Scientific Research Publishing

Psychology, 2021, 12, 1397-1409
https://www.scirp.org/journal/psych
ISSN Online: 2152-7199
ISSN Print: 2152-7180

The Integration of Classical Testing Theory and Item Response Theory

Zhongfeng Hu[1], Lin Lin[1], Youhan Wang[1*], Jiawei Li[2]

[1]South China Normal University, Guangzhou, China
[2]Michigan State University, Michigan, USA
Email: *2698260597@qq.com

How to cite this paper: Hu, Z. F., Lin, L., Wang, Y. H., & Li, J. W. (2021). The Integration of Classical Testing Theory and Item Response Theory. *Psychology, 12,* 1397-1409.
https://doi.org/10.4236/psych.2021.129088

Received: August 13, 2021
Accepted: September 12, 2021
Published: September 15, 2021

Abstract

The psychological and educational measurement theory, which was born in the early 20th century, has two critical theoretical schools: one is the Classical Testing Theory (CTT) that dominated in the first half of the 20th century, and the another is Item Response Theory (IRT), which developed from the 1950s to the 20s and reached its peak in the 1980s. Based on a prior study of key concepts of CTT and IRT, this article compared the two theories mentioned above and concluded that these two different theoretical schools are tending to be integrated which will be the direction of psychological and educational measurement in the future.

Keywords

Classical Testing Theory, Item Response Theory, Basic Hypothesis, Model, Comparison, Integration

1. What Are the Limitations of the Classical Test Theory?

1.1. The Basic Model of CTT

1.1.1. True Score Theory

CTT, also known as true score theory, is based on the true score. And the true score refers to the true value of the subject's measured traits, such as ability, knowledge, and personality, whereas CTT focuses on estimating the internal trait level of subjects based on their actual response performance. Furthermore, under the assumption that error scores and true scores are independent of each other with strict parallel tests, CTT calculates the difficulty of the subjects' pass rate, the discrimination between item score and sum score. Based on the follow-

ing assumptions of CTT, we can consider that the average observation score of multiple independent repeated measurements can be defined as the true score.

CTT puts forward a series of hypotheses: The relationship between the true score and the observed score is linear. The expected value of the error score is 0. The correlation between the true score and the error score is 0. The correlation between different measurement errors is 0. The error scores of different subjects on the same test are independent and identically distributed. Strict parallel testing exists. That is, the true scores of two parallel tests of any subject are equal, and the conditional variances of the error scores of the parallel tests are equal. The true score model has existed as early as the time of astronomer Galileo, whereas it was formally introduced in the psychological and educational assessment area relatively late (Schumacker, 1998). These basic assumptions are the foundations for the construction of classical measurement theory, without which the whole testing theory cannot be built. For example, the correlation between true and error scores is zero and the existence of strictly parallel tests is the requirements for CTT assumptions about the reliability of standardized concepts. Therefore, the authors argue that independent and equally distributed error scores of different subjects on the same test and the existence of strictly parallel tests are conditions to be satisfied by classical measurement theory, which need to be verified before used.

1.1.2. Reliability Theory

The reliability coefficient of a test is defined as the quotient of the test variance between the true score and the observed score. Under the premise of the satisfaction of strict parallel test existing in accordance with CTT hypothesis, it can be deduced that the reliability is the square of the correlation coefficient between the observed score and the true score. Since the true score cannot be accurately estimated, the reliability cannot be accurately calculated either. Psychometricians have had to seek other approximate estimation methods, resulting in numerous formulas with inconsistent results and confusing concepts for estimating credibility coefficients emerged. The fundamental reason for this situation is that the reliability coefficients involved in these concepts are different.

1.2. Advantages of CTT

In elaborate terms, CTT has the following merits (DeVellis, 2006). CTT is the basis for learning measurement which laid a solid foundation for the subsequent measurement theory. The methods used in CTT are mainly basic methods in algebra and statistics, which are relatively easy to master. Conditions are easily satisfied. Many practical assessments meet the conditions for the use of CTT. That is why it is very widely used. It is not necessary for every item to be optimal. As long as the item has a little correlation with the latent variables. Its shortcomings can be compensated by increasing the number of items and the quality

of each item does not need to be the best.

