MELTA

Malaysian Journal of ELT Research

Vol. 2, March 2006

Classical And Rasch Analyses Of Dichotomously Scored Reading Comprehension Test Items

Ainol Madziah Zubairi
Noor Lide Abu Kassim

International Islamic University Malaysia

Abstract

This study demonstrates the use of both the classical and the Rasch Model item analyses in an attempt to evaluate the quality of reading test items used in an English Language placement test context. Thirty-five dichotomously scored items from the reading comprehension section of the placement test administered to the Matriculation Centre students of the International Islamic University Malaysia (IIUM) were subjected to both the Classical and Rasch Model analyses using statistical packages appropriate for the two analyses *(SPSS and BIGSTEPS).* The discriminant and difficulty indices of the items and the classification of the items according to their item characteristics based on Classical Test Theory are reported along with the results of the Rasch analysis of the same set of items. The characteristics of the items based on the two analyses are compared in an attempt to demonstrate the amount of information gained and the usefulness of the more powerful Rasch Model analysis in investigating item quality and test reliability. Given the type of decisions made on the basis of test scores in high-stakes tests, this study demonstrates the need to apply not just the classical approach, but also the need to incorporate modern measurement theory in the evaluation of high-stakes tests.

INTRODUCTION

A major concern in test construction is ensuring the reliability of test items, and one typical step in investigating reliability has been classical theory (CT) item analysis. The classical item analysis essentially determines test homogeneity. That is, the more similar the items in a given test, the more likely they measure the same kind of intended ability and therefore, the higher the reliability (Davies, 1990). However, CT item analyses statistics do not provide the necessary information on how examinees at different ability levels on the latent trait measured have performed on an item (Crocker and Algina, 1986). Therefore, a more robust statistics based on modern test theory has also been widely used in more recent test reliability investigations.

CT Item Difficulty

Item difficulty is the first item characteristic in classical theory to be determined. This is a common practice as tests are often rejected as reliable measures of examinee performance due to the misfit of item difficulty to the ability of the examinees (Bachman, 1990). When tests are too easy or too difficult, the scoring distribution will tend to unnaturally concentrate at one end of the continuum (Henning, 1987). As a result, it is difficult to distinguish candidates' ability at the concentrated end. This, inevitably, also results in loss of person separability or reliability (Henning, 1987).

CT Item Discrimination

In addition to item difficulty, item discriminability is important in a reliability study. It provides information on how effectively the items in a given test discriminate between examinees who are relatively high in the ability measured and those who are relatively low. This information is useful as many tests are intended to provide information about individual differences on the ability that the tests aim to measure. For example, in a language placement test where the major decision to be made is whether to exempt students from language skill support classes or not, the items in the test should discriminate well between those who have the necessary language skills and those who do not. On the other hand, in achievement testing, where the aim is to assess how much learners have gained from the lessons provided, the items in the test should differentiate between students who know the material and those who do not.

One of the ways to obtain discrimination indices is by sample separation which essentially involves a computation procedure that separates the highest scoring group and the lowest scoring group from the entire sample on the basis of the total score in a test (Gronlund, 1976 and Henning, 1987). This will produce discrimination indices that range from zero to one. Ebel's (1972) criterion for item revision and test evaluation is widely referred to (see, for example, Gronlund, 1976; Crocker and Algina, 1986; Wong

et al, 1990 and Brown, 1996). The list of criteria (see below) is based on a range of DI that categorically defines the items in a test:

1. If DI ≥ 0.40, the item is functioning quite satisfactory.
2. If 0.30≤DI≤0.39, little or no revision is required.
3. If 0.20 ≤DI≤0.29, the item is marginal and needs revision.
4. If DI ≤ 0.19, the items should be eliminated or completely revised.

The list gives guidelines and possible steps that can be taken in the different categories. For example, the guideline suggests that an item having a DI of 0.35 needs little or no revision. Instead of excluding the 'weak' item and constructing a totally new one, test developers may, on the other hand, inspect the quality of the items again and do the necessary revisions. This information can be very useful in improving reliability. Instead of constructing new items for the test, it is more feasible and practical to revise the item as construction of new items has been shown to require a longer time than revision of existing ones (Crocker and Algina, 1986: 327).

