterising the (lower) asymptote at which even the least able persons will score due to guessing (for instance, 25% for pure chance on a 4-item multiple choice item).

Overview

The concept of the item response function was around before 1950. The pioneering work of IRT as a theory occurred during the 1950s and 1960s. Three of the pioneers were the Educational Testing Service psychometrician Frederic M. Lord,[2] the Danish mathematician Georg Rasch, and Austrian sociologist Paul Lazarsfeld, who pursued parallel research independently. Key figures who furthered the progress of IRT include Benjamin Drake Wright and David Andrich. IRT did not become widely used until the late 1970s and 1980s, when personal computers gave many researchers access to the computing power necessary for IRT.

Among other things, the purpose of IRT is to provide a framework for evaluating how well assessments work, and how well individual items on assessments work. The most common application of IRT is in education, where psychometricians use it for developing and refining exams, maintaining banks of items for exams, and equating for the difficulties of successive versions of exams (for example, to allow comparisons between results over time).[3]

IRT models are often referred to as *latent trait models*. The term *latent* is used to emphasize that discrete item responses are taken to be *observable manifestations* of hypothesized traits, constructs, or attributes, not directly observed, but which must be inferred from the manifest responses. Latent trait models were developed in the field of sociology, but are virtually identical to IRT models.

IRT is generally regarded as an improvement over classical test theory (CTT). For tasks that can be accomplished using CTT, IRT generally brings greater flexibility and provides more sophisticated information. Some applications, such as computerized adaptive testing, are enabled by IRT and cannot reasonably be performed using only classical test theory. Another advantage of IRT over CTT is that the more sophisticated information IRT provides allows a researcher to improve the reliability of an assessment.

IRT entails three assumptions:

1. A unidimensional trait denoted by θ;
2. Local independence of items;
3. The response of a person to an item can be modeled by a mathematical *item response function* (IRF).

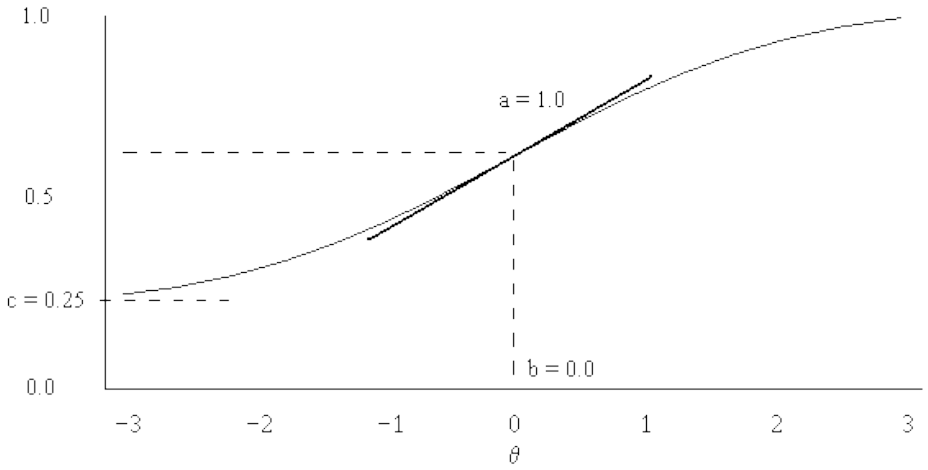
The trait is further assumed to be measurable on a scale (the mere existence of a test assumes this), typically set to a standard scale with a mean of 0.0 and a standard deviation of 1.0. 'Local independence' means that items are not related except for the fact that they measure the same trait, which is equivalent to the assumption of unidimensionality, but presented

separately because multidimensionality can be caused by other issues. The topic of dimensionality is often investigated with factor analysis, while the IRF is the basic building block of IRT and is the center of much of the research and literature.

The item response function

The IRF gives the probability that a person with a given ability level will answer correctly. Persons with lower ability have less of a chance, while persons with high ability are very likely to answer correctly; for example, students with higher math ability are more likely to get a math item correct. The exact value of the probability depends, in addition to ability, on a set of *item parameters* for the IRF.

