

Item Response Theory and Classical Test Theory: An Empirical Comparison of Item/Person Statistics in a Biological Science Test

Jimelo L. Silvestre-Tipay
 De La Salle-College of Saint Benilde

Abstract

Despite theoretical differences between item response theory (IRT) and classical test theory (CTT), there is a lack of empirical knowledge about how, and to what extent, the IRT- and CTT-based item and person statistics behave differently in a Biological Science test. This study examined the behaviors of the item and person statistics derived from these two measurement frameworks in a Biological Science Test designed for college freshman students. The study answered the following questions: (a) How consistent are the item difficulty levels across CTT-framework and IRT framework?; (b) How comparable are the CTT-based and IRT-based internal consistency measures?; (c) What is the dimensionality measure of items?; (d) How comparable are the differential item functioning of items across CTT and IRT frameworks? The findings indicate that the person and item statistics derived from the two measurement frameworks are quite comparable. The degree of difference of item and person statistics across samples, usually considered as the theoretical superiority IRT models, also appeared to be similar for the two measurement frameworks but essential areas of variation must be seriously considered and addressed.

Classical test theory (CTT) and item response theory (IRT) are widely perceived as representing two very different measurement frameworks. However, few studies have empirically examined the similarities and differences in the parameters estimated using the two frameworks. Prior to exploring this issue in some detail, a brief review of related theories may be helpful to the readers.

Brief Review of CTT and IRT

Although CTT has served the measurement community for most of this century, IRT has witnessed an exponential growth in recent decades. The major advantages of CTT are its relatively weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). Relatively weak theoretical assumptions not only characterize CTT but also its extensions (e.g., generalizability theory). Although CTT's major focus is on test-level information, item statistics (i.e., item difficulty and item discrimination) are also an important part of the CTT model. At the item level, the CTT model is relatively simple. CTT does not invoke a complex theoretical model to relate an examinee's ability to success on a particular item. Instead, CTT collectively considers a pool of examinees and empirically examines their success rate on an item (assuming it is dichotomously scored). This success rate of a particular pool of examinees on an item, well known as the p value of the item, is used as the index for the item difficulty (actually, it is an inverse indicator of item difficulty, with higher value indicating an easier item). The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as item discrimination, which is often expressed statistically as the Pearson product-moment correlation coefficient between the scores on the item (e.g., 0 and 1 on an item scored right-wrong) and the scores on the total test. When an item is dichotomously scored, this estimate is often computed as a point-biserial correlation coefficient. The major limitation of CTT can be summarized as circular dependency: (a) The person statistic (i.e., observed score) is (item) sample dependent,

and (b) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample dependent. This circular dependency poses some theoretical difficulties in CTT's application in some measurement situations (e.g., test equating, computerized adaptive testing).

Despite the theoretical weakness of CTT in terms of its circular dependency of item and person statistics, measurement experts have worked out practical solutions within the framework of CTT for some otherwise difficult measurement problems. For example, test equating can be accomplished empirically within the CTT framework (e.g., equipercentile equating). Similarly, empirical approaches have been proposed to accomplish item-invariant measurement (e.g., Thurstone absolute scaling) (Englehard, 1990). It is fair to say that, to a great extent, although there are some issues that may not have been addressed theoretically within the CTT framework, many have been addressed through ad hoc empirical procedures.

IRT, on the other hand, is more theory grounded and models the probabilistic distribution of examinees' success at the item level. As its name indicates, IRT primarily focuses on the item-level information in contrast to the CTT's primary focus on test-level information. The IRT framework encompasses a group of models, and the applicability of each model in a particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items. For test items that are dichotomously scored, there are three IRT models, known as three-, two-, and one-parameter IRT models. Although the one-parameter model is the simplest of the three models, it may be better to start from the most complex, the three-parameter IRT models; the reason for this sequence of discussion will soon become obvious.

