

## Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data

Carlo Magno  
De La Salle University, Manila

### Abstract

The present report demonstrates the difference between classical test theory (CTT) and item response theory (IRT) approach using an actual test data for chemistry junior high school students. The CTT and IRT were compared across two samples and two forms of test on their item difficulty, internal consistency, and measurement errors. The specific IRT approach used is the one-parameter Rasch model. Two equivalent samples were drawn in a private school in the Philippines and these two sets of data were compared on the tests' item difficulty, split-half coefficient, Cronbach's alpha, item difficulty using the Rasch model, person and item reliability (using Rasch model), and measurement error estimates. The results demonstrate certain limitations of the classical test theory and advantages of using the IRT. It was found in the study that (1) IRT estimates of item difficulty do not change across samples as compared with CTT with inconsistencies; (2) difficulty indices were also more stable across forms of tests than the CTT approach; (3) IRT internal consistencies are very stable across samples while CTT internal consistencies failed to be stable across samples; (4) IRT had significantly less measurement errors than the CTT approach. Perspectives for stakeholders in test and measurement are discussed.

Test developers are basically concern about the quality of test items and how examinees respond to it when constructing tests. A psychometrician generally uses psychometric techniques to determine the validity and reliability. Psychometric theory offers two approaches in analyzing test data: Classical test theory (CTT) and item response theory (IRT). Both theories enable to predict outcomes of psychological tests by identifying parameters of item difficulty and the ability of test takers. Both are concerned to improve the reliability and validity of psychological tests. Both of these approaches provide measures of validity and reliability. There are some identified issues in the classical test theory that concerns with calibration of item difficulty, sample dependence of coefficient measures, and estimates of measurement error which in turn is addressed by the item response theory. The purpose of this article to demonstrate the advantages and disadvantages of using both approaches in analyzing a given chemistry test data.

### Classical Test Theory

Classical test theory is regarded as the "true score theory." The theory starts from the assumption that systematic effects between responses of examinees are due only to variation in ability of interest. All other potential sources of variation existing in the testing materials such as external conditions or internal conditions of examinees are assumed either to be constant through rigorous standardization or to have an effect that is nonsystematic or random by nature (Van der Linden & Hambleton, 2004). The central model of the classical test theory is that observed test scores (TO) are composed of a true score (T) and an error score (E) where the true and the error scores are independent. The variables are established by Spearman (1904) and Novick (1966) and best illustrated in the formula:  $TO = T + E$ .

The classical theory assumes that each individual has a true score which would be obtained if there were no errors in measurement. However, because measuring instruments are

imperfect, the score observed for each person may differ from an individual's true ability. The difference between the true score and the observed test score results from measurement error. Using a variety of justifications, error is often assumed to be a random variable having a normal distribution. The implication of the classical test theory for test takers is that tests are fallible imprecise tools. The score achieved by an individual is rarely the individual's true score. This means that the true score for an individual will not change with repeated applications of the same test. This observed score is almost always the true score influenced by some degree of error. This error influences the observed to be higher or lower. Theoretically, the standard deviation of the distribution of random errors for each individual tells about the magnitude of measurement error. It is usually assumed that the distribution of random errors will be the same for all individuals. Classical test theory uses the standard deviation of errors as the basic measure of error. Usually this is called the standard error of measurement. In practice, the standard deviation of the observed score and the reliability of the test are used to estimate the standard error of measurement (Kaplan & Saccuzzo, 1997). The larger the standard error of measurement, the less certain is the accuracy with which an attribute is measured. Conversely, small standard error of measurement tells that an individual score is probably close to the true score. The standard error of measurement is calculated with the formula:  $Sm = S\sqrt{1-r}$ . Standard errors of measurement are used to create confidence intervals around specific observed scores (Kaplan & Saccuzzo, 1997). The lower and upper bound of the confidence interval approximate the value of the true score.

Traditionally, methods of analysis based on classical test theory have been used to evaluate tests. The focus of the analysis is on the total test score; frequency of correct responses (to indicate question difficulty); frequency of responses (to examine distracters); reliability of the test and item-total correlation (to evaluate discrimination at the item level) (Impara & Plake, 1997). Although these statistics have been widely used, one limitation is that they relate to the sample under scrutiny and thus all the statistics that describe items and questions are sample dependent (Hambelton, 2000). This critique may not be particularly relevant where successive samples are reasonably representative and do not vary across time, but this will need to be confirmed and complex strategies have been proposed to overcome this limitation.

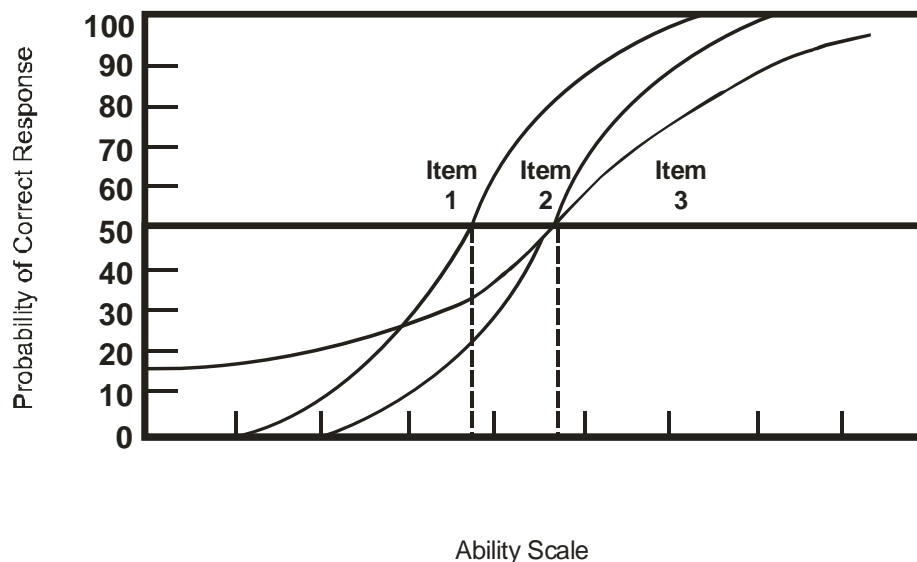
### Item Response Theory

Another branch of psychometric theory is the item response theory (IRT). IRT may be regarded as roughly synonymous with latent trait theory. It is sometimes referred to as the strong true score theory or modern mental test theory because IRT is a more recent body of theory and makes stronger assumptions as compared to classical test theory. This approach to testing based on item analysis considers the chance of getting particular items right or wrong. In this approach, each item on a test has its own item characteristic curve that describes the probability of getting each particular item right or wrong given the ability of the test takers (Kaplan & Saccuzzo, 1997). The Rasch model as an example of IRT is appropriate for modeling dichotomous responses and models the probability of an individual's correct response on a dichotomous item. The logistic item characteristic curve, a function of ability, forms the boundary between the probability areas of answering an item incorrectly and answering the item correctly. This one-parameter logistic model assumes that the discriminations of all items are assumed to be equal to one (Maier, 2001).

Another fundamental feature of this theory is that item performance is related to the estimated amount of respondent's latent trait (Anastasi & Urbina, 2002). A latent trait is symbolized as theta ( $\theta$ ) which refers to a statistical construct. In cognitive tests, latent traits are called the ability measured by the test. The total score on a test is taken as an estimate of that ability. A person's specified ability ( $\theta$ ) succeeds on an item of specified difficulty.

There are various approaches in the construction of tests using item response theory. Some approaches use the two-dimensions that plot item discriminations and item difficulties. Other approaches use a three-dimension for the probability of test takers with very low levels of ability getting a correct response (as demonstrated in Figure 1). Other approaches use only the difficulty parameter (one dimension) such as the Rasch Model. All these approaches characterize the item in relation to the probability that those who do well or poorly on the exam will have different levels of performance.

Figure 1  
Hypothetical Item Characteristic Curves for Three Items using a Three Parameter Model



The item difficulty parameter ( $b_1$ ,  $b_2$ ,  $b_3$ ) corresponds to the location on the ability axis at which the probability of a correct response is .50. It is shown in the curve that item 1 is easier and item 2 and 3 have the same difficulty at .50 probability of correct response. Estimates of item parameters and ability are typically computed through successive approximations procedures where approximations are repeated until the values stabilize.

The preset study focused on the one-parameter model or the Rasch model. The Rasch model is based on the assumption that both guessing and item differences in discrimination are negligible or constant. Rasch began his work in educational and psychological measurement in the late 1940's. Early in the 1950's he developed his Poisson models for reading tests and a model for intelligence and achievement tests which was later called the "structure models for items in a test" which is called today as the Rasch model.

Rasch's (1960) main motivation for his model was to eliminate references to populations of examinees in analyses of tests. According to him that test analysis would only be worthwhile if it were individual centered with separate parameters for the items and the examinees (van der Linden & Hambleton, 2004). His worked marked IRT with its probabilistic modeling of the interaction between an individual item and an individual examinee. The Rasch model is a probabilistic unidimensional model which asserts that (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 discard those items that do not meet these assumptions (Wright & Stone, 1979).

The Rasch model was derived from the initial Poisson model illustrated in the formula:

$\varepsilon = \frac{\delta}{\theta}$  where  $\varepsilon$  is a function of parameters describing the ability of examinee and difficulty of the test,  $\theta$  represents the ability of the examinee and  $\delta$  represents the difficulty of the test which is estimated by the summation of errors in a test. Furthermore, the model was enhanced to assume that the probability that a student will correctly answer a question is a logistic function of the difference between the student's ability [ $\theta$ ] and the difficulty of the question [ $\beta$ ] (i.e. the ability required to answer the question correctly), and only a function of that difference giving way to the Rasch model.

From this, the expected pattern of responses to questions can be determined given the estimated  $\theta$  and  $\beta$ . Even though each response to each question must depend upon the students' ability and the questions' difficulty, in the data analysis, it is possible to condition out or eliminate the student's abilities (by taking all students at the same score level) in order to estimate the relative question difficulties (Andrich, 2004; Dobby & Duckworth, 1979). Thus, when data fit the model, the relative difficulties of the questions are independent of the relative abilities of the students, and vice versa (Rasch, 1977). The further consequence of this invariance is that it justifies the use of the total score (Wright & Panchapakesan, 1969). In the current analysis this estimation is done through a pair-wise conditional maximum likelihood algorithm.

The Rasch model is appropriate for modeling dichotomous responses and models the probability of an individual's correct response on a dichotomous item. The logistic item characteristic curve, a function of ability, forms the boundary between the probability areas of answering an item incorrectly and answering the item correctly. This one-parameter logistic model assumes that the discriminations of all items are assumed to be equal to one (Maier, 2001).

According to Fischer (1974) the Rasch model can be derived from the following assumptions:

(1) Unidimensionality. All items are functionally dependent upon only one underlying continuum.

(2) Monotonicity. All item characteristic functions are strictly monotonic in the latent trait. The item characteristic function describes the probability of a predefined response as a function of the latent trait.

(3) Local stochastic independence. Every person has a certain probability of giving a predefined response to each item and this probability is independent of the answers given to the preceding items.

(4) Sufficiency of a simple sum statistic. The number of predefined responses is a sufficient statistic for the latent parameter.

(5) Dichotomy of the items. For each item there are only two different responses, for example positive and negative. The Rasch model requires that an additive structure underlies the observed data. This additive structure applies to the logit of  $P_{ij}$ , where  $P_{ij}$  is the probability that subject  $i$  will give a predefined response to item  $j$ , being the sum of a subject scale value  $u_i$  and an item scale value  $v_j$ , i.e.  $\ln(P_{ij}/1 - P_{ij}) = u_i + v_j$

There are various applications of the Rasch Model in test construction through item-mapping method (Wang, 2003) and as a hierarchical measurement method (Maier, 2001).

