Azim Premji University

Assessing the Equivalence between Classical Test Theory and Item Response Theory

Assessing the equivalence between Classical Test Theory and Item Response Theory[i]

Context

Classical Test Theory (CTT) and Item Response Theory (IRT) are the most widely used statistical frameworks for developing tests. The uses of CTT and IRT include identifying poorly functioning items, determining test reliability, and equating parallel test forms (Hambleton & Jones, 1993). Both statistical frameworks have advantages and disadvantages, as elaborated by Hambleton and Jones (1993).

CTT uses models with weak assumptions that test data can meet easily. Item parameters (difficulty and discrimination) and test statistics (reliability) can be easily computed using descriptive statistics under the CTT framework. CTT statistics are also suitable for small samples where the pilot sample is similar to the population being tested. However, the sample dependence and test dependence of CTT item parameters pose a disadvantage. CTT also does not provide sensitive estimates of test reliability across a range of ability levels.

The IRT framework addresses the CTT framework's flaws of sample dependence, test dependency, and lack of sensitive reliability estimates. Additionally, the IRT framework provides advantages in true score estimation, test equating, and adaptive testing. However, the IRT framework requires a good fit between the test data and the IRT model for obtaining reliable item parameters. Large sample sizes are required for reliable item parameters as well. Additionally, IRT calculations are mathematically complex and require comfort with programming.

Thus, the broad question is **whether the benefits of IRT for selecting test items outweigh the difficulties posed by its restrictiveness and mathematical complexity.** Another way of addressing this question is **whether the much-simpler CTT framework yields comparable item parameters as the IRT model.** To that end, this literature review aims to summarise the results of 16 studies that investigated the empirical comparability and invariance of CTT and IRT item parameters.

Comparison of CTT and IRT item parameters

Broadly, the studies indicate that the CTT and IRT item difficulty parameters are highly correlated for the IRT 1PL and 2PL models. The IRT 1PL item difficulty parameters are more strongly correlated than the IRT 2PL item difficulty parameters. The correlation between CTT and IRT item difficulty is generally poorer for the IRT 3PL model than for the 2PL or 1PL model. Only in a couple of studies were the IRT 3PL difficulty parameters more strongly correlated than the 2PL difficulty parameters (Fan, 1998; Hwang, 2002). CTT and IRT difficulty parameters are also more strongly correlated for random samples than non-random samples (Adedoyin, Nenty, & Chilisa, 2008; Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998).

CTT and IRT item discrimination parameters for the IRT 2PL model are more strongly correlated than for the IRT 3PL model (Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998; Hwang, 2002). As with item difficulty, CTT and IRT discrimination parameters are more strongly correlated for random samples than non-random samples (Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998). The correlations of CTT and IRT discrimination parameters are more sensitive to sample size than difficulty parameters. In general, the correlations between difficulty parameters are higher than discrimination parameters for CTT and IRT (Awopeju & Afolabi, 2016; Bichi, Embong, Talib, Salleh, & Ibrahim, 2019; Courville, 2004; Fan, 1998; Hwang, 2002; Kiany & Jalali, 2009; Nasir, 2014; Progar, Sočan, & Peč, 2008; Stage, 2003).

Summary of Results

Correlations between CTT and IRT Item Difficulty Parameters (for Samples except for Non-random and N < 100)		
IRT Model	**Strength of Correlation/Presence of Significant Differences**	**Studies Cited**
1 PL	Strong/No significant differences between CTT and IRT difficulty parameters	Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998; Hwang, 2002; Kiany & Jalali, 2009; Latif, Yusof, Amin, Libunao, & Yusri, 2016; Ly, Rakkapao, Nualtong, & Sumathakulawat, 2020

	Weak/Significant differences between CTT and IRT difficulty parameters	Adedoyin, Nenty, & Chilisa, 2008 (for some sampling plans; only difficulty and the IRT model not mentioned in the paper)
2PL	Moderate	Demaidi, Gaber, & Filer, 2017
	Strong	Awopeju & Afolabi, 2016; Bichi, Embong, Talib, Salleh, & Ibrahim, 2019; Courville, 2004; Fan, 1998; Hwang, 2002; Kiany & Jalali, 2009; Nasir, 2014; Progar, Sočan, & Peč, 2008; Umobong & Jacob, 2016
3PL	Strong	Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998; Hwang, 2002; Stage, 2003
Correlations between CTT and IRT Item Discrimination Parameters		
2PL	Moderate	Nasir, 2014 (for some samples)
	Strong	Awopeju & Afolabi, 2016; Bichi, Embong, Talib, Salleh, & Ibrahim, 2019; Courville, 2004; Fan, 1998; Hwang, 2002; Nasir, 2014; Progar, Sočan, & Peč, 2008; Umobong & Jacob, 2016
3PL	Weak-to-moderate	Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998; Hwang, 2002; Kiany & Jalali, 2009 (some sub-parts); Stage, 2003
	Strong	Kiany & Jalali, 2009