Theoretically, IRT overcomes the major weakness of CTT, that is, the circular dependency of CTT's item/person statistics. As a result, in theory, IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. This invariance property of item and person statistics of IRT has been illustrated theoretically (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991) and has been widely accepted within the measurement community. The invariance property of IRT model parameters makes it theoretically possible to solve some important measurement problems that have been difficult to handle within the CTT framework, such as those encountered in test equating and computerized adaptive testing (Hambleton et al., 1991). Because IRT differs considerably from CTT in theory, and commands some crucial theoretical advantages over CTT, it is reasonable to expect that there would be appreciable differences between the IRT- and CTT-based item and person statistics. Theoretically, such relationships are not entirely clear, except that the two types of statistics should be monotonically related under certain conditions (Crocker & Algina, 1986; Lord, 1980). But such relationships have rarely been empirically investigated, and, as a result, they are largely unknown.

The empirical studies available in this area have primarily focused on the application of the two methods in test equating (e.g., Becker & Forsyth, 1992; Harris, 1991). With regard to test equating, Hambleton et al. (1991) suggested that, theoretically, the invariance property of the IRT item statistics obviated the need of equating tests; instead, it is (linear) scaling, rather than equating, that is necessary within the framework of IRT. The discussion implies that IRT models handle equating tasks better than the CTT equating approaches. The empirical studies in this area, however, provide a mixed picture, with some indicating the superiority of IRT approaches (e.g., Peterson, Cook, & Stocking, 1983), some suggesting better results from CTT ad hoc approaches (e.g., Clemans, 1993; Kolen, 1981; Skaggs & Lissitz, 1986a), and still some finding that both CTT and IRT equating methods produce very comparable results (Skaggs & Lissitz, 1988). The mixed picture has prompted some researchers to suggest that it might be unrealistic to expect one method to provide the best equating results for all types of tests (e.g., Skaggs & Lissitz, 1986b).

A literature search revealed only one study that empirically examined the comparability of IRT-based and CTT-based item and person statistics. Lawson (1991) compared IRT-based (one-parameter Rasch model) and CTT-based item and person statistics for three different data sets,

and showed exceptionally strong relationships between the IRT- and CTF-based item and person statistics. The results of the study, although the study was based on somewhat small data sets and only examined the most restrictive one-parameter IRT model, suggest that information from the two approaches about items and examinees might be very much the same. Similarly, the invariance property of IRT item/person parameters has been little explored empirically, although invariance has been illustrated theoretically (e.g., Hambleton & Swaminathan, 1985; Rudner, 1983). However, Miller and Linn (1988), using an extant large data set, did report the results of a study examining the variations of item characteristic functions in the context of instructional coverage variations. They reported relatively large differences in item curve responses, suggesting lack of invariance of IRT item parameters. Lack of invariance was also reported by Cook, Eignor, and Taft (1988) for both CTT- and IRT-based item difficulty estimates.

Given the limited number of empirical studies directly or indirectly addressing the invariance issue, there is an obvious lack of systematic investigation about the absolute invariance of the item and person statistics obtained from either CTF or IRT frameworks and a lack of studies that empirically compare the relative invariance of item and person statistics obtained from CTT versus those from IRT. The major criticism for CTT is its inability to produce item/person statistics that would be invariant across examinee/item samples. This criticism has been the major impetus for the development of IRT models and for the exponential growth of IRT research and applications in the recent decades. It is somewhat surprising that empirical studies examining and/or comparing the invariance characteristics of item statistics from the two measurement frameworks are so scarce. It appears that the superiority of IRT over CTT in this regard has been taken for granted by the measurement community, and no empirical scrutiny has been deemed necessary. The empirical silence on this issue seems to be an anomaly. This lack of empirical investigation has prompted some researchers to state that item response modeling has been too focused on mathematical elaboration at the expense of empirical exploration (Goldstein & Wood, 1989).