Three parameter logistic model



For example, in the *three parameter logistic* (3PL) model, the probability of a correct response to an item *i* is:

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}}$$

where θ is the person (ability) parameter and a_i, b_i, and c_i are the item parameters. The item parameters simply determine the shape of the IRF and in some cases have a direct interpretation. The figure to the right depicts an example of the 3PL model of the ICC with an overlaid conceptual explanation of the parameters.

The item parameters can be interpreted as changing the shape of the standard logistic function:

$$P(t) = \frac{1}{1 + e^{-t}}$$

In brief, the parameters are interpreted as follows (dropping subscripts for legibility); *b* is most basic, hence listed first:

- *b* – difficulty, item location: $p(b) = (1 + c)/2$, the half-way point between c_i (min) and 1 (max), also where the slope is maximized.
- *a* – discrimination, scale, slope: the maximum slope $p'(b) = a \cdot (1 - c)/4.$
- *c* – pseudo-guessing, chance, asymptotic minimum $p(-\infty) = c.$

If $c = 0$, then these simplify to $p(b) = 1/2$ and $p'(b) = a/4$, meaning that *b* equals the 50% success level (difficulty), and *a* (divided by four) is the maximum slope (discrimination), which occurs at the 50% success level. Further, the logit (log odds) of a correct response is $a(\theta - b)$ (assuming $c = 0$): in particular if ability *θ* equals difficulty *b*, there are even odds (1:1, so logit 0) of a correct answer, the greater the ability is above (or below) the difficulty the more (or less) likely a correct response, with discrimination *a* determining how rapidly the odds increase or decrease with ability.

In words, the standard logistic function has an asymptotic minimum of 0 ($c = 0$), is centered around 0 ($b = 0$, $P(0) = 1/2$), and has maximum slope $P'(0) = 1/4$. The a parameter stretches the horizontal scale, the b parameter shifts the horizontal scale, and the c compresses the vertical scale from $[0, 1]$ to $[c, 1]$. This is elaborated below.

The parameter b_i represents the item location which, in the case of attainment testing, is referred to as the item difficulty. It is the point on θ where the IRF has its maximum slope, and where the value is half-way between the minimum value of c_i and the maximum value of 1. The example item is of medium difficulty since b_i=0.0, which is near the center of the distribution. Note that this model scales the item's difficulty and the person's trait onto the same continuum. Thus, it is valid to talk about an item being about as hard as Person A's trait level or of a person's trait level being about the same as Item Y's difficulty, in the sense that successful performance of the task involved with an item reflects a specific level of ability.

The item parameter a_i represents the discrimination of the item: that is, the degree to which the item discriminates between persons in different regions on the latent continuum. This parameter characterizes the slope of the IRF where the slope is at its maximum. The example item has a_i=1.0, which discriminates fairly well; persons with low ability do indeed have a much smaller chance of correctly responding than persons of higher ability.

For items such as multiple choice items, the parameter c_i is used in attempt to account for the effects of guessing on the probability of a correct response. It indicates the probability that very low ability individuals will get this item correct by chance, mathematically represented as a lower asymptote. A four-option multiple choice item might have an IRF like the example item; there is a 1/4 chance of an extremely low ability candidate guessing the correct answer, so the c_i would be approximately 0.25. This approach assumes that all options are equally plausible, because if one option made no sense, even the lowest ability person would be able to discard it, so IRT parameter estimation methods take this into account and estimate a c_i based on the observed data.[4]

IRT models

Broadly speaking, IRT models can be divided into two families: unidimensional and multidimensional. Unidimensional models require a single trait (ability) dimension θ. Multidimensional IRT models model response data hypothesized to arise from multiple traits. However, because of the greatly increased complexity, the majority of IRT research and applications utilize a unidimensional model.

IRT models can also be categorized based on the number of scored responses. The typical multiple choice item is *dichotomous*; even though there may be four or five options, it is still scored only as correct/incorrect (right/wrong). Another class of models apply to *polytomous* outcomes, where each response has a different score value.[5][6] A common example of this are Likert-type items, e.g., "Rate on a scale of 1 to 5."