Invariance of CTT and IRT parameters

The ability of the CTT and IRT frameworks to produce unbiased and consistent item parameters across different types of samples is called invariance (Fan, 1998). This property implies that the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameter that characterize an examinee does not depend on the set of items (Baker & Kim, 2017).

In general, CTT and IRT difficulty parameters are more invariant than CTT and IRT discrimination parameters (Courville, 2004; Fan, 1998; Progar, Sočan, & Peč, 2008). However, it is unclear whether IRT item parameters are more invariant than CTT item parameters under all conditions (such as variation in sample size, test length, geographical condition, models, types of samples, etc.). Thus, more research is required to compare the invariance of CTT and IRT item parameters across different conditions.

Invariance of CTT parameters across different conditions

- **Sample Size:** CTT difficulty parameters are less invariant for small sample sizes and non-random samples than for large sample sizes and random samples (Adedoyin, Nenty, & Chilisa, 2008; Courville, 2004). There is a total collapse of CTT discrimination invariance at a sample size of 100 (Courville, 2004).
- **Type of sample:** CTT difficulty and discrimination parameters are especially sensitive to the type of sample. Variations among candidates in different geographical locations (Adedoyin, Nenty, & Chilisa, 2008; Kunovskaya, Cude, & Alexeev, 2014; Progar, Sočan, & Peč, 2008), variations in grade and gender (Adedoyin, Nenty, & Chilisa, 2008; Progar, Sočan, & Peč, 2008), and variations in ability (Awopeju & Afolabi, 2016; Courville, 2004; Fan, 1998; Progar, Sočan, & Peč, 2008) can cause variance in CTT item parameters.

Invariance of IRT parameters across different conditions

- **Sample size:** IRT difficulty parameters are less invariant for small sample sizes than for large sample sizes (Courville, 2004). The invariance of IRT discrimination parameters collapses at small sample sizes of 100, especially for 3PL models, thus concurring with literature (Courville, 2004).
- **Type of sample:** IRT item parameters are more invariant for random than non-random samples (Courville, 2004; Fan, 1998). But there is some conflict in the literature. One study found that IRT item parameters are invariant across geographic locations (Adedoyin, Nenty, & Chilisa, 2008). Another study found that IRT item parameters are sensitive to variations in geographic locations (Progar, Sočan, & Peč, 2008). It is suggested that the invariance of item parameters as a function of sample size be explored further.

- **IRT Models:** IRT difficulty estimates are more invariant for the 1PL model, followed by the 2PL model and the 3PL model (Courville, 2004; Fan, 1998). IRT discrimination parameters are more invariant for the 2PL model than for the 3PL model (Courville, 2004; Fan, 1998). IRT item parameters are more invariant when there is a good fit between the data and the model (Progar, Sočan, & Peč, 2008).

Summary

This literature survey indicates that both CTT and IRT may be used interchangeably under the following conditions:

- Sample sizes are large
- Samples are random
- Samples are restricted to a relatively narrow geographical area

CTT is preferred for clinical samples, whether random or non-random. IRT is not suitable for small samples (<100), especially the 3PL model. IRT is preferred when the test population is spread across large and varied geographical areas, the sample varies in ability, and the sample size is large. More research is required on IRT item parameter stability as a function of sample size, test length, and type of questions.

References

Adedoyin, O., Nenty, H. J., & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review, 3*(2), 83-93.

Awopeju, O., & Afolabi, E. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, 12*(28), 263-284. doi:https://doi.org/10.19044/esj.2016.v12n28p263

Baker, F. B., & Kim, S.-H. (2017). *The Basics of Item Response Theory using R.* Cham, Switzerland: Springer.