1.3. Deficiencies of CTT

1.3.1. Item Parameters Are Subject to the Sample

Since the facility of the CTT question is equivalent to the average score rate of the number of people who passed the question, and the discrimination index is equivalent to the correlation coefficient between the individual scores and the overall score. Therefore, the item parameters of CTT will be affected by the ability level of different samples of subject groups. In order to make up for the shortcomings of true score theory and reliability coefficients, the Generalizability Theory and the Strong True Score Theory have been proposed. In contrast, these theories were merely modified from CTT and had not fundamentally overcome the shortcomings of CTT (Christophersen & Lund, 2008).

1.3.2. The ability Parameters Depend on Item Difficulty

CTT calculates the total score based on the raw score. The lower the difficulty is, the higher the total score is. Therefore, the subject's ability is not stable, which leads to the unfairness of the measurement (Henson, 1999).

1.3.3. Mismatch between Item Parameters and Ability Parameters

The item parameters and subject scores in CTT are derived on different data bases. Thus, it is impossible to establish a functional relationship between them. In other words, the estimation of the subject's ability will change due to changes in the test (Magno, 2009).

1.3.4. The CTT Assumes That Strictly Parallel Tests and Errors Are Not Correlated with True Scores and Challenging to Satisfy in Actual Tests

The reason is that many psychological factors such as subjects' memory, development of new skills, and forgetting may lead to the hypothesis being unfulfilled (Royal & Bradley, 2008).

1.3.5. There Are Also Problems with the Reliability and Validity of CTT

Strict parallel testing is not easy to satisfy in practice as it is not easy to ensure that the average and standard deviation of the scores between different measurements are equal. The random error defined in CTT is very general. It cannot explain which error source the measurement error comes from and the magnitude of the respective errors. CTT requires the measurement conditions to be fully standardized. There are strict and precise regulations from test instruction to test scoring, making the measurement target narrow. If the measurement conditions change slightly, it can do nothing (Liu & Zhang, 1998).

For the test to achieve excellent reliability and validity, the item parameters of the test must have an appropriate match with the trait level distribution of the subjects. The mathematical model of CTT does not involve the mathematical relationship between both, whereby it is difficult to solve this problem perfectly within its theoretical framework (Li & Wang, 1998).

2. What Are the Advantages of IRT?

Item Response Theory (IRT) is a new measurement theory developed after the middle of the 20th century and is hailed as one of the most critical advances in psychometric methodology in the second half of the 20th century (McKinley & Mills, 1989). Its arising background is based on the limitations of CTT and stems from the rapid development of computer science and statistics.

2.1. The Basic Hypotheses and Features of IRT

The true score model of CTT is mainly derived from physical measurement, which is more suitable for measuring the physical properties of objects, and the theoretical basis of IRT is the potential theory of traits. IRT believes that the underlying traits in cognitive measurement refer to the intrinsic ability to be measured. The latent theory of traits believes that people's behavior and behavior are closely related to their psychological qualities. Therefore, the intrinsic characteristics of the individual can be estimated by quantitatively measuring the behavior of the individual. Conversely, if the intrinsic quality of the individual is estimated, the individual's behavioral response in the corresponding situation can be predicted and explained.

