

An empirical evaluation and comparison of Classical Test Theory and Rasch Model

Mehtap ERGÜVEN*

Abstract

Although it is considered an interdisciplinary subject, theory of measurement is mostly based on the mathematical foundations. In the last quarter of the 19th century, various psychometric approaches have been developed on the strength of those scientific foundations. Those theories allow users to analyze and determine relationship among individuals' latent abilities and their response patterns with respect to surveys, aptitude tests, examinations or other educational measurement tools. Besides, these models of measurement are enhanced to estimate model parameters on the strength of features of items. Item-person statistics are used for the educational measurement and assessment purposes mostly. In that context, Classical Test Theory (CTT) and Rasch model are focused and elaborated to evaluate School Olympiad Examination. The way how to implement both theories and how to interpret results of the analyses is presented in the study. Comparison of both theories is done. This comparison gives chance to determine similarities and differences between them. Using the IRTPRO program, some notable features of the items are described and discussed in detail. According to gender, item-person statistics/parameters are determined and interpreted. Behaviors of Item Characteristics Curves are elaborated with respect to both genders. Reliability of the test is defined with Cronbach's alpha in the whole test and with respect to genders. As a conclusion it is suggested that SOE (School Olympiad Examination) has a high reliability as an educational test, groups of female and male students handled almost all of the items similarly. CTT and Rasch model give similar information with respect to different genders in general. However, Rasch model gives better and more informative results in the ability determination and comparison.

Keywords: Item response theory, Rasch model, Classical Test Theory, discrimination index, difficulty index, item-person statistics, item characteristic curve, reliability

Introduction

Since the beginnings of psychological measurement, Classical Test Theory (CTT) has been the dominant approach to the construction, analysis, and scoring of psychological scales. Although CTT methods dominate to this day, a second approach, Item Response Theory (IRT) is becoming more popular and better appreciated. (Embretson & Reise, 2000).

Main purpose of both CTT and Item Response Theory is determining numerical scores that approximate an individual's latent ability level. Although they have a common target, both primary theories of measurement differ significantly. They make fundamentally different assumptions about the nature of the construct being measured as well as about how individuals respond to test items (Sharkness & DeAngelo, 2011).

The traditional approach - "classical test theory or true score theory (Magno, 2009)" - has a simple implementation algorithm and has served measurement researchers successfully for many years. However, in 1960 Georg Rasch developed IRT models to measure reading ability and to devise tests for the military. His name was given to one of the best-known IRT models "Rasch Model".

In CTT a person's "true score" is entirely dependent on a particular set of items because the true score is defined

in relation to a specific test or scale. In IRT, a person's "true score" is entirely independent of items (Allen & Yen, 2002). IRT is theoretically and mathematically more sophisticated than CTT and can be used to obtain estimates of constructs and latent traits that have many desirable attributes (Sharkness & DeAngelo, 2011). It can be assumed that, Rasch model is the most preferable and easiest among IRT models for the items which have equal "good" measures of the trait and only different difficulty parameters.

CTT is based on relatively weak assumptions that are easy to meet with real data and modest sample sizes. Those above-mentioned models are simple to use and require little mathematical knowledge on the part of the user (De Champlain, 2010). The article gives opportunity to compare both models in whole and among female and male examinees. According to distinct genders item-person parameters (discrimination, difficulty and ability) are determined and interpreted. Reliability of the mathematical part of the School Olympiad Examination (SOE) (Georgia, 2013) is analyzed. Relation among ability, difficulty and probability of answering an item correctly is represented by item characteristic curves (ICC).

In the last decades there has been explosive growth in programs that can analyze tests on the basis of IRT as-

* Ph.D Student, Faculty of Computer Technologies and Engineering, International Black Sea University, Tbilisi, Georgia
E-mail: merguven@ibsu.edu.ge

sumptions. IRTPRO software is one of them. Identification of item-person parameters and representation of item characteristic curves is done by IRTPRO.

Methodology

CTT and Rasch models are analyzed from different perspectives. Mathematics category of the School Olympiad Examination (2013) is elaborated, which is administered in Georgia by the International Black Sea University. The examination is in multiple-choice format and responses of 523 students were converted to dichotomized answers (0/1). False answers were converted to 0 and true responses - to 1.