Bichi, A. A., Embong, R., Talib, R., Salleh, S., & Ibrahim, A. b. (2019). Comparative analysis of classical test theory and item response theory using chemistry set data. *International Journal of Engineering and Advanced Technology, 8*(5C), 1260-1266. doi:https://doi.org/10.35940/ijeat.E1179.0585C19

California State University Long Beach. (1998, 11 3). *PPA 696: Sampling*. Retrieved from California State University Long Beach: https://home.csulb.edu/~msaintg/ppa696/696sampl.htm

Cook, L. L., & Eignor, D. R. (1991). NCME instructional module: IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37–45.

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person characteristics.* Texas A&M University, Office of Graduate Studies. College Station: Texas A&M University.

Demaidi, M. N., Gaber, M. M., & Filer, N. (2017). Evaluating the quality of the ontology-based auto-generated questions. *Smart Learning Environments, 4*(7). doi:https://doi.org/10.1186/s40561-017-0046-6

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational & Social Sciences, 3*(1), 57-75.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381. doi:https://doi.org/10.1177/0013164498058003001

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice*, 38–47.

Hwang, D.-Y. (2002). Classical test theory and item response theory: Analytical and empirical comparisons. *Annual Meeting of the Southwest Educational Research Association.* Austin: Educational Resources Information Centre.

Kiany, G. R., & Jalali, S. (2009). Theoretical and practical comparison of classical test theory and item-response theory. *Iranian Journal of Applied Linguistics, 12*(1), 167-197.

Kunovskaya, I. A., Cude, B. J., & Alexeev, N. (2014). Evaluation of a financial literacy test using classical test theory and item response theory. *Journal of Family and Economic Issues, 35*, 516-531. doi:https://doi.org/10.1007/s10834-013-9386-8

Latif, A. B., Yusof, I. J., Amin, N. F., Libunao, W. H., & Yusri, S. S. (2016). Multiple choice items analysis using classical test theory and Rasch measurement model. *Man in India, 96*(1-2), 173-181.

lumen Boundless Statistics. (n.d.). *Correlation*. Retrieved from lumen Boundless Statistics: https://courses.lumenlearning.com/boundless-statistics/chapter/correlation/

Ly, C. Y., Rakkapao, S., Nualtong, K., & Sumathakulawat, W. (2020). Classical test theory and Rasch analysis of test of understanding of vectors. *Journal of Physics: Conference Series. 1719*, pp. 1-4. Siam: IOP Publishing. doi:https://doi.org/10.1088/1742-6596/1719/1/012089

Nasir, M. (2014). *Application of classical test theory and item response theory to analyze multiple choice questions.* Calgary: University of Calgary. doi:http://dx.doi.org/10.11575/PRISM/24958

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world, volume 1: Analysis.* OECD. Retrieved from https://www.oecd-ilibrary.org/pisa-2006_5l4gzqxtbtd6.pdf?itemId=%2Fcontent%2Fpublication%2F9789264040014-en&mimeType=pdf

Progar, Š., Sočan, G., & Peč, M. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology, 17*(3), 5-24.

Rust, J., Kosinksi, M., & Stillwell, D. (2021). *Modern psychometrics: the science of psychological assessment (4th ed.).* New York: Routledge.

Spector, D. (2014, May 9). *These Hilarious Charts Will Show You Exactly Why Correlation Doesn't Mean Causation*. Retrieved from Business Insider: https://www.businessinsider.in/These-Hilarious-Charts-Will-Show-You-Exactly-Why-Correlation-Doesnt-Mean-Causation/articleshow/34892968.cms

Stage, C. (2003). *Classical test theory or item response theory: The Swedish experience.* Santiago: Centro de Estudios Públicos,.

Umobong, M., & Jacob, S. S. (2016). A comparison of classical and item response theory person/item parameters of physics achievement test for technical schools. *African Journal of Theory and Practice of Educational Assessment, 4*(131).

Yale University. (1997). *Sampling*. Retrieved from Yale University Department of Statistics: http://www.stat.yale.edu/Courses/1997-98/101/sample.htm

Glossary

Correlation

Two variables that have a relationship or interdependence are said to correlate with each other (lumen Boundless Statistics, n.d.). For instance, consider an example of a positive correlation between the scientific research output of a country and its PISA science score. As the scientific research output of a country increases, its PISA score increases (OECD, 2007). An example of a negative correlation can be the association between poverty and test scores on a national school-leaving examination. As poverty increases, the mean test score of a region decreases. However, there is an important caveat in interpreting correlations. Correlations do not imply causation. For instance, just because poverty and test scores are correlated, we **cannot**, **with certainty**, **say** that poverty **causes** poor test scores. Other statistical tests would be required for more robust conclusions. There may also be spurious correlations between variables, for instance, correlations between the consumption of margarine and divorce rates (Spector, 2014).