Rasch Model Item Analysis

The one parameter or Rasch Model is one of the three models of the item response theory (sometimes called latent trait measurement) advanced by psychometricians as a new measurement system to address the limitations of CT measurement (Bachman, 1990, Hambleton et al, 1991 and McNamara, 1996). One key difference between the Rasch Model and CT is that Rasch analysis is probabilistic in nature (Hambleton, 1989, Henning, 1987 and McNamara, 1996). Items and persons parameters in Rasch analysis are estimated according to the '*probability or likelihood of their response patterns given the person ability and item difficulty*' (Henning, 1987). The underlying theory behind the Rasch Model is how examinees at different levels of ability for a particular trait should respond to an item (Crocker and Algina, 1986). Rasch analysis attempts to model the relationship between two common test facets: the ability of the candidates and the difficulty of the items (McNamara, 1996). Both the facets are jointly estimated using a mathematical procedure (called maximum likelihood) based on the probabilistic patterns of responses of all examinees to all items (McNamara, 1996 and Crocker and Algina, 1986). The estimations of item difficulty and person abilities are both expressed according to a common interval scale: the logit scale (McNamara, 1996).

Requirements For Rasch Analysis

Before the relevant aspects of Rasch measurement is further discussed, it is necessary to look at the three requirements for Rasch Model analysis:

Sample Size

The guidelines for sample sizes for proper parameter estimations in item response theory (IRT) vary (Hambleton, 1989). Rasch analysis however, requires fewer samples than the other two models. This is one of the reasons for the widespread use of the Rasch Model. Wright and Stone's (1979) recommendations of a minimum of 20 items and a sample size of 200 examinees are often referred to as a sample size for studies involving the Rasch Model.

Unidimensionality

Another requirement for Rasch analysis (and for all IRT models in general) is that it assumes the presence of a dominant ability or trait that influences test performance-unidimensionality (Hambleton et. al., 1991). Unidimensionality is not a new concept in language testing and in CT item analysis. For example, the same assumption is made in the practice of summing up scores from different parts of tests or across different items in a given test (Henning, 1992).

Locally Independent Items

The next assumption in the application of Rasch analysis is that the item responses of a given examinee on a given test should be statistically independent. This simply means that a person's response to one item should not affect his/her responses on other items in the test. This, therefore, requires that '*the content of one item must not provide any clues to the answer to another item*' (Hambleton and Swaminathan, 1985: 23).

Basic Concepts Of The Rasch Model

The first three concepts of Rasch analysis, which are relevant and useful in testing contexts, particularly in this research are: the estimates of item difficulty; person ability; and the relationship between item difficulty and person ability. Where relevant, these parameters are compared with the parallel parameters in CT in the discussions that follow.

Rasch item difficulty is analogous to CT item difficulty. However, the estimation procedures are essentially different. While CT item difficulty estimation of the dichotomously scored item is presented in terms of the proportion of the candidates getting an item correct (Heaton, 1979 and Henning, 1987), estimates of item difficulty in Rasch measurement theory is estimated from the responses of a set of candidates, by taking into account the ability of the candidates and the degree of match between the ability of the group and the difficulty of the items (McNamara, 1996). Rasch estimates of item difficulty are therefore expressed as the probability that a person of a given ability

will have a 50% chance of getting the item correct. Conventionally, the average difficulty of items in a test is set at 0 (zero) logit. Thus, items of above average difficulty will be at the positive end of the scale, while those items of below average difficulty will be at the negative end (Hambleton et al., 1991 and McNamara, 1996). Another important difference between CT and the Rasch item parameter is that no item parameter in Rasch corresponds to the CT item discrimination index. This is because the Rasch analysis assumes that all items are equally discriminating (Bachman, 1990; McNamara, 1996 and Baker, 1997).

Rasch estimates of person ability are equivalent to scoring in a test (McNamara: 1996). However, unlike in CT where ability estimates are limited to the particular characteristics of the test (their difficulty, for example), Rasch estimates of ability are based on the candidate's performance on a set of items, after giving allowance to the difficulty of the items and how well they match the candidate's ability level (Henning, 1987 and McNamara, 1996). Instead of representing total scores in a test (as in CT approach), a person's ability in Rasch analysis is defined as the probability of a person having a 50% chance of getting an item of a given difficulty right. Ability values (also known as theta values) can take both the positive and negative values in the logit scale, where negative scores represent the less able, and positive scores the more able (Wood and Baker, 1985 and McNamara, 1996). A computer programme that runs the Rasch analysis, such as *BIGSTEPS*, will produce the Rasch estimate of ability (theta), the corresponding CT ability (correct number score), and the standard error of measurement associated with each theta value.

One of the most important features of the Rasch approach is that students' scores and item difficulty are transformed onto one scale so that they are related to each other (Alderson et al., 1995 and McNamara, 1996). This allows item difficulty and person ability for a group of examinees on a group of items to be directly compared. This facility is known as 'mapping', where estimates of person ability and item difficulty are represented graphically in the form of an item-by-person map. The most basic usage of the mapping facility is in comparing the range of difficulty of the group of items with the ability of the group of students. Since both the items and persons are represented graphically on the same logit scale, it is possible to see if the items fit the ability of the students. That is, it is possible to determine whether the test given to the group of students is too easy or too difficult.