2.1.1. The Basic Hypothesis of IRT

Unlike CTT, IRT is based on strong hypothesis. There are four underlying hypotheses as following: The unidimensional hypothesis of Latent Space of the test.IRT assumes that if a subject's responses to all test items involve its n latent traits (or abilities), then these n latent traits constitute an n-dimensional space. IRT assumes that the potential space of the test it acts on has only one dimension, which means that the test can only measure one characteristic of the subject. This characteristic is called unidimensionality. There are many controversies about the unidimensionality assumption. First of all, what is unidimensionality? The concept is rather vague. Secondly, how to verify unidimensionality? A recognized method is absent. A proposed method existed previously, while it has not been fully approved by the academic community (Hu & Mo, 2002). Besides, some experts have developed a multidimensional IRT model. This model can be applied directly without verifying the satisfaction of unidimensionality. However, it is quite complicated to apply and difficult to understand. Thus, it is not widely used at present (McKinley & Reckase, 1983). Local independence hypothesis: This hypothesis refers to the responses of subjects with the same ability or trait level on different test items are independent of each other, and the responses of these subjects to one test item are not affected by their reactions to other items. With this assumption, it is possible to use conditional probabilities to estimate capacity and project parameters. Otherwise, parameter estimation cannot be performed as only the probability of the product of independent events is equal to the product of the probabilities of independent events (Gustafsson, 1980). From a formal perspective, local independence and unidi-

mensionality are completely different hypotheses, but most IRT theorists consider that these two hypotheses are equivalent. The reason is that if a test is unidimensional, then all its items assess the same ability or trait. Therefore, the probability of a subject's correct response to the item is only related to its corresponding ability or trait level. Moreover, it has nothing to do with its reaction to other test items. It is partial independence; conversely, if local independence is satisfied, the probability of correct response to the same test item is the same for subjects of the same ability or trait level, meaning the test is unidimensional (Reese, 1999). Although local independence and unidimensionality imply each other and they are proposed from two different perspectives. Unidimensionality is a property that test designers hope for the test, while local independence is proposed for parameter estimation purposes. Even the two are equivalent, they are two different concepts, both of which have been retained for historical reasons and convenience in practicum. The view is from Professor David Andrich, a well-known surveyor in the School of Education at the University of Western Australia. The author once asked about the relationship between unidimensionality and partial independence in his "Advanced Measurement Methods" class. He thinks so. Strictly speaking, if the local independence condition is not satisfied, the IRT model cannot be used. Thus, some schoolers have studied how to use the IRT model in local correlation and proposed corresponding methods (Thompson & Pommerich, 1996). Hypothesis about Item Characteristic Curve (ICC) of the test: The Item Characteristic Curve is a specific graphical representation of the relationship between item characteristic functions. The item characterization function refers to a functional relationship between the probability of correct response to a test item and the level of ability or trait corresponding to the item. In IRT, the probability of a subject's correct response to an item is determined by the assumed IRT function, the item parameters, and the level of the subject's corresponding ability or trait. The first and most crucial step in using IRT in a test is to make assumptions about the ICC of the test. Generally, a suitable model from the existing models that have been proven to be effective. Why do we need model hypotheses? The hypothesis of a model is the premise of measurement data analysis. Otherwise, the data cannot be analyzed. With a hypothetical model, the researcher can analyze the measurement results based on the relevant characteristics of the model. Just as in statistics, the data are assumed to obey a normal distribution and the relevant features of the normal distribution can be applied to analyze the data. For instance, it should be considered that 95% of the data are within 1.96 standard deviations of the average. The different assumptions about the models have led to the arising of various assessment theories and models. Simultaneously, the model's fitness test has also become an essential prerequisite for using the model (Kingston & Stocking, 1986). Again, multiple tests emerged, and some models do not even require testing, which can cause problems (Ackerman, 1987). Unspeedness hypothesis of tests. The last primary

hypothesis of IRT is the unspeedness hypothesis of the test. That is, the test is required to be conducted under unlimited time. In this case, if a subject does not respond to certain test items, it can be considered that it is due to its insufficient ability, and the item is dealt with incorrectly. The unspeedness hypothesis is the natural reasoning of the unidimensionality assumption. Without the unspeediness assumption, other factors (e.g., speed) affect the ability or trait to be measured. The unspeedness test is just an ideal test, all other tests too. As long as the time limit is reasonable, the same effect can be obtained if the test is not limited. Since all tests are time-limited, the unspeedness test is to test the reasonableness of the test's time limit. Then, what does it mean to have a reasonable time limit? There are various answers to this question. For example, some people take the criterion that most of them have enough time to finish the test they know how to answer, but the definition of percentage of completion is rather vague. Some believe that the criterion is that subjects with intermediate abilities can answer all questions. The relationship between capacity and speed is also involved here, which is more complicated. Therefore, many measurements are just empirically time-limited and more subjective. Given this situation, the unspeedness assumption is sometimes ineffective, couple with the unidimensionality hypothesis mentioned above. Some scholars suggest canceling this hypothesis, while some psychometricians propose it separately in the theories. For example, the hypothesis was not put forward in the lecture notes of Professor David Andrich. Also, some studies have shown that the impact of unspeedness on the different content and purposes is different (Yamamoto & Everson, 1994).