Random Sample

A random sample is a sample where each member of the population of interest has an equal chance of being selected. The purpose of choosing a random sample is to ensure that the sample is unbiased and representative of the population of interest (Yale University, 1997). Thus, the findings generated from a study with a random sample can be generalised to the entire population.

Non-random Sample

Non-random samples are those samples that have not been chosen using random sampling techniques. The members of a non-random sample are either chosen for convenience (they just happen to be available) or for some specific characteristic that they possess (e.g. gender, cultural identity, etc.). The findings generated from a study with a non-random sample cannot be generalised to an entire population (California State University Long Beach, 1998).

Test Reliability

The extent to which a test is free of error is called test reliability (Rust, Kosinksi, & Stillwell, 2021). Simplistically put, test reliability is often measured by correlating different test scores or sets of test scores. The higher the correlations between the test scores, the greater the consistency of the test results, the less the error in test score measurements, and hence, the greater the reliability. A good reliability measure is one of the characteristics of a good test.

Item Difficulty

In CTT, item difficulty or difficulty is the number of correct responses divided by the total number of responses for a question (Rust, Kosinksi, & Stillwell, 2021). The ranges of difficulty are as follows: >0.7, easy; 0.3–0.7, moderate, 0.0 to 0.3, hard. In IRT, the item difficulty is the ability level at which the probability of a correct response is 0.5. If the ability level is less than 0, then the question is easy. If it is greater than 0, the question is difficult (Baker & Kim, 2017).

Item Characteristic Curve

In IRT, the plot of the probability of getting a correct answer (Y-axis) versus student ability (X-axis) is called an item characteristic curve. The item characteristic curve is a logistic curve or an S-shaped curve (Baker & Kim, 2017).

Item Discrimination

In CTT, item discrimination or discrimination provides an estimate of how well a question discriminates among high-ability and low-ability examinees (Rust, Kosinksi, & Stillwell, 2021). The discrimination index is used to compute discrimination. In IRT, the slope of the item characteristic curve at the difficulty of the item is used to determine the discrimination. The steeper the slope, the better the item at discriminating between high-ability and low-ability examinees (Baker & Kim, 2017).

IRT 1PL Model

In the IRT 1PL model, also called the Rasch model, the probability of a student answering a question correctly is a function of the following (Baker & Kim, 2017):

a) the student's ability
b) the difficulty of the item

IRT 2PL Model

In the IRT 2PL model, the probability of a student answering a question correctly is a function of the following (Baker & Kim, 2017):

a) the student's ability
b) the difficulty of an item
c) the discrimination of an item

IRT 3PL Model

In the IRT 2PL model, the probability of a student answering a question correctly is a function of the following (Baker & Kim, 2017):

a) the student's ability
b) the difficulty of an item
c) the discrimination of an item
d) the guessing parameter

Parallel Forms

Different tests with similar items that test the same construct are called parallel forms of a test. Parallel forms of a test are systematically linked to each other so that they provide comparable estimates of student performance (Rust, Kosinksi, & Stillwell, 2021).

Test Equating

In many large-scale assessment and selection exams, parallel forms of a test are administered to make cheating harder. A problem with using parallel forms of a test is that the item parameters, especially difficulty and reliability, of parallel tests may differ slightly, thus affecting the equivalence of test scores. Examinees may be unduly advantaged or disadvantaged if differences in test scores because of test characteristics or item parameters. Thus, a test equating, a procedure for generating comparable test scores on different test forms, is carried out (Cook & Eignor, 1991).

Assessing the equivalence between Classical Test Theory and Item Response Theory

[i] [i] This article is authored by Madhumati Manjunath and Shilpi Banerjee.
Madhumati Manjunath is a research associate at the School of Continuing Education at Azim Premji University. She can be reached at madhumati.manjunath19_mae@apu.edu.in.
Shilpi Banerjee works as Assistant Professor in the School of Continuing Education at Azim Premji University. She can be reached at shilpi.banerjee@azimpremjifoundation.org.
This article can be cited as-
Assessing the equivalence between Classical Test Theory and Item Response Theory, Assessment resources, 2022, Azim Premji University