COMPARISON OF SEVERAL POPULAR DISCRIMINATION INDICES BASED ON DIFFERENT CRITERIA AND THEIR APPLICATION IN ITEM ANALYSIS

by

FU LIU

(Under the Direction of Seock-Ho Kim)

ABSTRACT

This study applied the classical item analysis to a regular classroom test in a high school. As the core of the classical item analysis, discrimination indices were the main focus of this study. The purpose of the study was to investigate the relatively dependable and simple discrimination index by comparing some popular discrimination indices based on different internal criteria (i.e., the subtest-total score and the entire-test-total score). The acquisition and comparison of the discrimination indices were fulfilled by means of modern computer programs—Lertap5 and SPSS. Additionally, the difficulty indices and the distractor analysis were also taken into account as a whole. The data set was collected from an English pretest of the third grade in a high school of China. 1059 students participated in the test, which included 45 multiple choices, 15 fill-in blanks, 5 matchings, and 2 essays. These questions were assigned to four subtests. The discrimination indices (r_{pbis}, r_{bis}, $D_{10\%}$, $D_{27\%}$, $D_{33\%}$, $D_{50\%}$) and their corresponding difficulty indices were compared based on the two different internal criteria. As a result, r_{pbis}, $D_{33\%}$ and the related *p's* were recommended to use in the classroom item analysis for the high school teachers.

INDEX WORDS: Classical test theory, Discrimination indices, Difficulty indices, College entrance examination in China, Item response theory

COMPARISON OF SEVERAL POPULAR DISCRIMINATION INDICES BASED ON DIFFERENT CRITERIA AND THEIR APPLICATION IN ITEM ANALYSIS

by

FU LIU

B.A., Northeast Forestry University, China, 1996

A Thesis Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2008

COMPARISON OF SEVERAL POPULAR DISCRIMINATION INDICES BASED ON DIFFERENT CRITERIA AND THEIR APPLICATION IN ITEM ANALYSIS

by

FU LIU

Major Professor: Seock-Ho Kim

Committee: Stephen Olejnik

Jonathan Templin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

This study mainly applied an item analysis procedure in one of the currently popular measurement frameworks—classical test theory (CTT) to an English test, which was one of the school pretests for the National College Entrance Examination (CEE) in a high school of China. Since the test was composed of several subtests and different types of items, the item analysis procedure appeared to be more complicated and time-consuming than we expected. This study aimed to take full advantage of the discrimination indices of CTT over item response theory (IRT) in attempt to find the relatively simple and dependable item statistics to check the item quality for the under-college school teachers. Particularly, as the core of the classical item analysis, discrimination indices can always provide the sufficient evidence on verifying the quality of test items. Thus, comparison of different measurements on various discrimination indices can lead us to get the more appropriate and satisfactory one for future item analysis taken in the high school classroom. Some item analysis examples, related to dichotomously-scored items such as multiple choices, were also provided by using the verified discrimination indices. Another concern of the research was to apply the classical item analysis studied mostly by American educators to an English exam taken in a Chinese high school in order to extend this technique to the wider domain concerning the cross-cultural setting.

As the foundation of measurement theory, CTT has been studied for more than 80 years and even today is still widely employed in the testing measurement field. Compared

to CTT, the initial concept of item response theory (IRT) was first generated by F. M. Lord in his doctoral thesis in 1952 and has sequentially encountered an exceptional development in the recent decades. However, depending on its particular advantages, CTT was not substituted by IRT, but was consistently rooted in many testing programs. Especially, in the latest years, more and more measurement specialists place their attention on the comparison in measuring the quality of test items between CCT and IRT (Tinsley & Dawis, 1977; Shannon & Cliver, 1987; Cook, Eignor, & Taft, 1988; Lawson, 1991; Hambleton & Jones, 1993; Ndalichako & Rogers, 1997; Fan, 1998; MacDonald & Paunonen, 2002). The respective results obtained from the preceding studies were introduced in the later section of this study. Throughout the overall accessible studies, IRT does not appear distinctively superior in item analysis to CTT in most situations, particularly for the classroom or school tests. Obviously, in the case of this study, the target users of the particular measurement technique fixed on the under-college school teachers, especially the high school teachers. Regarding the specific group, very complicated and tedious measurement techniques were not suitable due to the characteristics of their occupation. For one reason, most of them were rarely exposed to the training in classical or modern item analysis. The item analysis procedure was often conducted by some experts and the practitioners in the test-developing field. Another reason was that the majority of classroom teachers in high schools had to pay intensive attention on teaching, supervising or tutoring students. It was certain that the intricate and painstaking computation and understanding of item statistics would greatly limit the application amongst the classroom teachers. Thus, a simpler and more efficient measurement procedure was strongly called for in order to fit in the special school setting

and also benefit the classroom teachers. Considering the restrictions for the application of item analysis in the high schools, CTT presents its great advantages over IRT. Shannon and Cliver (1987) suggested:

> Conventional item discrimination indices usually do not require large samples of examinees, nor the making of unrealistic assumptions about the data. They can be calculated easily, quickly, and inexpensively and can be understood by most users. Moreover, some of these measures seem to satisfy Findley's (1956) requirement that an item discrimination measure be based on algorithms that seem meaningful for the purpose, thus making it easier for psychometricians to explain the item selection process to nontechnical audiences (e.g., classroom teachers). (p. 348)

In the meantime, given the intense competition in China's national College Entrance Examination (CEE), all of China's high schools focus more and more attention on their teaching and testing quality. However, there might not be, so far, any existing scientific item analysis system for these schools, so the school administrative committee and even the classroom teachers have a strong demand to the technique.

Item Analysis of CTT

CTT is a theory about test scores that introduces three concepts—test score (often called the observed score), true score, and error score. The most acknowledged relationship between these three scores is

$$X = T + E, \tag{1}$$

where X is the observable test score, T is the true score, and E is the error score. This model is based on the following assumptions: (a) true scores are independent of

error scores, (b) the sum of all the error scores in the population of examinees is zero, and (c) error scores are independent between parallel tests. Across parallel forms which cover the same content and in which the true score of the examinees remain the same, and where no difference in measurement errors occur across forms, true score can equivalently considered as the expected test score across parallel forms (Hambleton & Jones, 1993).

Under the framework of CTT, item analysis mainly refers to estimating the item difficulty and item discrimination indices and also includes distractor analyses in each item (Hills, 1981).

Item Difficulty Index

The classical item difficulty is defined as the proportion of examinees who get the right answer to an item, when the item is dichotomously scored (Crocker & Algina, 1986). It is also referred to as the item mean or item p value. Its value falls between 0 and 1, which 1 means too easy and 0 means too difficulty. That is,

$$p_i = E(X_i), \tag{2}$$

where p_i is the difficulty index of the item X_i and X_i is dichotomously scored as either 0 or 1. The overall item difficulty or mean item difficulty indicates the difficulty level of an average item on the test. It is formulated as

$$E(p_i) = \frac{1}{k}\sum_{i=1}^{k} p_i, \tag{3}$$

where k is the total number of items. The item variance is

$$\sigma_i^{\,2} = p_i q_i, \tag{4}$$

where $q_i = 1 - p_i$. Under the random guessing assumption, the optimal difficulty value

(p_o) for a multiple choice item with m options is

$$p_o = .5 + \frac{.5}{m}. \quad (5)$$

Based on simulated response data, Lord (1952) demonstrated that the optimal p_o values are .694 (m =5), .742 (m =4), .770 (m =3), and .846 (m =2), assuming the intercorrelations among items (i.e., the corrected tetrachoric intercorrelations) are .20. For a polytomously-scored item, the item difficulty can be represented as the average item score.

Item Discriminating Power

Item discrimination power indicates the extent of an item to differentiate the examinees with different ability levels (MacDonald & Paunonen, 2002). An ideal item should possess the function of distinguishing the more able from the less able examinees. That is, the high achieving students get the item correct while the low achieving students get it wrong. There are over twenty discrimination indices used as indicators of the item's discrimination effectiveness such as the index of discrimination (D), the point-biserial correlation coefficient (r_{pbis}), biserial correlation coefficient (r_{bis}), phi coefficient (ϕ), tetrachoric correlation coefficient (r_{tet}), and rank biserial correlation coefficient (rb), etc. The values of the discrimination indices stated above typically fall between -1 and 1, of which absolute values approaching 1 imply the items can differentiate the examinees in the largest degree. However, the items with negative discrimination values are never desirable. It indicates that these items are missed by many high-achieving examinees but are mostly chosen correctly by low-achieving examinees. For a polytomously-scored item,

the correlation between the item score and the total score is referred to as an item discrimination index.

Distractor Analyses

Normally, an item which is too easy or too difficult can not obtain a good discriminative power because most of or only a few of the examinees can get the item right, so the professional test constructors prefer to set the items as medium difficulty level. Ebel and Frisbie (1986) pointed out that "test composed of items of moderate difficulty stand the best chance of discriminating between levels of achievement and producing high score variability" (p. 225). However, the appropriate difficulty level of an item does not ensure that the item can discriminate well. For example, if half of the total examinees answer an item correctly, but most of them gather in the lower group, we can not say the item differentiates the examinees with high and low ability. Thus, a more specific analysis is required if the type of items is multiple choice, true-false or matching. We need to analyze the options of the item including the distractors. The well-functioning distractors should attract poor-ability examinees, but not good-ability examinees. Usually items with the appropriate *p* values and acceptable discrimination values can be viewed as sound. However, to further ensure the quality of items, the distractor analyses are also needed. The distractors which are not chosen by any examinees should be revised or replaced. Likwise, items with extreme *p* values and low or negative discrimination values can be improved by distractor analyses. As Nitko and Hsu (1984) suggested, we can use the similar formula for the discrimination index (*D*) to calculate a discrimination coefficient (d_{ij}) for each distractor, which measures the effectiveness of a distractor. That is,

$$d_{ij} = p_{uij} - p_{lij}, \quad (6)$$

where d_{ij} is the discrimination coefficient for the i_{th} distractor in the j_{th} item of the test ($i \neq$ correct answer), p_{uij} and p_{lij} respectively represent the proportions in the upper and lower groups who choose the i_{th} distractor in the j_{th} item of the test. The upper and lower groups refer to the groups which are the students with high scores and low scores. The discrimination coefficient for a distractor is expected to be negative, which implies that more low-scoring examinees choose the distractor than high-scoring examinees. Also, it comes to perfection if p_{uij} is equal to 0, but p_{lij} is not.

According to Hills (1981, p. 76), the best items have a pattern of responses in which (a) every option should be able to attract at least one examinee, (b) the right answer should be selected by more high-scoring students in the total test than by low-scoring students, (c)the distractors can attract more low-scoring students than the high-scoring students, (d) and it is the most desirable that the item difficulty index is close to the optimal proportions (p_o). In addition, any technique defects, unaware hints, wrong keys, or ambiguous expressions in building an item may cause high-achieving examinees to rule out the correct answer. Any problems mentioned above result in identification and revision or replacement of malfunctioning items.