Number of IRT parameters

Dichotomous IRT models are described by the number of parameters they make use of.[7] The 3PL is named so because it employs three item parameters. The two-parameter model (2PL) assumes that the data have no guessing, but that items can vary in terms of location (b_i) and discrimination (a_i). The one-parameter model (1PL) assumes that guessing is a part of the ability and that all items that fit the model have equivalent discriminations, so that items are only described by a single parameter (b_i). This results in one-parameter models having the property of specific objectivity, meaning that the rank of the item difficulty is the same for all respondents independent of ability, and that the rank of the person ability is the same for items independently of difficulty. Thus, 1 parameter models are sample independent, a property that does not hold for two-parameter and three-parameter models. Additionally, there is theoretically a four-parameter model (4PL), with an upper asymptote, denoted by d_i, where $1 - c_i$ in the 3PL is replaced by $d_i - c_i$. However, this is rarely used. Note that the alphabetical order of the item parameters does not match their practical or psychometric importance; the location/difficulty (b_i) parameter is clearly most important because it is included in all three models. The 1PL uses only b_i, the 2PL uses b_i and a_i, the 3PL adds c_i, and the 4PL adds d_i.

The 2PL is equivalent to the 3PL model with $c_i = 0$, and is appropriate for testing items where guessing the correct answer is highly unlikely, such as fill-in-the-blank items ("What is the square root of 121?"), or where the concept of guessing does not apply, such as personality, attitude, or interest items (e.g., "I like Broadway musicals. Agree/Disagree").

The 1PL assumes not only that guessing is not present (or irrelevant), but that all items are equivalent in terms of discrimination, analogous to a common factor analysis with identical loadings for all items. Individual items or individuals might have secondary factors but these are assumed to be mutually independent and collectively orthogonal.

Logistic and normal IRT models

An alternative formulation constructs IRFs based on the normal probability distribution; these are sometimes called *normal ogive models*. For example, the formula for a two-parameter normal-ogive IRF is:

$$p_i(\theta) = \Phi\left(\frac{\theta - b_i}{\sigma_i}\right)$$

where **Φ** is the cumulative distribution function (cdf) of the standard normal distribution.

The normal-ogive model derives from the assumption of normally distributed measurement error and is theoretically appealing on that basis. Here $b_i$is, again, the difficulty parameter. The discrimination parameter is σ_i, the standard deviation of the measurement error for item *i*, and comparable to 1/a_i.

One can estimate a normal-ogive latent trait model by factor-analyzing a matrix of tetrachoric correlations between items.[8] This means it is technically possible to estimate a simple IRT model using general-purpose statistical software.

With rescaling of the ability parameter, it is possible to make the 2PL logistic model closely approximate the cumulative normal ogive. Typically, the 2PL logistic and normal-ogive IRFs differ in probability by no more than 0.01 across the range of the function. The difference is greatest in the distribution tails, however, which tend to have more influence on results.

The latent trait/IRT model was originally developed using normal ogives, but this was considered too computationally demanding for the computers at the time (1960s). The logistic model was proposed as a simpler alternative, and has enjoyed wide use since. More recently, however, it was demonstrated that, using standard polynomial approximations to the normal *cdf*,[9] the normal-ogive model is no more computationally demanding than logistic models.[10]

The Rasch model

The Rasch model is often considered to be the 1PL IRT model. However, proponents of Rasch modeling prefer to view it as a completely different approach to conceptualizing the relationship between data and the theory.[11] Like other statistical modeling approaches, IRT emphasizes the primacy of the fit of a model to observed data,[12] while the Rasch model emphasizes the primacy of the requirements for fundamental measurement, with adequate data-model fit being an important but secondary requirement to be met before a test or research instrument can be claimed to measure a trait.[13] Operationally, this means that the IRT approaches include additional model parameters to reflect the patterns observed in the data (e.g., allowing items to vary in their correlation with the latent trait), whereas the Rasch approach requires both the data fit the Rasch model and that test items and examinees confirm to the model, before claims regarding the presence of a latent trait can be considered valid. Therefore, under Rasch models, misfitting responses require diagnosis of the reason for the misfit, and may be excluded from the data set if substantive explanations can be made that they do not address the latent trait.[14] Thus, the Rasch approach can be seen to be a confirmatory approach, as opposed to exploratory approaches that attempt to model the observed data. As in any confirmatory analysis, care must be taken to avoid confirmation bias.