Purpose of the Study

The present study focused on the issue: How comparable are the item and person statistics from the CTT framework with those from the IRT framework in a Biological Science test for college freshmen students? More specifically, the study addressed the following research questions:

1. How consistent are the item difficulty levels across CTT-framework and IRT framework?
2. How comparable are the CTT-based and IRT-based internal consistency measures?
3. What is the dimensionality measure of items ?
4. How comparable are the differential item functioning of items across CTT and IRT frameworks?

The need to construct and standardize the Biological Science Achievement Test was expressed by the Science and Math Department Chairperson since there is no test available in their university which they can claim as their own. It is the aim of the school to measure the Biological Science achievement of their freshmen college students before the students proceed to their major field of specialization. The test covers topics on (1) Characteristics of life and the levels of biological organizations; (2) Chemical Bases of Life; (3) Cell Structure and Function; and (4) Membrane Structure and Function based on the syllabus on Biological Sciences for Freshmen college students of Arellano University-Pasig.

Method

Participants

The participants were 326 college freshmen students of Arellano University-Pasig. The participants are freshmen college students taking up AB Psychology, Bachelor of Science in

Nursing which was classified as Science Related Courses and Bachelor of Science in Commerce and Bachelor of Science in Hotel and Restaurant Management classified as Non-Science Related Courses. There were 233 participants in the Science Related courses and 93 for the Non-Science Related Courses. When classified according to gender, the participants were made up of 92 males and 234 females. The decision to choose this group was in line with the purpose of constructing a valid and reliable Biological Science Achievement test for college freshmen of Arellano University-Pasig.

Instrument

The Biological Science Achievement Test was constructed to measure the competencies stated on the course syllabus. The test blue print or the table of specifications contained four major topics namely: (a) Introduction which covers Characteristics of life and levels of biological organization; (b) Chemical bases of life; (c) Cell structure and function; and (d) Membrane structure and function. The items were distributed as follows: Easy Level (50%) which is composed of Remembering (30%) and Understanding (20%); Moderate Level (30%) made up of Applying (20%) and Analyzing (10%); and Difficult level which is composed of Evaluating (10%) and Creating (10%). The test aims for the students to: (1) Identify the basic characteristics of life; (2) describe each level of biological organization and their emergent properties ; (3) differentiate elements and compounds; (4) discuss the role of water in life; (5) explain the chemistry of the element carbon; (6) discuss the structure and function of biomolecules; (7) demonstrate mastery of the cell theory; (8) distinguish between procaryotes and eukaryotes; (9) identify the structures and functions in a cell; (10) differentiate between plant and animal cell; (11) discuss the composition of the cell membrane; (12) discuss the model of the cell membrane; (13) identify the structure and function of the membrane; (14) identify the various mode of cell transport.

The Biological Science Achievement test is made up of 60 items. The item format was limited to multiple-choice for purposes of easy scoring and processing. The test was content validated and reviewed by two Biological Sciences teacher of the Science and Math Department of Arellano University-Pasig using the table of specifications. Prior to printing the final test questionnaire for administration, it was checked by the Department Chairperson of the Science & Math Department of the said university to insure its appropriateness and to check on some typographical errors and most of all content errors.

Procedure

The Biological Science Achievement Test was administered to college freshmen students of Arellano University-Pasig. The science teachers were given detailed instructions on how to administer the test. A copy of the instructions to be given to the students was provided so that the administration would be constant across situations. The test was then scored manually and scores were encoded to facilitate easy computation of the results and analysis using statistical softwares such as SPSS and Winsteps.

Data Analysis

In order to establish the reliability of the test, it was pilot tested to the 326 college students by their respective science teacher for about 50 minutes. The split-half method using Spearman-Brown formula was used to obtain the internal consistency. Further reliability test was conducted using Kuder-Richardson 20. These reliability procedures gave the reliability coefficients of 0.70 and 0.72 respectively which gave a high degree of consistency for the test with a sample size of 326 students.

In order to check the normality of the distribution across samples, skewness and kurtosis were computed.

Using the Classical Test Theory framework, a norm of performance was created and a cut off score was set using Angoff Method. The estimated cut off score was calculated by summing up the item difficulty estimates generated from an item analysis result. Item analysis was employed by computing the item difficulty index and discrimination index.