#### Issues in CTT

There are four main limitations in the CTT approach that will be demonstrated in the present study. First is that estimates of item difficulty are group dependent. A test item functions to be easy or difficult given a sample of examinees and these indices change when a different sample takes the test. Another problem is that the  $p$  and  $r$  values are also dependent on the examinee sample from which they are taken. This problem is similar with item difficulty estimates. The third is

that ability scores of examinees are entirely test dependent. The examinees ability change depending on different occasions they take the test which results to poor consistency of the test.

#### Advantages of the IRT

The benefit of the item response theory is that its treatment of reliability and error of measurement through item information function are computed for each item (Lord, 1980). These functions provide a sound basis for choosing items in test construction. The item information function takes all items parameters into account and shows the measurement efficiency of the item at different ability levels. Another advantage of the item response theory is the invariance of item parameters which pertains to the sample-free nature of its results. In the theory the item parameters are invariant when computed in groups of different abilities. This means that a uniform scale of measurement can be provided for use in different groups. It also means that groups as well as individuals can be tested with a different set of items, appropriate to their ability levels and their scores will be directly comparable (Anastasi & Urbina, 2002).

The present study demonstrates the difference between CTT and IRT approach based on estimates of item difficulty, internal consistency values, variation of ability, and measurement errors using a chemistry test for junior high school students.

### Method

#### Participants

The participants in the study are 219 junior high school students from a private school in the National Capital Region in the Philippines. These students were randomly selected from 8 sections to take two forms of the chemistry test. These junior students have completed their chemistry subject in the previous school year.

#### Instrument

A chemistry test was constructed by two science teachers who specialize in teaching chemistry with the help of their science coordinator. Two forms of the chemistry test were constructed following the same table of specifications. Each form was composed of 70 items. The test is in the form of a multiple choice for all 60 items for the two forms. The items in the chemistry test cover cognitive skills on understanding (20 items), applying (33 items), analyzing (16 items), and evaluating (1 item). The content areas includes are chemistry as a science (history, branches, scientific method, measurement), nature of matter (atomic models, states of matter, subatomic particles, classes of matter and separation techniques), trends, bonds, and changes (periodicity of elements, atomic trends, ionic, metallic, covalent bonds, chemical nomenclature, formula writing, intermolecular forces, balancing equations, types of chemical reactions/predicting, impact of chemical reactions), quantitative relationships in chemistry (empirical and molecular formulas, mole and mole ratio, percentage composition, percent yield, limiting and excess reactants), and nature of solutions (solubility, factors affecting solubility, acids, bases, and salts). The skills measured in the test were based on the following general objectives:

(1) Demonstrate understanding of the nature of chemistry, its historical development as a science, its requirements and tools in conducting scientific inquiry.

(2) Demonstrate understanding of how matter is classified; relate physical and chemical properties of elements to their atomic structure.

(3) Demonstrate recognition of patterns in periodic properties of elements through the use of the modern periodic table; relate the manner in which atoms combine to the physical and chemical properties of the substances they form and to the intermolecular forces that bind them; predict new substances formed from chemical changes.

(4) Demonstrate understanding of how the conservation of atoms in a chemical reaction leads to the conservation of matter and from this, calculate the masses of products and reactants.

(5) Demonstrate understanding of how characteristic properties of solutions are determined by the nature and size of dispersed particles and the changes in them.

(6) Demonstrate understanding of the nature and uses of acids and bases, their strength and effects on the environment.

The two forms, of the test were content validated in two stages. First, a testing consultant reviewed the objectives tested and the frame of items under each skill measured. In the second review, the items together with the table of specifications were shown to an expert in chemistry. The second review ensured whether the items are within the skills and content areas intended by the test. The items were revised based on the reviews provided.

#### Procedure

After the construction and review of the items, it was administered to 219 randomly selected junior high school students from 8 sections. During the test administration, the students were given one a half hour to complete the test. They were not allowed to use calculators and periodic tables to answer the test items. During the preliminary instructions, the students were requested to answer the test to the best of their ability. After the test, the examinees were debriefed about the purpose of the study.

#### Results

The results compares CTT and IRT approaches across two samples and two forms of the Chemistry test. Tests for difference of proportions, means, and correlation coefficients were used for the comparisons. CTT and IRT approaches across samples and forms were compared on difficulty estimates, internal consistencies, and measurement errors.

#### Comparison of Item Difficulty Estimates

To compare item difficulty estimates for two samples, the sample with  $N=219$  was split into two by equating their abilities based on the total scores of the chemistry test ( $N_1=110$ ,  $N_2=109$ ). The matching ensures that there is equality in terms of ability for both samples and this will not influence the results of item difficulty estimates. The total scores of two groups were tested and no significant difference was found on their chemistry scores for forms A and B (Form A:  $N_1$  Mean=25.22,  $N_2$  Mean=25.13, n. s.; Form B:  $N_1$  Mean=29.94,  $N_2$  Mean=31.00, n. s.).

Item difficulties were determined [ $di=(pH+pL)/2$ ] for  $N_1$  and  $N_2$  using both CTT and IRT. Items difficulty mismatch is when the item difficulty is not consistent for  $N_1$  and  $N_2$ , and item difficulty matching is when the item difficulty index is the same for  $N_1$  and  $N_2$ . The number of items that matched and did not match was expressed in percentage. These percentage of match and mismatch item difficulties were compared for Form A and Form B, and for CTT and IRT approach. Comparison of percentage of matching and mismatching determined across forms determines consistency of results across tests while comparison of matching and mismatching across approach (CTT and IRT) determines which approach is more consistent across samples.

The item difficulty index in the CTT between  $N_1$  and  $N_2$  were correlated to determine if the item difficulties are consistent across samples. The logit measures that indicates item difficulty in the IRT was also correlated between  $N_1$  and  $N_2$  for the same purpose. The same procedure is done for both Form A and Form B. These correlations were then compared (between Forms and between CTT and IRT) to determine which technique is more consistent for item difficulty estimates.

Table 1  
Difference of CTT and IRT on Item Difficulty for Two Samples

	CTT		Difference
	Form A N1 vs. N2	Form B N1 vs. N2	
Mismatch	17.14% (12 items)	12.86% (9 items)	$p=.51$
Match	82.86% (58 items)	87.14% (61 items)	$p=.75$
	$r=.82^*$	$r=.84^*$	$p=.78$
	IRT		
Mismatch	0% (0 items)	0% (0 items)	$p=1.00$
Match	100% (70 items)	100% (70 items)	$p=1.00$
	$r=.91^{**}$	$r=.92^{**}$	$p=.65$
Difference of r for CTT and IRT	$p=.03$	$p=.03$	
Mismatch Difference	$p=.00$	$p=.003$	
Matching Difference	$p=.00$	$p=.002$	

\*\* $p<.01$

When the item difficulties across samples were matched, there are significantly more items that mismatched in terms of their difficulty for the CTT approach,  $p=.00$  (for Form A 12 items were mismatched, for Form B 9 items were mismatched). All items were exactly matched for the IRT approach with no mismatch across the two samples (0 items mismatch for Forms A and B).

When the proportion of items for the mismatching and matching were compared, they were consistent across the two forms ( $p=n. s.$ ). However, the consistency of matching and mismatching are more stable across forms for the IRT approach with  $p=1.00$ .

Correlation of item difficulty using the CTT across the two samples are consistent for Form A ( $r=.82^*$ ) and Form B ( $r=.84^*$ ). These correlations were also consistent across the two forms of the test. However, more consistent results were obtained when item difficulty logit measures (IRT) were correlated across the two samples and even for both forms of the test ( $r=.91^{**}$  and  $r=.92^{**}$ ) as compared with the CTT approach.

#### Comparison of Internal Consistencies

The person and item reliabilities using the one-parameter Rasch model was used to estimate Form A and Form B versions of the chemistry tests. This procedure was done for  $N_1$  and  $N_2$ . For the CTT approach, the Cronbach's alpha and split half reliabilities were estimated for each form and each sample. The internal consistency estimates were compared across forms and across samples to determine if the coefficient values will be stable.

Two estimates of reliability were obtained in the one-parameter Rasch model because estimates for person and item measures are independent.

Table 2  
Difference of CTT and IRT on Internal Consistency Measures

	Form A			Form B		
	N <sub>1</sub>	N <sub>2</sub>	p	N <sub>1</sub>	N <sub>2</sub>	p
IRT						
Person	.66	.62	.62	.81	.77	.43
Reliability						
Item reliability	.90	.90	1.00	.93	.93	1.00
CTT						
Cronbach's alpha	.77	.63	.04	.81	.69	.04
Split half	.53*	.71*	.03	.67*	.50*	.04

All estimates of internal consistencies were adequate for both forms and both samples. The comparison of internal consistencies for the IRT approach remained stable across the two samples for both forms A and B of the test. This is especially true for estimates of item reliability where coefficients were exactly the same. This occurred for both forms A and B of the chemistry test. However, in the CTT approach both the Cronbach's alpha and split-half did not remain stable across two samples. This instability was consistent for both forms A and B of the test.

#### Comparison of Measurement Errors

Measurement errors were estimated using both IRT and CTT approach. For the IRT approach (one-parameter Rasch model), both standard errors for person and item measures were obtained given that their estimates are independent. These two standard errors were averaged in order to be compared with the standard errors of the mean for the CTT version. Standard errors for the two samples were compared to determine if they will remain stable. This comparison was done for both forms of the test.

Table 3  
Difference of CTT and IRT on Standard Error Estimates

	Form A			Form B		
	N1	N2	p	N1	N2	p
IRT						
Person SE	.04	.08	.76	.06	.05	.94
Item SE	.08	.04	.76	.10	.10	1.00
Ave. Person and Item SE	.06	.06	1.00	.08	.08	1.00
CTT						
SE of the M	.64	.60	.76	.83	.78	.71
Confidence Interval 95%	23.96-26.94	23.92-26.32		28.29-31.5	29.45-32.54	
Difference of CTT and IRT SE	p=.00	p=.00		p=.00	p=.00	

All measure of standard errors across the two samples remained to be stable. This is true for both CTT and IRT approaches. However, standard errors for the IRT are more stable across samples with a minimum SE difference of p=.76 and maximum of p=1.00 (SE difference for CTT is International Journal of Educational and Psychological Assessment 2009; Vol. 1(1)

$p=.71$ ). When the SE's were compared for the CTT and IRT, the SE's for the CTT were significantly higher than SE's for the IRT,  $p<.001$ .

### Discussion

The present study compared the difference between CTT and IRT approach across samples and test forms in chemistry. The difference is demonstrated on estimates of item difficulty, internal consistencies, and standard errors. It was found in the study that (1) IRT estimates of item difficulty do not change across samples as compared with CTT with inconsistencies; (2) difficulty indices were also more stable across forms of tests than the CTT approach; (3) IRT internal consistencies are very stable across samples while CTT internal consistencies failed to be stable across samples; (4) IRT had significantly less measurement errors than the CTT approach. These findings further support the marked difference between the CTT and IRT approaches pertaining to sampling and tests. Aside from demonstrating differences between IRT and CTT, the findings are helpful for measurement experts to decide on what approach to use in analyzing test data.

It was shown in the study that estimates of item difficulty in the IRT did not change across two samples. In the CTT approach there were some items that failed to have the same difficulty index across the two samples. These findings demonstrate that it is possible to maintain constant item difficulties across similar samples using the IRT approach. The same can also be assumed with the CTT given the high correlations of item difficulty index across the two samples (.82 and .84) but more consistent findings were obtained for the IRT. Some changes in the item difficulty index in the CTT approach is influenced by proportions included in the analysis (27%). Getting both extreme ends of a sample is relatively unstable causing inconsistencies in estimates of item difficulty. It can be noted that those who topped and got bottom ranks in the form A is not the same the ones in form B. This technique which causes changes in the sample involved in the analysis made the difference. In this case, relying on difficulty index using the CTT approach is problematic when test developers wanted to establish an item's identification when used for adaptive testing because the estimate changes depending on the sample. For the IRT, the entire sample is included in the analysis to estimate item difficulty. This is obtained by transforming the proportion of those who got the item correct into logarithm values. The log values estimates items within positive and negative integers within 50% chance of getting an answer correct which arrives with difficulty estimates relatively accurate. It was not only that IRT logit measures are stable across sample, it was also demonstrated that it can be stable across parallel forms of the test. Tests measuring the same construct, skills, and scope can be expected to have consistent item difficulties using the IRT approach.