Experimental studies are done with IRTPRO program to determine the ability levels of the examinees. IRTPRO is used to find out item difficulty, item discrimination indices (in CTT)/parameters (in Rasch) and to demonstrate ICCs as well.

Additionally, the ability levels description has been done among genders. 284 female and 239 male students were included in research. Hence, the comparison of different groups of students was done and similarities/differences among the groups were represented. For that reason reliability coefficient was described and behaviors of item characteristic curves were defined in both groups. Correlations among discrimination and difficulty parameters were evaluated and Microsoft Excel s used for the required statistical analyses.

Rasch Model

Since 1960, when Georg Rasch (1901–1981) produced his now well-accepted measurement model published as the “Probabilistic Models for Some Intelligence and Attainment Tests” (Rasch, 1960), there has been a quiet revolution in measuring variables in education, psychology, business and medicine. Rasch’s initial measurement model, now called the Simple Logistic Model of Rasch, is only applied to items in a dichotomous format, such as no/yes, disagree/agree, wrong/right (Cavanagh & Waugh, 2011).

In an IRT analysis, trait levels and item difficulties are usually scaled on a standardized metric, so that their means are 0 and the standard deviations are 1. The b parameter on the ability (theta) axis corresponds to the location where the probability of answering correctly for the focused item is 0.5 both in the Rasch and 2PL-IRT model (Furr & Bacharach, 2007).

Probability of answering an item correctly depends on the respondent’s ability level (θ) and difficulty of item “b”. Lower b values are indicators of more difficult items (Wauters, Desmet, & Van Den Noortgate, 2010) in Rasch model. *If an individual’s ability level is higher than an item’s difficulty parameter, probability of answering the item correctly would be higher than 0.5.* Correspondingly, if the value of b increases, required ability level for a test taker will be higher to have a 50% chance of getting the item correctly. As a simple summary:

- (1) The easier the question, the more likely the student will respond correctly to it, and
- (2) The more able the student, the more likely he/she will pass the question compared to a less able student.

In constructing tests using this model frequently discards those items that do not meet these assumptions (Wright & Stone, 1979).

It can be interpreted that one-parameter (Rasch) model includes only the difficulty parameter. Rasch model assumes that the score is a function of only the difficulty (McBride., 2001) and generally the item discrimination parameter is equaled to 1.

$$P_i(\theta) = \frac{e^{(\theta-b_i)}}{1 + e^{(\theta-b_i)}}$$

In the above equation, Pi (θ) is equal to the probability that an examinee chose at random with a given theta score answers item i in the scaled direction, bi is the item i difficulty parameter and e is equal to 2.718 (McBride., 2001). The model based on the exponential function yields measures of people and items on a natural scale, whose unit is called “logit” (Choppin, 1987).

According to IRT assumptions, ability level of a person can be evaluated with logit:

$$\theta_s = LN\left(\frac{P_s}{1-P_s}\right)$$

P_s is the proportion of correct answers for respondent “s”. The proportion correct can be simply found as the ratio of the number of correctly answered items and total number of items.

Table 1. The computation process of determining initial estimates of “ability levels and difficulty parameters” with the given formulas

examinees	item1	item2	item3	item4	item5	proportion correct for each respondent	ability level
1	0	1	1	1	1	1	0.8 1.39
2	1	1	0	0	1	1	0.6 0.41
3	1	0	0	0	0	0	0.2 -1.39
4	0	1	0	0	1	1	0.4 -0.41
5	1	1	0	1	1	1	0.8 1.39
6	1	0	0	1	1	1	0.6 0.41
7	1	1	1	0	1	1	0.8 1.39
proportion correct for each item	0.71	0.71	0.29	0.43	0.86		
difficulty level	-0.92	-0.92	0.92	0.29	-1.79		

Rasch measurement requires the researcher to design the items in a scale from easy to hard, but with certain conditions in mind. The conditions mean that the probability of answering positively must be related to the difference between the person’s measure and the item difficulty (Cavanagh & Waugh, 2011). In Rasch model, the following equation expresses what happens when a person meets an item in a test.

LOG (P/1-P) =person measure – item difficulty

According to the given equation, if an examinee’s personal measure is 2.3 and item’s difficulty is 1.2, then logit is

LOG (P/1-P) = 2.3-1.2=1.1

It can be interpreted that the examinee's probability of getting the item correctly on the Y-axis corresponds to the value of logit which is represented on the X-axis. Thus, it is obvious that Rasch equation is related to the graphic of item characteristic curve.