Item Analysis was conducted using both Classical Test Theory (CTT) and Item Response Theory (IRT). In the CTT the item difficulty and item discrimination were determined using the proportion of the high group and the low group. Item difficulty is determined by getting the average proportion of correct responses between the high group and low group. The Item discrimination is determined by computing for the difference between the high group and the low group. The estimation of Rasch item difficulty and person ability scores and related analyses were carried out using WINSTEPS. This software package begins with provisional central estimates of item difficulty and person ability parameters, compares expected responses based on these estimates to the data, constructs new parameter estimates using maximum likelihood estimation, and then reiterates the analysis until the change between successive iterations is small enough to satisfy a preselected criterion value. The item parameter estimates are typically scaled to have $M = 0$, and person ability scores are estimated in reference to the item mean. A unit on this scale, a logit, represents the change in ability or difficulty necessary to change the odds of a correct response by a factor of 2.718, the base of the natural logarithm. Persons who respond to all items correctly or incorrectly, and items to which all persons respond correctly or incorrectly, are uninformative with respect to item difficulty estimation and are thus excluded from the parameter estimation process.

Differential Item Functioning was done using the Statistical Package for Social Sciences to analyze the data. DIF was done in order to check fairness of the test items across samples of the same ability but with different gender and course. DIF made use of the Mantel-Haenzel Method wherein examinees were matched on their ability levels and then item performance on the two groups was compared in each score group.

Results

The data from the pilot test were used for reliability and item analysis. The Kuder-Richardson reliability was used to determine the internal consistency of the items. This method was used to be able to find the consistency of the responses on all the items in the test. The test's reliability was generated through the split-half method by correlating the odd numbered and even numbered items. The internal consistency arrived is 0.70. The other reliability procedure used in this test is the Kuder-Richardson 20 which gave the same reliability coefficient of 0.72, both indicating a high degree of consistency for the test with a sample size of 326 students.

Person performance in the Biological Science achievement test revealed that distribution is normal. Skewness gave a result of 0.05 which is almost zero while kurtosis gave a value of 0.26 which is also very close to 0.27. Both measures of normality proved that the distribution of scores of the 326 examinees is normal. Moreover, the normality was also supported by the values of the mean, median and mode which registered 26.11, 26.0 and 24.26 respectively. These values of the measures of central tendency further proved that the distribution is normal since they coincide with each other. A standard deviation of 6.69 means that the scores are dispersed.

Based on CTT, a norm was created for the performance of the 326 examinees. Cut off score was set using Angoff method. The sum of the difficulty estimates turned out to be 26.42, therefore the cut off is set at 26.

Table 1
Distribution of Test items According to Difficulty and Discrimination Indices

Item Discrimination	Easy	%	Item Difficulty Average	%	Difficult	%
Very Good			Items 7,12,15,37,45,52,53,54,58	15%		
Good			Items 3,4,6,17,19,23,24,26,27,28,30,36,40,42,48,49,55	28.33%		
Reasonably Good	Items 1 & 10	3.33%	Items 9,13,21,33,35,41	10%		
Marginal	Item 25	1.67%	Items 2,8,14,16,18,20,31,38,43,44,47,50,51,56,57,60	26.67%	Item 5	1.67%
Poor			Items 11,29,32,34,46,59	10%	Items 22 & 39	3.33%

Table 1 show that 3.33% of the items are easy and have reasonably good discrimination index while 1.67% are easy items and have marginal discrimination index. For the items with average level of difficulty, 15% have very good discrimination index, 28.33% have good discrimination index, 10% have reasonably good discrimination index, 26.67% have marginal discrimination index and 10% have poor discrimination index. For the items that are difficult, 1.67% have marginal discrimination index and 3.33% have poor discrimination index.

Furthermore, this result reveals that the Biological Science test based on the CTT framework is an acceptable test with 53% of its items average difficulty and good discrimination index, thus only 47% of the items would require revision to improve their discrimination indexes.