Reliability & Validity

Reliability refers to the accuracy and consistency of test scores. Validity indicates the degree to which we can justify the inferences drawn from test scores. Thus, during the process of the item analysis, any ignorance of verifying these two issues is very dangerous, no matter for item score or total test score. As we usually regard the total test

score as the internal criterion, the item reliability index is defined as $\sigma_i \rho_{iX}$, where ρ_{iX} is the correlation between item score and total test score. For dichotomously-scored items, it can be rewritten as $\sqrt{p_i q_i}\,\rho_{iX}$, where ρ_{iX} is the point biserial correlation between item and total test score. As we use the external criterion, the item validity index is defined as $\sigma_i \rho_{iY}$, where ρ_{iY} is the correlation between item score and the external criterion. If the total test reliability and validity is of interest, we can use the internal consistency coefficient, as measured by coefficient alpha, to estimate the test reliability. That is,

$$\rho_{\alpha} = \frac{k}{k-1}\left[1 - \frac{\sum \sigma_i^2}{(\sum \sigma_i \rho_{iX})^2}\right], \tag{7}$$

where *k* is the number of items in the test. Then, the validity coefficient can be estimated by

$$\rho_{XY} = \frac{\sum \sigma_i \rho_{iY}}{\sum \sigma_i \rho_{iX}}. \tag{8}$$

As Gronlund and Linn (1990) stated:

> Ideally, we would examine each test item in relation to some independent measure of achievement. However, the best measure of the particular achievement we are interested in evaluating is usually the total score on the achievement test we have constructed because each classroom or school test is related specific instructional objectives and course content. Even standardized tests in the same content area are usually inadequate as independent criteria, because they are aimed at more general objectives than those measured by a

classroom or school test in a particular course (p. 253),

the test in this study is a school test, so our concern should focus more attention on the test's internal consistency reliability in order to fulfil the goal of discriminating the high and low achievers. That is, the high test reliability normally implies the high discriminating power of items. Equation 7 also presents the positive correlation between the discriminating ability of an item and the total test reliability.

CTT versus IRT

IRT is a general statistical theory of evaluating the abilities of the examinees by measuring the performance of the examinees on the specific item in the test. The advantages of IRT include: (a) item statistics that are not determined by the groups that are selected to estimate them; (b) scores that are uncorrelated to the test difficulty; (c) test models that combine test items with ability levels; (d) test models where the rigid equivalent tests are not necessary to guarantee reliability (Hambleton & Jones, 1993). In contrast to IRT, the primary weakness of CTT is that the item statistics (i.e., discrimination index and difficulty index) are correlated to the person statistics (i.e., observed test scores). In other words, the values of the discrimination index and the difficulty index of an item are impacted by the specific measured examinees while the values of the person statistics are impacted by the selected items in the test as well. (MacDonald & Paunonen, 2002). However, measurement specialists have worked out some practical solutions such as test equating within the framework of CTT for some unexpected difficult measurement problems. Furthermore, considering the applicability of item analysis procedure among the classroom teachers in the under-college schools, CTT brings up the superior benefits as followed: (a) A large sample size is not necessary for

CTT model; (b) The mathematical computation in model parameter estimation is much simpler than item response theory; (c) The related concepts and analyses are straightforward; and (d) The CTT model is much easier to fit the test data (Hambleton & Jones, 1993). As we know, the number of the students in a high school is limited, especially in one grade or one classroom, in comparison with the number of the candidates in the national or interstate official exams. Moreover, straightforward computation and interpretation of the item statistics can make the item analysis procedure more applicable among the high school teachers. Additionally, the assumptions of the classical test models are fairly easily met by test data. All of the advantages of CTT mentioned above fit in the situation of a high school very well. Therefore, the classical item analysis is preferable to the high school teachers.

Discrimination Index of CTT

As we discussed previously, the discriminating power of an item can largely show the evidence for the quality of the item. Thus, measurement specialists have made a lot of efforts to develop the discrimination indices which represent the discriminating power of an item. So far, more than 20 item discrimination indices are proposed. However, based on the previous studies (Englehart, 1965; Guilford, 1965; Nunnally, 1967; Aleamoni & Spencer, 1969; Henrysson, 1971; Bowers, 1972; Hales, 1972; Oosterhof, 1976; Beuchert & Mendoza, 1979; Carroll, 1987; Shannon & Cliver, 1987; Millman & Greene, 1989; Fan, 1998; Attali & Fraenkel, 2000; MacDonald & Paunonen, 2002), only a few of them are widely used and compared for the dichotomously-scored items by psychometricians and item-analyzing software.

The Index of Discrimination (D)

D , which is only applied to dichotomously-scored items, is a recognized simpler discrimination parameter rather than the other discrimination indices. Before we compute D, we need to divide the examinees into the upper and lower groups. The examiners normally rank the test-takers from the top to the bottom according to their total test scores. They divide the total group into several subgroups by one or two cut-off points. As Kelley (1939) suggested, a more sensitive and stable cut-off point for D is 27% under certain conditions. That is, the top 27% of the examinee group is the upper group and the bottom 27% is the lower group. When the sample size is reasonably large, 30% and 50% can also yield the similar results as 27% (Beuchert and Mendoza, 1979; Englehart, 1965). After the upper and lower groups are generated, the computation of D can be conducted through the formula below:

$$D = p_u - p_l , \tag{9}$$

where p_u is the proportion in the upper group who get the item right and p_l is the proportion in the lower group who get the item right. The following guidelines were provided by Ebel (1965), based on his own practical experience when he selected items that discriminated on the internal criterion of total test score:

- If $D \geq .40$, very well-functioning items.
- If $.30 \leq D < .40$, reasonably well-functioning items.
- If $.20 \leq D < .30$, marginal items which need revised.
- If $D < .20$, poorly-functioning items which need eliminated or fully revised.

Although D has no particular sampling distribution, so we can not statistically judge its significance, it is still widely utilized, especially by classroom teachers on account of its

simple computation and straightforward interpretation. Meanwhile, it is the most appropriate when a computer is unavailable.

Point-biserial Correlation Coefficient

As the special case of the Pearson product moment correlation, the point-biserial correlation (r_{pbis}) reflects the correlation between the dichotomously-scored item score and the total test score which is continuously distributed. It is defined as:

$$r_{pbis} = \frac{(\mu_+ - \mu_X)}{\sigma_X}\sqrt{p/q} \quad , \tag{10}$$

where μ_+ is the mean criterion score for the examinees who get the item right, μ_X is the mean criterion score for the entire group, σ_X is its standard deviation, p is item difficulty index and q is $(1-p)$. Since an item score accounts for part of the total score, when a small number of items are included in the test, another formula is strongly recommended:

$$r_{i(X-i)} = \frac{r_{Xi}\sigma_X - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_X^2 - 2r_{Xi}\sigma_X\sigma_i}}, \tag{11}$$

where $r_{i(X-i)}$ is called the corrected point-biserial correlation between an item score and the item-excluded total score, and σ_X and σ_i are separately the total and item standard deviations. Also, we need to note that the point-biserial correlation is impacted by changing the difficulty level of the item. It favors the items with modest difficulty.

Biserial Correlation Coefficient

The biserial correlation (r_{bis}) is the correlation between the assumed normally-distributed latent variable underlying item performance and the continuously-distributed total test score. It is denoted as

$$r_{bis} = \frac{(\mu_+ - \mu_X)}{\sigma_X}(p/Y), \quad (12)$$

where some denotations mean the same as above in r_{pbis} in addition that Y is the Y ordinate of the normal distribution corresponding to the *p* value. The biserial correlation, which differs from the point-biserial correlation, can reflect the discriminating power, independently of difficulty level. Thus, when we select items at one extreme of the difficulty range, r_{bis} is recommended (Crocker & Algina, 1986, p. 319). However, Adams (1960) and Richardson (1936) demonstrated that the value of the biserial correlation coefficient may exceed 1 when the distribution of the criterion scores is bimodal or skewed. The solution to the problem is to transform criterion scores into a normally-distributed scale (e.g., stanines) (Henrysson, 1971, p. 141) or that the coefficient of selective efficiency (*S*), proposed by Brogden (1949), can be used in this situation. Additionally, if we doubt the ability level of the future samples differs from the currently analyzed sample, the biserial correlation is superiorly recommended over the point-biserial correlation (Crocker & Algina, 1986, p. 319). Otherwise, if we want to select the items with high internal consistency, the point-biserial correlation is preferable (Lord & Novick, 1968, p. 341).

Note that the biserial correlation can be expressed by another formula

$$r_{bis} = \frac{\sqrt{pq}}{Y} r_{pbis}, \quad (13)$$

which presents the relationship between the biserial and point-biserial correlation. From the formula, we can also obtain the corrected biserial correlation by replacing r_{pbis} by $r_{i(X-i)}$. Lord and Novick (1968, p. 340) deduced through the formula that the value of a

biserial correlation is always at least one-fifth larger than the point-biserial correlation for the same variables. Likewise, Magnusson (1967) indicated that the biserial correlation may be four times larger than the point-biserial correlation when the item difficulty extremely ranges. Thus, it is worth remembering that some large differences between the item discrimination indices may be caused by the alternative of correlational formulas rather than by the actual differences in the discriminating ability between items.

Phi Coefficient

The phi coefficient (ϕ) is used when the measure falls to the correlation between the dichotomously-scored item and a dichotomized criterion (e.g., success and failure). It is also called fourfold point correlation coefficient and can be generally expressed as:

$$\phi = \frac{p_{ij} - p_i p_j}{\sigma_i \sigma_j}, \tag{14}$$

where p_{ij} is the proportion of examinees who get the item correct and simultaneously get successful in the criterion, p_i and p_j are respectively the proportions of examinees who get successful in the item and in the criterion, and σ_i and σ_j correspondingly stand for the item and criterion standard deviation. This formula is perfectly applied when the criterion scores are dichotomized into success and failure. The phi correlation can be used to measure the consistency of the responses of the same examinees to the same dichotomous item in different situations. Another popular usage of the phi correlation is to determine the relationship between two items by analyzing their respective responses of examinees. Note, however, that some quantitative information in the difference between the dichotomized-criterion groups may be missing if the criterion scores are originally continuously distributed. Since the phi correlation is also a special

case of the Pearson product moment correlation, its value can only be 1 when the *p*-values for the item and the criterion are equal.

Tetrachoric Correlation Coefficient

Tetrachoric correlation (r_{tet}) shows the correlation of two dichotomized variables which underlying distributions are assumed normal. The calculation and the formula of the tetrachoric correlation coefficient is very complicated. Although the Camp's (1931) approximation was recognized as an excellent one in computing r_{tet} , researchers, nowadays, try to avoid employing r_{tet} unless they can not find other appropriate discrimination indices in some specific situations (Crocker & Algina, 1986, p. 319). For example, when the *p*-values for the two variables are not equal, the tetrachoric correlation is more appropriate for factor analysis than the phi correlation. In this case, some computer software fortunately provides the estimation of the tetrachoric correlation (see, for example, Dixon et al., 1981).

Rank Biserial Correlation Coefficient

The rank biserial correlation (*rb*) is appropriate in the situation that the criterion scores are reported as consecutive, untied ranks and the item scores are dichotomous. It is denoted as

$$rb = (2/n)(\overline{Yr} - \overline{Y_w}), \tag{15}$$

where n is the number of ranked examinees, $\overline{Yr}$ is the average rank of the examinees who get the item right and $\overline{Y_w}$ is the average rank of the examinees who get the item wrong. The rank biserial correlation is algebraically equivalent to Spearman's rho, which is, in fact, the Pearson product-moment correlation between two sets of ranking variables.