Using the above-given example and logit 1.1, we can convert the logit scale back to probability (Revelle, 2005) by taking $\exp(1.1)/(1 + \exp(1.1))$ and we get 0.75. As a conclusion, the person has a 75% chance of getting the item correctly. This valuable information is accessible through the IRT analysis.

Studies in various parts of the world indicate that in a given subject area, the typical child's achievement level would rise by rather less than half a logit in a typical school year (Choppin, 1987). For instance, if an individual's change is from -1.94 logits to -1.22 logits, this means too much positive change for him/her.

Findings and Interpretations

According to the IRT assumptions, "difficulty values less than -1.0 indicate to fairly easy items, whereas items with difficulty greater than 1.0 indicate to rather difficult items". (Zumbo, 1999).

Items that are very difficult and very easy for a particular group of examinees usually have substantially lower point biserials for that group of examinees than do items of medium difficulty (Lord & Novick, 1968). That situation is clearly met in the following table. On the strength of CTT assumptions, item 2 is the most difficult item and the corresponding discrimination parameter coefficient of point biserial correlation is 0.05. It is lower than average magnitude (0.51) of discrimination. Similarly, the easiest questions are determined as item 14 and item 16. The corresponding point biserial for the item 16 is lower than average.

Table 2. Representation of discrimination and difficulty indices of CTT and difficulty parameter of Rasch model in the SOE

items	CTT Discrm. index : point biserial correlation	CTT difficulty index "p"	Rasch difficulty parameter "b"
i1	0.52	0.45	0.196
i2	0.05	0.11	2.101
i3	0.43	0.28	0.958
i4	0.54	0.65	-0.603
i5	0.48	0.57	-0.297
i6	0.43	0.49	0.034
i7	0.47	0.39	0.447
i8	0.50	0.63	-0.553
i9	0.55	0.58	-0.336
i10	0.51	0.73	-1.006
i11	0.55	0.55	-0.188
i12	0.37	0.34	0.679
i13	0.52	0.37	0.537
i14	0.44	0.78	-1.277
i15	0.42	0.42	0.312
i16	0.32	0.77	-1.200
i17	0.48	0.52	-0.096

IRT methods are used because researchers want to (a) more rigorously study how items function differently in different groups;

(b) place individuals from different groups onto a common scale, even if they have responded to different items (Reise, Ainsworth & Haviland, 2005).

According to Hambleton, Swaminathan & Rogers (1991), except for measurement errors, evaluated item parameters will be the same in different groups. Using the mentioned idea, to check difficulty parameter invariance, the relations among difficulty parameters were determined by coefficient of correlation. Correlation coefficient is evaluated as 0.934172 between difficulty "b" parameters of female and male students. Obviously, the difficulty parameter invariance is observed with regard to a very high correlation. Unfortunately, to compare discrimination parameters is not possible in Rasch model, since they are considered as fixed values. In the study those values were equaled to 1 to have a very good discrimination among low ability and high ability groups of students.

Table 3. Rasch model's difficulty parameters and CTT's discrimination indices presentation with respect to genders in the SOE

items	Rasch Difficulty parameters			CTT Discrimination Indices	
	a	female b	male b	Item-total correlation	Item-total correlation
				female	male
i1	1	0.37	0.1	0.55	0.48
i2	1	2.46	2.75	0.04	0.07
i3	1	1.32	1.1	0.37	0.50
i4	1	-0.61	-0.98	0.50	0.58
i5	1	-0.08	-0.77	0.45	0.51
i6	1	0.08	-0.01	0.42	0.43
i7	1	0.85	0.25	0.44	0.49
i8	1	-0.86	-0.54	0.48	0.54
i9	1	-0.15	-0.79	0.56	0.54
i10	1	-1.32	-1.23	0.52	0.50
i11	1	-0.26	-0.23	0.54	0.56
i12	1	1.14	0.56	0.36	0.37
i13	1	0.6	0.79	0.53	0.51
i14	1	-1.25	-2.12	0.46	0.40
i15	1	0.49	0.29	0.39	0.46
i16	1	-1.76	-1.26	0.32	0.34
i17	1	-0.04	-0.23	0.51	0.44

Item discriminating power is defined as correlation between the item scores and the total test scores (McDonald, 1999, p. 231). The higher the value of coefficients, the better discrimination power the item has. High discrimination value shows that students with high scores answered the item correctly whereas students with low test scores responded incorrectly. If the item-total correlation is negative or close to zero, the item should be eliminated from the test.