Table 2
Distribution of Items According to Difficulty Level for CTT and IRT

Content	No. of Items	CTT			IRT			Hit Rate	Missed Rate
		Easy	Average	Difficult	Easy	Average	Difficult		
A. Introduction	9 (item #1-9)	1	2,3,4, 6,7,8,9	5	1,2,3,4, , 6,7,9		5, 8	2	7
B. Chemical Bases of Life	26 (item #10-35)	10, 25	11,12,13 , 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35	22	10,12, 15,17, 18 19,25	14	11, 13, 16, 20, 21, 22, 23, 24, 26, 27, 28,29, 31, 32, 33, 34, 35	4	22
C. Cell Structure & Function	16 (item # 36-51)		36, 37, 38, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51	39	36,37, 38, 43,45, 49,		39, 40, 42, 41, 44, 46, 47, 48, 50, 51,	1	15
D. Membrane Structure and Function	9 (item # 52-60)		52, 53, 54, 55, 56, 57, 58, 59, 60		52,53, 54,55,		56, 57, 58, 59, 60	0	9
Total	60	3	54	3	24	2	34	7	53

After calculation, the chi-square value for the hit rate and missed rate is .000. This means that at $\alpha = .05$ there is a significant mismatched among items in this Biological Science Test in terms of difficulty between the CTT and the IRT frameworks. Moreover, this result reveals that the difficulty level of items significantly differ when analyzed using the CTT framework and using the IRT framework. This means that items classified as easy, average or difficult in CTT may not necessarily have the same difficulty level when classified based on IRT framework.

Table 3
Comparison of Internal Consistency by CTT and IRT

CTT	IRT	
Cronbach's Alpha	Person Reliability	Item Reliability
0.72	0.70	0.98

The sample produced a reliability of 0.72 for Cronbach's alpha which denotes a high internal consistency for the Biological Science Test. For the IRT framework, person reliability gave a result of 0.70 which indicates that the items are working well to consistency reproduce a participant's score.

Furthermore, the result reveals that there is an agreement in the internal consistency result of the test for both frameworks. Therefore, it can be inferred that internal consistency of the Biological Science test remain stable across the CTT and the IRT framework, thus making the entire test highly reliable. This high reliability means that the sample is big enough to precisely locate the items on the latent variable for the Biological Science Test.

Table 4
Dimensionality of Items in the Biological Science Test

Person Separation	Item Separation
1.52	6.50
No. of Strata (HP) = 2.36	No. of Strata (HP) = 9.0

Table 4 reveals that the item separation of 1.52 will give 2.36 or two strata using the formula $HP = [(4 * \text{item/person separation}) + 1] / 1$. This means that there are two ability groups that can be generated from the sample. This statistic also represents how well the sample separates into distinct performance levels.

For the item separation, the value of 6.5 gives nine strata which means that the items in the test can still be classified into nine sub groups.

Furthermore, in terms of the person separation measure, the result reveals that the sample can be separated into two distinct ability groups i.e. gender (male versus female); course (students taking Science-Related courses versus students taking Non-Science Related courses). In terms of the item separation, the result reveals that the entire test can be divided into nine subtests. This means that the table of specifications must be reviewed in order to identify possible classification of test items either by content, competency, level of difficulty, etc.

Table 5
Comparison of CTT and IRT Differential Item Functioning on Course

Item Number	CTT Chi-Square	IRT Chi-Square	Hit	Missed
1	.365	.2209	1	
2	.000*	4.9302*	1	
3	.027*	.1131		1
4	.391	1.1188	1	
5	.201	.0000	1	
6	.204	.2567	1	
7	.002*	1.5322		1
8	.734	2.4898	1	