The presence or absence of a guessing or pseudo-chance parameter is a major and sometimes controversial distinction. The IRT approach includes a left asymptote parameter to account for guessing in multiple choice examinations, while the Rasch model does not because it is assumed that guessing adds randomly distributed noise to the data. As the noise is randomly distributed, it is assumed that, provided sufficient items are tested, the rank-ordering of persons along the latent trait by raw score will not change, but will simply undergo a linear rescaling. Three-parameter IRT, by contrast, achieves data-model fit by selecting a model that fits the data,[15] at the expense of sacrificing specific objectivity.

In practice, the Rasch model has at least two principal advantages in comparison to the IRT approach. The first advantage is the primacy of Rasch's specific requirements,[16] which (when met) provides *fundamental* person-free measurement (where persons and items can be mapped onto the same invariant scale).[17] Another advantage of the Rasch approach is that estimation of parameters is more straightforward in Rasch models due to the presence of sufficient statistics, which in this application means a one-to-one mapping of raw number-correct scores to Rasch θ estimates.[18]

Analysis of model fit

As with any use of mathematical models, it is important to assess the fit of the data to the model. If item misfit with any model is diagnosed as due to poor item quality, for example confusing distractors in a multiple-choice test, then the items may be removed from that test form and rewritten or replaced in future test forms. If, however, a large number of misfitting items occur with no apparent reason for the misfit, the construct validity of the test will need to be reconsidered and the test specifications may need to be rewritten. Thus, misfit provides invaluable diagnostic tools for test developers, allowing the hypotheses upon which test specifications are based to be empirically tested against data.

There are several methods for assessing fit, such as a chi-square statistic, or a standardized version of it. Two and three-parameter IRT models adjust item discrimination, ensuring improved data-model fit, so fit statistics lack the confirmatory diagnostic value found in one-parameter models, where the idealized model is specified in advance.

Data should not be removed on the basis of misfitting the model, but rather because a construct relevant reason for the misfit has been diagnosed, such as a non-native speaker of English taking a science test written in English. Such a candidate can be argued to not belong to the same population of persons depending on the dimensionality of the test, and, although one parameter IRT measures are argued to be sample-independent, they are not population independent, so misfit such as this is construct relevant and does not invalidate the test or the model. Such an approach is an essential tool in instrument validation. In two and three-parameter models, where the psychometric model is adjusted to fit the data, future administrations of the test must be checked for fit to the same model used in the initial validation in order to confirm the hypothesis that scores from each administration generalize to other administrations. If a different model is specified for each administration in order to achieve data-model fit, then a different latent trait is being measured and test scores cannot be argued to be comparable between administrations.

Information

One of the major contributions of item response theory is the extension of the concept of reliability. Traditionally, reliability refers to the precision of measurement (i.e., the degree to which measurement is free of error). And traditionally, it is measured using a single index defined in various ways, such as the ratio of true and observed score variance. This index is helpful in characterizing a test's average reliability, for example in order to compare two tests. But IRT makes it clear that precision is not uniform across the entire range of test scores. Scores at the edges of the test's range, for example, generally have more error associated with them than scores closer to the middle of the range.

Item response theory advances the concept of item and test information to replace reliability. Information is also a *function* of the model parameters. For example, according to Fisher information theory, the item information supplied in the case of the 1PL for dichotomous response data is simply the probability of a correct response multiplied by the probability of an incorrect response, or,

$$I(\theta) = p_i(\theta) q_i(\theta).$$

The standard error of estimation (SE) is the reciprocal of the test information of at a given trait level, is the

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

Thus more information implies less error of measurement.

For other models, such as the two and three parameters models, the discrimination parameter plays an important role in the function. The item information function for the two parameter model is

$$I(\theta) = a_i^2 p_i(\theta) q_i(\theta).$$

The item information function for the three parameter model is

$$I(\theta) = a_i^2 \frac{(p_i(\theta) - c_i)^2}{(1 - c_i)^2} \frac{q_i(\theta)}{p_i(\theta)}$$

[19]

In general, item information functions tend to look bell-shaped. Highly discriminating items have tall, narrow information functions; they contribute greatly but over a narrow range. Less discriminating items provide less information but over a wider range.

