THE PHILIPPINE APTITUDE CLASSIFICATION TEST: WHY SHIFT FROM CLASSICAL TEST THEORY TO ITEM RESPONSE THEORY?

Ma. Lourdes M. Franco
Center for Educational Measurement, Inc.
Philippines

Abstract

The Philippine Aptitude Classification Test (PACT) is an instrument developed by the Center for Educational Measurement, Inc. in response to the need for a comprehensive system of identifying specific abilities of high school students for the purpose of educational and vocational guidance. The present study was designed to investigate which of the two theories, classical test theory (CTT) or item response theory (IRT), would best improve the quality of the PACT in terms of item development, test design, and scoring; and to look into the implications of using IRT for making decisions about career choices. It specifically aimed to (a) evaluate which item response theory model among the one-, two- and three-parameter logistic models was most suitable in evaluating the PACT, and (b) compare the chosen IRT model with the indices of the CTT.

The study involved responses on the PACT of 1,023 third year high school students from 28 schools. The findings indicated that the three-parameter model is the most suitable for PACT. Furthermore, the IRT model yielded more accurate estimates over CTT, which in turn would lead to more reliable test-takers' prediction of success in the appropriate educational courses.

Background

The Philippine Aptitude Classification Test (PACT) is an instrument developed by the Center for Educational Measurement, Inc. in response to the need for a comprehensive system of identifying specific abilities of high school students for the purpose of educational and vocational guidance. It attempts to predict a student's probable performance in various courses of study. It measures a number of dimensions that have been found to be useful in the classification of students into different fields of study. Being locally normed, it provides a profile of aptitudes for several Philippine educational programs in order to assist students in the choice of their careers. (TDD, 2007).

The PACT is a battery of aptitude tests with multiple-choice items which are largely dependent on innate abilities and minimally on academic experience. It has eighteen scorable subtests which underwent *factor analysis* and yielded eight aptitude factor scores (Iledan & Franco, 2003). The eighteen subtests are listed in Appendix A with the corresponding number of items in each subtest. The eight aptitude/factor scores are also given in the table together with their corresponding reliabilities (r_{tt}).

The final form of the PACT consists of two parts. Part I comprises of two speeded tests with a total of 30 items and a testing time of 4 minutes. Part II consists of sixteen power tests with a total of 210 items and 1 hour and 45 minutes

of testing time. The whole test has a total of 240 items and a total testing time of 1 hour and 49 minutes. The various items composing the battery measure a range of aptitudes deemed relevant to selected college and vocational courses. None of the subtests are curriculum-bound. There are verbal and numeric items but their dependence on particular subjects in school is very minimal.

The Center, in a move to upgrade its statistical procedures, shifted from procedures based on classical test theory (CTT) to procedures based on item response theory (IRT). IRT supposedly yields test-free examinee ability scores and sample-free item statistics. However, in order to ascertain that the decision to shift from CTT to IRT is sound and appropriate, an empirical verification should be undertaken. Consequently, this study was carried out to examine which of the two theories, classical test theory (CTT) or item response theory (IRT), would best improve the quality of the PACT in terms of item development, test design, and scoring.

Brief Review of CTT and IRT

The measurement of learning outcomes by means of testing is one of the most fundamental issues in education. Results obtained from tests help educators know how much students learn and provide feedback for shaping the teaching-learning process. One important goal in measurement is to design tests with minimum errors so that the information received from the tests has high validity and reliability. In the course of development in educational measurement, there are two popular frameworks that have been used widely: classical test theory (CTT) and item response theory (IRT) (Hambleton & Jones, 1993). Both theories, when appropriately applied, optimize the validity and reliability of an instrument. Consequently, it is the challenge posed to test developers to choose and employ the framework that best fits the instrument, thus realizing the goal of providing accurate and reliable student information.

CTT has dominated the area of standardized testing and is based on the assumption that a test-taker has an observed score and a true score. The observed score of a test-taker is usually seen as an estimate of the true score of that test-taker plus/minus some unobservable measurement error (Crocker & Algina, 1986; Hambleton & Swaminathan, 1985). An advantage with CTT is that it relies on weak assumptions and is relatively easy to interpret. However, CTT can be criticized since the true score is not an absolute characteristic of a test-taker because it depends on the content of the test. If there are test-takers with different ability levels, a simple or more difficult test would result in different scores. Another criticism is that the difficulty of the items could vary depending on the sample of test-takers that take a specific test. Therefore, it is difficult to compare test-takers' results between different tests. In the end, good techniques are needed to correct for the errors of measurement (Hambleton, Robin, & Xing, 2000).

IRT was originally developed in order to overcome the problems with CTT. A major part concerning the theoretical work was produced in the 1960's (Birnbaum, 1968; Lord & Novick, 1968) but the development of IRT continues (van der Linden &

Glas, 2000). IRT is theory grounded and models the probabilistic distribution of test-takers' 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. One of the basic assumptions in IRT is that the latent ability of a test-taker is independent of the content of a test. The relationship between the probability of answering an item correctly and the ability of a test-taker can be modeled in different ways depending on the nature of the test (Hambleton et al., 1991). It is common to assume unidimensionality, i.e. that the items in a test measure one dominant latent ability. According to IRT, a test-taker with high ability should have a high probability of answering an item correctly.