2.1.2. Characteristics of IRT

The most distinguishing feature of IRT is item characteristic functions to establish a relational equation between subjects' ability and item parameters (e.g., facility value, discrimination index, and guessing coefficient). It establishes an equation containing the subjects' potential ability and the test item parameters. Through the relationship with the subject's response, the project parameters can be estimated through the iterative method in modern mathematics, and then the subject's potential ability can be estimated. It could theoretically overcome the deficiency of CTT tests to establish a functional relationship between subjects' scores and test item parameters.

2.2. Model of IRT

There are several models of item response theory to choose from, the most famous of which are the Logistic model and the Lacy model. The scoring model of IRT is primarily a two-tier scoring model, and later developed into the multi-tier scoring system. The two-tier scoring refers to the situation where there are only two possible test results, right or wrong. There are several types of it, two of which are discussed here.

2.2.1. Logistic Model

Logistic model was proposed by Birnbaum in 1957, and this model is classified into single-parameter, two-parameter and three-parameter. See Table 1. The forms of it are as follows.

The parameters a_i, b_i, c_i denote as the discrimination index, facility value and guessing coefficient of item i respectively. The normal range of values is $0 \le a \le 2.0$, $-3.0 \le b \le 3.0$, and $0 \le c \le 1$. Generally, $D = 1.704$ and e is the base of the natural logarithm.

The logistic model is currently recognized as the most effective and widely used two-tier scoring IRT model, and this model matches the actual test results quite well.

2.2.2. Rasch Model

The Rasch model looks exactly the same as the one-parameter model of the logistic model, but its assessment theory and assumptions are completely different. Its assessment model is.

$$P\left(X_{vi}=1\right)=\mathrm{e}^{(\beta v-\delta i)}\Big/\left(1+\mathrm{e}^{(\beta v-\delta i)}\right)$$

In the above equation, βv is the ability parameter, δi is the facility value parameter, and the left-hand side of the equation indicates the probability that subject V answered correctly on item i. The Rasch model, although simple, requires a fit test of the model, invariance of variance to be satisfied, and the data must conform to conditions such as the Guttman model, otherwise the model cannot be applied.

There have been many debates about the above two models. Some people argued that the Rasch model is a special case of Logistic Model, while others disagreed (Cantrell, 1997). The author also believes that although it appears to be a special case of the logistic model formally, they are completely different essentially. Under relatively independent conditions and assumptions, the Rasch model is a separate model in a mathematical point of view, and many psychometricians hold this view, especially those of the Chicago school, who are strongly committed to the Rasch model and have also developed corresponding software. For example, Professor David Andrich, one of the author's supervising professors at the University of Western Australia, and his mentor at the University of Chicago, Wright, are staunch advocates of the Lacy model, and Professor David and others have developed the corresponding software (RUMM). In summary, the single-parameter model and the multi-parameter model each have their own pros and cons (Custer, Sharairi, Yamazaki, Signatur, Swift, & Sharon, 2008).

Table 1. Unidimensional dichotomous response models.