Secondly, it is also possible to find out whether the distribution of items along the difficulty continuum is appropriate and sufficient. For example, if a language test is used for exemption purposes (to exempt students from any language support course), discrimination will need to be at the highest ability levels. This is because the purpose of the test is to carefully select students who are at the advanced level. Thus, only a small proportion of the examinees are expected to be exempted. We would, therefore, expect to see a sufficient number of items at the difficult end of the continuum. In contrast, in a context where the language test is used to upgrade students from one level to another, the

mapping can be used to see if the there are sufficient items around the decision point or the threshold level.

Two other aspects of Rasch measurement that are very useful are the evaluation of fit (of persons and items) and the test characteristic curve (TCC). The first, the evaluation of fit, is yet another essential part of the application of Rasch analysis to test evaluations. The evaluation of fit concerns the degree of correspondence between what is predicted and what is observed (in the model of relationship between person ability and item difficulty) (McNamara, 1996 and Baker, 1997). Therefore, it provides two essential pieces of information about the item and difficulty parameters: person fit and item fit. This evaluation is essential as it can provide a check on the adequacy of the application of the model for the set of data used (Baker, 1997). In addition, when the goodness of fit between the model and data has been ascertained, the investigation of fit helps to identify test takers and items that do not fit the model.

Properties Of Rasch Analysis Useful In A Reliability Study

Having provided an overview and brief descriptions of the features of Rasch measurement in the above sections, it is necessary to look at how these features are useful in this research. As Rasch analysis is utilised along with CT analysis in the reliability study of the reading test items in this research, it is necessary to discuss how the two approaches can complement each other in a reliability investigation.

Firstly, as reliability concerns investigations of possible errors that may affect test scores and the decisions made based on test scores, like in CT analysis, investigations of Rasch item and ability parameters are useful in a reliability study. Useful item parameters in Rasch analysis include both difficulty estimates and the item fit statistics. As in CT item analysis, difficulty estimates can be used to indicate whether the items are of an appropriate level for the group of students. However, in contrast to CT analysis where item difficulty indices are sample dependent, Rasch estimates of item difficulty are independent of the particular sample of individuals whose responses are used in the estimate (Bachman, 1990; McNamara, 1996 and Alderson et al., 1995). Therefore, estimates of the item difficulty based on Rasch analysis can be generalised across different test takers. This is a very useful property for practical applications in testing such as item banking and test equating.

The item fit statistics in Rasch analysis offer information about items that do not contribute much to the reliability of the test. If an item is found to be misfitting in Rasch analysis it may be (1) an indication of flawed item construction, or (2) an indication that the item is tapping some other ability other than the one measured by the test. In this sense, misfitting items (as in CT item analysis) are 'problematic' items (for example, in CT problematic items can be due to poor discrimination). Therefore, information on misfitting items provides yet another means of detecting poor items that may not contribute much to the total score, thus affecting the test score reliability.

Secondly, another feature of the Rasch Model that is useful in this research is the mapping facility. The first advantage is it is useful in making judgements about the suitability of the items (in terms of difficulty) for the group of students. The mapping facility can be very useful in an inspection and identification of items in the different positions along the common ability/difficulty continuum. This facility is of great importance in a study of a high-stakes test (such as a placement test) as it can help to identify whether there are enough items at the important points of the continuum (cut-off point, for example).

CONTEXT OF STUDY

At the International Islamic University, Malaysia, a tailor-made standardized English Language Placement Test (EPT) is administered to all incoming students to ascertain the level of their English language ability, and to place them at the appropriate language support courses for remediation. The placement test consists of two papers. Paper one consists of four sections of 4-option multiple-choice items assessing reading and grammatical abilities. Paper 2, on the other hand, consists of an essay question which aims to assess students' writing skill.

As the purpose of the EPT is to place students at different levels of ability, the items in the test are expected to discriminate well between students in terms of their level of English language proficiency. Therefore, as part of test evaluation practice, it is vital to evaluate the quality of the test items in terms of (1) how well they are able to make distinctions between levels of proficiency and (2) how far they are measuring the same trait, i.e., English language proficiency.