The problem of coefficient measures using the CTT was demonstrated in the findings. Estimates of Cronbach's alpha and split-half reliability did not remain the same across the two samples. This is problematic in the case of researchers using an instrument from a past research and claiming its internal consistency which is actually consistency for the sample of the past research sample. This suggests the necessity to estimate internal consistencies for every study using the sample obtained. It is difficult to rely on internal consistencies reported by previous researchers because these estimates are sample-dependent. On the other hand, estimates of reliability in the IRT can be more consistent than CTT approaches. This is especially true for item reliability measures. Using IRT estimates of person and item reliability can be more useful for researchers when reporting internal consistencies of tests because they are more stable and not sample-dependent. Majority of researchers are accustomed in relying at CTT approaches such as the Cronbach's alpha, item-total correlation etc because of their availability in statistical packages. There should be an increased demonstration how estimates of reliability using the IRT approach can be more advantageous in research articles. It is also recommended that statistical packages provide alternative estimates of internal consistencies such as the IRT to users.

Estimates of standard errors are remarkably larger in the CTT as compared to the IRT approach. Standard error estimates are conceptually considered as chance factors that confound test results. One of the goals of a test developer is to control measurement errors to a minimum. One of the ways to handle measurement errors is to have an independent calibration of person and items so that one does not influence the other. This independent calibration is made possible in the IRT. The independent calibration makes the items not influenced so much by person differential characteristics making standard error estimates at a minimum value. It should be importantly acknowledged that large standard errors cause invalidity of the test. This implies that test developers need to carefully select techniques that will control standard errors. It was also found in the study that standard errors can be present and remain stable across different tests and samples. They only remain different by changing the approach used in analyzing test data. This indicates that standard errors are present across samples and test forms and one way to minimize this is the independent calibration of item and persons taking the test.

The findings of the study provide perspectives for test developers, researchers, statisticians, psychometricians, statistical software developers, and test users. First is the use of better approaches of estimating item difficulties, internal consistencies, and standard errors that will result to consistent results. On this account, stakeholders in testing and measurement should be made aware of the advantages of using IRT approaches as compared to CTT. These advantages solve problems on repeated analysis of data sets every time a test is administered due to consistent efforts of establishing better findings for a test to be useful. Second is the reliance of findings on more stable estimates of test and scale reliabilities and item difficulties in publications. Researchers publishing in journal articles using CTT should not only rely on previous reliability estimates but to estimate their own and report noted differences. A better approach is the reliance of findings on solid approaches like using IRT estimates of person and item reliability. Third is the need to make available ways to use IRT approaches that are accessible. In order to accomplish the first two perspectives provided, IRT software packages should be made available to users easily. Available software packages are still difficult to use and it should be made more user friendly. Experts should start sharing free softwares that can be readily used by test specialists. In order to achieve consistency in theories, access to such tools as IRT should be made easy.

#### References

- Anastasi, A. & Urbina, S. (2002). *Psychological testing*. Prentice Hall: New York.
- Andrich, D. (1998). *Rasch models for measurement*. Sage University: Sage Publications.
- Dobby J, & Duckworth, D (1979): Objective assessment by means of item banking. *Schools Council Examination Bulletin*, 40, 1-10.
- Fischer, G. H. (1974) Derivations of the Rasch Model. In Fischer, G. H. & Molenaar, I. W. (Eds) *Rasch Models: foundations, recent developments and applications*, pp. 15-38 New York: Springer Verlag.
- Hambelton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38, 60-65.
- Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69-81.

- Kaplan, R. M. & Saccuzo, D. P. (1997). *Psychological testing: Principles, applications and issues*. Pacific Grove: Brooks Cole Pub. Company.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, 26, 307-331.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of mathematical psychology*, 3, 1 – 18.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In G. M. Copenhagen (ed.). *The Danish yearbook of philosophy* (pp.58-94). Munksgaard.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72 – 101.
- Van der Linden, A., & Humbleton, R. (1980). *Introduction to scaling*. New York: Wiley.
- Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement*, 29, 23-48.
- Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

#### Author Notes

Special thanks to Dr. Ma. Alicia Bustos-Orosa and Ms. Ma. Socorro Diesta for allowing me to use the achievement test results of De La Salle Santiago Zobel School.

Further correspondence can be addressed to the author at the Counseling and Educational Psychology Department, De La Salle University-Manila, 2401 Taft Ave. Manila. E-mail: carlo.magno@dlsu.edu.ph

## Comparison of the Item Discrimination and Item Difficulty of the Quick-Mental Aptitude Test using CTT and IRT Methods

Royce Hernandez  
 De La Salle-College of Saint Benilde

### Abstract

The purpose of this research is to compare the item difficulty and item discrimination of the Quick-Mental Aptitude Test (Q-MAT) using Classical Test Theory and Item Response Theory (IRT) methods across 1, 2, and 3 parameters. The developed instrument was administered to a college sample of N=229. The data gathered was analyzed for possible relationship of the item characteristics using CTT and IRT methods. Results indicate that the 2 parameter IRT model closely resembles CTT of the verbal and non-verbal test in terms of item discrimination ( $R^2_{\text{verbal}}=.891$ ,  $p<.01$ ;  $R^2_{\text{nonverbal}}=.945$ ,  $p<.01$ ) and item difficulty ( $R^2_{\text{verbal}}=.896$ ,  $p<.01$ ;  $R^2_{\text{nonverbal}}=.984$ ,  $p<.01$ ).

With the emerging trend of developing local instruments more and more research has been developed in producing psychological tests. Most often, these researchers rely on Classical Test Theory (CTT) to develop these instruments in spite of the strong presence of Item Response Theory in the recent decades.

As part of the test development process, analysis of the items is a crucial part. Two prevailing methods, both with strengths and weaknesses, are predominantly used. In the Classical Test Theory, its ease of use and adaptability in analyzing practically all kinds of tests renders it a popular choice. However, its strong dependence on the kind of sampling required often limits its applicability. Hence, CTT developed tests would see the need for bigger sampling every now and then which in the long run renders it expensive. On the other hand, the emerging Item Response Theory (IRT) seems to have found a way to avoid the pitfalls of CTT. It is said to be sample free or sample independent. The only drawback is the cumbersome statistical analysis required which other test developers would shy away from. Nevertheless, IRT is slowly gaining momentum in the field of psychology (Andrade, Tavares & Valle, 2000).

In CTT, test scores are said to be composed of three components: test score, true score, and error score. The invariance is brought about by differences contributed by the sample from which the scores were derived. Again, here lies the dependence of CTT on the sample the scores were taken from. However, IRT addresses this by disregarding the sample and instead looking at the characteristics of the item or item parameters. By focusing on the items, the issue of sampling becomes negligible. One can now generalize better item-generated scores across samples and person abilities (Hambleton, Swaminathan, & Rogers, 1991).

Studies linking CTT and IRT item characteristics have been done and have shown signs of positive indications of a relationship that exists (Adedoyin, Nenty, and Chilisa, 2008; Nukhet, 2002; Fan, 1998). However, local literature has yet to replicate the studies and results. It is then the goal of this paper is to analyze the item characteristics of a newly developed test using both CTT and IRT methods and to check if both methods are comparable and can used independently or interchangeably.

## Method

### Participants

A total of 400 college students in Metro Manila were targeted as participants in this study. With a return rate of 74.25%, 297 were able to answer the instrument. Of this number, only 229 respondents were included. The rest of the responses were discarded from the analysis because the respondents failed to answer all of the items. The final sample consists of 76% females and 24% males. Age ranges from 16 to 26 years old (average age = 18.76 and SD = 1.23).

### Materials

The Quick-Mental Aptitude Test (Q-MAT) was developed as part of this study. The 40-item instrument consists of two parts – Verbal and Non-Verbal tests. Psychometric properties of the test reveal some items needing revision. Nonetheless, reliability is reported KR-20 indices to be  $r_{\text{verbal}}=.39$ ,  $r_{\text{nonverbal}}=.69$ , and  $r_{\text{total}}=.71$ . Spearman-Brown Correction on split-half reliabilities for odd-even comparison also show similar results  $r_{\text{verbal}}=.57$ ,  $r_{\text{nonverbal}}=.79$ , and  $r_{\text{total}}=.77$ . Validity of the instrument was shown using inter-correlation of the sub scales (-.055 to .855). Confirmatory Factor Analysis reveals that the data obtain fits the model. However, some items do not significantly contribute to each test part necessitating revision.

### Procedure

Permission was sought from professors coming from 3 Private-Catholic institutions. An easy to follow test administration guide was prepared to aid the examiner/proctor (refer to appendices). Packets of scannable answer sheets and re-usable test booklets were also given to the professors. The instrument was answered in 15 minutes. The instruction specifically states that the respondent should do all computations and analysis mentally (without the aid of external mechanics such as calculators, rulers, and scratch papers). Data gathered were then analyzed using SPSS version 15, Winsteps (Linacre, 2007), Item and Test Analysis Package (ITAP) (Assessment Systems Corporation, 2007) software, and Microsoft Excel version 2002.

### Data Analysis

Classical Test Theory analysis was done using the ITAP software's ITEMAN program module. The software automatically generated the following: item difficulty (diff), item discrimination (disc), and point biserial correlation ( $r_{pb}$ ) to also denote item discrimination. To prepare the data for correlation with the IRT parameters, diff and  $r_{pb}$  had to be transformed into a Z (normal) distribution,  $\Delta$  and Z respectively (Fan, 1998; Anastasi, 1988; Holland and Thayer, 1985).

IRT parameters were obtained using the ITAP software's RASCAL and XCALIBRE program modules. RASCAL (Rasch Item Calibration) program provided the item difficulty parameter. On the other hand, XCALIBRE (Marginal Maximum-Likelihood Estimation) program generated the item difficulty (b - parameter) and item discrimination (a - parameter) for both 2 and 3 parameter logistic.

Pearson product moment correlation was then used to determine the relationship between the variables being studied. CTT diff was correlated with the b parameters of IRT (1-pl, 2-pl, and 3-pl). CTT  $r_{pb}$  (used to denote disc) was correlated with the a parameters of IRT (2-pl and 3-pl). It should be noted that in 1-pl (Rasch), discrimination is set to a fixed value; hence it is not included in the analysis. Item difficulty and item discrimination indices were then graphed versus their IRT parameter counterparts using MS Excel. The coefficient of determination was obtained by squaring the value of the r obtained.

## Results

Table 1 shows the mean and standard deviation values of the Verbal and Non-Verbal test when classified into CTT and IRT. Comparison of CTT Diff and Disc scores show that the item difficulty index of both test are of average difficulty with the non-verbal test slightly higher or easy than the verbal test and the non-verbal test item difficulty indices as more dispersed. The CTT item discrimination values for both test indicates their reasonable discrimination between high and low scorers. The Non-verbal test also shows better discrimination compared to the Verbal test.

The IRT Difficulty parameters for the 1-parameter logistic or Rasch generally have the lowest values (Mean and SD) for both Verbal and Non-Verbal Test. This indicates that the Rasch provides the lowest possible item difficulty index. Conversely, the 3-pl has the highest values. On the other hand, item discrimination as measured in IRT reveal that the 2-pl provides the lowest parameter values.