Another assumption is that it does not matter which items are used in order to estimate the test-takers' ability. This assumption makes it possible to compare test-takers' results despite the fact that they have taken different versions of a test (Hambleton & Swaminathan, 1985). IRT has been the preferred method in standardized testing since the development of computer programs which could perform the complicated calculations that IRT requires (van der Linden & Glas, 2000).

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.

The three-parameter logistic IRT model has three item parameters: the difficulty parameter, *b*, the discrimination parameter, *a*, and the psuedo-guessing parameter, *c*, (Birnbaum, 1968). For the three-parameter IRT model, the probability of success (x = 1) for person *j* with an ability level, q, on item *i* is denoted

$$P_{ij}(x_i = 1 \mid q_j) = c_i + (1 - c_i)\left[\frac{\exp\left(a_i\left(q_j - b_i\right)\right)}{1+\exp\left(a_i\left(q_j - b_i\right)\right)}\right] \quad (1)$$

where b_i is the difficulty parameter for item *i*, a_i is the discrimination parameter for item *i*, and *c*, is the psuedo-guessing parameter for item *i*.

The two-parameter logistic IRT model has two item parameters: the difficulty parameter, *b*, and the discrimination parameter, *a* (Birnbaum, 1968). For the two-parameter IRT model, the probability of success (x = 1) for person *j* with an ability level, q, on item *i* is denoted

$$P_{ij}(x_i = 1 \mid q_j) = \left[\frac{\exp\left(a_i\left(q_j - b_i\right)\right)}{1+\exp\left(a_i\left(q_j - b_i\right)\right)}\right] \quad (2)$$

where b_i is the difficulty parameter for item *i*, and a_i is the discrimination parameter for item *i.* The two-parameter model assumes that guessing does not exist.

The one-parameter logistic IRT model, also known as the Rasch model, estimates a person's ability based on the person's responses to items that have been calibrated for one item parameter (Rasch, 1960). It is the most parsimonious of the IRT models. The difficulty parameter, *b*, is the item parameter included in the model. For the one-parameter IRT model, the probability of success (x = 1) for person *j* with an ability level, q, on item *i* is denoted

$$P_{ij}(x_i = 1 \mid q_j) = \left[\frac{\exp(q_j - b_i)}{1 + \exp(q_j - b_i)} \right] \quad (3)$$

where b_i is the difficulty parameter for item *i*. The one-parameter model assumes that the items in the test discriminate equally well and that guessing is nonexistent.

Purpose of the Study

The purpose of this study was (1) to examine which IRT model is the most suitable for use when evaluating the Philippine Aptitude Classification Test, and (2) to use this model to compare the indices generated by classical test theory and item response theory.

More specifically, the study addressed the following questions:

1. How comparable were the CTT-based and IRT-based examinee ability estimates?
2. How comparable were the CTT-based and IRT-based item difficulty estimates?
3. How comparable were the CTT-based and IRT-based item discrimination estimates?
4. How invariant were the CTT-based and IRT-based difficulty estimates across different participant samples?
5. How invariant were the CTT-based and IRT-based item discrimination estimates across different participant samples?

Method

Sample

A sample of 1,023 third year high school students coming from 28 schools took the PACT test. There was an equal percentage of male and female examinees. The average age of the students was 15.

The high- and low-ability group samples were generated with the following sampling scheme. The high-ability group was defined as those whose scores fell within the 15th to 100th percentile range (Fan, 1998) on each of the eight (8) aptitude tests. The low-ability group was defined as those whose scores fell with the 0 to 85th percentile range on each of the eight (8) aptitude tests.

Comparability of IRT and CTT Person Statistics

The comparability of IRT- and CTT-based person statistics (ability [θ] in IRT vs. obtained score T in CTT) was assessed by correlating the [θ] and T estimates obtained from the same sample of participants. [θ] values were obtained through the IRT program of *MicroCAT* (PC Version for one-, two-, and three-parameter IRT models, respectively), and the obtained score T in CTT was simply the raw score. CTT obtained score T was correlated with IRT ability [θ] estimated through one-, two- and three-parameter IRT models. All IRT estimations were carried out using the marginal maximum likelihood (MML) method. Analyses were done for the eight aptitude factors with 30 dichotomously scored items in each test.

Comparability of IRT and CTT Item Statistics

Two types of IRT- and CTT-based item statistics were compared using correlations obtained from the same sample participants: (a) item difficulty parameter *b* (item location parameter) from IRT models with CTT item difficulty *p* value and (b) IRT item discrimination parameter *a* (item slope parameter from two- and three-parameter IRT models) with CTT item discrimination index [r_{pb}] (item-test, point-biserial correlation). The [r_{pb}] for CTT was bias correlated (i.e., the contribution of an item score to the total score was removed before calculating the [r_{pb}] for the item).