Plots of item information can be used to see how much information an item contributes and to what portion of the scale score range. Because of local independence, item information functions are additive. Thus, the test information function is simply the sum of the information functions of the items on the exam. Using this property with a large item bank, test information functions can be shaped to control measurement error very precisely.

Characterizing the accuracy of test scores is perhaps the central issue in psychometric theory and is a chief difference between IRT and CTT. IRT findings reveal that the CTT concept of reliability is a simplification. In the place of reliability, IRT offers the test information function which shows the degree of precision at different values of theta, θ.

These results allow psychometricians to (potentially) carefully shape the level of reliability for different ranges of ability by including carefully chosen items. For example, in a certification situation in which a test can only be passed or failed, where there is only a single "cutscore," and where the actually passing score is unimportant, a very efficient test can be developed by selecting only items that have high information near the cutscore. These items generally correspond to items whose difficulty is about the same as that of the cutscore.

Fisher Information and the Hessian of Log Likelihood

I've been taking some tentative steps into information geometry lately which, like all good mathematics, involves sitting alone in a room being confused almost all the time.

I was not off to a very good start when a seemingly key relationship between Fisher information and the second derivative of the log likelihood eluded me, despite being described as "obvious" or "simple" in several books. I finally figured out the main trick and thought I'd share it here in case someone else has trouble with it (e.g., me in six months).

Fisher Information

Fisher information is a quantity associated with parametric families of probability distributions. Let X be a set of outcomes and for each parameter θ in some set $\Theta \subset R^d$ let $p_\theta(x)$ be the distribution over X associated with θ. The *Fisher information* for the family $P=\{p_\theta:\theta\in\Theta\}$ is the matrix valued function where the entry[1] at the ith row and jth column is

$$I_{i,j}(\theta)=\mathrm{E}_X[(D_i \log p_\theta(X))(D_j \log p_\theta(X))]$$

where the expectation is over the random variable X drawn from the distribution p_θ, and D_i denotes the partial derivative $\partial/\partial\theta_i$. The Fisher information is always symmetric and positive semi-definite and can be seen as measuring the "sensitivity" of the *log likelihood* $\log p_\theta(x)$ on the outcomes in a neighbourhood of θ.

… and the Hessian of log likelihood

The result that had me puzzled for some time was the "obvious" fact that

$$I_{i,j}(\theta)=-\mathrm{E}_X[D_{i,j} \log p_\theta(X)]$$

where $D_{i,j}$ denotes the second-order partial derivative $\partial^2/\partial\theta_i\partial\theta_j$. What this says is that the Fisher information is closely related to the curvature of the log likelihood function, as measured by its *Hessian* — that is, the matrix of its second derivatives $H[\log p_\theta(x)]=(D_{i,j}\log p_\theta(x))_{d_{i,j}=1}$.

After much head-scratching, I realised that the "trick" I was missing was the observation that (under some mild conditions) the second derivatives and integrals can be switched so

$$\int_X D_{i,j} p_\theta(X)dx = D_{i,j}\int_X p_\theta(X)dx = D_{i,j}1 = 0$$

since each p_θ is a distribution.

With the above identity in hand, establishing the relationship between Fisher information and the Hessian of log likelihood is just an application of the chain and product rules and noting that $D_i \log p_\theta(x) = D_i p_\theta(x) p_\theta(x)$. Thus,

$$D_{i,j} \log p_\theta(x) = D_i (D_j p_\theta(x) p_\theta(x)) = D_{i,j} p_\theta(x) p_\theta(x) - D_i p_\theta(x) p_\theta(x) D_j p_\theta(x) p_\theta(x).$$

Taking expectations and using the aforementioned trick gives the result since $\mathrm{E}_X[D_{i,j} p_\theta(x) p_\theta(x)] = \int_X D_{i,j} p_\theta(x) dx = 0$.

Everything is obvious in hindsight!

Scoring

The person parameter θ represents the magnitude of *latent trait* of the individual, which is the human capacity or attribute measured by the test.[20] It might be a cognitive ability, physical ability, skill, knowledge, attitude, personality characteristic, etc.