METHODOLOGY

The Reading Comprehension Test Items

The 35 reading comprehension test items used in this study were taken from the EPT used in the 2003 test administration at the Matriculation Centre of the international Islamic University Malaysia. The sub skills assessed by the test items include:

- Deducing the meaning and use of unfamiliar lexical items.
- Understanding relations between parts of a text through lexical cohesive devices.
- Understanding information that is explicitly stated.
- Understanding information when not explicitly stated, through figurative language.
- Understanding information: inference.
- Transcoding information in diagrammatic display involving completing a diagram/ tables/ graph.
- Distinguishing the main idea from supporting details
- Selective extraction of relevant points from a text to summarize similar information.
- Evaluating and challenging evidence (evaluating difference between fact and opinion).

Data

The data used in the two analyses were the dichotomous responses of 2485 students on the reading comprehension test items. Most of these students completed form five in the Malaysian School System and attained first grade in the Malaysian School Certificate, a high-stakes national level standardized test, which is used for certification purposes and selection into tertiary educational institutions.:

Analyses:

CTT

The item characteristics that were examined in the classical item analyses were the difficulty and the discrimination indices. Students' responses to the test items were analysed using the *SPSS* statistical package in order to produce the indices. After the indices were produced, the next step was to decide on the criteria in order to categorically describe the items based on the difficulty and discrimination indices.
Ebel's (1972: 399) criteria and guidelines for categorizing discrimination indices is a widely quoted set of guidelines and therefore is used in this study to characterize the 35 reading test items. However, the difficulty indices analysis uses Henning's (1987: 52-54) guidelines. A summary of the criteria for describing the items based on the three classical item indices is given in the table below.

Table 1: Criteria For Defining Items Based On Classical Indices

Item difficulty	Item Discrimination
ID ≥ 0.67, too easy.	If DI ≥ 0.40, the item is functioning quite satisfactorily.
ID ≥ 0.33, too difficult.	If 0.30 ≤ DI ≥ 0.39, little or no revision is required.
	If 0.20 ≤ DI ≥ 0.29, the item is marginal and needs revision**.**
	If DI **≤** 0.19, the item should be eliminated or completely revised.

Rasch Model

The one parameter model of the IRT was chosen for the analysis of the reading test because the research was only concerned with the item difficulty and person ability parameters. The results based on Rasch analysis were divided into two areas. The first concerns the items in the test. Two steps were taken: (1) identifying the range of difficulty measures of the items and (2) examining the item fit statistics (mean square and standardized infit and outfit statistics). The recommended guideline in categorizing items based on fit statistics is summarized in Appendix 1.

The results of item analyses using the Rasch Model analysis were later compared with those of the classical method. The purpose of the comparison was to see if the same items were identified by the two methods. Based on this comparison, the characteristics of 'problematic' items were investigated.

RESULTS

The results are divided into two sections. The first reports on the CT item analysis and the second on the Rasch Model item analysis.

CTT Analysis

CT Item Difficulty: How Do The Items Spread In Terms Of Their Difficulty Values?

The results indicate that the difficulty level of the items ranged from 0.12 to 0.95 (A summary of the CT item analysis is given in Appendix 2). However, the 35 reading items can further be categorized as follows:

Table 2: Categories Of Items Based On CT Item Difficulty Values

	EASY ITEMS (ID ≥ 0.67)	**ITEMS OF AVERAGE DIFFICULTY**	**DIFFICULT ITEMS (ID ≥ 0.33)**
ITEMS	51, 58, 60, 61, 62, 64, 65, 69, 75, 81, 82	52, 56, 57, 63, 66, 67, 68, 70, 71, 73, 79, 83, 85	53, 54, 55, 59, 82, 74, 76, 77, 78, 84, 85
TOTAL	11	13	11

There seems to be an even distribution of easy as well as difficult items in the reading section of the placement test.

Discrimination Index: How Good Are The Items?

The items were further characterized based on Ebel's criteria. A summary of the categories is given in Table 3.

Table 3: Categories Of Items Based On CT Item Indices Based On Ebel's (1972) Guidelines

	Category 1 **DI ≤ 0.19**	Category 2 **0.20 ≤ DI ≥ 0.29,**	Category 3 **0.30 ≤ DI ≥ 0.39**	Category 4 **DI ≥ 0.40**
Items	51, 53, 55, 59, 60, 72, 74, 76, 77, 83, 84	54, 58, 62, 66, 73, 78, 85	52, 56, 57, 61, 65, 69, 71, 75, 79, 80, 81, 82	63, 64, 67, 68, 70
TOTAL	11	7	12	5

These results indicate that about one third of the reading test items are weak in discriminating between the 'good' and 'weak' students. If these items are to be used for future purposes they need to be looked at very closely as they may need considerable revision, if they are not eliminated altogether.

Rasch Model Analysis

Item Difficulty: How Do The Items Spread In Relation To The Abilities Of The Students Who Took The Test?