Table 1  
Mean and Standard Deviation

		CTT		IRT Difficulty			IRT Discrimination	
		Diff	Disc	1-pl	2-pl	3-pl	2-pl	3-pl
Verbal	Mean	.468	.242	0.00008	.380	1.354	0.422	0.748
	SD	.226	.148	1.092	1.696	1.552	0.081	0.079
Non-Verbal	Mean	.524	.213	0.118	-0.054	0.563	0.699	0.895
	SD	.269	.165	.866	0.953	1.017	0.137	0.093

Table 2 reveals that generally, there is a significant and high correlation that exists between CTT and IRT in terms of item difficulty (diff) and item discrimination (disc). However, there is no significant correlation between Disc and both of the discrimination as measured by the 3-pl models in Verbal and Non-Verbal test.

Table 2  
Correlations of Difficulty and Discrimination on Logistic Parameters (N=229)

	Number of Items	Verbal Test			Non-Verbal Test		
		1-pl	2-pl	3-pl	1-pl	2-pl	3-pl
Diff	24	.857**	.896**	.902**	.820**	.984**	.974**
Disc	16	NA	.891**	-.197	NA	.945**	.373

\*\*  $p < .01$

A look at figure 1 shows that there exists a variation in the coefficient of determination values of the three IRT models when graphed versus the CTT item difficulty. Results suggests an increasing value (slope) across the three IRT models with the 3-pl having the largest  $R^2$  value of 0.81. This graph suggests that there exists a positive relationship between CTT and IRT item difficulty of the items found in the Verbal test.

Figure 1  
Scatter plot of Verbal test Item Difficulty (CTT vs IRT) showing Coefficient of Determination Trend line

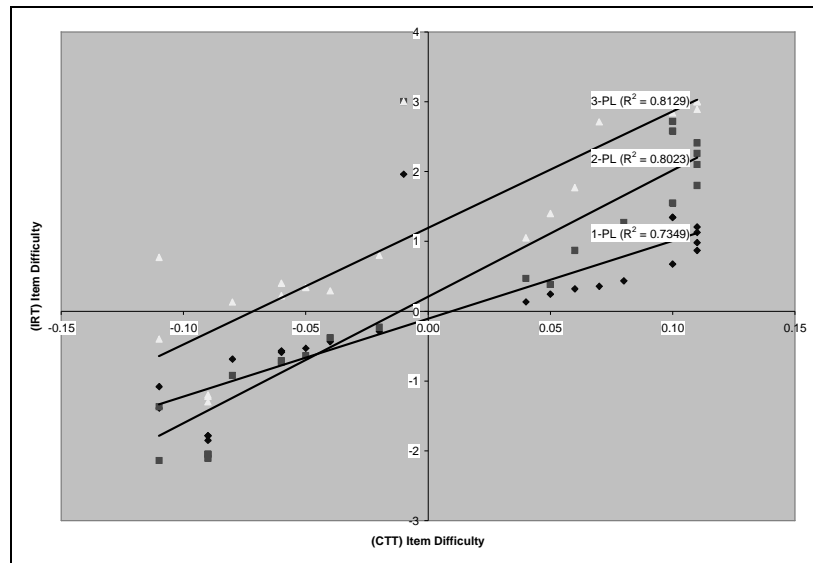


Figure 2 almost reveals a similar pattern with figure 1. The only difference lies in the 2-pl which has the largest R<sup>2</sup> value of 0.96 compared to the R<sup>2</sup> value of 3-pl which is 0.95. Nonetheless, this graph also suggests that there exists a positive relationship between CTT and IRT item difficulty of the items found in the Non-Verbal test.

Figure 2  
Scatter plot of Non-Verbal test Item Difficulty (CTT vs IRT) showing Coefficient of Determination Trend line

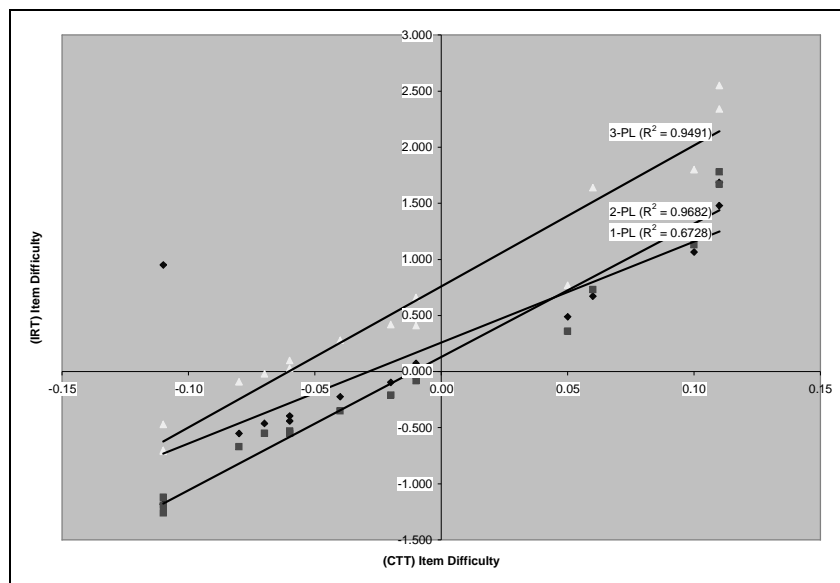


Figure 3 and 4 shows that as far as item discrimination is concerned, there exists a positive relationship between CTT and IRT in both the Verbal and Non-Verbal test items. However, there is very low variance observed in the 3-pl of both tests compared with the 2-pl. This indicates that the 2-pl is closely resembles the item discrimination as measured by the CTT compared to the 3-pl wherein guessing is also considered in the estimation parameters.

Figure 3  
Scatter plot of Verbal test Item Discrimination (CTT vs IRT) showing Coefficient of Determination Trend line

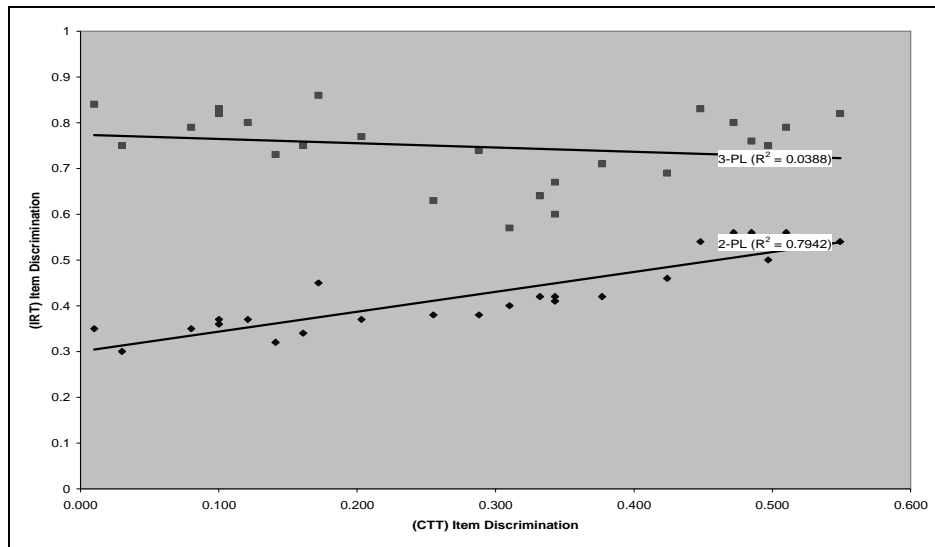
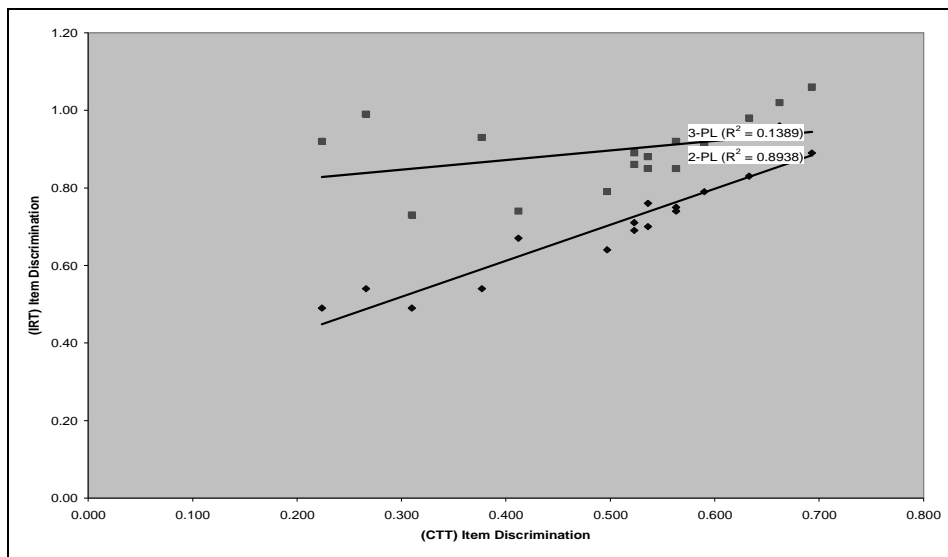


Figure 4  
Scatter plot of Non-Verbal test Item Discrimination (CTT vs IRT) showing Coefficient of Determination Trend line



## Discussion

Based on the results, it is evident that there is a relationship between the CTT and IRT approaches in analyzing the item characteristics of the Q-MAT with the Non-Verbal Test showing higher correlation values than the Verbal Test. This observed difference is an indication of better items found in the Non-Verbal test as reported in the item analysis and content validity of the instrument.

Results further reveal that when items are categorized from easy to hard item difficulty (in CTT), it would also correspond to almost the same IRT classification of item difficulty. The same can be said for item discrimination categorization between CTT and IRT. The chi-square statistic could have been used to establish such relationship but given the small category samples, the Pearson product-moment correlation was used instead.

A closer look into CTT versus IRT as represented by 1-, 2-, and 3-parameters reveal that the 2 parameter logistic (2-pl) shows higher significant relationship to CTT in both item discrimination and item difficulty. It can be noted that the 1-pl or the Rasch model lacks the capacity to distinguish item discrimination since it is held constant. As for the 3-pl wherein guessing is considered, the presence of such parameter significantly reduces the variance that relates both CTT and IRT. Hambleton and colleagues (1992) points out this commentary in the Rasch Transactions on the presence of a pseudo-guessing parameter:

The inclination to guess is an idiosyncratic characteristic of particular low ability examinees. Lucky guessing is a random event. Neither feature contributes to valid measurement of a latent trait. Parameterizing guessing penalizes the low performer with advanced special knowledge and also the non-guesser. Rasch flags lucky guesses as unexpected responses. They can either be left intact which inflates the ability estimates of the guessers, or removed which provides a better estimate of the guessers' abilities on the intended latent trait. In practice, 3-P guessing parameter estimation is so awkward that values are either pre-set or pre-constrained to a narrow range.

As such, this account for the probable better-fit or correlation of the CTT with the 2-pl IRT. After all, guessing is not directly measured or accounted for in CTT whereas in the 3-pl it forms part of the formula in obtaining the difficulty and discrimination parameter; while the guessing parameter, together with item discrimination is does not form part of the Rasch formula. Moreover, although the 1-pl is the simplest IRT method, studies have shown that items do have variations across item discrimination. Thus, this favors the use of a 2- or 3-parameter IRT model (Adedoyin, Nenty, & Chilisa, 2008; Nukhet, 2002; Fan, 1998). Hence, the almost similarity between the derivations of CTT and 2-pl IRT in item difficulty and item discrimination.

The foregoing results resemble that of previous studies (Adedoyin, Nenty, & Chilisa, 2008; Nukhet, 2002; Fan, 1998). However, the difference lies in the choice of a 2-pl or 3-pl. Nukhet (2002) reports 3-pl as having the most comparable indices with CTT. Whereas Fan (1998) indicates that all three are comparable with CTT. Perhaps, similar results would have been obtained had the sample used been large enough to prompt multiple and randomized sample selection.

The results further reflect the need to further improve the items found in the verbal component. This indicates that the items of the non-verbal portion of the Q-MAT is more stable than the verbal test as far as item discrimination and item difficulty indices are concerned in both CTT and IRT methods.