In table 3, item-total correlations were calculated for female and male test takers separately. Item-total correlations were the lowest values for both female and male students in the item 2. Those discrimination indices were 0.04 for female and 0.07 for the male students. Both of item-total correlations were very low. Hence, item 2 does not differentiate well between high ability and low ability groups among female and male examinees as well.

Similarly, discrimination index of item 16 is also low with regards to female and male students (0.32 and 0.34). Al-

though item 16 cannot differentiate among different ability groups well, the item is better than item 2 in both cases. Corrected item-total correlations are satisfactorily high for the item 1 (0.55) and item 9 (0.56) among females. In addition, item 4 (0.58) and item 11 (0.56) best discriminate the high ability male students from the low ability male students.

Ability Parameters Determination and Comparison for Female and Male Students

The maximum ability level was 2.24 among the male students while it was found 2.35 for the female students with respect to Rasch Model. On the other hand, the minimum ability levels were found -2.27 and -2.39 for female and male students respectively.

According to the classical theory, similar application can be done using total test scores. In that case, maximum score was 17 and minimum score was 0 among both female and male students. It is clear that this information is not enough to demonstrate differences among genders. However, previously given ability levels give much more idea about identification of capabilities of different genders. Therefore it can be judged that IRT gives much more information about the focused situation. Eventually, it can be interpreted that most successful student was a female (2.35) and lowest ability student was a male (-2.39), according to Rasch model.

Besides those outputs, the average score was calculated as 8.98 and 8.35 among male and female students respectively. According to average scores it can be interpreted that average score of males is greater than females.

On the other hand, mean value of the abilities were -0.00042 and -0.00049 for male and female test takers. This result is similar to the CTT outputs and correspondingly it can be denoted that average of the males is greater than females.

Table 4. General statistics results

General Statistics Results					
total size	523	mean (CTT female)	8.35	mean (Rasch female)	-0.0005
female	284	mean (CTT male)	8.98	mean (Rasch male)	-0.0004
male	239	max female	17	max female	2.35
mean	8.64	min female	0	min female	-2.27
variance	18.17	max male	17	max male	2.24
std	4.26	min male	0	min male	-2.39

According to Zumbo, (1999), items with “difficulty values less than -1.0 indicate fairly easy items whereas items with difficulty values greater than 1.0 indicate rather difficult items”.

In the light of the mentioned information, since difficulty parameters of item 14 and item 16, -1.277 and -1.20, were less than -1, they can be interpreted as easy items. Obviously the situation coincides with CTT results. Because difficulty values close to zero are considered as difficult and close to one are considered easier items. Thus, item 14 and item 16 can be interpreted as easy questions. In CTT, their difficulty “p” indices of those items are 0.78 and 0.77 respectively (see table 2). Besides, item 2 is very difficult, since its p index 0.05 was very close to zero and its b parameter 2.101 exceeded 1.

In Rasch model, the easiest item for male students was “item 14” and its difficulty was -2.12. On the other hand,

item 16 was the easiest for the female examinees and difficulty parameter was -1.76 in item 16.

Reliability of the Test with respect to Genders

Being able to make important and correct decisions in the educational process, as in other branches of science, depends on reliable and valid measurement which is described as the results of any observation that are matched with numbers or other symbols (Baykul, 2000).

Classical test theory’s reliability coefficients are widely used in behavioral and social research. Each provides an index of measurement consistency ranging from 0 to 1.00 and their interpretation (Webb, Shavelson & Haertel, 2006). A value of 0.7-0.8 is an acceptable value for Cronbach’s α ; values substantially lower indicate an unreliable scale (Field, 2005). According to Varma, non-reliable and non-valid test scores are simply meaningless numbers (Varma, 2013).

Marginal reliability for response pattern scores was found 0.75 for the both female and male students. In general reliability analysis, reliability coefficient Cronbach’s alpha was found 0.84 with respect to whole data. On the strength of the female test takers, coefficient alpha was 0.8394. In the male group of students the level of internal consistency was increased to 0.8504. In all cases, SOE had a high reliability coefficient as an educational test.