The estimate of the person parameter - the "score" on a test with IRT - is computed and interpreted in a very different manner as compared to traditional scores like number or percent correct. The individual's total number-correct score is not the actual score, but is rather based on the IRFs, leading to a weighted score when the model contains item discrimination parameters. It is actually obtained by multiplying the item response function for each item to obtain a *likelihood function*, the highest point of which is the *maximum likelihood estimate* of θ. This highest point is typically estimated with IRT software using the Newton-Raphson method.[21] While scoring is much more sophisticated with IRT, for most tests, the (linear) correlation between the theta estimate and a traditional score is very high; often it is .95 or more. A graph of IRT scores against traditional scores shows an ogive shape implying that the IRT estimates separate individuals at the borders of the range more than in the middle.

An important difference between CTT and IRT is the treatment of measurement error, indexed by the standard error of measurement. All tests, questionnaires, and inventories are imprecise tools; we can never know a person's *true score*, but rather only have an estimate, the *observed score*. There is some amount of random error which may push the observed score higher or lower than the true score. CTT assumes that the amount of error is the same for each examinee, but IRT allows it to vary.[22]

Also, nothing about IRT refutes human development or improvement or assumes that a trait level is fixed. A person may learn skills, knowledge or even so called "test-taking skills" which may translate to a higher true-score. In fact, a portion of IRT research focuses on the measurement of change in trait level.[23]

A comparison of classical and item response theories

Classical test theory (CTT) and IRT are largely concerned with the same problems but are different bodies of theory and entail different methods. Although the two paradigms are generally consistent and complementary, there are a number of points of difference:

- IRT makes stronger assumptions than CTT and in many cases provides correspondingly stronger findings; primarily, characterizations of error. Of course, these results only hold when the assumptions of the IRT models are actually met.
- Although CTT results have allowed important practical results, the model-based nature of IRT affords many advantages over analogous CTT findings.
- CTT test scoring procedures have the advantage of being simple to compute (and to explain) whereas IRT scoring generally requires relatively complex estimation procedures.
- IRT provides several improvements in scaling items and people. The specifics depend upon the IRT model, but most models scale the difficulty of items and the ability of people on the same metric. Thus the difficulty of an item and the ability of a person can be meaningfully compared.
- Another improvement provided by IRT is that the parameters of IRT models are generally not sample- or test-dependent whereas true-score is defined in CTT in the context of a specific test. Thus IRT provides significantly greater flexibility in situations where different samples or test forms are used. These IRT findings are foundational for computerized adaptive testing.

It is worth also mentioning some specific similarities between CTT and IRT which help to understand the correspondence between concepts. First, Lord[24] showed that under the assumption that θ is normally distributed, discrimination in the 2PL model is approximately a monotonic function of the point-biserial correlation. In particular:

$$a_i \cong \frac{\rho_{it}}{\sqrt{1-\rho_{it}^2}}$$

where ρ_{it} is the point biserial correlation of item *i*. Thus, if the assumption holds, where there is a higher discrimination there will generally be a higher point-biserial correlation.

Another similarity is that while IRT provides for a standard error of each estimate and an information function, it is also possible to obtain an index for a test as a whole which is directly analogous to Cronbach's alpha, called the *separation index*. To do so, it is necessary to begin with a decomposition of an IRT estimate into a true location and error, analogous to decomposition of an observed score into a true score and error in CTT. Let

$$\hat{\theta} = \theta + \varepsilon$$

where θ is the true location, and ϵ is the error association with an estimate. Then $\mathrm{SE}(\theta)$ is an estimate of the standard deviation of ϵ for person with a given weighted score and the separation index is obtained as follows

$$R_\theta = \frac{\mathrm{var}[\theta]}{\mathrm{var}[\hat{\theta}]} = \frac{\mathrm{var}[\hat{\theta}] - \mathrm{var}[\epsilon]}{\mathrm{var}[\hat{\theta}]}$$

where the mean squared standard error of person estimate gives an estimate of the variance of the errors, , across persons. The standard errors are normally produced as a by-product of the estimation process. The separation index is typically very close in value to Cronbach's alpha.[25]

IRT is sometimes called *strong true score theory* or *modern mental test theory* because it is a more recent body of theory and makes more explicit the hypotheses that are implicit within CTT.