Degree of Invariance of IRT and CTT Item Statistics

The degree of invariance of item statistics was assessed by correlating item parameter estimates of two different samples within each measurement framework. The sampling plan allowed the assessment of item statistics invariance between dissimilar samples: high and low ability samples. This facilitated the assessment of the degree of invariance of item statistics for the two measurement frameworks.

In CTT, the item difficulty index *p* (*p* value) - the proportion of examinees passing an item, expresses item difficulty on an ordinal scale - not on an interval scale. This *p* value, however, can easily be transformed to an interval scale so that it is more appropriate for statistical analyses. Transformation simply requires the assumption that the underlying trait being measured by an item is normally distributed. The transformation is achieved by finding the *z* score corresponding to the (1-p) percentile from the *z* distribution. This normalization removes the curvilinearity in the relationship between two sets of item *p* values (Anastasi, 1988).

This transformation of the CTT item difficulty index has been widely used in different measurement situations, such as in Thurstone absolute scaling (Donlon, 1984; Thurstone, 1974) and in research related to item bias detection (Angoff, 1982; Cole & Moss, 1993).

In CTT, item discrimination is expressed as the item-test, Pearson product-moment correlation (point biserial correlation). Because the correlation coefficient is not linearly scaled (Hinkle, Weirsma, & Jurs, 1988), the Fisher z transformation is usually recommended before statistical analyses are applied to correlation coefficients. For this reason, in the assessment of the invariance characteristics of

the CTT item discrimination index, correlation analyses were applied to the Fisher z-transformed point-biserial between two samples of examinees.

Results and Discussions

The results of the study are discussed as responses to the five research questions presented previously. Whenever appropriate, relevant interpretation and discussion about the meaning and implications are presented together with the results. But before the results related to the research questions are presented, the question of IRT model fit is addressed.

IRT Model Fit Assessment

In any application of the IRT model, it is important to assess to what extent the IRT model assumptions are valid for the given data and how well the testing data fit the IRT model selected for use in the particular situation. The violation of IRT model assumptions, misfit between the IRT model used and the testing data, may lead to erroneous or unstable IRT model parameter estimates.

Unidimensionality is the important assumption common for all IRT models. This assumption is sometimes empirically assessed by investigating whether a dominant factor exists among all items of the test (Hambleton et al., 1991). There are several empirical tests for unidimensionality found in the literature. This study uses the eigenvalue test.

The eigenvalue test makes use of graphical representation. The eigenvalues of the factors extracted from a test are plotted against their factor ranks. The first three eigenvalues for the eight aptitude factors are given in table 1. Figure 1 contains the graphical representation of the result of the eigenvalue test for aptitude factor *verbal english*. When the first factor is very large compared to the second factor, and the magnitudes of the other factors do not vary largely from the second factor eigenvalue, an angular trace results. Such a graph indicates that the items in the test are unidimensional (Lord and Novick, 1968). However, when the trace of the eigenvalues plotted forms a curve, the items in the test are not unidimensional. Based on the results presented at table 1, it appears reasonable to conclude that the unidimensionality assumption for the IRT models holds for the data used in the study.

Table 1.
The First Three Eigenvalues of the Eight Aptitude Factors

Aptitude	1st Eigenvalue	2nd Eigenvalue	3rd Eigenvalue
Perceptual Speed	12	4	4
Verbal English	9	1	1
General Reasoning	8	2	1
Flexibility of Closure	6	1	1
Verbal Filipino	7	1	1
Spatial Closure	15	1	1
Visualization	5	2	1
Perceptual Acuity	9	3	1

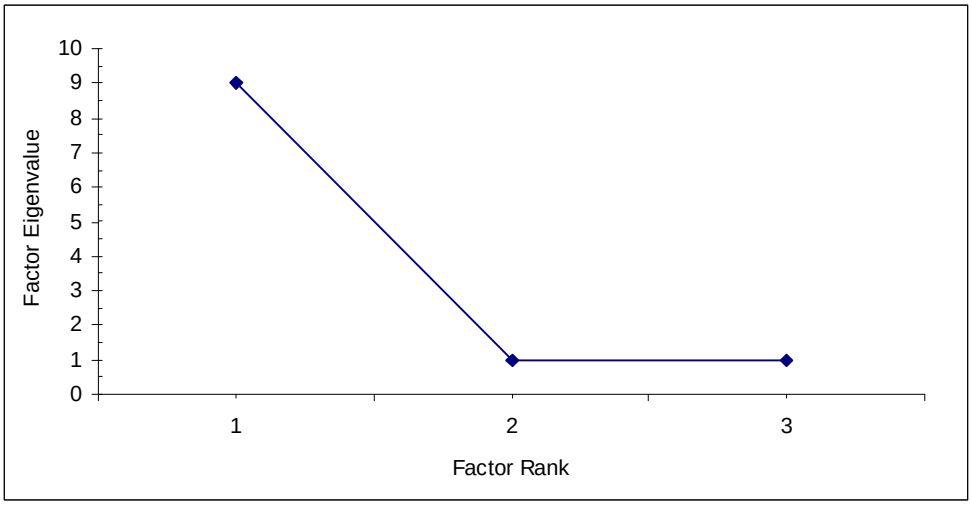

Figure 1. Plot of Factor Eigenvalues and Factor Ranks of Verbal English test.