In addition, the paper was able to establish that CTT and IRT can be used independently or altogether to describe the nature of the items. Test developers can bank on time-tested CTT methods to establish item difficulty and item discrimination characteristic of items. In the absence of sophisticated software and a big sample to derive IRT parameters, the test developer can be

theoretically assured of congruence of test item difficulty and discrimination for both methods provided enough sampling is done in CTT (between 200 to 500; CTT requires 200 minimum while IRT is 500 to 1000 N)). On the other hand, those employing IRT, whether 2- or 3-parameter, would also be able to infer congruence of CTT and IRT item characteristics, provided goodness of fit of the data is established. What's important is the emphasis or need for the pseudo-guessing parameter if it is needed in the analysis.

And for those able to do both methods can empirically say that using both methods can in fact address the issue of sampling dependence in CTT and the complications of IRT in order to provide two ways of seeing item characteristics and in improving items whether it is going to be sample free (CTT) or an objective measure of items (IRT). Likewise, using both methods will greatly improve characterization of items, item selection, and in turn lead to improved measures which are the aim of test developers.

### References

- Adedoyin, O. 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.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.
- Andrade, D. F., Tavares, H. R., & Valle, R. C. (2000). *Teoria da Resposta ao Item: conceitos e aplicações*. São Paulo: ABE.
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.
- DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park, CA: Sage.
- Duncan, P. W., Bode, R. K. Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The Stroke Impact Scale. *Archives of Physical Medicine and Rehabilitation*, 84, 950-963.
- Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58, 357-385.
- Hambleton, R. K. & et. Al. (1992). *Rasch Measurement Transactions*, 1992, 6:2 p. 215-7. Website URL: <http://www.rasch.org/rmt/rmt62d.htm> accessed April 20, 2009.
- Hambleton, R. K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Holland, P. W. & Thayer, D. T. (1985). An alternative definition of the ETS delta scale of item difficulty. Educational testing Service, Technical report (85-64)/ Research Report (85-43).
- Nukhet, C. (2002) A Study of Raven Standard Progressive Matrices test's item measures under classic and item response models: An empirical comparison. Ankara University, Journal of Faculty of Educational Science, 35 (1-2), 71-79.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

## Evaluation of Mathematics Achievement Test: A Comparison between CTT and IRT

Romel A. Morales

University of Eastern Philippines, Northern Samar

### Abstract

When evaluating the quality of an educational assessment, reliability, validity and item bias are critical to the process. This study applied the Classical Test Theory and Item Response Theory to evaluate the quality of an assessment constructed to measure college students' achievement in mathematics. The sample for this study consisted of the junior and senior Mathematics and English major teacher-education student from the University of Eastern Philippines in Northern Samar. A sample of 80 students was drawn for this study. The Mathematics Achievement Test for college students developed by the author was used. Data were analyzed in two dimensions: First, psychometric properties were analyzed using CTT and IRT; Second, detection of item biased was performed using a method for Differential Item Functioning (DIF).

A mathematics achievement test was developed in an effort to overcome the dismal performance of the teacher-education graduates in the mathematics portion of the Licensure Examination for Teachers (LET). This assessment is in line with the objectives of the mathematics curriculum of the college and in consonance with the mathematics ability required in the LET. It is hoped that the performance of our students in this achievement test will predict their performance in the mathematics portion of the LET.

Classical test theory and item response theory are commonly perceived as representing two very different measurement frameworks. Although CTT has been used for most of the time by the measurement community, in recent decades IRT has been gaining ground and is becoming to be the favorite measurement framework. The major advantage of CTT is its rather weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). Its major limitations are: The person statistic is item dependent and the item statistics such as item difficulty and item discrimination are sample dependent. On the other hand, IRT is more theory grounded and models the distribution of examinees' success at the item level. As its name implies, IRT mainly focuses on the item-level information in contrast to CTT's principal focus on test-level information. The IRT framework includes a group of models, and the applicability of each model in a particular situation depends on the nature of the test items and the practicality of different theoretical assumptions about the test items.

Measurement is central to the construction of a quality student assessment, even in the case of classroom-designed or non-standardized assessments. Measuring variables is one of the necessary steps in the research process. What follows are the statistical tools to analyze the data. Thus, the interpretation of data analyses can only be as good as the quality of measures (Bond & Fox, 2001). Although many testing and measurement textbooks present classical test theory as the only way to determine the quality of an assessment, the IRT offers a sound alternative to the classical approach. Because CTT is rooted in a process of dependability rather than measurement, it does not rely on item difficulty variable for precision and calibration nor on total score for indicating the measured ability (Sirotnic, 1987). Thus, the weaknesses of CTT have caused IRT to gain the attention of researchers because it makes allowances where CTT does not (De Ayala, 1993; Welch & Hoover, 1993).

The IRT is based on two basic assumptions. First, a more able person should have greater probability of success on assessment items than a less able person. Second, any person

should always be more likely to do better on an easier item than on a more difficult one. IRT assumes item difficulty is the characteristics influencing person responses and person ability is the characteristics influencing item difficulty estimates (Linacre, 1999). Thus, careful considerations should be given to the construction of assessments. Items should be written clearly and concisely such that they are not vulnerable to guessing.

In evaluating the quality of an assessment tool, a discussion of reliability and validity is essential. The reliability is the degree to which an instrument consistently measures the ability of an individual or group. Validity is the degree to which an instrument measures what it is intended to measure. The CTT gives a very simple way of determining the validity and reliability of a test. The classical item analysis provides us a way of doing this. By subjecting the whole test results to simple statistical tests, one can determine the validity and reliability of the test. On the other hand, IRT offers a more complex but more reliable way of determining validity and reliability of test. If the focus of CTT is on the test as a whole, IRT focuses on each item and each individual test taker.

Latent trait models in test construction are utilized for purposes of constructing equivalent test forms, developing tests that discriminate between ability levels, and improving customized test system. IRT can also be used to investigate item bias. A set of items is considered unbiased if all subpopulations are equally affected by the same sources of variance, thus producing similar ICCs for both groups (Cole & Moss, 1985). If a test item has different connotative meanings for different groups, then examinees' performance on that item may be subject to sources of variation that are unrelated to ability level. This refers to differential item function and can cause item bias (Crocker & Algina, 1986). Also, a set of items is considered unbiased if a source of irrelevant variance does not give an unfair advantage to one group over another (Scheuneman, 1979).

Unfortunately, the investigation of item bias is not that clear cut. IRT, as well as chi-square and item difficulty, can flag items as biased even if they are not (Park & Lautenschlager, 1990). Also, multidimensionality can be mistaken for item bias with IRT as a result of differences among ICCs. ICC differences can occur even when item bias does not exist. This distinction can indicate that items are not unidimensional.

DIF detection procedures can investigate the effects achievement tests have on different subpopulations (Zwick, Thayer, & Mazzeo, 1997). Some research has evaluated DIF analysis methods that involve matching examinees' test scores from two groups and then comparing the item's performance differences for the matched members (Zwick et al., 1997; Ackerman & Evans, 1994). Such nonparametric detection methods include the Mantel-Haenzsel procedure and Shealy and Stout's simultaneous item bias (SIBTEST) procedure. These procedures, however, lack the power to detect nonuniform DIF which may be even more important when dealing with polytomous items due to the multiple ways in which item scores can interact with the total score (Spray & Miller, 1994). There is also the newer procedure of detecting item bias, the Item Response Theory Likelihood-Ratio Test for Differential Item Function (IRTLRDIF). Of all of the procedures available for DIF detection and measurement, IRT-LR procedure posits several advantages over its rivals. IRT-LR procedures involve direct tests of hypotheses about parameters of item response models, they may detect DIF that arises from differential difficulty, differential relations with the construct being measured, or even differential guessing rates (Thissen, 2001). This is the reason why the author used this method in the detection of item bias.

The main objectives of the present study was to analyze the psychometric properties of the instrument administered on two different groups of students, i.e. established the validity and reliability of the instrument using CTT and IRT framework, and determined the Differential Item Functioning (DIF) for each item. The test that measured achievement in college mathematics is criterion-referenced so that test scores directly conveyed level of competence in defined mathematics domain.

## Method

### Participants

A total of 80 students (34 mathematics majors, 46 English majors) completed the mathematics achievement test during the ending period of the 2nd semester, school year 2008-2009.

### Measure

The mathematics achievement test, a multiple-choice assessment designed to measure college students' mathematics ability was administered. The author constructed a mathematics achievement test comprised of 40 multiple choice items with five answer choices. The achievement test was piloted with two groups of junior and senior teacher-education students. Mathematics majors comprised the first group while the second group was all English majors. The items on the achievement test were categorized into five content domains: Patterns and relations, equations and distances, geometric and trigonometric, shapes, areas and volumes and combinatory and probability. For all domains, the underlying construct of teacher-education mathematics remains the same; thus, the theoretical framework of unidimensionality is upheld. The test was content validated by a mathematics professor in the college of education in the same university. Suggestions were taken and the test was revised accordingly.

### Procedure

Cooperating teachers administered the test for the senior students while teachers of the junior students were the ones who conducted the test for the junior level. The students were given the test after receiving specific instruction for the test. The test was administered simultaneously for the two groups of students. The students completed the test for two hours under the lookout of their teachers. The purpose of the teacher-proctors monitoring of the test was to minimize measurement errors that could arise during the actual test.

### Data Analysis

Two sections of analysis were done to establish psychometric properties. First is using the classical test theory steps which include the item analysis. Microsoft Excel was used for the analyses and computations involved in the CTT analysis. SPSS software was also used to determine reliability of the test. Second, item response theory method was employed to calibrate for item and person difficulties. WINSTEPS' Bigsteps software was used for this analysis. To detect for item biased with regards to different groups of students, DIF test was conducted using software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio Tests. The software was downloaded from the website of L.L. Thurstone Psychometric Laboratory based at University of South Carolina, Chapel Hill.

## Results

This section is divided into three parts. First is the presentation of the psychometric properties of the mathematics achievement test. The validity and reliability analyses presented here were done following both Classical Test Theory (CTT) and Item Response Theory (IRT). The statistical package for Social Sciences (SPSS 15) was used to perform the analyses according to CTT. Second, the presentation of IRT analyses, were the software WINSTEPS' BIGSTEPS was utilized to estimate students' abilities and item difficulty for the test as well as the goodness of fit of

the items. The third and last part is the presentation of the Differential Item Functioning as a result of the IRTLRFID analysis.

### Reliability

The internal consistency of the test was found to be high with a Cronbach's alpha value of .77. This value indicates a good reliability for the achievement test. Aside from internal consistency, Split-half method was also performed resulting to a Guttman coefficient of .72, a value that indicates internal consistencies of the responses in the test. Finally, Kuder-Richardson, KR20 was also used to determine internal consistency with a value of .90.

### Item Difficulty and Discrimination

Each item's difficulty and discrimination index were determined using the classical test theory. It shows that 27 (73%) of the items are average items. The remaining 27% belong to difficult and easy items. It could be implied from this result that the achievement test was fairly difficult because more than half of the students got the most of the items correctly. But, considering that the examinees were mathematics and English majors, the result could also mean that they really have the ability to answer even difficult items. English and Math majors have rigid qualifying test to proceed with their field of specialization. Thus, to be able to major in Math or English, the students must have attained an above average score in the university entrance examination test.

Of the 37 items considered in the test, only 3 or 8% come up to be poor items. These items were be rejected. Only two items (marginal items) need to be improved. Thirty or 81% of the items were either good or very good items. This only means that generally, the items for the achievement test truly represent the learning ability of the test takers. Most of the items can discriminate well the high and low performing groups.