Evaluation of Item Characteristic Curves’ Behaviors with respect to Different Genders

Each item has its own ICC in IRT. Item characteristic curves are used to represent probability of a correct response and ability (θ) for each item, on a test. In the graphical representation of the ICC, the probability of correct response exists on the Y-axis and, range from 0.0 to 1.0. Ability (θ) is along the X-axis.

One of the main characteristics of the item is difficulty and it is accessible from the graph of ICC. ICCs in the research were obtained for the male and female students by using IRTPRO. Trace lines ICCs are all identically shaped since Rasch model assumes all items are equally discriminating. Slopes of the lines are same. According to the group of male students, in the table 3 it is observed that difficulty parameter of the item 2 was 2.75. This result is depicted in the below ICC which belongs to item 2. In the curve, 2.75 is represented on the X-axis and corresponding probability 0.5 is placed on the Y-axis. Meaning of the information is that if the ability level of a student was greater than 2.75, there is 50% chance of responding the item 2 correct. Clearly, item 2 can be interpreted as a very difficult item.

In the same group of male students, difficulty parameter of item 14 was the lowest “-2.12” and this is visible in figure 2.

According to Zumbo (1999), item discrimination values of 1.0 or greater are considered very good. Items with very low “a” values are useless to distinguish among individuals. Therefore, item discrimination parameter is fixed to the “1” to get a good discrimination. Hence, the slope of the curves are same in all figures. It is not possible to talk about discrimination property of the items with respect to separate groups of genders in the Rasch model. Similar to the figure 1, item 2 had the largest difficulty parameter “2.46” in the female group as represented in figure 3. However, unlike males’ item 16 was the easiest for the females and its

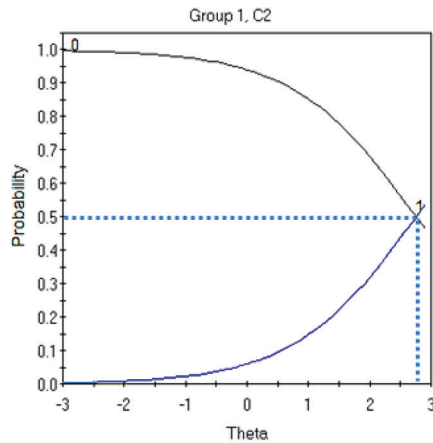


Figure 1. ICC of the item 2 in the group of male students (the most difficult item)

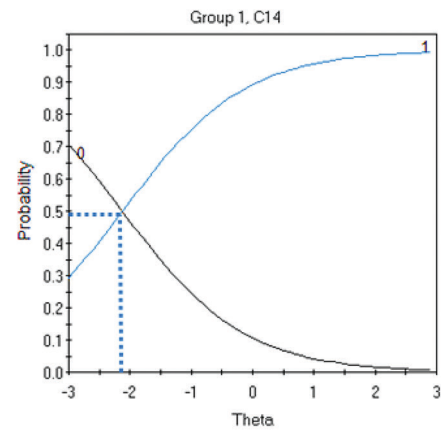


Figure 2. ICC of the item 14 in the group of male students (easiest item)

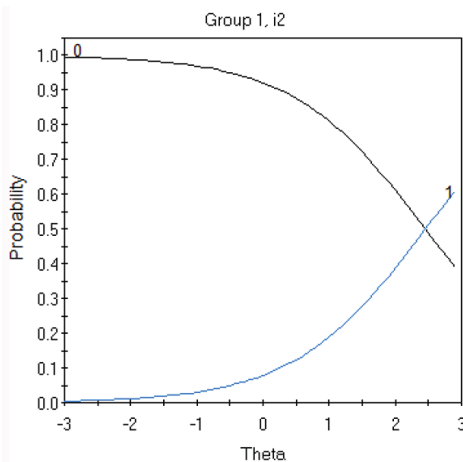


Figure 3. ICC of the item 2 in the group of female students

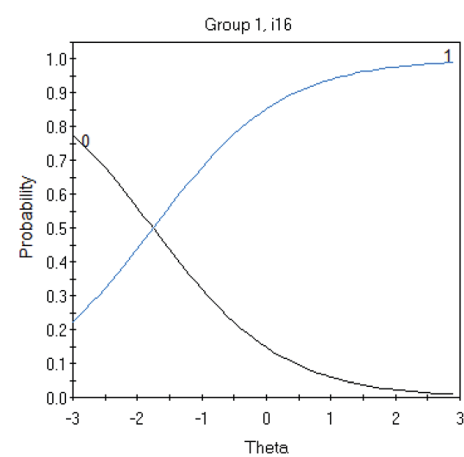


Figure 4. ICC of the item 16 "easiest item in the group of female students"

difficulty level was found as -1.76 (see figure 4).