References

1. ^ A. van Alphen, R. Halfens, A. Hasman and T. Imbos. (1994). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing*. **20**, 196-201
2. ^ ETS Research Overview
3. ^ Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.
4. ^ Bock, R.D.; Aitkin, M. (1981). "Marginal maximum likelihood estimation of item parameters: application of an EM algorithm". *Psychometrika* **46** (4): 443–459. doi:10.1007/BF02293801.
5. ^ Ostini, Remo; Nering, Michael L. (2005). *Polytomous Item Response Theory Models*. Quantitative Applications in the Social Sciences **144**. SAGE. ISBN 978-0-7619-3068-6.
6. ^ Nering, Michael L.; Ostini, Remo, eds. (2010). *Handbook of polytomous item response theory models*. Taylor & Francis. ISBN 978-0-8058-5992-8.
7. ^ Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & Wainer, H. (Eds.), *Test Scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
8. ^ K. G. Jöreskog and D. Sörbom(1988). *PRELIS 1 user's manual, version 1*. Chicago: Scientific Software, Inc.
9. ^ Abramowitz M., Stegun I.A. (1972). *Handbook of Mathematical Functions*. Washington DC: U. S. Government Printing Office.
10. ^ Uebersax, J.S. (December 1999). "Probit latent class analysis with dichotomous or ordered category measures: conditional independence/dependence models". *Applied Psychological Measurement* **23** (4): 283–297. doi:10.1177/01466219922031400.
11. ^ Andrich, D (1989), Distinctions between assumptions and requirements in measurement in the Social sciences", in Keats, J.A, Taft, R., Heath, R.A, Lovibond, S (Eds), *Mathematical and Theoretical Systems*, Elsevier Science Publishers, North Holland, Amsterdam, pp.7-16.
12. ^ Steinberg, J. (2000). Frederic Lord, Who Devised Testing Yardstick, Dies at 87. New York Times, February 10, 2000
13. ^ Andrich, D. (January 2004). "Controversy and the Rasch model: a characteristic of incompatible paradigms?". *Medical Care* **42** (1): I–7. doi:10.1097/01.mlr.0000103528.48582.7c. PMID 14707751.
14. ^ Smith, R.M. (1990). "Theory and practice of fit". *Rasch Measurement Transactions* **3** (4): 78.
15. ^ Zwick, R.; Thayer, D.T.; Wingersky, M. (December 1995). "Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests". *Journal of Educational Measurement* **32** (4): 341–363. doi:10.1111/j.1745-3984.1995.tb00471.x.
16. ^ Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.
17. ^ Wright, B.D. (1992). "IRT in the 1990s: Which Models Work Best?". *Rasch measurement transactions* **6** (1): 196–200.
18. ^ Fischer, G.H. & Molenaar, I.W. (1995). *Rasch Models: Foundations, Recent Developments, and Applications*. New York: Springer.
19. ^ de Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory*, New York, NY: The Guilford Press. (6.12), p.144
20. ^ Lazarsfeld P.F, & Henry N.W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

21. ^ Thompson, N.A. (2009). "Ability estimation with IRT".
22. ^ Kolen, Michael J.; Zeng, Lingjia; Hanson, Bradley A. (June 1996). "Conditional Standard Errors of Measurement for Scale Scores Using IRT". *Journal of Educational Measurement* **33** (2): 129–140. doi:10.1111/j.1745-3984.1996.tb00485.x.
23. ^ Hall, L.A., & McDonald, J.L. (2000). *Measuring Change in Teachers' Perceptions of the Impact that Staff Development Has on Teaching.* Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24–28, 2000).
24. ^ Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
25. ^ Andrich, D. (1982). "An index of person separation in latent trait theory, the traditional KR.20 index, and the Guttman scale response pattern". *Education Research and Perspectives* **9**: 95–104.

External links

- "HISTORY OF ITEM RESPONSE THEORY (up to 1982)", University of Illinois at Chicago
- A Simple Guide to the Item Response Theory(PDF)
- Psychometric Software Downloads
- flexMIRT IRT Software
- IRT Tutorial
- IRT Tutorial FAQ
- An introduction to IRT
- The Standards for Educational and Psychological Testing
- IRT Command Language (ICL) computer program
- IRT Programs from SSI, Inc.
- IRT Programs from Assessment Systems Corporation
- IRT Programs from Winsteps
- Latent Trait Analysis and IRT Models
- Rasch analysis
- Free IRT software
- IRT Packages in R