As Glass (1965) suggested, *rb* is easy to compute in contrast to r_{bis}. It is also more stable than the upper and lower 27% *D* because it is computed on all the data.

Brief Introduction of Some Other Discrimination Indices

Phi-Over-Phi-Max, phi/phi max, is a solution to the restriction in range of phi when the *p*-values of the two variables are not equal. It can range from -1 to +1 (Cureton, 1959).

The B-Index, B, is the difference in the item difficulties between examinees with high and low ability. It ranges from -1 to +1. Like phi, *B* is also restricted when the *p*-values of the two variables differ largely (Brennan, 1972).

The Agreement Statistic, A or $p(X_c)$, is to measure the agreement between results on a given item and results on the mastery test. For an item where both item and test results are consistent, the agreement statistic would equal the maximum value of 1. Otherwise, Harris and Subkoviak (1986) used an equation to estimate the practical lower bound.

Davis Discrimination Index is to convert the tetrachoric coefficient for the highest and lowest 27% split to a scale with equal units based on Fisher's *z* (Davis, 1949).

Flanagan's (1939) Correlation Coefficient, r_F, shows "the estimation of the product-moment correlation coefficient for various proportions of success in the upper and lower 27% of the criterion group" (Hales, 1972, p. 929).

Flanagan's (1939) Corrected Correlation Coefficient, is "computed from the proportions which have been corrected for chance success and having corrected indices of difficulty falling within the range .15-.75" (Hales, 1972, p. 931).

The College Entrance Examination (CEE) of China

The College Entrance Examination (CEE) is the most important written examination for the under-college students in China. The official CEE website (www.gaokao.edu.cn) provides substantive updated information related to CEE. Currently, in Guangdong province CEE is made up of four sections—Mathematics, Chinese, English, and Comprehensive including politics, history, geography, physics, chemistry, and biology. CEE is the only measure which determines if students can have an opportunity to receive the advanced education. High school GPA and school reports are not taken into account for the acceptance by colleges and universities. The universities in different levels have different criteria in students' selection depending on CEE scores. Thus, CEE is a norm-referenced achievement test which attempts to differentiate students with high and low abilities corresponding to the different criteria of universities. Nowadays, an increasing number of educators doubt the function of CEE in students' selection. Although China has conducted the CEE reform and the New Curriculum Reform in recent years, the importance of CEE is never changed. Because of the limitation of resources in education, only half of the candidates can get the chance to go to college. Considering the rigorous situation, all of the high schools make their utmost to improve the teaching quality in order to send more of their students to colleges. Therefore, any applicable technique which can benefit school teaching will be greatly accepted and employed. Exams and tests are the most efficient and straightforward means to know about students' learning status and assess teachers' or schools' overall instructional quality. Only high-qualified tests can guarantee the fairness and preciseness of the statistical estimation and assessment to students, teachers and schools. However, no

feasible technique is widely used in China's high school, which can scientifically measure the quality of items in the tests. Most of the assessment to the tests depends on the teachers' experience, which is not reliable and consistent. The English test utilized in this study was one of the pretests for CEE in a high school in Guangdong province of China. The test result provided referential evidence in estimating the performance of students in the later CEE. Meanwhile, it offered important suggestive information on teachers' instruction and students' learning orientation. Thus, this test was a typical test in China's high schools, which represents the common item types and test content.

CHAPTER 2

LITERATURE REVIEW

Since CTT was introduced nearly a century ago, it has received large numbers of psychometrician's attention. Especially during the 1930s to 1960s, CTT was extensively studied and applied by the researchers. Some researchers also tried to apply CTT to the ordinary classrooms. Thereafter, IRT with more advanced techniques and wider applications generally became the highlight of the research domain although CTT is still widely used. In the recent decades, an increasing number of researchers got engaged on the comparison between CTT and IRT and even the combination of the two frameworks in item analysis. The points mentioned above have fortunately been well-documented in the measurement literature.

CTT versus IRT

As early as 1977, Tinsley and Dawis mentioned that the Rasch model (i.e., the one-parameter IRT model) produced relatively test-free person measurement in ability estimates compared to the classical psychometric theory. Also, Hambleton and Jones (1993) revealed some superior advantages of IRT over CTT which has been presented in the prior section. However, very few studies have compared CTT and IRT for item analysis and test design. In the point, Fan (1998) indicated in his paper:

> It is somewhat surprising that empirical studies examining and comparing the invariance characteristics of item statistics from the two measurement frameworks are so scarce. It appears that the superiority of

> IRT over CTT in this regard has been taken for granted in the measurement community, and no empirical scrutiny has been deemed necessary. (p. 357)

Among the limited number of the studies in comparison between CTT and IRT, Shannon and Cliver (1987) presented that conventional indices such as phi, other than phi-over-phi-max, *B*-index, and the agreement statistic, may be comparable to IRT-derived item parameter estimation of the discriminating power at the passing score for criterion-referenced tests. Moreover, in the study of Cook, Eignor, and Taft (1988), they stated:

> it was equally apparent that neither CTT nor IRT was sufficiently robust to provide viable item analysis or equating results when faced with a lack of parallelism such as that exhibited by the 58 common items given to the spring and fall groups. It is interesting to note how similarly the classical and IRT approaches behaved, particularly the very parallel effect that choice of common items had on the results produced by the two methods. (p. 43)

Lawson (1991) also drew an identical conclusion that person statistics and item difficulty estimates were found to be very similar from CTT and IRT although the correlation coefficient on item discrimination between CTT and IRT did differ across three data sets and the data sets with different sample sizes were recommended to use in order to justify the specific results drawn in his study. He emphasized that the Rasch latent trait model which is complicated to compute and understand seems not more superior over classical

test theory for teachers, measurement experts, test constructors, and developers. In 1997, Ndalichako and Rogers built five scoring models for scoring multiple-choice items to estimate person ability, based on CTT, IRT and finite state score theory (García-Pérez, 1987, 1993; García-Pérez & Frary, 1989, 1991). They found that the rankings of the examinees in the five scoring models were very similar. Furthermore, the scores yielded by the five scoring models had the very close values of the mean absolute differences. These findings provided the sufficient evidence for the necessity of the application of classical test theory for test scoring and item analysis due to its simple analyzing and understanding procedures. Later, Fan (1998) constructively replicated (Gall, Borg, & Gall, 1996; Lykken, 1968) Lawson's study (1991) by using a large-scale test database from a statewide assessment program. The findings from his empirical investigation failed to support the assumed perspectives that IRT was superior over CTT in consistently estimating item statistics and person statistics. All the studies stated above employed archival real data sets. Instead, MacDonald and Paunonen (2002) investigated the comparability, invariance, and accuracy of IRT and CCT parameter estimates under a variety of testing conditions through Monte Carlo simulations. They found that the CTT-based item discrimination statistic yielded less accuracy estimates in some conditions such as in the test with a wide range of item difficulties, but the item difficulty and person statistics from IRT and CTT frameworks were strikingly similar in all conditions and even the item difficulty and item discrimination statistics yielded by CTT were more consistent. Overall, the above findings generally reiterated a popular quote by Thorndike (1982) on the future of IRT:

> For the large bulk of testing, both with logically developed and with standardized tests, I doubt that there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting tests will continue to have much the same properties. (p. 12)

Shannon and Cliver (1987) also pointed out, "For many users who must continue to assemble and analyze CRTs under practical restraints (e.g., limited funding), the technical advantages of IRT-derived item discrimination measures may not be sufficient to justify their choice over simpler, less expensive conventional measures" (p. 348).

Comparison Between Classical Discrimination Indices

Dating back to 1965, Engelhart first made an influential comparison between several item discrimination indices—r_{tet}, ϕ, r_{bis}, r_{pbis}, Davis discrimination indices, D_1 with the proportions split at the median, and D_2 with the proportion of the highest and lowest 33% in terms of total scores. He used a data set which contained two samples of 210 students. These students had to complete Forms A and B with 60-item Constitution Test for the Illinois State High School Equivalency Testing Program. The main finding of his study presented the evidence that the items in the test which required to be revised or be removed could be efficiently identified by means of the discrimination indices D_1 and D_2. He also pointed out that "where item-analysis data are obtained largely for use by teachers and for tests locally constructed, the indices D_1 and D_2 should suffice, especially if they can be obtained for all answers to each item" (Engelhart, 1965, p. 75).

In the study of Aleamoni and Spencer (1969), they compared r_{bis}, r_{pbis} and the difficulty index, using the sampling data set of the Modern Language Association-Foreign Language Tests in reading and listening comprehension which included 4,300 participants. These participants took the first four semester courses in French, German, Russian, and Spanish at the University of Illinois. His study found that r_{bis} and r_{pbis} were highly correlated and their relationship remained consistent to the difficulty index. He also indicated that r_{bis} or r_{pbis} was not the appropriate one used for item analysis when the item difficulty index fell in a large range, but there was an exception when the statistical significance of the indices are taken into consideration. In 1972, Bowers made the similar research which compared r_{bis} and r_{pbis} from the technical and theoretic point. He eventually made the similar conclusion that there was little difference between r_{pbis} and r_{bis} in selecting a subset of best items when the selection criterion is symmetric. He also indicated if only a small number of items were selected from the tryout items, r_{bis} would be more dependable in selecting better items than r_{pbis} when a test aimed to reject the bottom five percent of a population. In Hales' (1972) study, he employed the data set obtained from "Test 1: Social Studies" of the Tests of Academic Progress for the ten-grade, eleven-grade and combined-grade students. The test included 50 items for each test and 83 students in the tenth grade and 82 for each of the remaining tests. Flanagan's *r*, Flanagan's corrected *r*, and *D* with 27% of upper and lower groups were investigated in the study. The conclusion drawn from the study was that *D* statistically functioned as well as Flanagan's *r* and Flanagan's corrected *r* in selecting items in test construction. Hales (1972) also suggested that "*D* should be appropriate

index of discrimination in item selection of a test for classroom teachers to use, in conjunction with the index of difficulty, in the selection of items for inclusion on a test" (p. 936). The same suggestion was given by Oosterhof (1976) in his study. He investigated 19 different discrimination indices. They included r_{pbis}, r_{bis}, ϕ, r_{tet}, Findley's (1956) *D*, r_F, and Gulliksen's (1950) item reliability index. In addition, they included *S* which is the adjusted r_{pbis} which is independent of item difficulty, , r_{nb} which is the adjusted r_{bis} by normalizing the continuous variable, both adjusted r_{bis} and r_{pbis} for spuriousness suggested by Henrysson (1963), and also ϕ and r_{tet} with various proportions students in the upper and lower groups, that is, 0.10, 0.27,0.33 and 0.50. The data set was obtained from the results of the Verbal Reasoning subtest of Form M in the Differential Aptitude Tests, which was composed of 50 items and sampled 1,000 students being randomly selected from 2,311 ten-grade students. As a result, the study suggested that the specific index with an ease of computation would be preferred compared to the other discrimination indices. Meanwhile, the study indicated that the *S* index proposed by Brogden (1949) would be more appropriate than the biserial correlation when the difficulty levels in the test largely range. In the late 1970s, Beuchert and Mendoza (1979) generalized the previous studies and first employed the Monte Carlo approach to avoid the shortcomings yielded by the previous studies. Through the Monte Carlo, test validity can be more precisely estimated when using a larger computer-simulated sample. Secondly, the item parameter estimation can be manipulated in order to extensively apply this study to various item analysis situations. 10 discrimination indices were compared: r_{pbis}, r_{bis}, ϕ, r_{tet}, rb, Ivens' (1971) index, the

B index (Brennan, 1972), and *D* with 50%, 33%, and 10% of the upper and lower groups. By computer simulation, 16 different item analysis situations, each containing 100 items, were generated on the basis of two sample sizes—60 and 200. They eventually draw the similar conclusion as the previous studies that there were little differences among the 10 discrimination indices no matter what situation they were employed in.

