The International Journal of Educational and Psychological Assessment
April 2009, Vol. 1, Issue 1, pp. 1-11

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.

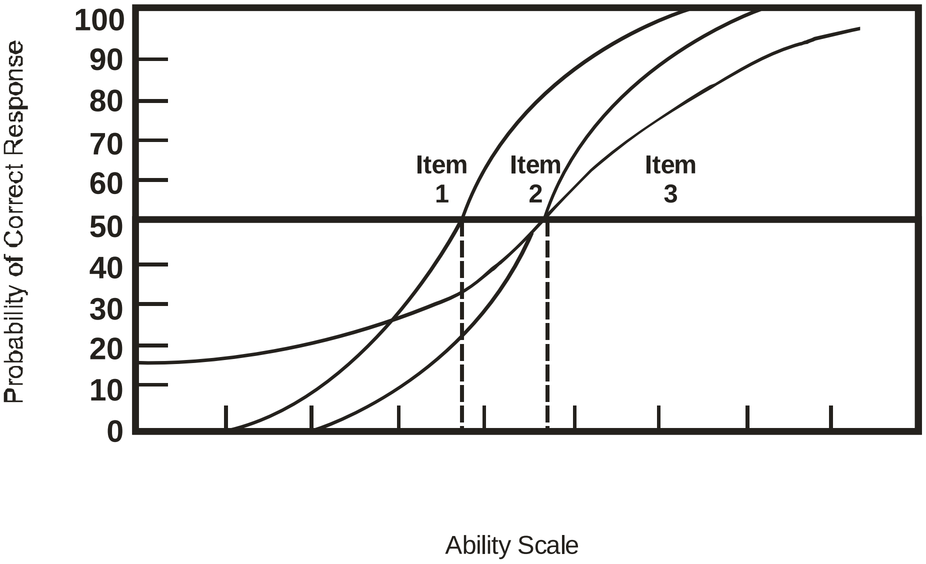
Item Response Theory

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 (θ) 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 (θ) 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 (b1, b2, b3) 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 ε is a function of parameters describing the ability of examinee and difficulty of the test, θ represents the ability of the examinee and δ 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 [θ] and the difficulty of the question [β] (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 θ and β. 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 Pij, where Pij is the probability that subject i will give a predefined response to item j, being the sum of a subject scale value ui and an item scale value vj, i.e. In (Pij/1 - Pij) = ui + vj

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	
	Form A	Form B	
	N1 vs. N2	N1 vs. N2	Difference
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.

Tow 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_1	*N_2*	*p*	*N_1*	*N_2*	*p*
IRT						
Person Reliability	.66	.62	.62	.81	.77	.43
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 *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 results 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. Soccorro 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