### One Parameter-Rasch Model

The Rasch model was applied to the responses of 80 students to the achievement test in its original form of forty multiple-choice items. First, the item and person separation and reliability were examined prior to any interpretations of the data. The person separation and reliability values for the pilot data were 1.84 and 0.77 respectively. This person separation indicates the number of groups the students can be separated into according to their abilities. So, in this case there are approximately two different levels of performance in the sample. Likewise, the item separation and reliability for the pilot data was 4.4 and 0.95, respectively (Table 1). Considering the moderate sample size, person and item reliabilities are acceptable for the analysis to continue.

Table 1  
Summary of Measure Persons

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	18.6	37.0	.01	.40	1.00	.0	.97	-.2
S.D.	5.4	.0	.83	.02	.20	1.1	.33	1.0
MAX.	30.0	37.0	1.88	.47	1.49	2.5	2.11	2.5
MIN.	7.0	37.0	-1.88	.38	.59	-2.5	.41	-2.2
REAL RMSE	.41	ADJ.SD	.72	SEPARATION	1.75	PERSON RELIABILITY	.75	
MODEL RMSE	.40	ADJ.SD	.73	SEPARATION	1.84	PERSON RELIABILITY	.77	
S.E. OF PERSON MEAN	.09							
SUMMARY OF 37 MEASURED (NON-EXTREME) ITEMS								
	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	40.1	80.0	.00	.28	1.00	.1	.97	-.1
S.D.	18.4	.0	1.26	.04	.07	.7	.15	.8
MAX.	72.0	80.0	2.82	.44	1.16	1.9	1.50	1.9
MIN.	6.0	80.0	-2.43	.24	.84	-1.3	.74	-1.6
REAL RMSE	.28	ADJ.SD	1.23	SEPARATION	4.33	ITEM RELIABILITY	.95	
MODEL RMSE	.28	ADJ.SD	1.23	SEPARATION	4.40	ITEM RELIABILITY	.95	
S.E. OF ITEM MEAN	.21							
WITH 3 EXTREME ITEMS	=		40 ITEMS	MEAN	.41	S.D.	1.88	
REAL RMSE	.47	ADJ.SD	1.82	SEPARATION	3.82	ITEM RELIABILITY	.94	
MODEL RMSE	.47	ADJ.SD	1.82	SEPARATION	3.84	ITEM RELIABILITY	.94	
MAXIMUM EXTREME SCORE: 3 ITEMS								

All items fit the expectations of the Rasch model. In other words, all items had ZSTD infit and/or outfit statistics between -2 and 2 (Table 2).

The item map [on which stems are indicated on the left side and students are indicated by their number] was examined for gaps where a number of students were located along the continuum without items targeted at that ability level (see Figure 1 for circles indicating gaps). Inserting items reflecting corresponding levels of difficulty provides more accurate measures of student abilities at these levels. Notice there are gaps between item 27 and item 23 with five students falling in this ability range. Similarly, 12 students fall in the gap between items 31 and 38, and so on. Addition of items at these difficulty levels will provide more precise measures for students at these ability levels.

Table 2  
Item Statistics Misfits Order

ENTRY NUMBR	RAW SCORE	COUNT	MEASURE	ERROR	INFIT		OUTFIT		PTBIS CORR.	ITEMS
					MNSQ	ZSTD	MNSQ	ZSTD		
7	72	80	-2.43	.38	1.10	.3	1.50	1.0	A-.01	Item 7 ; Item 7 : 7-7
5	45	80	-.29	.24	1.16	1.9	1.22	1.9	B .08	Item 5 ; Item 5 : 5-5
26	11	80	2.09	.34	1.12	.5	1.19	.5	C .08	Item 26 ; Item 26 : 26-26
10	39	80	.05	.24	1.09	1.1	1.14	1.4	D .17	Item 10 ; Item 10 : 10-10
35	36	80	.23	.24	1.09	1.0	1.14	1.3	E .18	Item 35 ; Item 35 : 35-35
16	67	80	-1.84	.31	1.05	.3	1.13	.4	F .11	Item 16 ; Item 16 : 16-16
8	20	80	1.26	.28	1.10	.7	.98	-.1	G .21	Item 8 ; Item 8 : 8-8
11	39	80	.05	.24	1.06	.7	1.09	.9	H .21	Item 11 ; Item 11 : 11-11
25	52	80	-.71	.25	1.06	.7	1.05	.4	I .18	Item 25 ; Item 25 : 25-25
32	44	80	-.23	.24	1.06	.8	1.03	.2	J .21	Item 32 ; Item 32 : 32-32
12	31	80	.53	.25	1.06	.6	1.05	.4	K .22	Item 12 ; Item 12 : 12-12
30	54	80	-.84	.25	1.02	.2	1.05	.3	L .21	Item 30 ; Item 30 : 30-30
38	47	80	-.41	.24	1.04	.5	1.01	.1	M .23	Item 38 ; Item 38 : 38-38
20	53	80	-.77	.25	1.04	.4	1.01	.1	N .21	Item 20 ; Item 20 : 20-20
36	39	80	.05	.24	1.03	.4	1.03	.3	O .25	Item 36 ; Item 36 : 36-36
17	56	80	-.97	.26	1.03	.3	.99	.0	P .21	Item 17 ; Item 17 : 17-17
34	35	80	.29	.24	1.01	.1	.99	-.1	Q .28	Item 34 ; Item 34 : 34-34
6	25	80	.91	.26	1.01	.1	.99	-.1	R .28	Item 6 ; Item 6 : 6-6
31	52	80	-.71	.25	1.00	.0	.97	-.2	S .26	Item 31 ; Item 31 : 31-31
23	58	80	-1.10	.26	1.00	.0	.94	-.3	r .25	Item 23 ; Item 23 : 23-23
15	17	80	1.50	.29	1.00	.0	.87	-.6	q .30	Item 15 ; Item 15 : 15-15
9	71	80	-2.29	.36	.99	.0	.93	-.2	p .16	Item 9 ; Item 9 : 9-9
40	22	80	1.11	.27	.99	-.1	.94	-.3	o .30	Item 40 ; Item 40 : 40-40
29	24	80	.97	.26	.97	-.3	.90	-.7	n .34	Item 29 ; Item 29 : 29-29
14	47	80	-.41	.24	.93	-.9	.97	-.3	m .35	Item 14 ; Item 14 : 14-14
33	27	80	.78	.25	.96	-.4	.95	-.3	l .33	Item 33 ; Item 33 : 33-33
22	15	80	1.68	.30	.95	-.3	.83	-.7	k .32	Item 22 ; Item 22 : 22-22
27	64	80	-1.56	.29	.95	-.3	.89	-.4	j .27	Item 27 ; Item 27 : 27-27
21	15	80	1.68	.30	.95	-.3	.85	-.6	i .33	Item 21 ; Item 21 : 21-21
2	30	80	.59	.25	.92	-.8	.94	-.5	h .37	Item 2 ; Item 2 : 2-2
3	15	80	1.68	.30	.94	-.3	.77	-.9	g .36	Item 3 ; Item 3 : 3-3
4	66	80	-1.74	.31	.94	-.3	.75	-.9	f .30	Item 4 ; Item 4 : 4-4
24	6	80	2.82	.44	.93	-.2	.80	-.4	e .26	Item 24 ; Item 24 : 24-24
19	52	80	-.71	.25	.91	-1.1	.84	-1.2	d .38	Item 19 ; Item 19 : 19-19
13	65	80	-1.65	.30	.90	-.6	.76	-1.0	c .34	Item 13 ; Item 13 : 13-13
1	52	80	-.71	.25	.90	-1.1	.81	-1.4	b .40	Item 1 ; Item 1 : 1-1
37	22	80	1.11	.27	.84	-1.3	.74	-1.6	a .49	Item 37 ; Item 37 : 37-37
MEAN	40.	80.	.00	.28	1.00	.1	.97	-.1		
S.D.	18.	0.	1.26	.04	.07	.7	.15	.8		

The item map was also used to examine whether the difficulty of items were spread across all five content domains: Patterns and relations, equations and distances, geometric and trigonometric, shapes, areas and volumes and combinatory and probability. It can be deduced from the resultant item map that the difficulty of the items are well-distributed across the domains.

#### Differential Item Functioning Analysis

The result of the IRTLRF procedure for all the achievement test items is shown in Table 2. The significant tests for items 3, 4, 7, 8, 11, 36, 38 and 40 indicated DIF. English majors are more likely to respond in the lower score categories of item 3 as evidenced by the chi-square value ( $\chi^2 = 5.5$ ,  $df = 3$ ) greater than the critical value of  $X^2 = 3.84$ . Similar significant values can be observed on items 4, 7, 8, 11, 36, 38 and 40 with computed chi-square values of 6.4, 4.9, 4.6, 3.9, 8.1, 14.11, and 5.5, respectively. This result indicates that the difficulty of the items function differentially across the two groups, and as a result, the English and mathematics major examinees may have different probabilities of getting the same scores. On the other hand, upon close examination of the items, it could possibly mean that these particular items' concepts were not discussed in depth for the English majors. Nevertheless, all of these items with DIF are flagged for revision or rephrasing in a way that should be balanced for both groups of students.

## Discussion

Based on the test results, the author revisited all items flagged for review in the IRT analysis. Item 9 on the achievement test belong to the easiest items, yet, no students were able to answer it. The item will be rejected or it will be revised thoroughly and make it the first item in an effort to place an easier item first on the student assessment. The item will be reworded because the author felt students were overanalyzing the question. The item with the negative item-total correlation (item 7) will be deleted because the item in general was confusing.

Overall, result of analysis could deduce that the achievement test in general is a good test. Although there are items removed, revised, and rephrased, most of the items came out to be good items. While classical test theory (CTT) and item response theory (IRT) methods are different in a so many ways, result of the analyses using these two methods does not say so. Items which were found to be "bad items" in CTT came out be not fitting also in the Rasch Model. Items 7, 9, 16, 24 and 26 were found to be marginal if not poor items in CTT. These were also the items that turned out to have extreme logit measures qualifying it to be unfitting in the latent trait model.

Surprisingly, some of the items came out to be biased as detected in the DIF analysis. Items 3, 4, 7, 8, 11, 36, 38 and 40 will be subjected to revision to remove its bias that is in favor to mathematics majors (Table 2). Although it could be said that math majors have the advantage in taking the test, it should not stop there. The test was made to measure the knowledge that was supposedly acquired by a student regardless of his/her field of specialization. Besides, most of the items were patterned from the mathematics items in the General Education part of the LET where there is no biasness in its items.

## References

- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.
- Cole, N. S., & Moss, P. A. (1993). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 201-219). Phoenix, AZ: Oryx Press.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Spray, J., & Miller, T. (1994). Identifying nonuniform DIF in polytomously scored test items. (RR 941). Iowa City, IA: American College Testing Program.
- 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*, 12(3), 535-556.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Linacre, J.M. (2002) 'What do Infit and Outfit, Mean-Square and Standardized Mean?', *Rasch Measurement Transactions*, 16 (2), 878.
- Spray, J.A., & Miller, T.R. (1992). Performance of the Mantel-Haenszel statistic and the standardized difference in proportions correct when population ability distributions are incongruent. (Research Report 92-1). Iowa City, Iowa: ACT, Inc.
- Thissen, D. (2001). *IRTLRDIF user's guide: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning [computer program]*. L.L. Thurstone Psychometric Laboratory. University of North Carolina at Chapel Hill.

## CutOff Scores: The Basic Angoff Method And the Item Response Theory Method

Niclie L. Tiratira

University of Rizal System, Morong, Rizal Campus

### Abstract

The objective of this study is to identify if there is a significant difference using the basic Angoff method and the IRT method in the setting of cutoff scores. Since the College Aptitude Test will be used for the study, it was tested for goodness of fit using the Bigstep software. chi-square using the SPSS software was also used to test the significant difference in setting of cut -off scores using basic Angoff method and the IRT method. The comparative analysis made between the setting of cutoff score using the Angoff and IRT has no significant difference. This means that using the basic Angoff method as well as the IRT method will give us the same cutoff scores.