Previous Studies for the Classroom Classical Item Analysis

As far back as the early 1980s, some researchers made efforts to apply the classical item analysis to the regular classrooms (Hills, 1981; Ebel & Frisbie, 1986; Hopkins, Stanley, & Hopkins, 1986; Gronlund & Linn, 1990, etc.). Their common principle was to simplify the analysis procedure on the basis of ensuring the validity of item analysis. As suggested, teachers can get the assist from students to tabulate the number of students who selected each alternative. A test item card can also be employed for convenience of calculating the item statistics. In all the studies mentioned above, *D* for the upper and lower groups was recommended for the classroom item analysis due to the ease of computation. Hills (1981, p. 75) pointed out the item analysis can also be conducted through computer. However, slow-developing computer technology limited the popularity of computer, so the classical item analysis through computer did not sufficiently developed.

Overall, some studies related to comparing CTT and IRT, in some degree, proved that IRT had no superior advantages in estimating the personal statistics and item statistics over CTT. Moreover, large numbers of comparisons among the classical discrimination indices were also conducted. The conclusion obtained from most of these studies was very similar that there were no distinct differences among the compared

discrimination indices in some specific testing situations through different experimental methods. Moreover, some previous studies related to the classroom item analysis were mostly based on the mental computation. Today, modern computer facilities are available at each of schools. Computers can analyze tests and items very efficiently at a low cost. Thus, the item analysis can be achieved more easily by means of modern computer software.

CHAPTER 3

PROCEDURE

Instrumentation

In this study, the data set was originally obtained from a high school in Guangdong province of China. It was the record of scores on an English exam which was one of the school pretests for the National College Entrance Examination (CEE) in 2008. Every year, approximately 10,000,000 candidates take the CEE. Before the CEE, all of the high schools normally make several pretests in all the subjects to investigate the students' learning status and teachers' instructions and also predict the students' performance in the following CEE.

The data set included the original responses of each student to the multiple choices and matching questions. As for the items of fill-in blanks and writing, the score of each of fill-in blanks and each scoring aspect of writing section were provided. Through the original responses, the distribution of students who chose each of the options was easily obtained. The score of each of fill-in blanks told us if the students got the blank correct, partially correct, or not correct at all. The score of each scoring aspect in writing section let us know the distribution of scores in every aspect among all the students. All the information provided by these responses and scores totally contributed to estimating the item statistics.

The test used in the study fell into the subject of English. It was constructed to evaluate some key English skills of the students and to predict how prepared the students

were for the English exam of CEE. The key English skills included listening, grammar, reading, and writing. The format of all the questions in the English test corresponded with the counterpart of the English exam in CEE, composed of four subtests (listening, language application, reading comprehension, and writing). The total full score was 150. The listening portion involved 15 multiple choices (2 points each) and 5 fill-in blanks (1 point each). Each of the 15 multiple choices had three response options. For the fill-in questions, the students can get the partial point (0.5) for some incomplete response but very close to the right answer. Considering these 5 fill-in blanks were not dichotomously-scored, they were analyzed as the three-level Likert items. That is, 15 multiple choices and 5 fill-in blanks were analyzed separately as two parts—Lis1 and Lis2. The portion of language application included 10 multiple choices (2 points each) and 10 fill-in blanks (1.5 each). The 10 multiple choices each have four options. No partial credits were given for the 10 fill-in blanks, so they were scored dichotomously and were treated as true-or-false questions. Reading was made up of four short passages. Five multiple-choice questions followed each of the first three passages and accounted for 2 points for each question. Also four options were provided. As for the last passage, the question type was matching question. It included five questions. Students needed to choose five out of six options. Each was awarded two points. Since matching is a special case of the multiple-choice questions, it was considered as the multiple choices with six options in the analyses. The last portion is writing, composed of two essays. The first essay was only the basic writing based on the given topic. The second essay needed students to write based on the information given by a short English passage. They altogether covered 40 points. Generally speaking, the item types included 45 multiple

choices, 15 fill-in blanks, and two essays. The students needed to complete all the questions in two hours. The test was constructed by a classroom teacher of a high school in Guangdong province of China. The general exam format is shown in Table 1.

Table1

General Format of the CEE

	Listening (Sub1)		Language Application (Sub2)		Reading Comprehension (Sub3)		Writing (Sub4)	
	Lis1	Lis2	LanApp		Reading		Writing	
Type	Multiple choice	Fill-in blank	Multiple choice	Fill-in blank	Multiple choice	Matching	Essay1	Essay2
Number	15	5	10	10	15	5	1	1
Options	3		4		4	6		
Points	2 points each	1 point each; partial point (.50) available	2 points each	1.5 points each	2 points each	2 points each	15 points	25 points
	35 points		35 points		40 points		40 points	
	150 points							
Time	Two hours							

Sample

A high school in Guangdong province of China provided the data set. This school is one of the regional best high schools which have large numbers of high-ability students. Therefore, it is a typically comparable high school for the CEE among all the national high schools. The English test was administered among all the third-grade students of this school. There are 1059 students in the third grade. Most of them were selected from all the city middle schools three years ago when they were graduated from the middle school. They were assigned to 20 classes according to their interest. Ten English teachers were in charge of English teaching of the grade, two classes for each teacher. All the 1059 students participated in the exam. The range of the score was 111.50, from 27 to 138.50. The mean total score of the test was 103.27 with the standard deviation of 16.33. Since the objective of the test was to investigate the students' different skills in English, the coefficient alpha values for all the subtests were separately reported as .41 (Lis1), .55 (Lis2), .52 (Sub1), .71 (LanApp), .73 (Reading), and .62 (Writing). The coefficient alpha value for the overall test is .86. The correlations between subtests and between subtests and the total test can be found in Table 2.

Table2

Intercorrelations Between Subtest Scores and the Total Test Score

Subtest	1	2	3	4	5	6	7
1. Lis1	—	.44	.98	.48	.46	.36	.70
2. Lis2		—	.60	.58	.52	.47	.68
3. Sub1			—	.55	.51	.42	.76
4. LanApp				—	.63	.51	.86
5. Read					—	.44	.84
6. Writing						—	.70
7. Total							—

Note. Sub1=Lis1+Lis2.

Computer Program

The version 5.6.3 of Lertap (Curtin University of Technology, 2007) is a recently upgraded version as an Excel-based classical item and test analysis computer program. The first version of Lertap appeared in 1971, which was called “DIEitem” at that time and initially used for analyzing conventional achievement tests and Kuhlmann-Anderson aptitude, or “IQ”, tests. In 1973, the second version of Lertap employed free-form control cards, which were thought of as being more advanced than any other similar program at its time. Lertap2 also introduced the methods of handling affective tests (e.g., Likert items). In this regard, it received great support and encouragement from Ken Hopkins and Gene Glass. By the end of 1974, Lertap2 was widely used in many centers in Canada and the United States (Nelson, 2001). Later, Lertap developed the related versions corresponding to the improvement of microcomputer operating systems. The latest version 5.6.3 of Lertap is of enormous

benefit and convenience because of the application of running within Microsoft's Excel system. Its default operating mode is derived from classical test theory for cognitive test data (e.g., multiple-choice item). Some other advantages of Lertap5 mainly include that: (a) it reports both correlational indices (point-biserial and biserial) and upper-lower groups indices to identify the quality of an item; (b) it reports simple summary item functioning tables in terms of item difficulty, discrimination, and distractor performance; (c) it can change an item's response weights according to the flexible test scoring policy in reality; (d) it can analyze the items with different numbers of options at one whole; and (e) the other criterion measures can be imported in Lertap's item analyses along with the sub-test total score as the default internal criterion (Nelson, 2001). The outputs produced by Lertap5 were mainly used in estimation of item statistics to solve the research question of this study: "Which discrimination index is relatively dependable and simple among the most popularly used ones (r_{pbis}, r_{bis}, $D_{10\%}$, $D_{27\%}$, $D_{33\%}$, and $D_{50\%}$) based on different internal criteria (the subtest-total score and the entire-test-total score)? And how does it apply in the item analysis along with the other item statistics (p and d_{ij})?". Additionally, Lertap5 was compared with SPSS in this study, which is also widely used for item statistical analyses. They both yielded the same item statistics such as the coefficient alpha values for each subtest and the correlations between the item and the different internal criteria.

CHAPTER 4

RESULTS

Dichotomously-scored Items

As the main body of this study, the measures on the dichotomously-scored items were achieved by estimating and analyzing the discrimination indices, difficulty indices, and distractor statistics. These items were assigned into three subtests—Listening, Language Application, and Reading Comprehension, to measure students' different skills in English. Based on the previous studies and the available computer resources, the discrimination indices including the point-biserial correlation, biserial correlation, *D* for 10%, 27%, 33%, and 50% upper and lower groups and their respective difficulty indices were reported by Lertap5, considering the subtest-total score and entire-test-total score as two separate internal criteria. The reports were shown in Table 3 through Table 8. Simultaneously, by means of Figure 1 through Figure 12, straightforward pictures were depicted for convenience of comparing the discrimination indices and difficulty indices for each item in the three subtests corresponding to different internal criteria.

Table 3

The Item Statistics Involving Discrimination Indices and Difficulty Indices Based on the Sub1-total Score for Lis1

Item	p	r_{pbis}	r_{bis}	$p_{10\%}$ [1]	$D_{10\%}$ [2]	$p_{27\%}$ [3]	$D_{27\%}$ [4]	$p_{33\%}$ [5]	$D_{33\%}$ [6]	$p_{50\%}$ [7]	$D_{50\%}$ [8]
Q1	.78	.40	.56	.68	.60	.74	.40	.76	.36	.78	.26
Q2	.84	.37	.55	.75	.49	.81	.31	.82	.27	.84	.19
Q3	.96	.27	.60	.90	.19	.94	.10	.94	.10	.96	.06
Q4	.31	.33	.43	.36	.57	.32	.35	.33	.36	.31	.24
Q5	.96	.27	.60	.89	.22	.94	.12	.95	.10	.96	.08
Q6	.95	.22	.47	.91	.16	.94	.09	.94	.09	.95	.06
Q7	.27	-.07	-.09	.31	-.07	.28	-.06	.28	-.08	.27	-.05
Q8	.73	.19	.25	.74	.25	.73	.19	.73	.19	.73	.16
Q9	.84	.37	.56	.78	.42	.81	.33	.82	.30	.84	.23
Q10	.72	.44	.59	.67	.63	.70	.53	.71	.46	.72	.34
Q11	.45	.38	.47	.56	.71	.52	.52	.50	.45	.45	.26
Q12	.47	.42	.53	.55	.76	.50	.53	.49	.48	.47	.35
Q13	.82	.36	.53	.73	.44	.78	.33	.80	.30	.82	.22
Q14	.81	.43	.63	.73	.53	.78	.40	.79	.36	.81	.27
Q15	.56	.49	.61	.54	.78	.55	.63	.55	.56	.56	.42

Note. [1] is the item difficulty index based on the 10% upper and lower groups;
[2] is the item index of discrimination based on the 10% upper and lower groups;
[3] and [4] are for the 27% upper and lower groups;
[5] and [6] are for the 33% upper and lower groups;
[7] and [8] are for the 50% upper and lower groups.