Standardized residuals were used to assess the fit of the items to the IRT model. The residuals indicate how well the response data fit the selected IRT model (2PL or 3PL) for the item parameters estimated. The use of standardized residuals was selected to avoid the problems with the use of a chi-square item fit statistics with large samples, where many or all items are statistically but not significantly different from the predicted IRT model. The rule of thumb for using standardized residuals is that a value > 2 is considered bad or identifies a model "misfit." Table 2 summarizes the number of items identified as misfitting the given IRT model at the [Alpha] = .01 level.

Table 2.
Number of Misfitting Items Identified for the Eight Tests

Aptitude	Number of Items	IRT Models		
		1P	2P	3P
Perceptual Speed	30	18	1	0
Verbal English	30	20	14	0
General Reasoning	30	11	11	0
Flexibility of Closure	30	10	16	0
Verbal Filipino	30	16	18	0
Spatial Closure	30	14	12	0
Visualization	30	9	15	0
Perceptual Acuity	30	14	9	0

Note: IRT= Item Response Theory; 1P= one parameter; 2P= two parameter; 3P= three parameter

It is worth pointing out that the statistical test for identifying misfitting items has a lot of statistical power. Even with the powerful statistical test, there are no items identified in all the eight aptitude tests as misfitting in the three-parameter IRT model. The results indicate that the data fit the three-parameter IRT model exceptionally well. The fit of the data for the one- and two-parameter models, however, is obviously very questionable, with the number of items identified as misfitting ranging from 9 (30%) to 20 (60%). Since there is the obvious misfit

between the data and the one- and two-parameter IRT models, and because the consequences of such misfit are not entirely clear (Hambleton, 1991), the results related to the one- and two-parameter IRT models presented in later sections should be viewed with extreme caution.

1. How comparable were the CTT- and IRT-based examinee ability estimates

Table 3 presents the results for the eight aptitudes. Two steps were involved in arriving at each entry in Table 3: (a) from each sample of examinees, both CTT- and IRT-based (one-, two- and three-parameter IRT models, respectively) ability estimates were obtained; and (b) the CTT- and IRT- based ability estimates from the same sample were correlated.

Table 3.
Comparability of Person Statistics From the Two Measurement Frameworks:
Correlations Between CTT- and IRT-Based Person Ability Estimates

Sample	Test	IRT Models		
		1P	2P	3P
Total Sample				
	Perceptual Speed	.951 (.000)	.944 (.000)	.916 (.000)
	Verbal English	.992 (.000)	.988 (.000)	.846 (.000)
	General Reasoning	.970 (.000)	.965 (.000)	.806 (.000)
	Flexibility of Closure	.991 (.000)	.981 (.000)	.897 (.000)
	Verbal Filipino	.997 (.000)	.986 (.000)	.892 (.000)
	Spatial Closure	.874 (.000)	.839 (.000)	.816 (.000)
	Visualization	.989 (.000)	.986 (.000)	.885 (.000)
	Perceptual Acuity	.966 (.000)	.961 (.000)	.886 (.000)
High Ability Samples				
	Perceptual Speed	.993 (.000)	.936 (.000)	.921 (.000)
	Verbal English	.992 (.000)	.988 (.000)	.852 (.000)
	General Reasoning	.966 (.000)	.961 (.000)	.833 (.000)
	Flexibility of Closure	.982 (.000)	.981 (.000)	.893 (.000)
	Verbal Filipino	.996 (.000)	.986 (.000)	.880 (.000)
	Spatial Closure	.854 (.000)	.821 (.000)	.827 (.000)
	Visualization	.988 (.000)	.985 (.000)	.916 (.000)
	Perceptual Acuity	.960 (.000)	.957 (.000)	.910 (.000)
Low Ability Samples				
	Perceptual Speed	.987 (.000)	.976 (.000)	.909 (.000)
	Verbal English	.993 (.000)	.988 (.000)	.832 (.000)
	General Reasoning	.984 (.000)	.980 (.000)	.797 (.000)
	Flexibility of Closure	.984 (.000)	.989 (.000)	.865 (.000)
	Verbal Filipino	.884 (.000)	.986 (.000)	.882 (.000)
	Spatial Closure	.884 (.000)	.855 (.000)	.809 (.000)
	Visualization	.991 (.000)	.990 (.000)	.880 (.000)
	Perceptual Acuity	.982 (.000)	.975 (.000)	.868 (.000)

Note: Significance level of correlation is enclosed in parenthesis

Table 3 shows that the CTT- and IRT-based examinee ability estimates correlate most highly with each other for all the eight tests, for the different samples using the one-parameter IRT model, with correlations between CTT- and IRT-based ability estimates greater than .874 for all conditions. These very high correlations indicate that CTT- and one-parameter IRT-based person ability estimates are closely comparable with each other. In other words, regardless of which measurement framework we rely on, the same or very similar conclusions will be drawn regarding the ability levels of individual examinees.