Model type	Mathematical Forms	Item Parameter
One-parameter logistic	$P_i(\theta)=1\Big/\left(1+\mathrm{e}^{-D*(\theta-b_i)}\right)$	Difficulty (b)
Two-parameter logistic	$P_i(\theta)=1\Big/\left(1+\mathrm{e}^{-D*a_i*(\theta-b_i)}\right)$	Difficulty (b), discrimination (a),
Three-parameter logistic	$P_i(\theta)=c_i+(1-c_i)\Big/\left(1+\mathrm{e}^{-D*a_i*(\theta-b_i)}\right)$	Difficulty (b), discrimination (a), guessing (c)

2.3. Advantages of IRT

1) Invariance of item parameter estimation. In IRT, the item parameters are invariant regardless of the ability distribution of the subject group. Of course, there is a prerequisite assumption that is the capacity of the sample is large enough (Zhang & Liu, 1998).

2) The potential abilities or traits estimated by the IRT are highly stable, being independent of the topic being tested and does not change with any modification of the test. This provides a theoretical and methodological basis for subjects at different levels to be assessed with different items or adaptive tests (Thornton, 2002).

3) Subjects with different levels varying from different measurement errors. In IRT, item characteristic functions can be used to estimate the ability of each subject, and the measurement error is generally not the same as per subject.

2.4. Limitations and Weaknesses of IRT

1) A high fit of the collected data to the item characteristic function is required. Since the sample size and the number of items will restrict the fit between the measured data and the model. It is necessary to check the fit between the measured data and the selected characteristic function before using IRT to analyze the items. Otherwise, the model cannot be utilized. Therefore, if the fit of one model is not satisfactory, it is necessary to try another model, while one of the tasks of the assessment specialists is to develop a suitable new model to be chosen or as a backup.

2) There is a lack of recognized methods for testing the hypothesis of unidimensionality. Although there are several methods to test unidimensionality, none of them has been accepted by everyone. Some scholars even believe that the unidimensionality assumption cannot be satisfied, and there is no need for this hypothesis. Therefore, in some IRT models, there is no need to test unidimensionality. For example, the author studied ConQuest (a program for IRT models) to visit the Assessment Research Centre at the University of Melbourne in 2006. There was no need to verify unidimensionality when using this model.

3) The assumption of local independence and the assumption of unidimensionality are equivalent, as mentioned earlier, so it is also difficult to verify.

3. Integration of CTT and IRT

3.1. A Comparison of CTT and IRT

There are many comparative studies on CTT and IRT varying from dimensions and perspectives (Magno, 2009). This essay is a comparison of the following four aspects (Hwang, 2002).

1) The model of CTT is simple, i.e., the true score model whereas the model of IRT is complex, which has been introduced in detail previously. In addition, the assumptions of CTT are mainly weak assumptions, which can be easily satisfied, while those of IRT is relatively difficult and mainly strong assumptions and lack

standardized methods to verify them. Also, the assumptions of different schools of IRT are not the same. In short, the hypotheses of IRT are much more complex than those of CTT.

2) Test scores or item responses. The CTT calculates the subject's total score (observed score) by adding the scores of the items together, and the true score uses the mean of multiple measurements. On the other hand, IRT does not give scores directly but instead assesses the potential ability of the subject based on the subject's responses on each item.

On the one hand, the CTT calculates scores based on the items that subjects answer correctly on the test, and subjects with the same score may differ in their ability since the difficulty of the items varies from test to test, requiring subjects to answer different items correctly with different knowledge and abilities. Even so, they may get the same score on different items. On the other hand, instead of scoring subjects according to how many questions they answered correctly, IRT scoring depends on the responses of all subjects on all questions at first, and then estimates the item parameters of each question, then estimates the ability parameters of the subjects based on the item parameters in the end. In short, the ability reflected by the questions that subjects answered correctly with different item parameters (facility value, discrimination index, and guessing coefficient, etc.) are varied. Thus, even though the subjects completed the same number of items, it is possible to obtain different ability scores if the item content differs, i.e., the item parameters are diverse (French, 2001).

3) Item parameters. The item parameters of CTT vary with the sample and are not stable. Namely, the calculated item parameters have different values for separate sample groups of subjects, whereas the item parameters of IRT are independent of the sample and have stability (Henson, 2000).