There have been several controversies in the use of cut off scores in the setting of passing scores especially in the admission tests of colleges and universities. What could be the most appropriate method to use in the setting of cutoff scores? This question makes us realize the importance of making empirical study to answer this kind of inquiry. The purpose of the study is to make a comparative analysis of establishing cut off scores using the Basic Angoff method and the Item Response Theory (IRT) theory method.

A present issue we can cite is that which involves the United States, with the advent of the Public Law 107-110, No Child Left Behind Act of 2001 (United States Government, 2002). The act includes assessment and measurement of student progress as one of four main pillars. As a result, each state is federally mandated to develop state-wide tests of student achievement in core (mathematics and reading currently, science by 2005-2006) areas of school curriculum in grades 3 to 8. Over \$770 million in funding has been allocated to aid states in the development and administration of these tests for 2002-2003 alone.

In connection with this Public Law 107-110, No Child Left Behind Act of 2001 (United States Government, 2002), standardized tests is made to increase "accountability" among educators and students. As such, students are expected to meet some standard of proficiency that the tests are designed to assess (Ricker, 2002). Ideally, this standard will be the embodiment of the learning objectives. The standard should represent "mastery" of the learning objectives, or some level of basic proficiency necessary to move on to the next level, or to function in the real world (Van der Linden, 1982). In effect, establishing a standard can be conceptualized as policy making that has an impact on everyone involved in the testing procedure (Kane, 2001).

In the Philippine scenario, admission to top universities like the University of the Philippines, De La Salle, and Ateneo University has become a highly competitive and difficult because of the entry level requirements such as that of admission test based on norm cut off scores.

One of the testing procedures that we need to do is that which involves cutoff scores to determine who will be qualified to enter a certain college or university after taking the admission test. Students are expected to meet some standard of proficiency that the tests are designed to assess. Ideally, this standard will be the embodiment of the learning objectives. The standard should represent "mastery" of the learning objectives, or some level of basic proficiency necessary to move on to the next level, or to function in the real world (Van der Linden, 1982). In effect,

establishing a standard that can be conceptualized as policy that has an impact on everyone involved in the testing procedure (Kane, 2001).

In establishing a standard, we need to establish the possible and appropriate cutoff scores to give a fair stance between the student's ability and prediction of their performance. There are several ways of establishing cut off scores. One way is through the basic Angoff method and another is the IRT method.

Angoff (1971) inadvertently introduced a method for standard setting that is, using the amount of attention devoted to it in the research context as an indicator, one of the most commonly used method of setting standards today. The original method has been modified in different ways by researchers (e.g., Hambleton & Plake, 1995; Impara & Plake, 1997; Taube, 1997) in an attempt to improve it. In 1986, Berk published a "consumer's guide" to standard setting techniques, which included a set of criteria to be used to assess standard setting methods. He also the assessed various cut score setting procedures, including five Angoff-type methods.

The variants in Angoff methods can be classified as item-judgment methods. Each item on a test is assessed in terms of how likely minimally acceptable or competent candidates (those who would barely meet mastery standards) are to answer that item correctly (Ricker, 2002).

The Angoff method, in its most basic form, is seemingly a very simple process. Perhaps its simplicity should not be surprising, given that it arose from footnote in a book chapter (Angoff, 1971, p.515). A group of judges are each asked to (independently) think of a group of minimally competent candidates who would border on the mastery/non-mastery cutoff. The most typical instruction is for judges to think of a pool of 100 candidates who would "just barely" meet the performance criteria. When Angoff first proposed the method, his instruction was to think of only one candidate. However, with the exception of Impara and Plake (1997), the hypothetical pool of candidates is used.

The judges, working independently, then estimate what proportion of that sample of minimally acceptable candidates would answer each item in the test correctly. These p-values are summed and usually denoted as the minimum passing level for judge ( $mpl_j$ ). The  $mpl$  represents an individual judge's cut score for the test. The mean of these cut scores is the final cut score for the test. The standard error can also be calculated for the cut score. a lower standard error is desirable since it denotes better agreement among the judges (and less uncertainty about where the "true" cut score should lie).

This method does not just apply to minimally competent candidates, but could also be used to create a cut score for any grouping within the population. For example, Angoff methods could be used to set a cut score for a standard of excellence on a test. In this case, judges would be required to conceptualize a group of 'minimally excellent' examinees.

However, another way of establishing cutoff scores is through the IRT. When judges are asked to assess the probability of candidates correctly answering an item, they are in essence determining the difficulty of the item. In effect, the Angoff rating estimates the ability level denoted as  $\theta$  in Item Response Theory (IRT) of a minimally acceptable examinee (Kane, 1987). Taube (1997) extended this idea by using judge's ratings to work backwards to calculate b- (difficulty) parameters for each item using a Rasch IRT model, given by:

$P(\theta)_i = 1 / (1 + \exp(-D(\theta - b_i)))$  where  $P(\theta)_i$  is the probability of an examinee with a given  $\theta$  correctly answering item  $i$ , and  $D$  is a scaling constant equal to 1.7. Instead of calculating the sum of the item probabilities as the cut score, the mean item difficulty was calculated.

Methods for setting cut-scores have suffered severe attack historically, mainly because regardless of continuous efforts to improve standard-setting methodology, deciding what is appropriate remains very much a matter of subjective judgment. Attacks also happen because finding the appropriate balance between passing those who should fail and failing those who should pass continues to haunt people involved in setting cut-scores (Zieky, 2001). On the other hand, there are specific situations in which there is no other choice, but to establish cut-scores.

According to Zieky (2001), an example of such situation is the competency tests used for licensing professionals in which cut-score setting is mandated by law.

Some authors in the measurement literature make a distinction between two different notions of cut-scores: The cutoff scores and the critical scores. According to Maurer, Raju, and Collins (1998), a cutoff score, contrarily to critical score, depends on the number of openings and the number of applicants (i.e., for a specific job position). Therefore, it does not necessarily depend on considerations of the criterion and are not necessarily criterion-referenced. A critical score, on the other hand, is the specific point in the criterion that is considered minimally acceptable with respect to some definition of success or competency, and does not take into consideration the number of examinees nor the number of openings. Another distinction between cutoff score and critical score is that the latter one is the same for all applicant groups in every version of the test. The definition of a critical score is much closer to the concept intended here; being so, it is the one adopted in this paper and referenced generally as a cut-score (D'Almeida, 2006).

A cut-score takes into consideration different levels of performance. Thus, by definition, it is criterion-referenced. Because this standard corresponds to a measure of what would be considered as a minimally acceptable performance, it can vary widely depending on the job and/or on the specified criterion of performance levels (D'Almeida, 2006). The objective of the study is to make a comparative analysis of setting cutoff scores using the basic Angoff method and the IRT method.

## Method

### Participants

There were one hundred and sixty seven (167) participants from the College of Science of the University of Rizal System who belongs to various courses and year levels of the college; specifically, third year BS Psychology students, second and third year BS Math students, second year BS Biology students and second year BS Guidance and Counseling students. The responses were used to establish the norm for the College Aptitude test using the Item Response Theory of cutoff scores. However, to establish cut off score for the Angoff method, there were four professors from the same college who were ask as judge of the College Aptitude test.

### Instruments

The College Aptitude Test (CAT) was utilized for the comparison of the cutoff scores using the Basic Angoff method and the IRT method. The CAT is a test that measures a person's ability to acquire learning using specific skills in comprehension, inductive reasoning or general reasoning, understanding of relational concepts, and figuring out a rule or principle that explains the relational concepts. The components of the test were based on verbal schemes of the taxonomy such as: Verbal analogy, syllogism, and letter series. The reading comprehension component was derived from the Wikipedia Swedish Scholastic Aptitude Test (modified on 18 January 2009). The College Aptitude Test was reliable with Cronbach alpha of .59 and .51 with Guttman Split-half reliability coefficient. However, Spearman rho convergent validity coefficient was .97 using Self-efficacy Scale by Schwarzer correlated with the College Aptitude test.

Aside from the CAT, various statistical softwares were utilized like the Bigstep software to compute for the goodness of fit of the items and the ability of the persons taking test.

### Procedure

Prior to the setting of cutoff scores, the College Aptitude Test was first administered to 167 college students. Item analysis, validity, reliability, and norm were established. Thus, a sample

test questionnaire was given to four of the professors in the College of Science of the University of Rizal System, specifically two of them were teaching courses in English, one teaching Psychology, and the other one teaching Guidance and Counseling courses. The professors who were judges of the test were given instruction to work independently, and then estimate what proportion of that sample of minimally acceptable candidates would answer each item in the test correctly.

The items were then tabulated as to columns of Rater (Easy and Difficult), IRT (Easy and Difficult), Hit, Miss, Cutoff raters, and Cutoff IRT. The column for Rater was derived from the four professors who judged the items. The item that they judged was categorized as easy and difficult. They indicated a check on the item that was difficult. So therefore, an item with no check is considered easy and those that have a check mark are difficult. However, for IRT column of easy and difficult, the final item calibration index was used to identify the difficult and easy item. Easy items have a negative sign before its value, and those that are difficult have none. The number of negative value was counted and was categorized as easy. For the "Hit" column, a validation of items between rater and IRT was compared and the same item number who falls under rater and IRT was counted as one item for "Hit" column and vice versa applies for "Miss" column. Number of item counted in difficult under rater becomes the cutoff for rater and the number of difficult items under IRT becomes IRT cutoff.

Chi-square for each column in the table was computed using the SPSS software. The Winstep software was also used to save file and then imported and computed in the Bigsteps software for the goodness of fit. The tables was analyzed and given interpretation.

#### Data Analyses

The goodness of fit was used in the study to identify the items that are not fitting for the test because we want to make sure that before we use the test for comparison of cut off scores it has established the goodness of fit. The goodness fit is a statistical output that will help us to see the match between the ability of students to the items as to difficult or easy. The person taking the test might have a high ability but the test item is easy, so there was no matching between the ability of the person who took the test and the test item. The item must be revised in case we see that it does not match the ability of the person. To arrive at the data for the goodness of fit, there was a need to use the Bigstep software. The Infit Zstd column of the Bigstep data output must first be analyzed. A value of the IN Zstd which is more that 2.0 means that the item does not have a good fit or "misfitting". Also, we can take a look at the MSQ, if it is above 1.3 the item is also "misfitting." Thus, looking at both the IN Zstd and MSQ having the specified value for misfitting will give us the final analyses whether that item has a good fit or not. Table 2 lists the Infit MSQ and Infit ZStd.

The analyses of the IRT items if it's easy or difficult were done by computing for the corrected item calibration of the final estimates of item difficulties using the IRT method. The corrected item calibration can be obtained by multiplying the initial item calibration with the special spread expansion factor. The product that will appear will be in positive and negative signs. The value with negative signs of corrected item calibration will be easy items and those with no negative signs will be difficult items. The items are counted and categorized as easy or difficult.

The value of the rater column for the easy and difficult was obtained through counting of the frequencies of responses of raters who judge the items as easy or difficult (refer to table 2). The items that the raters have checked was classified as difficult and those that don't have any check marks were classified as easy. The tabulated value was considered in the basic Angoff method in deciding for the cutoff.

The "Hit and Miss" columns were just a tally the frequency of an item that appears easy or difficult on both the raters and IRT (refer to table 1). Similar category of a single item as easy or difficult was counted and considered a "hit" item because both rater and IRT classified that item as

easy or difficult. Thus, items that do not fall on any category of both Rater and IRT difficult or easy item was considered "Miss" items because the items have different category on both rater and IRT.

The cut off score columns for Angoff and IRT was obtained through counting of the frequencies of difficult items. The number of items categorized as difficult on rater and IRT column was counted and then value was reflected over the total number of items per component (reading comprehension, syllogism, verbal analogy, and letter series).