However, the same strong relationship between CTT- and IRT-based ability estimates is not seen with the two-parameter model and much weaker with the three-parameter IRT and CTT-based examinee estimates, with the lowest correlation at $p = .797$. This weaker relationship could be due to the discrimination and guessing parameters, which are not present in the one-parameter model but are components of the ability estimates in the three-parameter model. Though significantly correlated at the .01 level, CTT and IRT ability relationship does not warrant the same or similar conclusions regarding the ability levels of individual examinees.

2. How comparable were the CTT- and IRT-based item difficulty estimates

Table 4 presents the results associated with comparability between CTT and IRT difficulty estimates. Again, from the same sample, CTT-based item difficulty estimates were correlated with IRT item difficulty estimates derived from IRT models (one-, two- and three-parameter IRT models).

The CTT *p* values were reversed in direction so that the higher the value, the more difficult the item. This linear reversal of *p* value direction had no statistical effect other than to make the correlations in Table 4 positive in sign.

As the tabled results indicate, the relationship between CTT- and IRT-based item difficulty estimates for all the three IRT models is almost perfect. The correlations range from .993 to 1.00, indicating strong comparability between CTT- and IRT-based item difficulty estimates.

Table 4.
Comparability of Item Statistics From the Two Measurement Frameworks:
Correlations Between CTT- and IRT-Based Item Difficulty Indexes

Sample	Test	IRT Model CTT *p* Values		
		1P	2P	3P
Total Sample				
	Perceptual Speed	.995 (.000)	.989 (.000)	.982 (.000)
	Verbal English	.999 (.000)	.988 (.000)	.961 (.000)
	General Reasoning	.999 (.000)	.987 (.000)	.951 (.000)
	Flexibility of Closure	.999 (.000)	.994 (.000)	.980 (.000)
	Verbal Filipino	.998 (.000)	.987 (.000)	.971 (.000)
	Spatial Closure	.998 (.000)	.987 (.000)	.992 (.000)
	Visualization	.999 (.000)	.986 (.000)	.933 (.000)
	Perceptual Acuity	.996 (.000)	.995 (.000)	.988 (.000)
High Ability Samples				
	Perceptual Speed	.987 (.000)	.990 (.000)	.975 (.000)
	Verbal English	1.00 (.000)	.989 (.000)	.969 (.000)
	General Reasoning	.999 (.000)	.992 (.000)	.960 (.000)
	Flexibility of Closure	.999 (.000)	.995 (.000)	.984 (.000)
	Verbal Filipino	.998 (.000)	.987 (.000)	.968 (.000)
	Spatial Closure	.997 (.000)	.984 (.000)	.990 (.000)
	Visualization	.999 (.000)	.984 (.000)	.955 (.000)
	Perceptual Acuity	.994 (.000)	.994 (.000)	.988 (.000)
Low Ability Samples				
	Perceptual Speed	.993 (.000)	.993 (.000)	.985 (.000)
	Verbal English	.999 (.000)	.988 (.000)	.960 (.000)
	General Reasoning	.999 (.000)	.988 (.000)	.940 (.000)
	Flexibility of Closure	.999 (.000)	.994 (.000)	.978 (.000)
	Verbal Filipino	.998 (.000)	.968 (.000)	.971 (.000)
	Spatial Closure	.999 (.000)	.988 (.000)	.991 (.000)
	Visualization	.999 (.000)	.987 (.000)	.932 (.000)
	Perceptual Acuity	.997 (.000)	.994 (.000)	.987 (.000)

Note: Significance level of correlation is enclosed in parenthesis

Since item difficulty parameter estimates of the one-, two- and three-parameter IRT models were almost perfectly related to the CTT-based item difficulty indices, it appears that the three IRT models provide almost the same information as CTT with regard to item difficulty. However, unless the three IRT model estimates could show superior performance in terms of invariance across different samples over the CTT item difficulty indices, the results here would not suggest that the IRT models offer empirical advantages over the much simpler CTT framework. The degree of invariance of item statistics of the two measurement frameworks will be discussed in the succeeding sections.

3. How comparable were the CTT- and IRT-based item discrimination estimates

Table 5 presents the results associated with the third research question. Each table entry is the correlation between CTT item point-biserial correlations and IRT discrimination estimates (IRT item slopes). Since the one-parameter model assumes fixed item discrimination for all items, no correlation coefficient could be computed, thus, N/A (not applicable) is entered for this column in the table.

Table 5.
Comparability of Item Statistics From the two measurement Frameworks:
Correlations Between CTT- and IRT- based Item Discrimination indexes.

