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

Royce Hernandez
De La Salle-College of Saint Benilde

Abstract

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

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

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

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

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

Participants

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

Materials

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

Procedure

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

Data Analysis

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

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

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

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

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

Table 1
Mean and Standard Deviation

<table>
<thead>
<tr>
<th></th>
<th>CTT</th>
<th>IRT Difficulty</th>
<th>IRT Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diff</td>
<td>Disc</td>
<td>1-pl</td>
</tr>
<tr>
<td>Verbal</td>
<td>Mean</td>
<td>.468</td>
<td>.242</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>.226</td>
<td>.148</td>
</tr>
<tr>
<td>Non-Verbal</td>
<td>Mean</td>
<td>.524</td>
<td>.213</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>.269</td>
<td>.165</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Number of Items</th>
<th>Verbal Test</th>
<th>Non-Verbal Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff</td>
<td>1-pl</td>
<td>2-pl</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>.857**</td>
</tr>
<tr>
<td>Disc</td>
<td>16</td>
<td>NA</td>
</tr>
</tbody>
</table>

** p<.01

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

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

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

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

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

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

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

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

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

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

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

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

In addition, the paper was able to establish that CTT and IRT can be used independently or altogether to describe the nature of the items. Test developers can bank on time-tested CTT methods to establish item difficulty and item discrimination characteristic of items. In the absence...
of sophisticated software and a big sample to derive IRT parameters, the test developer can be theoretically assured of congruence of test item difficulty and discrimination for both methods provided enough sampling is done in CTT (between 200 to 500; CTT requires 200 minimum while IRT is 500 to 1000 N)). On the other hand, those employing IRT, whether 2- or 3-parameter, would also be able to infer congruence of CTT and IRT item characteristics, provided goodness of fit of the data is established. What’s important is the emphasis or need for the pseudo-guessing parameter if it is needed in the analysis.

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

References