4) Item Information Index. The two concepts of test information function and item information function in IRT are not found in CTT. As we know, the purpose of the test is to obtain information about the subject's ability or potential traits and the amount of information provided by different quality items (Hobart., 2003). It has also been shown that CTT and IRT can provide complementary items information to each other (Hays, Brown, Brown, Spritzer, & Crall, 2006).

5) For subjective questions, CTT can also calculate the facility value and discrimination index of the subjective questions and plot the difficulty curve of the questions for item analysis. However, IRT was initially designed for objective tests, such as multiple-choice questions, and was overwhelmed when analyzing subjective questions. Models are even forced to be used when they do not meet the conditions for their use or abandon their initial conditions of use. IRT is now fully mature in analyzing polytomous data, such as Likert scale data (Li, Li, & Wang, 2010), but further improvements are needed in analyzing qualitative data such as essay scores (Cai & Monroe, 2014).

6) Others. CTT and IRT are associated with each other in estimating item parameters, despite their considerable differences. F.M. Lord and M.R. Novick

have proven an approximate relationship between them if the subjects' abilities obey a normal distribution and the item characteristic curves are two-parameter normal-ogive functions (Guthrie, 2000).

In conclusion, IRT is conceptually more rigorous. The item parameters are not dependent on the subject samples, but the plotting on the characteristics of the items is more reasonable and profound than CTT. However, using IRT requires stronger assumptions that are not easily satisfied. The CTT, on the other hand, requires only weaker assumptions and is simpler and easier to understand. It is the reason why most of the tests are still using CTT for item analysis.

3.2. How Can CTT and IRT Be Integrated?

1) Changing in perception. First, we need to discard the prejudice between different portals and schools, significantly changing the concept of superiority or inferiority between different schools. For instance, it is wrong to believe that IRT is an advanced measurement method, and CTT is an inferior measurement method, or multi-parameter is an advanced method, and a single-parameter is an inferior method.

2) Clarifying concepts. There is an urgent need for a systematic, complete, and standardized conceptual system for modern measurement theory. The current conceptual system of IRT is still the same as that of CTT, with some concepts that are not available in CTT. However, IRT uses concepts in a different range than CTT, such as facility value and discrimination index, which can be taken arbitrarily in IRT and not necessarily within [0, 1] or [−1, 1]. Moreover, the IRT is not a new approach but builds on the CTT. Furthermore, IRT has not made substantial progress on the most critical issue, i.e., validity regarding psychological and educational assessment (Hu & Mo, 2007).

3) Clarifying the conditions. Each measurement model has its conditions, boundaries, prerequisites, and corresponding assumptions that must meet to use. Therefore, the conditions of each model for adoption need to be well defined, and the prerequisites are supposed to be standardized. Therefore, the conditions of each model need to specify in detail, and the prerequisites should be standardized. In addition, the researcher cannot decide to adopt a particular method based on personal interest or preference. On the contrary, it is crucial to develop new methods when administration conditions do not meet, rather than using a specific method hastily (Rimen, 2009).

4) Standardization of methods is desirable. There are numerous measurement methods and even the same batch of data being processed in different schools with diverse methods (Reckase & McKinley, 1982). We believe that the corresponding analysis methods should be standardized strictly according to the conditions. For example, it cannot be straightforward to apply the traditional IRT method for data that do not satisfy the assumption of unidimensionality, which gave rise to the creation of MIRT (Multidimensional Item Response Theory) and its application in practice (Yen & Leah, 2007). Also, some scholars have devel-

oped a method that combines CTT and MIRT in a good way (Ourania, Elmore, & Headrick, 2001). Some experts believe that the only way to analyze data effectively is to integrate CTT and IRT (Crislip & Chin-Chance, 2001).

5) Integration among various schools. The integration of different schools is a challenging thing to do, involving four factors. There is the question of the beliefs of measurement scientists. There is the issue of traditional measurement research. Another is the difference in the way of thinking. The problem of technology development. Therefore, it is almost impossible to rely on western psychometricians to integrate different measurements of these schools, while eastern psychometricians can accomplish this challenging task. Because integration is an important way of thinking in the East, and because research traditions do not limit the East, it is most likely that a new integrated modern theory of measurement will emerge from the East.