Sample	Test	IRT Model CTT *p* Values		
		1P	2P	3P
Total Sample				
	Perceptual Speed	N/A	.861 (.000)	.679 (.000)
	Verbal English	N/A	.939 (.000)	.813 (.000)
	General Reasoning	N/A	.938 (.000)	.794 (.000)
	Flexibility of Closure	N/A	.902 (.000)	.569 (.001)
	Verbal Filipino	N/A	.861 (.000)	.615 (.000)
	Spatial Closure	N/A	.880 (.000)	.751 (.000)
	Visualization	N/A	.890 (.000)	.934 (.000)
	Perceptual Acuity	N/A	.868 (.000)	.470 (.009)
High Ability Samples				
	Perceptual Speed	N/A	.872 (.000)	.773 (.000)
	Verbal English	N/A	.920 (.000)	.790 (.000)
	General Reasoning	N/A	.920 (.000)	.795 (.000)
	Flexibility of Closure	N/A	.888 (.000)	.517 (.000)
	Verbal Filipino	N/A	.843 (.000)	.530 (.000)
	Spatial Closure	N/A	.850 (.000)	.727 (.000)
	Visualization	N/A	.875 (.000)	.931 (.000)
	Perceptual Acuity	N/A	.835 (.000)	.386 (.000)
Low Ability Samples				
	Perceptual Speed	N/A	.784 (.000)	.374 (.042)
	Verbal English	N/A	.948 (.000)	.792 (.000)
	General Reasoning	N/A	.913 (.000)	.760 (.000)
	Flexibility of Closure	N/A	.902 (.000)	.436 (.016)
	Verbal Filipino	N/A	.530 (.003)	.502 (.005)
	Spatial Closure	N/A	.897 (.000)	.765 (.000)
	Visualization	N/A	.856 (.000)	.912 (.000)
	Perceptual Acuity	N/A	.862 (.000)	.466 (.009)

Note: Significance level of correlation is enclosed in parenthesis

In contrast to the overwhelmingly strong relationships between CTT- and IRT-based estimates of item difficulty, the relationship between CTT and IRT item discrimination indices appear to be weaker. Majority of the correlations fell within the

range of .374 to .948. Furthermore, this relationship shows considerable variation across tests, across sampling conditions, and across the IRT models.

Although the relationship between the CTT-based and IRT-based item discrimination indices in Table 5 could be considered strong or somewhat strong under some conditions (.80s to .90s), the relationship is precariously low (.30s to .40s) in a few cases. Almost all the extremely low correlations occurred for the *Perceptual Acuity* and *Flexibility of Closure* test items under the two- and three-parameter models. However, *Perceptual Speed*, for the *Low Ability Group Sample* had the lowest correlation between CTT and IRT item discrimination (p = .374). In general, the item discrimination indices using the IRT three-parameter model correlated somewhat less with CTT point-biserials than did those using the IRT two-parameter model.

The results in table 5 show that the item discrimination indices from the CTT and IRT frameworks tend to be less comparable than the person ability estimates and the item difficulty estimates presented previously. The lower comparability between the discrimination indices derived from CTT and IRT implies that, in some cases, CTT and IRT may yield noticeable discrepancies with regard to which items have more discrimination power, which, in turn, may lead to the selection of different items for a test, depending on which framework is used in the estimation of item discrimination.

Up to this time, we have solely focused on the question of comparability between estimates derived from the two measurement frameworks. Low comparability between item discrimination indices of CTT and IRT in some cases does not suggest which measurement framework provides more stable, or more invariant, item parameter estimates across different samples. To understand the invariance characteristics of the item statistics of the two measurement frameworks, we turn now to Research Questions 4 and 5.

4. How invariant were the CTT- and IRT-based item difficulty estimates across different participant samples

The fourth research question addresses one crucial question about CTT and IRT. As discussed previously, the assumption of item parameter invariance across different participant samples has played the most important role in the development and application of IRT models.

Table 6 presents the results for this research question. Notice that the correlations in this table (and, similarly, in Table 7) are correlations between item estimates from two different samples derived from the same measurement framework. It is important to note that the invariance property of item parameters can only be investigated by administering the same items to different samples and then comparing item parameter estimates obtained across samples.

Table 6.
Invariance of Item Statistics From the Two Measurement Frameworks:
Between-Sample Correlations of CTT and IRT Item Difficulty Indexes

CTT: Invariance across		P Value		
High-low ability samples				
	Perceptual Speed	.993 (.000)		
	Verbal English	.990 (.000)		
	General Reasoning	.990 (.000)		
	Flexibility of Closure	.995 (.000)		
	Verbal Filipino	.994 (.000)		
	Spatial Closure	.996 (.000)		
	Visualization	.982 (.000)		
	Perceptual Acuity	.994 (.000)		
IRT: Invariance Across		**1P**	**2P**	**3P**
High-low ability samples				
	Perceptual Speed	.995 (.000)	.985 (.000)	.983 (.000)
	Verbal English	.990 (.000)	.994 (.000)	.994 (.000)
	General Reasoning	.990 (.000)	.994 (.000)	.995 (.000)
	Flexibility of Closure	.995 (.000)	.996 (.000)	.996 (.000)
	Verbal Filipino	.994 (.000)	.995 (.000)	.997 (.000)
	Spatial Closure	.997 (.000)	.998 (.000)	.998 (.000)
	Visualization	.983 (.000)	.984 (.000)	.976 (.000)
	Perceptual Acuity	.996 (.000)	.997 (.000)	.998 (.000)