Table 4

The Item Statistics Involving Discrimination Indices and Difficulty Indices Based on the Entire-test-total Score for Lis1

Item	p	r_{pbis}	r_{bis}	$p_{10\%}$	$D_{10\%}$	$p_{27\%}$	$D_{27\%}$	$p_{33\%}$	$D_{33\%}$	$p_{50\%}$	$D_{50\%}$
Q1	.78	.28	.38	.72	.39	.77	.29	.78	.27	.78	.20
Q2	.84	.29	.44	.78	.38	.82	.26	.83	.22	.84	.15
Q3	.96	.26	.59	.89	.21	.94	.09	.94	.08	.96	.06
Q4	.31	.17	.22	.36	.32	.32	.21	.32	.18	.31	.13
Q5	.96	.26	.57	.91	.18	.94	.10	.95	.09	.96	.07
Q6	.95	.18	.37	.92	.13	.94	.09	.95	.07	.95	.06
Q7	.27	-.27	-.36	.31	-.44	.28	-.27	.28	-.27	.27	-.19
Q8	.73	.06	.07	.68	.10	.71	.04	.73	.02	.73	.02
Q9	.84	.32	.48	.80	.38	.83	.27	.84	.26	.84	.20
Q10	.72	.31	.42	.71	.43	.73	.36	.72	.34	.72	.26
Q11	.45	.28	.35	.61	.48	.52	.38	.50	.36	.45	.26
Q12	.47	.29	.36	.60	.56	.50	.37	.49	.34	.47	.22
Q13	.82	.31	.46	.76	.39	.80	.27	.81	.25	.82	.20
Q14	.81	.41	.60	.71	.55	.78	.37	.80	.34	.81	.22
Q15	.56	.43	.54	.58	.72	.55	.55	.55	.49	.56	.35

Table 5

The Item Statistics Involving Discrimination Indices and Difficulty Indices Based on the Sub2-total Score for LanApp

Item	p	r_{pbis}	r_{bis}	$p_{10\%}$	$D_{10\%}$	$p_{27\%}$	$D_{27\%}$	$p_{33\%}$	$D_{33\%}$	$p_{50\%}$	$D_{50\%}$
Q21	.95	.24	.53	.87	.26	.93	.13	.94	.10	.95	.08
Q22	.92	.34	.61	.77	.44	.87	.24	.89	.21	.92	.15
Q23	.53	.35	.43	.52	.78	.52	.63	.52	.58	.52	.42
Q24	.44	.34	.43	.49	.75	.45	.64	.45	.58	.44	.42
Q25	.63	.27	.34	.57	.60	.62	.52	.63	.48	.63	.34
Q26	.84	.16	.24	.78	.42	.83	.24	.83	.20	.84	.13
Q27	.73	.13	.18	.68	.49	.70	.31	.71	.27	.73	.17
Q28	.19	.14	.20	.36	.44	.23	.24	.22	.22	.19	.16
Q29	.65	.17	.22	.63	.47	.65	.38	.67	.36	.65	.26
Q30	.11	-.03	-.05	.24	.13	.14	.05	.13	.03	.12	.02
Q31	.35	.44	.57	.47	.90	.40	.68	.38	.58	.35	.45
Q32	.91	.30	.53	.79	.42	.89	.21	.88	.21	.91	.14
Q33	.39	.40	.51	.45	.79	.42	.65	.40	.58	.39	.44
Q34	.37	.28	.36	.67	.67	.38	.46	.38	.44	.37	.33
Q35	.86	.45	.70	.58	.66	.79	.41	.81	.36	.86	.26
Q36	.53	.28	.35	.61	.59	.58	.50	.55	.46	.53	.39
Q37	.62	.43	.54	.60	.78	.60	.66	.61	.62	.62	.47
Q38	.87	.39	.61	.69	.61	.81	.35	.84	.31	.87	.21
Q39	.47	.47	.59	.50	.90	.47	.73	.46	.66	.47	.52
Q40	.82	.27	.40	.75	.47	.77	.34	.79	.28	.82	.22

Table 6

The Item Statistics Involving Discrimination Indices and Difficulty Indices Based on the Entire-test-total Score for LanApp

Item	p	r_{pbis}	r_{bis}	$p_{10\%}$	$D_{10\%}$	$p_{27\%}$	$D_{27\%}$	$p_{33\%}$	$D_{33\%}$	$p_{50\%}$	$D_{50\%}$
Q21	.95	.31	.69	.88	.22	.92	.14	.94	.11	.95	.08
Q22	.92	.41	.75	.80	.40	.88	.25	.89	.21	.92	.14
Q23	.53	.41	.52	.55	.70	.52	.53	.52	.46	.53	.33
Q24	.44	.42	.52	.48	.70	.45	.51	.46	.48	.44	.38
Q25	.63	.31	.40	.61	.53	.62	.35	.63	.32	.63	.26
Q26	.84	.22	.34	.83	.28	.84	.16	.84	.16	.84	.12
Q27	.73	.19	.26	.67	.33	.72	.19	.72	.16	.73	.13
Q28	.19	.18	.26	.33	.31	.21	.18	.20	.15	.19	.14
Q29	.65	.23	.29	.66	.37	.66	.28	.65	.24	.65	.19
Q30	.11	.02	.03	.23	.07	.15	.01	.14	.00	.11	.01
Q31	.35	.50	.64	.46	.88	.40	.63	.38	.57	.35	.41
Q32	.91	.35	.61	.82	.35	.88	.22	.89	.19	.91	.14
Q33	.39	.46	.59	.45	.79	.41	.59	.41	.54	.39	.37
Q34	.37	.33	.42	.35	.53	.35	.43	.35	.36	.37	.24
Q35	.86	.52	.82	.65	.71	.79	.39	.82	.34	.86	.24
Q36	.53	.33	.41	.61	.59	.57	.47	.54	.41	.53	.30
Q37	.62	.48	.61	.60	.72	.60	.61	.60	.55	.62	.45
Q38	.87	.52	.82	.67	.66	.81	.37	.83	.32	.87	.22
Q39	.47	.54	.68	.48	.87	.47	.69	.45	.65	.47	.51
Q40	.82	.34	.49	.75	.42	.78	.33	.79	.29	.82	.22

Table 7

The Item Statistics Involving Discrimination Indices and Difficulty Indices Based on the Sub3-total Score for Reading

Item	p	r_{pbis}	r_{bis}	$p_{10\%}$	$D_{10\%}$	$p_{27\%}$	$D_{27\%}$	$p_{33\%}$	$D_{33\%}$	$p_{50\%}$	$D_{50\%}$
Q41	.90	.24	.41	.86	.27	.87	.24	.88	.21	.90	.16
Q42	.93	.31	.60	.84	.32	.90	.19	.91	.15	.93	.10
Q43	.80	.41	.58	.68	.64	.74	.49	.77	.43	.80	.30
Q44	.95	.21	.45	.90	.20	.95	.09	.95	.08	.95	.06
Q45	.83	.25	.38	.73	.54	.79	.31	.80	.26	.83	.18
Q46	.62	.25	.33	.64	.70	.62	.47	.62	.44	.62	.30
Q47	.97	.35	.86	.88	.25	.95	.11	.95	.09	.97	.06
Q48	.90	.21	.36	.82	.35	.88	.20	.89	.16	.90	.10
Q49	.74	.43	.58	.60	.78	.67	.57	.69	.50	.73	.38
Q50	.49	.24	.30	.58	.68	.53	.52	.52	.49	.49	.36
Q51	.82	.26	.38	.76	.47	.81	.37	.81	.33	.82	.24
Q52	.80	.29	.42	.74	.50	.78	.41	.79	.37	.80	.29
Q53	.61	.29	.37	.64	.71	.64	.57	.63	.51	.61	.36
Q54	.51	.24	.30	.56	.73	.51	.50	.51	.48	.51	.33
Q55	.80	.37	.53	.72	.56	.74	.47	.77	.42	.80	.31
Q56	.90	.32	.55	.77	.46	.87	.26	.88	.22	.90	.15
Q57	.91	.30	.53	.82	.37	.88	.22	.89	.21	.91	.15
Q58	.76	.34	.46	.71	.58	.74	.48	.75	.45	.76	.37
Q59	.71	.42	.56	.60	.79	.68	.60	.69	.56	.71	.44
Q60	.89	.29	.48	.78	.43	.86	.26	.87	.22	.89	.16

Table 8

The Item Statistics Involving Discrimination Indices and Difficulty Indices Based on the Entire-test-total Score for Reading

Item	p	r_{pbis}	r_{bis}	$p_{10\%}$	$D_{10\%}$	$p_{27\%}$	$D_{27\%}$	$p_{33\%}$	$D_{33\%}$	$p_{50\%}$	$D_{50\%}$
Q41	.90	.26	.45	.87	.26	.89	.21	.90	.18	.90	.12
Q42	.93	.36	.70	.82	.35	.90	.17	.91	.14	.93	.10
Q43	.80	.48	.69	.66	.68	.74	.46	.76	.42	.80	.29
Q44	.95	.27	.58	.90	.21	.94	.09	.94	.09	.95	.06
Q45	.83	.33	.49	.71	.49	.79	.26	.79	.24	.83	.16
Q46	.62	.31	.39	.62	.55	.58	.36	.59	.32	.62	.23
Q47	.97	.35	.88	.89	.22	.95	.10	.96	.09	.97	.06
Q48	.90	.22	.38	.83	.18	.87	.16	.88	.13	.90	.08
Q49	.74	.52	.70	.61	.75	.68	.58	.69	.53	.74	.36
Q50	.49	.34	.43	.59	.51	.52	.49	.50	.45	.49	.34
Q51	.82	.29	.42	.82	.34	.83	.28	.83	.27	.82	.19
Q52	.80	.37	.53	.78	.42	.76	.40	.78	.36	.80	.23
Q53	.61	.33	.42	.69	.58	.64	.40	.63	.39	.61	.30
Q54	.51	.33	.41	.57	.47	.54	.42	.53	.42	.51	.34
Q55	.80	.44	.63	.72	.55	.76	.43	.77	.41	.80	.29
Q56	.90	.33	.57	.81	.39	.87	.23	.88	.19	.90	.12
Q57	.91	.32	.57	.83	.33	.89	.20	.90	.18	.91	.12
Q58	.76	.33	.46	.74	.48	.77	.35	.78	.30	.76	.22
Q59	.71	.40	.53	.64	.67	.70	.45	.72	.38	.71	.27
Q60	.89	.31	.52	.80	.38	.87	.21	.88	.19	.89	.14

Figure 1. The distribution of discrimination indices for each item in Lis1 based on the Sub1-total score.