Parallel to CT analysis of item difficulty, Rasch analysis also produces the difficulty values of each of the reading test item. However, instead of looking at the range and categorizing the items based on the difficulty measures, Rasch analysis provides a better view of what is more relevant to what we would like to find out: How do the items spread along the ability range of the group of students they are given to? The mapping facility provides the answer to the question (see Figure 1).

As mentioned earlier the mapping facility is very useful as it allows for the examination of item and person distributions. From the map it is evident that a large number of items can be found along the continuum on which the majority of students' abilities fall. However, it is also interesting to find that there are items at the difficult and easy ends where a minimal number of students could be found.

Fit Statistics: How Good Are The Items?

Two types of fit statistics are used in the analysis. The first is the mean square value. Based on expert's guidelines in diagnosing misfitting items, items with mean square values of between 0.5 – 1.5 are considered productive for measurement. It was found that all the items fall within this range (e.g. Wright & Stone, 1979; Linacre, 2002).

However, we further referred to the guidelines that specify the standardized fit statistics (zstd). The proposed guideline by Bond and Fox (2001) was followed. The commonly accepted interpretation of *z* values, infit and outfit *z* values of greater than +2 or less than -2 generally are interpreted as having less compatibility with the model than expected ($p < .05$). Negative values indicate less variation than modelled: The response string is closer to the Guttman-style response string (all easy items correct then all difficult items incorrect). Positive values indicate more variation than modelled: The response string is more haphazard than expected.

Based on this criterion, a lot more items were identified as 'problematic' as compared with what was found in the CT analysis. The only good items found are items 62, 63, 65, 66, 67, 69, 70, 78, 81, 82 and 85. A summary of the 'problematic' items as identified from the standardized fit statistics (of both the infit and outfit values) is given in Table 4.

Table 4: Misfitting Items Based On Rasch Fit Statistics

Item No	51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 64, 68, 71, 72, 73, 74, 75, 76, 77, 79, 80, 83, 84
TOTAL	24 items