Note: Significance level of correlation is enclosed in parenthesis

A comparison of the CTT with IRT entries shows that CTT item difficulty estimates are closely comparable with IRT item difficulty estimates in terms of their invariance properties. This is indicated by the high between-sample correlations coefficient of item difficulty estimates. If there is any trend at all, it appears that the IRT two- and three-parameter item difficulty estimates are slightly more invariant than CTT item difficulty estimates in almost all conditions. The between-sample correlations of *p* values appear to be slightly higher than the between-sample correlations of CTT location parameters in seven (7) out of eight (8) and six (6) out of the eight (8) aptitude tests, respectively. This empirical observation about the invariance property of the item difficulty indices of the two measurement frameworks supports the argument in favor of the IRT framework with respect to invariance.

5. How invariant were the CTT- and IRT-based item discrimination indices across different participant samples

Table 7 presents the results of the correlation analyses of the CTT and IRT item discrimination indices. As explained earlier, because the IRT one-parameter (Rasch) model does not provide item discrimination estimates for individual items, and instead assumes a fixed item discrimination for all items, no correlations could be computed for the one-parameter model. Hence, N/A is listed under the one-parameter IRT column in the table. It should be reiterated that each table entry is the correlation of point-biserial of CTT between two samples or the correlation of item slopes of IRT between two samples.

Table 7.
Invariance of Item Statistics From the Two Measurement Frameworks:
Between-Sample Correlations of CTT and IRT Item Discrimination Indexes

CTT: Invariance across		P Value
High-low ability samples		
	Perceptual Speed	.039 (.838)
	Verbal English	.970 (.000)
	General Reasoning	.950 (.000)
	Flexibility of Closure	.971 (.000)
	Verbal Filipino	.993 (.000)
	Spatial Closure	.966 (.000)
	Visualization	.970 (.000)
	Perceptual Acuity	.973 (.000)

IRT: Invariance Across		1P	2P	3P
High-low ability samples				
	Perceptual Speed	N/A	.108 (.570)	.792 (.000)
	Verbal English	N/A	.976 (.000)	.977 (.000)
	General Reasoning	N/A	.956 (.000)	.956 (.000)
	Flexibility of Closure	N/A	.932 (.000)	.962 (.000)
	Verbal Filipino	N/A	.987 (.000)	.970 (.000)
	Spatial Closure	N/A	.970 (.000)	.944 (.000)
	Visualization	N/A	.968 (.000)	.978 (.000)
	Perceptual Acuity	N/A	.969 (.000)	.913 (.000)

Note: Significance level of correlation is enclosed in parenthesis

The item discrimination indices of both CTT and IRT were to some extent less invariant across participant samples than the item difficulty indices presented in Table 6. This result parallels what was observed about comparability between CTT and IRT item statistics in Tables 4 and 5. Also, with higher correlations of CTT point-biserials in some cases and higher correlations of IRT item slopes in others, no systematic advantage of one framework over another is obvious. In most cases, the between-sample correlations of item discrimination indices of CTT and those of IRT were highly comparable (.90s), indicating reasonable invariance across samples.

The CTT- and IRT-based discrimination correlation estimates across samples for *Perceptual Speed* are quite interesting to note. It is in this test where CTT-based discrimination correlation index across different samples was not significantly correlated implying considerable variation across high and low performing samples. Among the different measurement frameworks, it was only in the three-parameter model where invariance across different samples for discrimination estimates was observed.

Summary and Conclusion

The present study empirically examined the behavior of item and person statistics obtained from the CTT and IRT measurement frameworks. The study focused on two main issues: (a) How comparable are the item and person statistics from CTT with those from IRT?, and (b) How invariant are the CTT item statistics and the IRT item statistics, respectively? The test item pool was composed of eight aptitude tests with 30 dichotomously scored items in each and the participant pool had more than a thousand examinees who took the 240 item test.

The major findings were as follows:

1. The person statistics (examinee ability estimates) from CTT were comparable with those from IRT for all three IRT models with correlation coefficients ranging from 0.797 to 0.997.

2. The item difficulty indices from CTT were highly comparable with those from all IRT models with correlation coefficients ranging from .932 to 1.00.

3. Compared with item difficulty indices, the item discrimination from CTT were somewhat less comparable with those from IRT. Although under majority of the conditions, the comparability was moderately high to high (.813 - .948), there were a few cases where the comparability was very low (.30s).