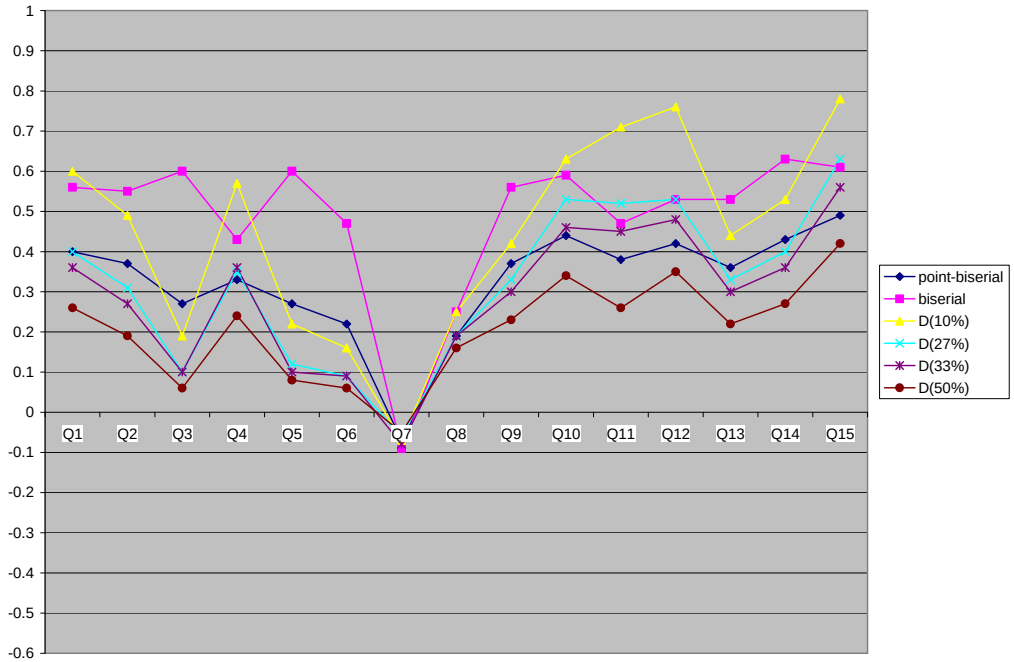


Figure 2. The distribution of difficulty indices for each item in Lis1 based on the Sub1-total score.

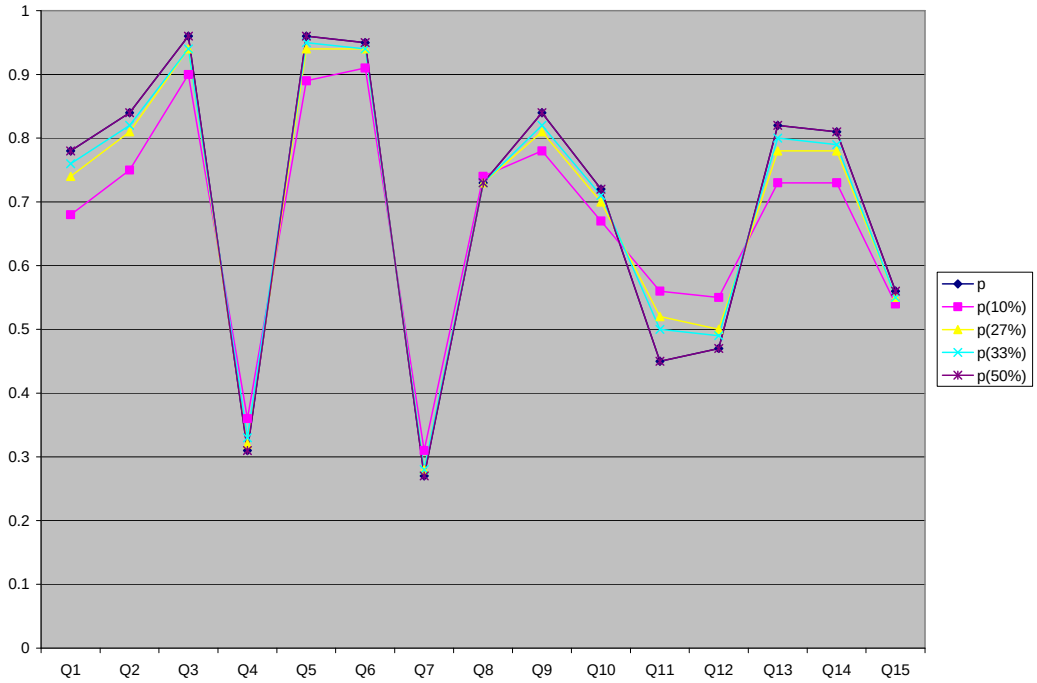


Figure 3. The distribution of discrimination indices for each item in Lis1 based on the entire-test-total score.

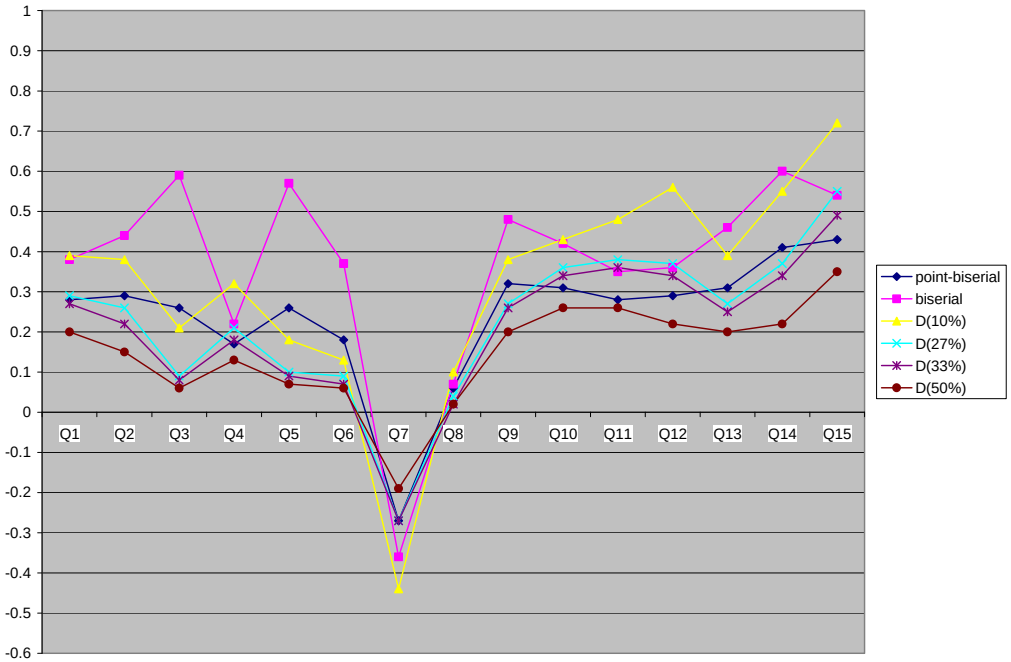


Figure 4. The distribution of difficulty indices for each item in Lis1 based on the entire-test-total score.

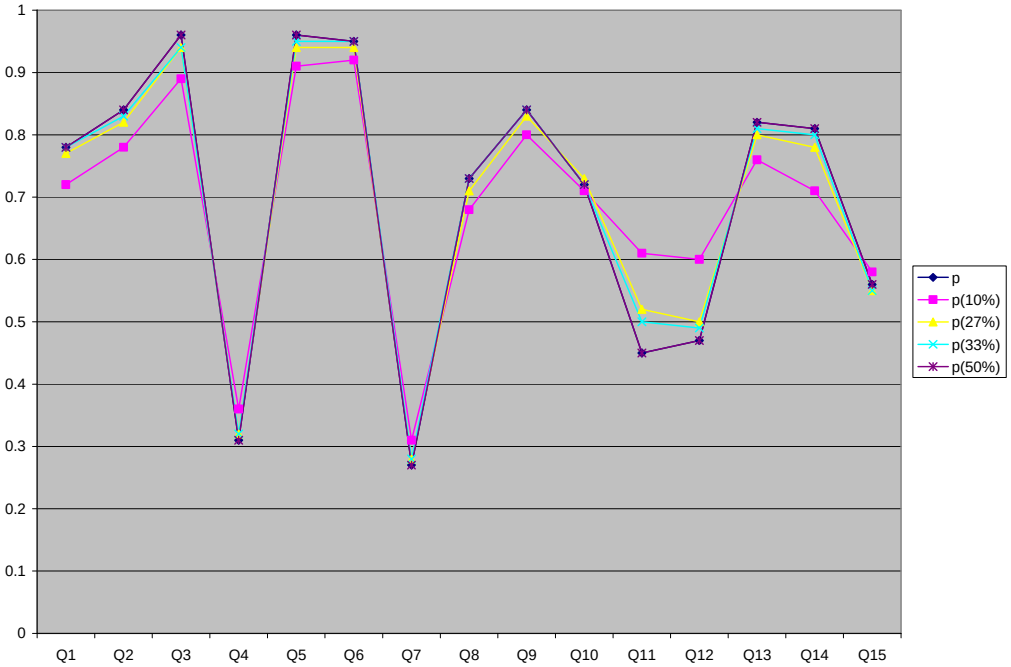


Figure 5. The distribution of discrimination indices for each item in LanApp based on the Sub2-total score.

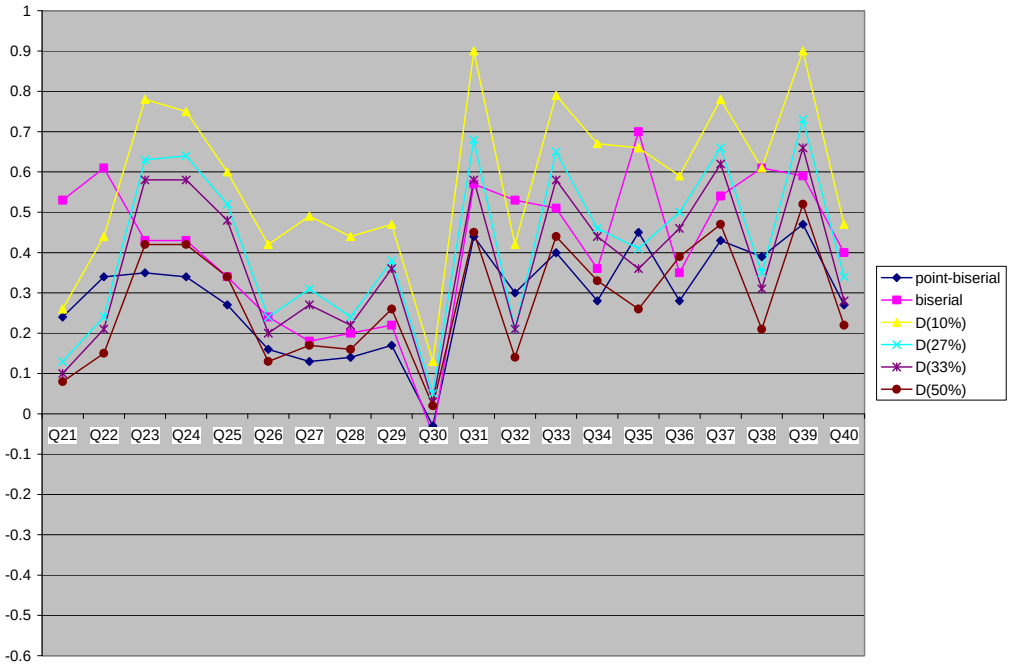


Figure 6. The distribution of difficulty indices for each item in LanApp based on the Sub2-total score.