Supported

National Education and Science Program 2021 National General Project: Research on the theory and practice of value-added Evaluation based on students' development (Grant No. BFA210064); Major bidding projects for Educational Science Planning in Guangzhou: Research on the Evaluation Standard of school-running quality of Primary and Middle Schools in Guangzhou (Grant No. 2017-01).

The psychological and educational theories were born at the beginning of the 20th century, and the influential ones were Classical Test Theory (CTT), Item Response Theory (IRT) and Generalizability Theory. During this period, the CTT occupied a dominant position by the first half of the 20th century; after the 1950s, people gradually realized the limitations of CTT then the IRT derived, reaching its heyday in the 1980s; currently, the two theories are in a stage of integration. This article made an in-depth reflection and comparison of the two theories, expounded the basic principles of modern education measurement and proposed that the integration of the CTT and IRT is the future development direction of psychology and education measurement.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Ackerman, T. A. A. (1987). Comparison Study of the Unidimensional IRT Estimation of Compensatory and Noncompensatory Multidimensional Item Response Data. *The Annual Meeting of the American Educational Research Association,* Washington DC, 20-24 April 1987.

Cai, L., & Monroe, S. (2014). *A New Statistic for Evaluating Item Response Theory Models for Ordinal Data (CRESST Report 839).* University of California.

Cantrell, C. E. (1997). Item Response Theory: Understanding the One-Parameter Rasch

Model. *The Annual Meeting of the Southwest Educational Research Association,* Austin, 23 January 1997.

Christophersen, K.-A., & Lund, H. T. (2008). A Generalizability Study of the Norwegian Version of KINDLR in a Sample of Healthy Adolescents. *Quality of Life Research, 17,* 87-93. https://doi.org/10.1007/s11136-007-9289-y

Crislip, M. A., & Chin-Chance, S. (2001). Using Traditional Psychometric Methodologies and the Rasch Model in Designing a Test. *The Annual Meeting of the American Educational Research Association,* Seattle, 10-14 April 2001.

Custer, M., Sharairi, S., Yamazaki, K., Signatur, D., Swift, D., & Frey, S. (2008). A Paradox between IRT Invariance and Model-Data Fit When Utilizing the One-Parameter and Three-Parameter Models. *The Annual Meeting of the American Educational Research Association,* New York, 28 March 2008.

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care, 44,* S50-S59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

French, C. L. A. (2001). Review of Classical Methods of Item Analysis. *The Annual Meeting of the Southwest Educational Research Association,* New Orleans, 1-3 February 2001.

Gustafsson, J.-E. (1980). *A Introduction to Rasch's Measurement Model. ERIC Clearinghouse on Tests, Measurement, and Evaluation.*

Guthrie, A. C. (2000). A Review of Coefficient Alpha and Some Basic Tenets of Classical Measurement Theory. *The Annual Meeting of the Southwest Educational Research Association,* Dallas, 27-29 January 2000.

Hays, R. D., Brown, J., Brown, L. U., Spritzer, K. L., & Crall, J. J. (2006). Classical Test Theory and Item Response Theory Analyses of Multi-Item Scales Assessing Parents' Perceptions of Their Children's Dental Care. *Medical Care, 44,* S60-S68. https://doi.org/10.1097/01.mlr.0000245144.90229.d0

Henson, R. K. (1999). Understanding the One-Parameter Rasch Model of Item Response. *The Annual Meeting of the Southwest Educational Research Association,* San Antonio, 21-23 January 1999.

Henson, R. K. (2000). A Primer on Coefficient Alpha. *The Annual Meeting of the Mid-South Educational Research Association,* Bowling Green, 16 November 2000.

Hobart, J. C. (2003). An Empirical Demonstration of the Limitations of Classical Test Theory and the Advantages of New Psychometric Methods. *Quality of Life Research, 12,* 744.