As a result, item 2 is the most problematic question in the examination because it does not differentiate well and its difficulty is very high, it should be eliminated from the examination.

Conclusion

In the research, traditional approach and Rasch model were elaborated and compared from different perspectives. The way how to implement both CTT and Rasch model and how to interpret results of the analyses is presented. Comparison of both theories is done for the whole group and among genders. Therefore, main assumptions and principles of both theories are empirically presented; similarities and differences with respect to different genders were detected.

Several notable features of the items were described and discussed with the help of IRTPRO program. According to different genders, item-person statistics/parameters were determined and interpreted. Behaviors of Item Characteristic Curves were elaborated with respect to female and male students. Reliability of the test was detected with Cronbach's alpha for the whole test and for genders.

Eventually, it is suggested that, SOE has a high reliability as an educational test; groups of female and male students handled almost all of the items similarly. Although male students more successfully endorsed the items in whole, the highest ability level was observed among female examinees. CTT and Rasch model give similar information with respect to different genders in general. Besides that, Rasch model gives better and more informative results in the ability determination and comparison.

References

- Allen, M., Yen, W. (2002). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press. (Original Work Published 1979).
- Baykul, Y. (2000). "Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması". Ankara: ÖSYM.
- Cavanagh, R. F., Waugh, R. (2011). *Applications of Rasch Measurement in Learning Environments Research*. Netherlands: Sense Publishers.
- Choppin, B. (1987). The Rasch Model for Item Analysis. In D. L. McArthur, *Alternative Approaches to the Assessment of Achievement*. LA: Springer.
- De Champlin, A. F. (2010). A Primer on Classical Test Theory and Item Response Theory for Assessments in Medical Education. *Medical Education*, 44, pp. 109-117.
- Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.
- Field, A. (2005). *Reliability Analysis*. In *discovering statistics using SPSS*. London.
- Furr, R. M. and Bacharach, V. R. (2007). Item Response Theory and Rasch Models. In R. M. Furr, & V. R. Bacharach, *Psychometrics, an Introduction* (pp. 314-334). Sage Publications.
- Hambleton, R.K.; Swaminathan, H.; Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. (D. S. Foster, Ed.) USA: Sage Publications.
- Lord, F. M.; Novick, M. R. (1968). Normal Ogive Parameters: Item Difficulty. In F. M. Lord, & M. R. Novick, *Statistical Theories of Mental Test Scores*. New York, USA: Addison-Wesley Publishing Company, Inc.
- 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), 1-11.
- McBride, N. (2001). *An Item Response Theory Analysis of the Scales from the International Personality Item Pool and the NEO Personality Inventory-Revised*. Blacksburg, VA: Virginia Polytechnic Institute and State University.
- Rasch, G., (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen, Danish Institute for Educational Research.
- Reise, S.P.; Ainsworth, A. T.; Haviland, M.G. (2005). Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research. *Current Directions in Psychological Science*, 14(2), 95-101.
- Revelle, W. (2005). A short guide to R. retrieved from <https://personality-project.org>
- Sharkness, J.; De Angelo, L. (2010). *Measuring Student Involvement: A Comparison of Classical Test Theory and Item response theory in the Construction of Scales from Student Surveys*. Higher Education Research Institute, 52, 480-507.
- Wauters, K., Desmet, P., Van Den Noortgate, W. (2010). *Acquiring Item Difficulty Estimates: a Collaborative Effort of Data*. Belgium.
- Webb, N.M.; Shavelson, R.J.; Haertel, E.H. (2006). *Reliability Coefficients and Generalizability Theory* (Elsevier B.V ed., Vol. 26). USA.
- Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.
- Zumbo, B. D. (1999, April). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary*. Ottawa, Canada.
- Varma, S. (2013). Preliminary Item statistics Using Point Biserial Correlation and p-Values. Retrieved October 2013, from Educational Data Systems: http://www.eddata.com/resources/publications/EDS_Point_Biserial.pdf