The tabulated values on columns of Raters, IRT, Hit and Miss, and cut off scores of Angoff and IRT were further analyzed through the use of chi-square. The data was first encoded in the excel program and then imported in the SPSS program. Using the chi-square statistic and its associated degrees of freedom, the software reports the probability that the differences between the observed and expected frequencies occurred by chance. Generally, a probability of .05 or less is considered to be a significant difference. Table 1 list the Rater, IRT, Hit, Miss, Angoff Cutoff, and IRT Cutoff.

### Result

The College Aptitude was also tested in terms of its goodness of fit as well as the fit of the item to ability of the person taking the test. Thus, item numbers 6, 27, 32, and 36 are not fitting. Majority of the test items (36 items) have a good fit. This is important in the setting of cutoff scores because it assures that the test is good for the purpose of setting cutoff scores.

Table 1 shows the goodness of fit of the College Aptitude Test. There are three columns in the table which are the Entry, the IN MSQ, and IN ZStd. The Entry column pertains to the Item number of the Test. The IN MSQ pertains to the Mean Square Fit and ZStd pertains to Standardized Fit. An item which has a Mean Square fit of less than 1.3 and Standardized fit of less 2.0 is a good fit. In terms of the IN MSQ we can see that all the items are in good fit, while in terms of ZStd item numbers 6, 27, 32, and 36 are not in good fit.

Table 1  
Goodness of fit using the IN MSQ and IN ZSTD

ENTRY	IN.MSQ	IN ZSTD
Item 1	1.0	.00
Item 2	.96	-.22
Item 3	1.0	.00
Item 4	1.4	.71
Item 5	1.0	.00
Item 6	1.11	2.19
Item 7	1.0	1.00
Item 8	1.16	.98
Item 9	1.0	.00
Item 10	1.11	1.33
Item 11	1.0	.00
Item 12	1.0	.00
Item 13	1.0	.00
Item 14	1.03	.41
Item 15	1.0	.00
Item 16	.93	-1.33
Item 17	1.0	.00
Item 18	1.05	.75
Item 19	1.02	.25
Item 20	1.0	.00

Cont. Table 1

Item 21	1.08	1.11
Item 22	1.0	.00
Item 23	1.06	1.17
Item 24	1.0	1.0
Item 25	.99	-.18
Item 26	1.0	.00
Item 27	.87	-2.77
Item 28	1.0	.00
Item 29	1.0	.00
Item 30	.96	-.76
Item 31	1.00	.00
Item 32	.84	-2.90
Item 33	1.0	.00
Item 34	.88	-1.77
Item 35	1.0	.00
Item 36	.87	-2.39
Item 37	1.00	.00
Item 38	1.01	.22
Item 39	1.0	.00
Item 40	1.05	.35

Note. IN MSQ = Mean Square Fit, IN ZStd = Standardized Fit

Table 2 are the columns for rater's (for Angoff method) judgment of easy and difficult items, IRT method easy and difficult, Hit, Miss, Angoff cutoff Scores, and ITR cutoff scores based on the four components of the College Aptitude Test. The rater column pertains to the categorization of items as to easy or difficult using the Angoff method. Here the raters are the judges of items taking in consideration the probability that the students can answer the item correctly. A total of 21 items specifically 6 items for reading comprehension, 6 items for syllogism, 5 items for verbal analogy, and 4 items for letter series were categorized as easy, while 19 items specifically 6 items for reading comprehension, 4 items for syllogism, 7 items for verbal analogy, and 2 items for letter series were categorized as difficult items for rater column. Thus, using the IRT column was derived by computing for the logits and getting the corrected item calibration of positive and negative, a total of 18 items were categorized as easy items specifically 6 items for reading comprehension, 7 items for syllogism, 3 items for verbal analogy, and 2 items for letter series; and a total of 22 items for difficult specifically 6 items for reading comprehension, 3 items for syllogism, 9 items for verbal analogy, and 3 items for letter series. There is also the Hit column wherein similar item number categorized as both easy and difficult for both rater and IRT method was tallied. The Hit column has a total of 24 items that have similar category in easy and difficult of both rater and IRT method specifically 9 items for reading comprehension, 4 items for syllogism, 8 items for verbal analogy and 2 items for letter series. The Miss column wherein the items left after getting the Hit items was counted. The Miss items has a total of 16 items specifically 2 items for reading comprehension, 6 items for syllogism, 4 items for verbal analogy, 4 items for letter series. Moreover, there is also the Angoff Cutoff score column wherein 6/12 is the cut –off or passing score for reading comprehension, 4/10 for syllogism, 7/12 for verbal analogy, and 2/6 for letter series, the whole test total passing score is 19/40. The IRT Cutoff score column shows that the cutoff or passing score for reading comprehension was 6/12, syllogism 3/10, verbal analogy 7/12 and 4/6 for letter series, so the total passing score for the test based on IRT method was 20/40. Thus after these categorizing the chi-square analyzes was done.

Table 2  
 Rater, IRT, Hit, Miss, Angoff Cutoff, and IRT Cutoff

Components of the College Aptitude Test	Rater		IRT		Hit	Miss	Angoff Cutoff	IRT Cutoff
	Easy	Difficult	Easy	Difficult				
Read Comp	6	6	6	6	9	2	6/12	6/12
Syllogism	6	4	7	3	4	6	4/10	3/10
Verbal Ana	5	7	3	9	8	4	7/12	7/12
Letter Series	4	2	2	3	2	4	2/6	4/6
Total	21	19	18	22	24	16	19/40	20/40

The chi-square analyzed showed that the columns on rater easy and IRT easy has an asymptotic significance of .238,  $df=6$  ( $p<.05$ ). Also, the chi-square analyses of rater D and IRT D has an asymptotic significance (2-sided) .213,  $df = 6$  ( $p<.05$ ), Hit and Miss with asymptotic significance (2-sided) .238,  $df = 6$  ( $p<.05$ ), Angoff cutoff scores and IRT cutoff scores with asymptotic significance (2-sided) .213,  $df = 6$  ( $p<.05$ ). The interpretation is that there was no difference between the Angoff method and IRT method of item classification as to easy and difficult. There was no difference between the "hit" items, meaning items categorized as easy and difficult based from comparison between Angoff method and IRT method is the same, thus no significant difference was also found in the "miss" items. Moreover, the result showed that there was no significant difference in the cut off scores using Angoff method and IRT method. However, the probability of the no significant differences between the observed and expected frequencies occurred by chance by 21.3 percent for Rater D and IRT D; 23.8 percent for Hit and Miss; and 21.3 percent for Angoff cutoff score and IRT cutoff score.

### Discussion

The result of the goodness of fit test showed that only four items are misfitting. This indicates that majority of the College Aptitude Test items fit or matches the ability of the person who took the test. This is important in this empirical study since it will take away the doubt that the comparison will not be appropriate for a reason that the ability of the person who took the test does not fit the item. Therefore, the goodness of fit test will strengthen the College Aptitude Test to be a good instrument to compare the difference in setting of cutoff scores using the basic Angoff method and the IRT method.

The comparison made between the setting of cutoff score using the Angoff and IRT has no significant difference shown in the chi-square analyses between Angoff cut off score and IRT cut off score. We can do another chi-square test by using the modified approach of the Angoff method and then test it again with the IRT method. Nonetheless, the explanation of the no significant difference of the cutoff scores is due to the process of deriving the cutoff score for Angoff method. The use of basic Angoff method in the setting of cutoff scores involves a great consideration of the judges as participants who rates the item whether there is probability that students can answer it correctly. However, the result is advantageous to those who are using the Angoff method because the IRT takes a lot of computation before arriving at a cut off score while we can arrive at the same result with the simple process of using the basic Angoff method.

Another explanation of the no significant difference in identifying difficult items between the basic Angoff method and the IRT method is the influence of the rater's mastery of a subject matter. The raters who are oriented with such kind of test or item content find that particular item as

easy and if not used in one item type of test find it difficult. Thus, mastery of the subject matter influences the rater's judgment of the item. Compared with IRT, those students who are also not familiar with such kind of test can have the probability of answering the test items incorrectly. An interview with the students can be done to correlate the probability of getting correct response having the conception that the student is familiar with the kind of test given.

However, the IRT method can affirm the result of the Angoff method because IRT method can derive the frequency of responses for an item for it to be considered difficult. If the Angoff method identifies the items as difficult, we can use IRT method to affirm the Angoff method because the frequency of the responses will tell us if that item was really difficult because only few students can answer a particular item. Thus, the use of basic Angoff method categorized items as easy or difficult and IRT method affirms it. This argument can also be the same for the result of the chi-square analysis of the non-significant difference of the Hit and Miss columns.

The congruence of the results using the chi-square for the rater easy and IRT easy columns, rater difficult and IRT difficult columns, and Hit and Miss columns made the result of the cutoff scores for both the Angoff method and IRT method to be non significantly different. It follows the explanation that if the result of the chi-square for any columns mentioned was significantly different, it can change the result of the chi-square for the cut off score.

Moreover, the results also showed that there are probabilities that the no significant differences between the observed and expected frequencies occurred by chance. We can lengthen the number of items of the test and increase the number of raters as participants for the Angoff method for setting cut off scores to change the result to significant difference. The limitation of the empirical study is that there were only four participants for the Angoff method and there were only forty items comprising the College Aptitude Test that was used in the study. The use of the modified Angoff method can change the result of the study since the procedure for the modified method is more objective than the basic method.

Further empirical study to test the cut off scores for modified Angoff method and the IRT method can be done for comparative analysis to test the significant difference between the cut off score using any Angoff method and the IRT method.

## References

Anastasi, A., & Urbina, S. (2000). *Psychological testing* (7th ed). New York: Macmillan Publishing Company.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ed.), *Educational measurement* (2nd ed.) (pp. 508-600). Washington, DC: American Council on Education.

D'Almeida, M. (2006). *Standard-Setting Procedures To Establish Cut-Scores For Multiple-Choice Criterion Referenced Tests In The Field Of Education: A Comparison Between Angoff And Idmatching Methods*. University Of British Columbia, 2006.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-55.

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 353-366.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.

- Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, 83, 693-702.
- Pabico, A. P. (2008). Improved English proficiency among Filipino adults 'surprising', Available on-line: <http://www.pcij.org/blog/?p=2330>.
- Ricker, K. L. (2002). Setting Cut Scores: Critical Review of Angoff and Modified-Angoff Methods. Unpublished manuscript, University of Alberta, Canada.
- Taube, K. T. (1997). The incorporation of empirical item difficulty data in the Angoff standard-setting procedure. *Evaluation and the Health Professions*, 20, 479-498.
- Van der Ven, A. H. G. S. (1980). *Introduction to scaling*. New York: Wiley.
- Zieky, M. (2001). So much has changed: How the setting of courses has evolved since the 1980s. In G. Izek (ed.), *Setting performance standards: Concepts, methods, and practices*. NJ: Lawrence Erlbaum.

## Editors

### Consulting Editor

Dr. Wai Chan  
The Chinese University of Hong Kong  
wchan@psy.cuhk.edu.hk

Dr. Stephen Sireci  
University of Massachusetts  
sireci@acad.umass.edu

### Senior Associate Editor

Dr. Carlo Magno  
De La Salle University-Manila  
carlo.magno@dlsu.edu.ph

### Editorial Board

Dr. Alexander Davies  
Social Survey Methods Division, Statistics Canada  
Alexander.davies@statcan.ca

Dr. April L. Zenisky  
University of Massachusetts  
azenisky@educ.umass.ed

Dr. Karin M. Butler  
University of New Mexico  
kmbutler@unm.edu

Dr. Harold D. Delaney  
University of New Mexico  
hdelaney@unm.edu

Dr. Jan Armstrong  
University of New Mexico  
jka@unm.edu

Dr. Donald Yeo Hong Huang  
National University of Singapore  
psyhhhd@nus.edu.sg

Dr. Chang Lei  
Chinese University of Hong Kong  
leichang@cuhk.edu.hk