4. Both CTT and IRT item difficulty indices exhibited very high invariance across samples. The degree of invariance of the IRT two- and three-parameter item difficulty indices were slightly better than that of CTT item difficulty parameter estimates.

5. Both the CTT and IRT item discrimination estimates were somewhat less invariant than their item difficulty estimates. The degree of invariance of CTT item discrimination estimates was highly comparable with that of IRT item discrimination estimates except in *Perceptual Speed* where the correlation coefficient of CTT item discrimination estimates between high and low group was very low ($p = .039$).

Overall, the findings from this empirical investigation failed to completely discredit the CTT framework with regard to the alleged inability to produce person-invariant item statistics; however, the findings likewise supported the IRT framework for its superiority over CTT in producing person-invariant item statistics.

The results suggest that the three-parameter model is, in general, to be preferred over the two- and one-parameter models. The three-parameter model produced better item and person statistics both in terms of the comparability of item and person statistics between the two frameworks, and in terms of the degree of invariance of item statistics between the two competing measurement frameworks.

The test evaluated, the Philippine Aptitude Classification Test, is a multiple-choice item test which attempts to predict a student's probable performance in various courses of study. It measures eight aptitude factors that have been found to

be useful in the classification of students into different fields of study. Though the PACT results may be interpreted more effectively in light of other information such as family background, educational and socio-economic background, etc., its validity and reliability should be unquestionable. Based on the findings, the three-parameter IRT appears to be the most suitable model for the PACT, making it a more appropriate instrument in helping students define their career options.

Of course, the present study, like many other research studies, has its share of limitations that may potentially undermine the validity of its findings. One shortcoming could be the limited examinee pool used in the study. Ideally, the examinee pool should be large in the sense that a variety of different samples can be drawn from it. Although, the test item pool is large and more diverse in terms of item characteristics, the same cannot be said about the examinee pool. Future studies may benefit from using a larger examinee database from which a variety of different samples can be drawn.

Appendix A

Philippine Aptitude Classification Test:
Subtest Composition, Number of Items,
Name of Aptitude/Factor Scores and Reliability Estimates

Subtest	No. of Items	Aptitude/Factor Score	r_{tt}
Part I			
1. Matching Letters/Numbers	15	1. Perceptual Speed	.879
2. Form Matching	15		
Part II			
1. Vocabulary	15	2. Verbal English	.855
2. Analogies	15		
3. Numeric	10	3. General Reasoning	.844
4. Number Series	10		
5. Figural Reasoning	10		
6. Paper Form Board	15	4. Flexibility of Closure	.790
7. Hidden Figure	15		
8. Talasalitaan	15	5. Verbal Filipino	.775
9. Mga Salitang Magkaugnay	15		
10. Hidden Blocks I	15	6. Spatial Closure	.927
11. Hidden Blocks II	15		
12. Patterns	10	7. Visualization	.765
13. Mechanical Motion	10		
14. Assembly	10		
15. Figure Series	15	8. Perceptual Acuity	.848
16. Proofreading	15		
T O T A L	240		

Test Length: 240 Items
Testing Time: Part I – 4 minutes
Part II – 1 hour and 45 minutes
Intended User: Second year high school students, but not lower. May be administered to the same purpose to students in the higher levels up to at most first year college. The test is most recommended for third year high school students.

References

Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York: Macmillan.

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: John Hopkins University Press.

Assessment Systems Corporation. (1996). MicroCAT User's Manual. ST. Paul. Minnesota, USA.

Birnbaum, A. (1968). In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Cole, N. S., & Moss, P. A. (1993). Bias in test use. In R. L. Linn (Ed.), *Educational Measurement*. (3rd ed., pp. 201-219). Pheonix, AZ: Oryx Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston, Inc.

Donlon, G. (1984). *The college board technical handbook for the Scholastic Aptitude Test and Achievement test.* New York: College Entrance Examination Board.

Fan, Xitao. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and psychological measurement: Issues and Practice.* 58 (3), 357-382.

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., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H. Tinsley & S. Brown (Eds.), *Handbook of applied multivariate statistics and modeling.* San Diego, CA: Academic Press.

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.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1988) *Applied statistics for the behavioral sciences.* (2nd ed.). Boston: Houghton Mifflin.

Iledan, B. R. & Franco, M. L. (2003). Factor structure of the revised PACT. *Test development technical report.* CEM TDD TR-0603002. Center for Educational Measurement, Inc., Makati City.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Test Development Division. (2005). *Scaling and equating of the two parallel forms of the revised PACT.* CEM TDD TR-0105006. Center for Educational Measurement, Inc., Makati City.

Test Development Division. (2006). *Development of the PACT course norms.* CEM TDD. Center for Educational Measurement, Inc., Makati City.

Test Development Division. (2007). *Philippine Aptitude Classification Test. Score interpretation guide.* Center for Educational Measurement, Inc., Makati City

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Thurstone, L. L. (1947). The calibration of test items. *American Psychologist*, 2, 103-104.

Van der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: theory and practice.* Dordrecht: Kluwer Academic Publisher.