Cont. Table 5

9	.397	.5979	1	
10	.075	10.191*		1
11	.700	5.1067*		1
12	.006*	.7724		1
13	.017*	.3463		1
#14	.050	15.3797*		1
15	.000*	7.9532 *	1	
16	.932	3.9876*		1
17	.012*	.6612		1
18	.872	2.7174	1	
19	.164	.2151	1	
20	.394	7.7066*		1
21	.068	.0000	1	
22	.017*	16.6440*	1	
23	.004*	1.2421		1
24	.001*	2.9142		1
25	.005*	2.0919		1
26	.000*	5.8366*	1	
27	.092	.0000	1	
28	.200	.0933	1	
29	.535	5.6725*		1
30	.001*	1.7830		1
31	.869	3.9365*		1
32	.931	3.7470	1	
33	.028*	.1892		1
34	.518	6.4152*		1
35	.002*	1.5580		1
36	.000*	5.6490 *	1	
37	.000*	17.0446*	1	
38	.001*	1.9609		1
39	.960	3.1515	1	
40	.004*	1.7483		1
41	.685	1.9524	1	
42	.000*	5.3240*	1	
43	.316	.8216	1	
44	.003*	1.8766		1
45	.000*	7.5257*	1	
46	.885	2.6959	1	
47	.537	1.0306	1	
48	.062	.0000	1	
49	.001*	2.3553		1
50	.921	3.1053	1	
51	.181	.1101	1	
52	.001*	2.3781		1
53	.000*	15.5710*	1	
54	.026*	.1341		1
55	.000*	4.0169*	1	
56	.524	6.0246*		1
57	.481	1.0071	1	
58	.000*	6.1365*	1	
#59	.700	5.1067*		1
60	.471	1.2651	1	

* with differential item functioning (for CTT $p < .05$; for IRT chi-square > 3.841)

misfitting item (exceeded both the MS and ZSTD criteria: MS = 1.30 and ZSTD = 2.0)

Table 5 shows that using CTT method in determining differential item functioning there are 28 items with DIF or 46.67% of the items have course bias. This means that there 28 items in which significantly more students taking Science Related courses can answer correctly than those taking Non-Science related courses.

Table 6
Summary of CTT and IRT DIF results

Content	No. of Item	Hit	Missed
A. Introduction	9 (item #1-9)	7	2
B. Chemical Bases of Life	26 (item #10-35)	9	17
C. Cell Structure & Function	16 (item # 36-51)	12	4
D. Membrane Structure and Function	9 (item # 52-60)	5	4

Table 6 gives the summary of hits and missed in terms of DIF grouped according to content domain. When subjected to chi-square computation, the result gave a value of .000 which is significant at $\alpha = .05$. This result reveals that the number of items with DIF analyzed within the CTT framework is significantly different from the number of items with DIF analyzed within the IRT framework. Therefore, it can be inferred that items identified to have DIF or biased items in CTT may not necessarily be classified right away to be biased under IRT. This implies that for DIF computation, the test developer must check which items are biased for both IRT and CTT.

Discussion

The purpose of this study was to compare CTT and IRT results of the Biological Science Test. The study investigated on the consistency of the item difficulty levels across CTT-framework and IRT framework, compared the CTT-based and IRT-based internal consistency measures, identified the dimensionality measure of items and compared the differential item functioning fit items across CTT and IRT frameworks.

Overall the Biological Science test demonstrated good psychometric properties but it is worth noting that the test items also demonstrated different behaviors across CTT and IRT frameworks.

Moreover, the item fit statistics were analyzed to determine the dimensionality and results reveal that two of the 60 items are misfit based on the IRT while in CTT, 27 items must undergo revision due poor discrimination index. These 27 items may be further subjected to option analysis to determine the effectivity of its distracters.

Furthermore, analysis result reveals that the difficulty level of items significantly differ when analyzed using the CTT framework and using the IRT framework which means that items classified as easy, average or difficult in CTT may not necessarily have the same difficulty level when classified based on IRT framework. This implies that CTT and IRT should be used independently when checking the difficulty level of the items considering that both frameworks have different assumptions at the start.