Hu, Z. F., & Mo, L. (2002). On the Integration of Factor Analysis Methods. *Psychological Science, No. 4,* 474-475.

Hu, Z. F., & Mo, L. (2007). Reconstruction of the Theory of Validity in Psychological and Educational Measurement. *Journal of South China Normal University (Social Science Edition), No. 6,* 82-90.

Hwang, D.-Y. (2002). Classical Test Theory and Item Response Theory: Analytical and Empirical Comparisons. *The Annual Meeting of the Southwest Educational Research Association,* Austin, 14-16 February 2002.

Kingston, N. M., & Stocking, M. L. (1986). Psychometric Issues in IRT-Based Test Construction. *The Annual Meeting of the American Psychological Association,* Washington DC, 22-26 August 1986.

Li, J. B., & Wang, Q. (1998). Research on Simulate Trial for the Optimal Match between the Distribution of Ability and the Distribution of Item Difficulty. *Acta Psychologica Sinica, No. 2,* 197-202.

Li, Y. M., Li, S. H., & Wang, L. (2010). *Application of a General Polytomous Testlet Model to the Reading Section of a Large-Scale English Language Assessment*. ETS, Research Report. https://doi.org/10.1002/j.2333-8504.2010.tb02228.x

Liu, Y., & Zhang, H. C. (1998). Application of Generality Theory in Composition Scoring. *Acta Psychologica Sinica, No. 2,* 211-217.

Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment, 1,* 1-11.

McKinley, R. L., & Reckase, M. D. (1983). *An Extension of the Two-Parameter Logistic Model to the Multidimensional Latent Space*. Research Report ONR83-2 August 1983, ACT.

McKinley, R., & Mills, C. (1989). Item Response Theory: Advances in Achievement and Attitude Measurement. In B. Thompson (Ed.), *Advances in Social Science Methodology* (p. 71). JAI Press.

Ourania, R., Elmore, P. B., & Headrick, T. C. (2001). Number Correct Scoring: Comparison between Classical True Score Theory and Multidimensional Item Response Theory. *The Annual Meeting of the American Educational Research Association,* Seattle, 10-14 April 2001.

Reckase, M. D., & McKinley, R. L. (1982). Some Latent Trait Theory in a Multidimensional Latent Space. *Item Response Theory and Computerized Adaptive Testing Conference Proceedings,* Wayzata, 27-30 July 1982.

Reese, L. M. (1999). *A Classical Test Theory Perspective on LSAT Local Item Dependence*. LSAC Research Report Series. Statistical Report. Law School Admission Council.

Rimen, F. (2009). *Three Multidimensional Models for Testlet-Based Tests: Formal Relations and an Empirical Comparison*. Educational Testing Service.

Royal, K. D., & Bradley, K. D. (2008). Rethinking Measurement in Higher Education Research. *The 2008 Mid-Western Educational Research Association*.

Schumacker, R. E. (1998). Comparing Measurement Theories. *The Annual Meeting of the American Educational Research Association,* San Diego, 13-17 April 1998.

Thompson, T. D., & Pommerich, M. (1996). Examining the Sources and Effects of Local Dependence. *The Annual Meeting of the American Educational Research Association,* New York, 8-12 April 1996.

Thornton, A. (2002). A Primer on the 2- and 3-Parameter Item Response Theory Models. *The Annual Meeting of the College of Education, University of North Texas, Educational Research Exchange,* Denton, 1 February 2002.

Yamamoto, K., & Everson, H. T. (1994). Modelling the Mixture of IRT and Pattern Responses by a Modified HYBRID Model. *A Symposium Titled "Applications of Latent Trait and Latent Class Models in the Social Sciences",* Hamburg, May 1994.

Yen, S. J., & Leah, W. (2007). Multidimensional IRT Models for Composite Scores. *The 2007 Annual Meeting of the National Council of Measurement in Education,* Chicago, 10-12 April 2007.

Zhang, M. Q., & Liu, X. Y. (1998). A Study on the Applying of Item Response Models. *Acta Psychologica Sinica, No. 4,* 436-441.