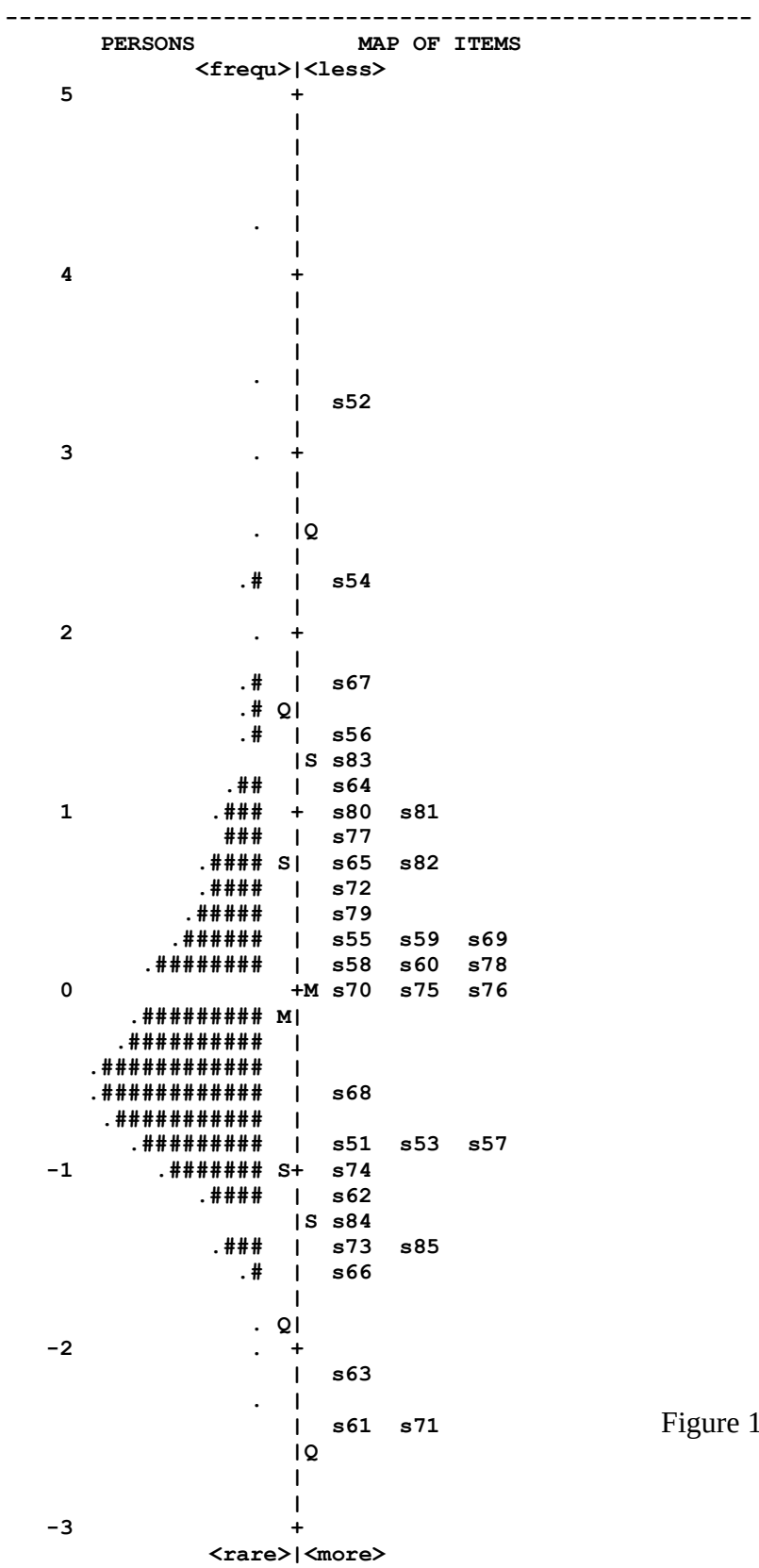

Figure 1: Item By Person Mapping

As the items are found to be misfitting as a result of unexpected and haphazard patterns of responses, it is necessary that response patterns of candidates or person fit is checked before any conclusions of item fit can be drawn conclusively. Routine and explicit tests of person fit are recommended procedures in evaluating test items (Wright & Stone, 1979).

Item 64 is one example of how to explain why an item is identified as 'problematic' by Rasch analysis. The most unexpected response output as shown in Figure 2 (item string no 14) indicates that there were some students whose abilities were higher (Person id: 570 and 1064) than the difficulty of the item (1.11 logit) got the item wrong, but students whose ability was lower than the difficulty level of the item (Person id: 2432, 80, 1641, 1945, 2105 and 462) got it right.

Comparison Between CT And Rasch Analysis

It was initially found in CT item analysis that eleven of the reading comprehension items were 'problematic' in that the DI values were lower than 0.19. However, besides these eleven items, the Rasch analysis also found 13 other items (see Table 5) that possibly were 'problematic' and therefore could contribute to the reliability of the test.

Table 5: Comparison Between 'Problematic' Items By CT And Rasch

CT 'problematic' items **DI ≤ 0.19**	Rasch analysis 'problematic' items
51, 53, 55, 59, 60, 72, 74, 76, 77, 83, 84	51, **52**, 53, **54**, 55, **56, 57, 58,** 59, 60, **61, 64, 68, 71,** 72, **73**, 74, **75,** 76, 77, **79, 80**, 83, 84
11 items	24 items

```
MOST UNEXPECTED RESPONSES
PERSON      MEASURE  | ITEM
                     |121132312   1222 21  122132331 3 1
                     |113653424731860588095992527104 36742
                   high-----------------------------------
  326 20400    3.49 Y|........0.........................
 1302 20475    3.49 M|........0..........0..............
  570 20220    2.98 Q|....0.....................0...0.....
   19 22690    2.60  |....0.....................0.........
  291 23472    2.60  |...0.....................0...0.......
 1064 22968    2.60  |........0..............0.....0.....
 1087 20599    2.60  |....0..............0..............
  613 23030    2.29 W|...0..0.................0...........
 1276 20243    1.79 H|..0.00.......00.....................
  116 21183    1.57 P|..00...0.....0.....................1
 1587 20833    -.23  |0.............................1..1
 2253 20148    -.23  |..................................11
   31 22163    -.38  |...................................1
   77 22108    -.38  |.............................1.....1
  493 22094    -.38  |...................................1
  495 22425    -.38  |.............................1.....1
 1590 22578    -.38  |.0..................................1
 1673 22729    -.38  |...................................1
 2138 20745    -.38  |.............................1......1
 2432 20130    -.38  |...............................1....1
   80 23646    -.54  |...............................1....1
 1010 23011    -.54 V|..................................11
 1477 20632    -.54  |.............................1.....1
 1641 21572    -.54  |...............................1....1
 1838 20441    -.54  |.............................1.....1
 1945 23119    -.54 R|.0...........................11.1..1
 2052 21407    -.54  |..0................................1
 2105 22514    -.54 J|.0.........................1.1.1.11
 2323 21343    -.54  |0..................................1
   82 20885    -.70 U|0.0........................1.......1
  175 21764    -.70 X|..........................1.1......1
  462 22377    -.70 O|.00.........................1.11...1
 1321 21277    -.70  |.............................1.......1
 1340 21747    -.70 Z|..0.........................11.....1
 2203 22726    -.70  |.............................1.......1
 1443 20549    -.86 T|................................1..1
 1145 23561   -1.03 E|................................1.11
 1234 22601   -1.03 K|.....................1........1..1
 1000 22132   -1.21 N|...................................1
 1233 22623   -1.21 L|...................................1
```

Figure 2: Most Unexpected Responses

DISCUSSION

The conventional CT item analysis found a number of items that did not work well and therefore did not possibly contribute much to the reliability of the reading test. An example is item 53 with a discriminant index of 0.10, which indicated that it did not discriminate between the good and the weak students well.