Regarding the internal consistency of the Biological Science test, the analysis revealed a certain degree of stability across the CTT and the IRT framework in terms of the internal consistency of the items and the test as a whole. The consistent high reliability measure of the test across two frameworks implies that the sample is large enough to precisely locate the items on the latent variable which is achievement in Biological Science. High reliability (of persons or items) means that there is a high probability that persons (or items) estimated with high measures actually do have higher measures than persons (or items) estimated with low measures.

In terms of the person separation as well as the item separation, the result reveals that the sample can be separated into two distinct ability groups such as gender and course (in terms of International Journal of Educational and Psychological Assessment 2009; Vol. 1(1)

relation to Science) and entire test can be divided into nine subtests which would require a review of the content domain stated in the table of specifications.

Also, comparison of the DIF for CTT and IRT frameworks revealed that items identified to have DIF or biased items in CTT may not necessarily be classified right away to be biased under IRT. This result requires that differential item functioning be treated differently by CTT and IRT thus, test developer may use both to on items that are totally free of bias.

Overall, the findings from this empirical investigation failed to discredit the CTT framework with regard to its alleged inability to produce wide range of difference in item statistics and psychometric properties of test. On the other hand, the findings failed to support the IRT framework for its superiority over CTT in producing variance in internal consistency statistics. The findings here simply show that the two measurement frameworks produced very similar item and person statistics both in terms of the comparability of item and person statistics, difficulty level of items, internal consistency and differential item functioning between the two frameworks.

These findings pose some interesting questions about how to view the differences between IRT and CTT models both in theory and in testing practice. It is my view that in psychological measurement, as in any other areas of science, theoretical models are important in guiding our research and practice. But the merits of a theoretical model should ultimately be validated through rigorous empirical scrutiny.

Of course, the present empirical study, like many other research studies, had its share of limitations that may potentially undermine the validity of its findings. First of all, the characteristics of the test items used in the study may be somewhat unique. Although it is unclear what systematic impact this characteristic of the data may have had on the results, it would be desirable in future studies to replicate the present study using data from norm-referenced testing, which usually involves items varying more in item difficulty and in item discrimination. The second shortcoming of the investigation is the somewhat limited item pool used in the study. Although the examinee pool is quite adequate in the sense that a variety of different samples can be drawn from it, the same cannot be said about the item pool. Ideally, the test item pool should be larger and more diverse in terms of item characteristics so that items can be sampled from the pool to study the behaviors of CTT and IRT item statistics under different conditions of item characteristics.

Recommendations

Further research is warranted to validate the unidimensionality of the Biological Science test by conducting a principal components analysis (PCA). A PCA was not conducted for this study due to limited time.

Next, it is recommended that a more heterogeneous sample be used in further analysis to reexamine the reliability and separation statistics.

Finally, it is recommended that the psychometrics of the test be reevaluated after biased items have been removed and replaced by new items prior to the finalization of the Biological Science Achievement test.

References

- Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore: Johns Hopkins University Press.
- Becker, D. F., & Forsyth, R. A. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement*, 29, 341-354.

- Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, 25, 31-45.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.
- Englehard, G., Jr. (1990, April). Thorndike, Thurstone and Rasch: A comparison of their approaches to item-invariant measurement. Paper presented at the annual meeting of the American Educational Research Association, Boston. (ERIC Document Reproduction Services No. ED 320 921)
- Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology*, 42, 139-167.
- 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), 3847.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications*. Boston: Kluwer.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.),
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, 25, 205-219.
- Rudner, L. M. (1983). A closer look at latent trait parameter invariance. *Educational and Psychological Measurement*, 43, 951-955.
- Skaggs, G., & Lissitz, R. W. (1986a). An exploration of the robustness of four test equating models. *Applied Psychological Measurement*, 10, 303-317.
- Skaggs, G., & Lissitz, R. W. (1986b). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research*, 56, 495-529.
- Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement*, 12, 69-82.
- Wright, B & Stone, M. (1999). *Measurement Essentials (2nd ed.)*. Wilmington, Delaware: Wide Range Inc.