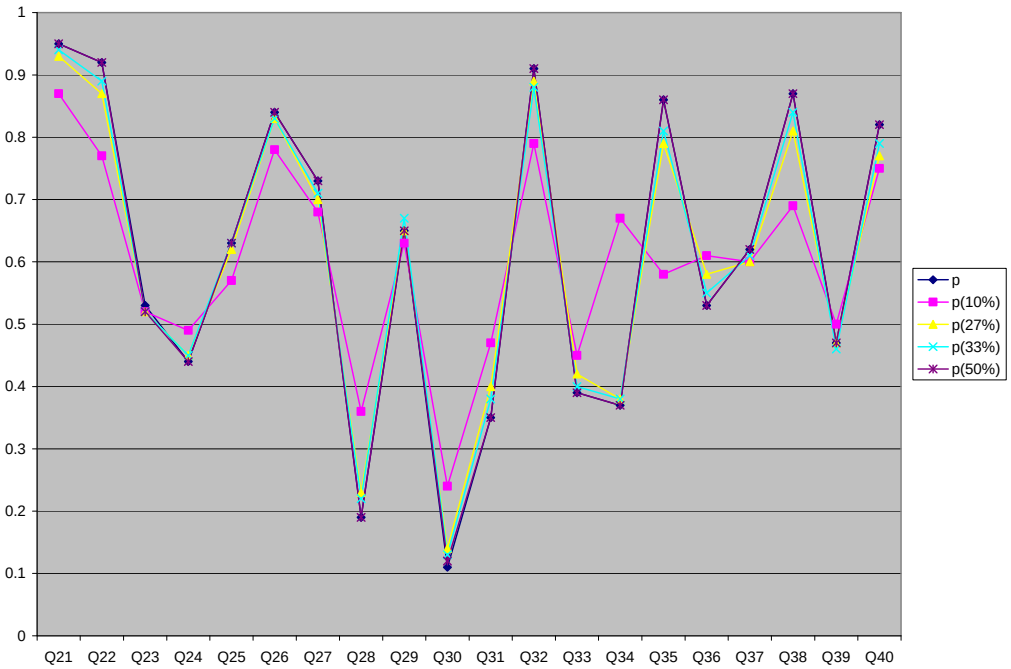


Figure 7. The distribution of discrimination indices for each item in LanApp based on the entire-test-total score.

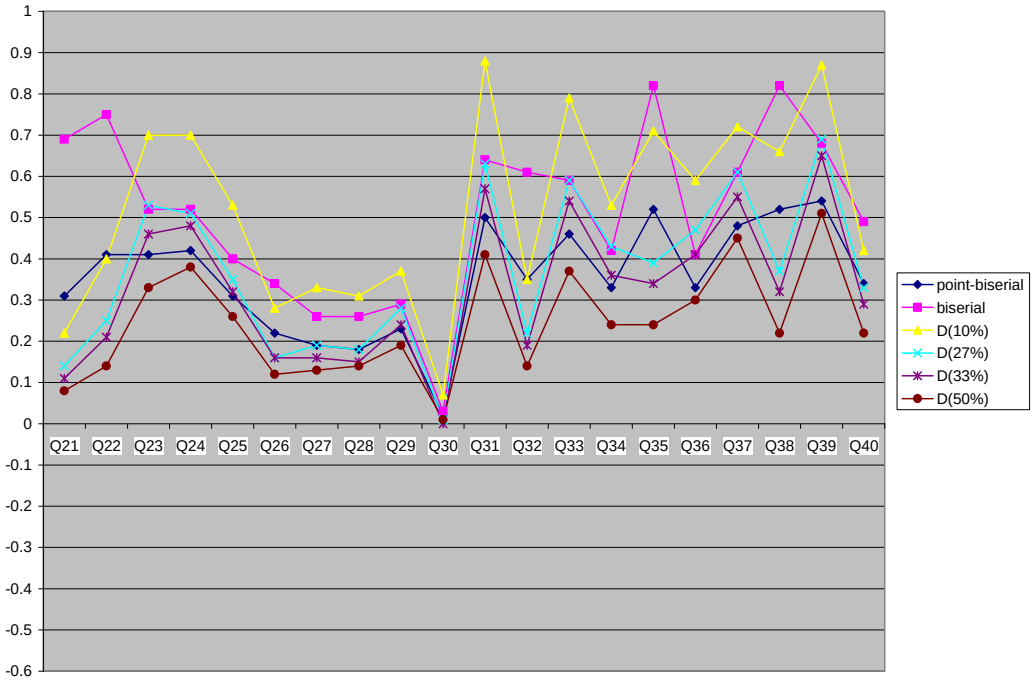


Figure 8. The distribution of difficulty indices for each item in LanApp based on the entire-test-total score.

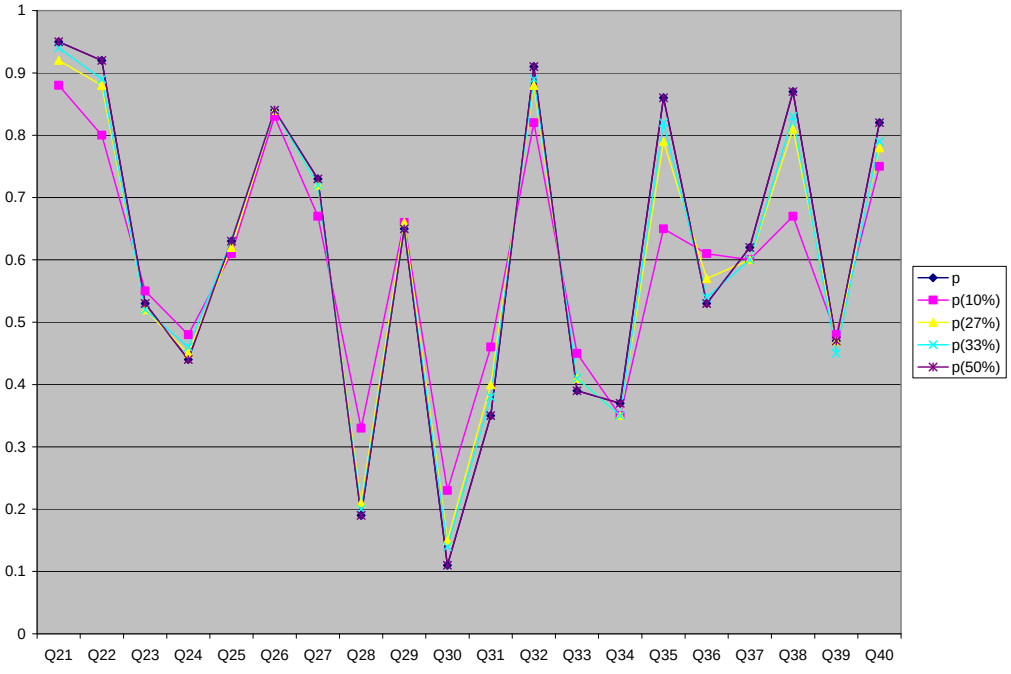


Figure 9. The distribution of discrimination indices for each item in Reading based on the Sub3-total score.

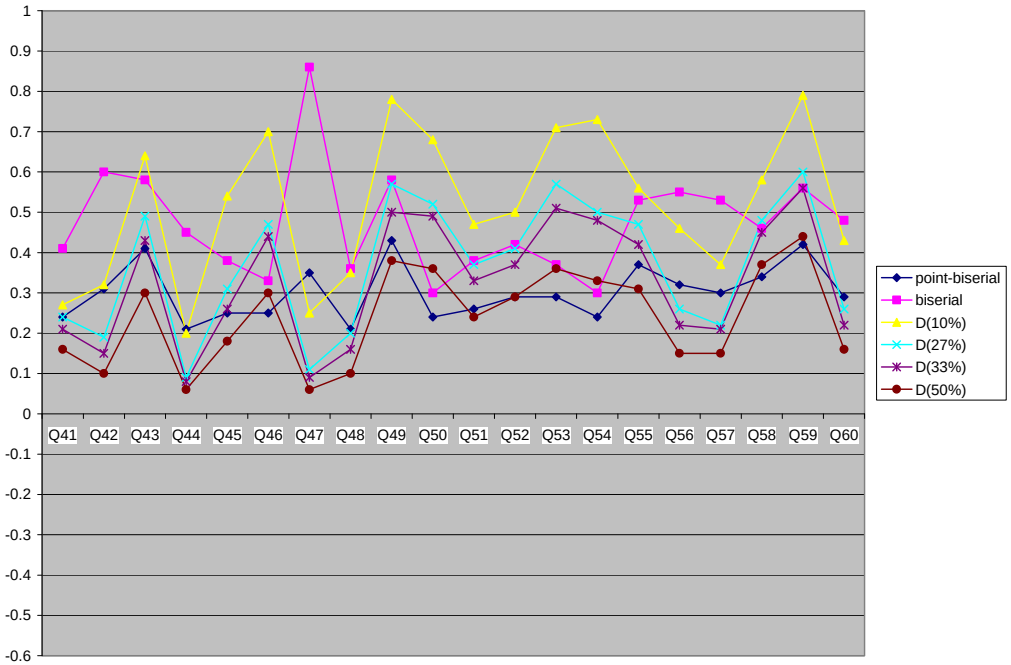


Figure 10. The distribution of difficulty indices for each item in Reading based on the Sub3-total score.

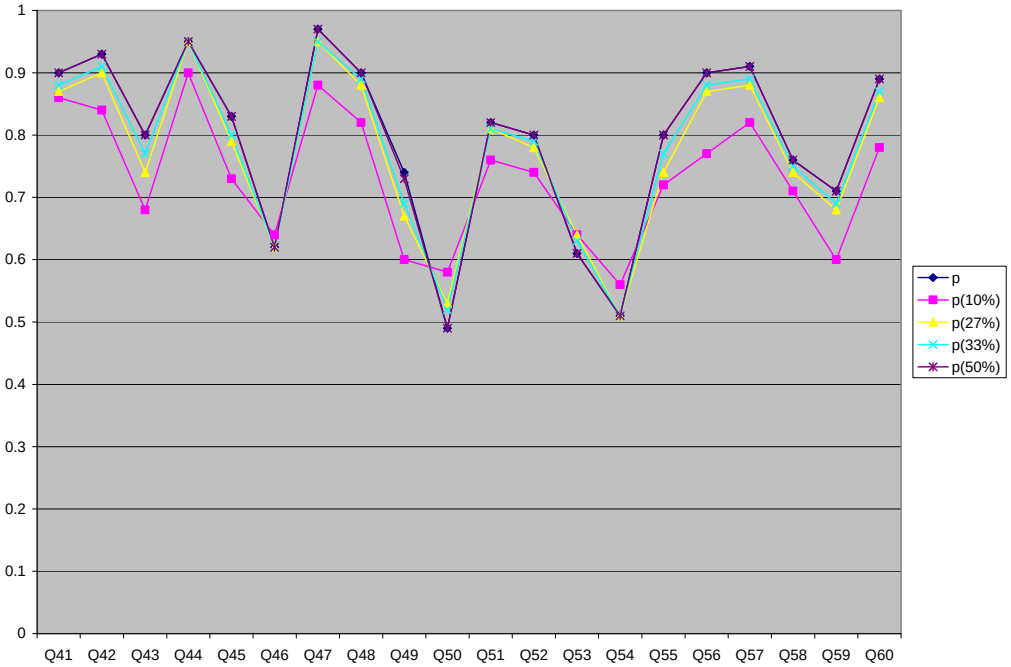


Figure 11. The distribution of discrimination indices for each item in Reading based on the entire-test-total score.