Example:

ITEM 53

'trepidation' (para. 4) means all of the following EXCEPT

A. fear (47%)
B. anxiety (14%)
C. excitement (7.2%)
D. happiness (31.8%)

TEXT:

> Meanwhile, he married Dina, a city girl. She came with **trepidation**, to the Pixuna Dutapara community, where life is hard and the annual floods spill hundreds of kilometres over the land, forcing families to move upstairs in their stilt houses for months.

There were more items that could be identified as 'problematic' when the same set of items were analysed through Rasch analysis. Item 64 is one item that was not identified through CT but was characterized as misfitting through the Rasch analysis.

Item 64

…. his piracucu breeding and observation centre, a large (64) _____ which the ministry of …….

A. pit (10.1%)
B. lake (68.1%)
C. river (10.3%)
D. flood plain (11.5%)

As indicated by the response strings (see Figure 2a and 2b), there were good students who did not get the item right while a number of weak students got it right.

CONCLUSION

The superiority of Rasch analysis in test development and evaluation has been stressed in the literature. However, the usefulness of CT analysis in test evaluation should not be dismissed altogether. The usefulness of the CT and Rasch analyses as complementary approaches in test evaluation can be summarised as below:

1. While CT item difficulty values give an indication of how difficult or easy items in a test are for a group of students, Rasch analysis gives a better view of the spread of the difficulty of items in relation to the test takers' ability levels. This is made possible by the mapping facility.

2. Identification of poor or problematic items is one important step in test evaluation in that it provides insights as to how items performed in the test, and how much they contribute to the reliability of the test. In order words, item analysis is crucial in investigating how well items in a test function in tapping the intended abilities. When both the CT and Rasch analyses are utilized, a greater number of possible 'problematic' items can be found. This can better facilitate item revision, item banking and test improvement.

3. Explanations for problematic items are not limited to the item parameters (as in CT) when Rasch analysis is used. While CT distractor efficiency analysis may explain the reason for a 'problem' item, Rasch's most unexpected responses and misfitting responses strings may explain the reasons for the identification of the 'problematic' items by Rasch.

The classical approach was the standard approach in the investigation of reliability and in minimising measurement errors long before the introduction of modern measurement theory. Given the type of decisions to be made from the test scores of high-stakes tests (such as a language placement test), the investigation of reliability should apply not just the classical approach, but also incorporate modern measurement theory (the Rasch Model analysis in this case). Even though the use of the Rasch Model usually needs more time and resources, and the involvement of test developers and teachers, the amount of information gained and the usefulness of the more powerful information will justify the need. Therefore, the design of the reliability study in this research utilises both classical and modern measurement methods.

References

Alderson, J.C.; Clapham, C. & Wall, D. (1995) *Language Test Construction And Evaluation*. Cambridge: Cambridge University Press.

Bachman, L.F. (1990) *Fundamental Considerations In Language Testing*. Oxford: Oxford University Press.

Baker, R. (1997) *Classical Test Theory And Item Response Theory In Test Analysis Extracts From An Investigation Of The Rasch Model In Its Application To Foreign Language Proficiency Testing.* Special Report No. 2: Language Testing Update 1997. Lancaster: Department Of Linguistics And Modern English Language, Lancaster University.

Bond, T.G. & Fox, C.M. (2001) *Applying The Rasch Model: Fundamental Measurement In The Human Sciences*. London: Lawrence Erlbaum Associates.

Brown, J. D. (1996) *Testing In Language Programs*. New Jersey: Prentice Hall Regents.

Crocker, L. & Algina, J. (1986) *Introduction To Classical And Modern Test Theory*. London: Holt, Rinehart And Winstion, Inc.

Davies, A. (1990) *Principles Of Language Testing. Cambridge*. Oxford: Basil Blackwell Ltd.

Ebel, R.L. (1972) *Essentials Of Educational Measurement* (1st Edition). New Jersey: Prentice Hall .

Gronlund, N.E. (1976) *Measurement And Evaluation In Teaching*. New York: Macmillan Publishing .

Hambleton, R.K., Swaminathan, H. & Roger H.A. (1991) *Fundamentals Of Item Response Theory*. London: SAGE Publications .

Hambleton, R.K. (1989) Principles And Selected Applications Of Item Response Theory,In R.L. Linn, (Ed.) *Educational Measurement*. New York: Mcmillan Publishing Company , 147-200.

Hambleton, R.K. & Cook, L.L. (1977) Latent Trait Models And Their Use In The Analysis Of Educational Test Data, *Journal Of Educational Measurement*, 14(2), 75-96.

Hambleton, R.K. & Swaminathan, H. (1985) *Item Response Theory*. Boston: Kluwer-Nijhoff .

Heaton, J.B. (1979) *Writing English Language Tests: A Practical Guide For Teachers Of English* (5th Edition). London: Longman.

Henning, G. (1987) *A Guide To Language Testing- Development, Evaluation, Research*. London: Newbury House Publisher.

Henning, G. (1992) Dimensionality And Construct Validity Of Language Tests. *Language Testing*, 9(1), 1-11.

Hughes, A. (1989) *Testing For Language Teachers*. Cambridge: Cambridge Univ. Press Linacre, J. M. (2002) Diagnosing Misfit. Retrieved On 16 Nov 2002. Available At http://www.rasch.org/rmt/rmt162f.htm.

Mcnamara, T. (1996) *Measuring Second Language Performance*. London: Longman.
Wong, Harriet; Hazita Azman & Lee Siew Chin. (1990) *English Language Proficiency, Monograph 3*. Kuala Lumpur: Language Centre, Universiti Kebangsaan Malaysia
Woods, A. & Baker, R. (1985) Item Response Theory. *Language Testing*, 2(2), 119-140.
Wright, B.D. & Stone, M.H. (1979) *Best Test Design*. Chicago, ILL: MESA Press .

Appendix 1: Summary Of Guidelines For Categorising Items Based On Fit

Mean-square fit statistics show the size of the randomness, i.e., the amount of distortion of the measurement system. 1.0 is their expected values. Values less than 1.0 indicate observations are too predictable (redundancy, data overfit the model). Values greater than 1.0 indicate unpredictability (unmodeled noise, data underfit the model). Statistically, mean-squares are chi-square statistics divided by their degrees of freedom. Mean-squares are always positive. In general, mean squares near 1.0 indicate little distortion of the measurement system, regardless of the standardized value. Mean-squares high above 1.0 should be evaluated before mean-squares below 1.0, because the average mean-square is usually forced to be near 1.0.

Mean-square Value (MNSQ)	**Implication for Measurement**
>2.0	Distorts or degrade the measurement system. May be caused by only one or two observations.
1.5 – 2.0	Unproductive for construction of measurement, but not degrading.
0.5 – 1.5	Productive for measurement.
<0.5	Less productive for measurement, but not degrading. May produce misleadingly high reliability and separation coefficients.

Appendix 2: Summary Of CT Item Analysis

ITEM NO	CT DISCRIMINANT INDEX	CT DIFFICULTY VALUE	LOGIT DIFFICULTY	INFIT/ OUTFIT ITEMS *Zstd*
51	0.15	0.84	-.84	*
52	0.32	0.38	3.34	*
53	0.10	0.32	-.87	*
54	0.24	0.31	2.35	*
55	0.17	0.19	.34	*
56	0.37	0.58	1.42	*
57	0.38	0.62	-.88	*
58	0.28	0.70	.07	*
59	0.003	0.12	.33	*
60	0.09	0.95	.21	*
61	0.34	0.76	-2.41	*
62	0.21	0.72	-1.12	
63	0.50	0.45	-2.09	
64	0.44	0.68	1.11	*
65	0.37	0.81	.67	
66	0.27	0.39	-1.58	
67	0.60	0.45	1.67	
68	0.46	0.34	-.53	*
69	0.38	0.67	.35	
70	0.47	0.66	-.03	
71	0.35	0.51	-2.38	*
72	0.08	0.22	.61	*
73	0.29	0.39	-1.45	*
74	-0.002	0.19	-.99	*
75	0.32	0.71	.00	*
76	0.17	0.25	-.05	*
77	0.17	0.29	.81	*
78	0.23	0.25	.12	
79	0.33	0.59	.39	*
80	0.39	0.40	1.05	*
81	0.35	0.71	1.04	
82	0.36	0.71	.75	
83	0.16	0.42	1.28	*
84	0.002	0.19	-1.24	*
85	0.23	0.27	-1.48	

* Misfitting items (by infit, outfit standardized values or both)