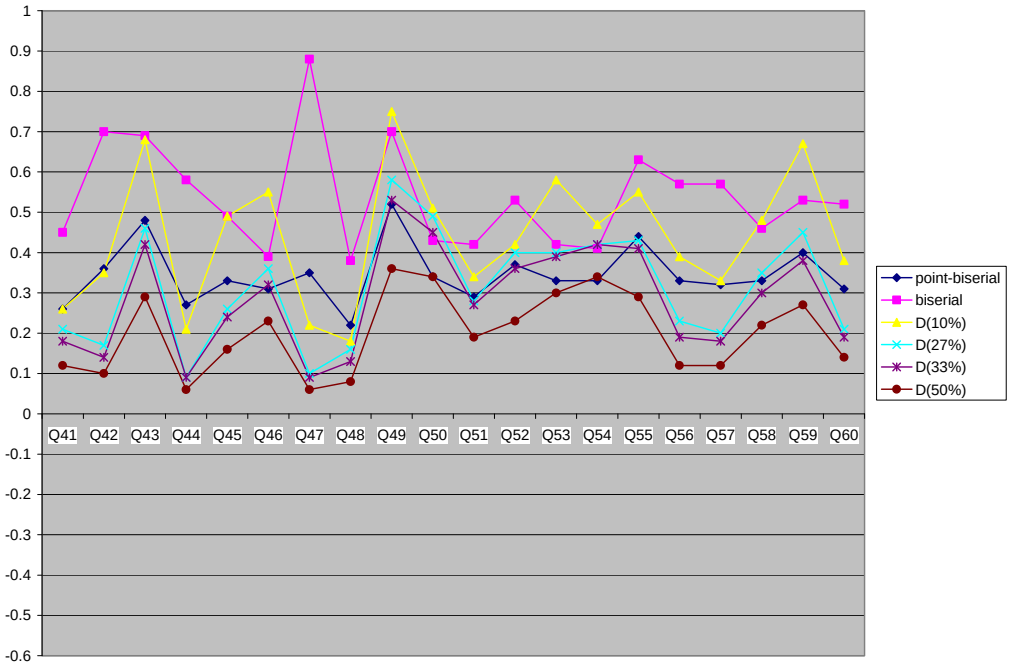
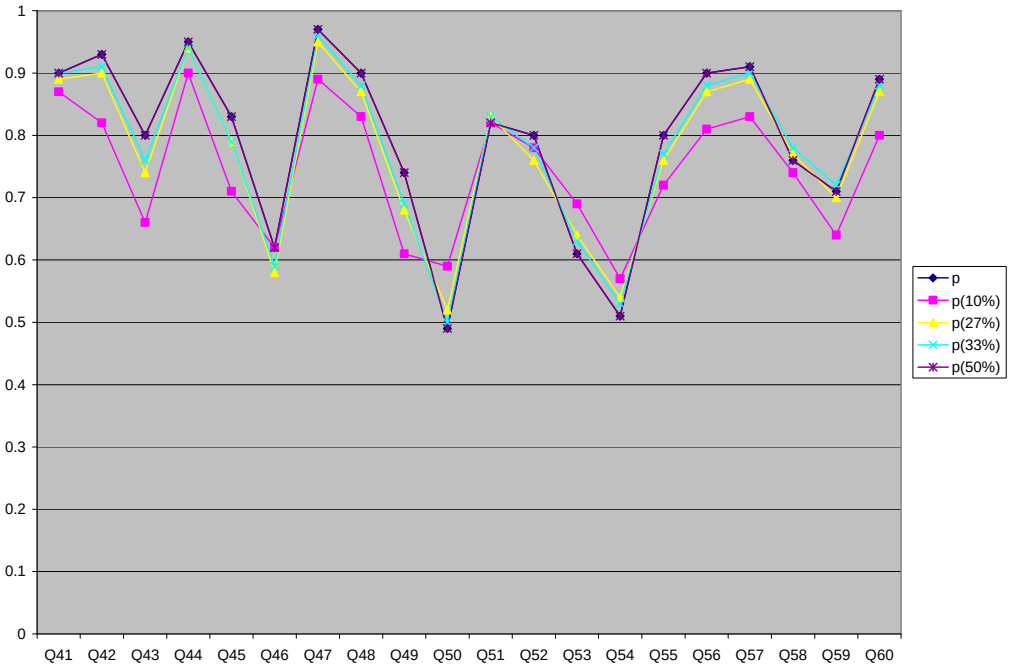


Figure 12. The distribution of difficulty indices for each item in Reading based on the entire-test-total score.



From Table 3 to Table 8, it was obvious that the *p* based on the subtest-total score for each item was identical to the *p* based on the entire-test-total score because values of the single *p* were decided by the number of the students who got the item right in the whole group. It was not impacted by changing the internal criteria. Likewise, the *p* for the 50% upper and lower groups was also the same as the *p*'s mentioned above because the 50% upper and lower groups included all the students. Although the *p* for 50% based on the subtest-total score had a few exceptions for Item 23, 30, and 49, the small differences were ignored because they were simply caused by the rounding error. The *p*'s for other proportions—10%, 27%, and 33% were also very close to the single *p*, no matter whether they were based on the subtest-total score or the entire-test-total score. By contrast, the *p* for 10% was a little deviated from the other *p*'s. Especially, for extreme difficult or easy items, it was much farther from the other *p*'s. That is, for extreme difficulty items, the *p* for 10% seemed to largely underestimate the difficult degree of items while for extreme easy items, the *p* for 10% seemed to largely overestimate the difficult degree in comparison with the other difficulty indices. Actually, through the mathematical deduction, it could be easily understood that for extreme difficult items, the *p* for 10% captured almost the same number of students as the other proportions who got the item right as its numerator, but it has a relatively small number as its denominator compared with other proportions. As such, for the extreme easy items, the *p* for 10% captured much fewer students who got the answer right from the top and bottom 10% of all the students than 27%, 33%, and 50% although it has a small-number denominator. Moreover, the *p*'s for different proportions based on the subtest-total score were very similar with the counterparts based on the entire-total-test score. It implied that the first three subtests contributed the similar information as the entire test on estimating

the difficulty indices at the different-proportion upper or lower group levels. This point could also be easily detected in Figures 2, 4, 6, 8, 10, and 12. Also, given the guidelines (Lord, 1952) of the appropriate *p*'s for items with different options, Items 4, 7, 28, and 30 are too difficult while Items 3, 5, 6, 21, 22, 32, 41, 42, 44, 47, 48, 56, 57, and 60 are too easy.

Furthermore, Figures 1, 3, 5, 7, 9, and 11 showed that all the discrimination indices were roughly close to each other based on either of the internal criteria. By comparison, the biserial correlation and *D* for 10% appeared to depart from the other discrimination indices. As mentioned in the previous chapter, the discrepancy between the biserial correlation and the other discrimination indices was caused by the alternative of correlational formulas rather than the actual differences in the discriminating power between items. The results in the graphs further showed that the value of a biserial correlation is always at least one-fifth larger than the point-biserial correlation and it may be four times larger when the item difficulty extremely ranges. In this study, some items with extreme difficulty indices factually yielded larger values of the biserial correlation such as Items 3, 5, 6, 21, 22, 35, 41, 42, 44, 47, 56, 57, and 60. As to *D* for 10%, it always overestimated the discriminating power of almost all the items no matter what internal criterion was used. It can also be explained through the mathematical logic that *D* for 10% always captured larger values in difference between the upper and lower groups as its numerator but it has a relatively smaller value as its denominator. It was more obvious when the item difficulty indices were extremely large or small. Thus, for a big sample size, 10% was so small that it could not reflect a whole picture to estimate the item statistics. On the contrary, the graphs also showed another rule that *D* for 50% seemed constantly to underestimate the discriminating power of the items based on either of the internal

criteria although the values of *D* for 50% appeared very close to the other discrimination indices. Likewise, it demonstrated the same mathematical logic that *D* for 50% always has the largest value as its denominator. However, since it was not very far from the other discrimination indices, it was not a bad referenced index. In addition, in Figures 1, 3, 5, 7, 9, and 11, the general trend yielded by the six discrimination indices based on the subtest-total score was very consistent to the one based on the entire-test-total score. Thus, generally speaking, the first three subtests contributed the similar information as the entire test on estimating the discrimination indices at the different-proportion upper or lower group levels. According to the rule of thumb (Ebel, 1965), the items with poor discriminating power were listed as followed: Items 3, 4, 5, 6, 7, 8, 21, 22, 26, 27, 28, 30, 32, 41, 42, 44, 47, 48, 56, 57, and 60. These items required further carefully revised. Especially, Item 7 might have to be eliminated after careful check.

It is known that an item with too high or too low difficulty level often yields very poor discriminating power. This study reiterated the relationship between the difficulty index and the discrimination index. Most of the items with low discrimination indices fell into the extreme difficulty range. This was clearly shown in Figures 13 and 14. As for the items with poor discriminating power, it is often recommended to carefully revise them or even remove the ones with serious problems. Distractor analyses can conduce to revising these items in some degree. In virtue of the formula for the discrimination index (*D*), the discrimination coefficient (d_{ij}) for each distractor can also be obtained, which value is normally negative for well-functioning items, but values close or equal to 0 or positive are never desirable. Table 9 showed the discrimination index for 33% and the discrimination coefficient of each distractor for the items with poor discriminating power.

Figure 13. Scatterplot of the relationship between the 33% discrimination index and the 33% difficulty index based on the subtest-total score for the items with poor discriminating power.

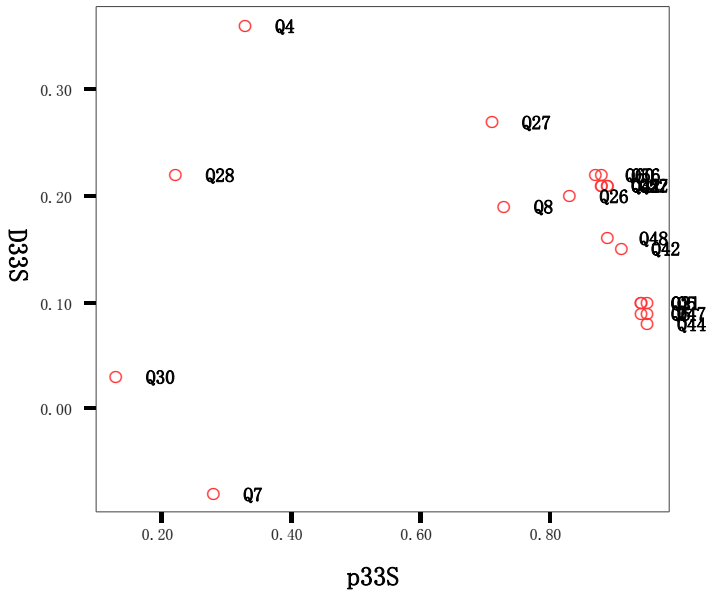


Figure 14. Scatterplot of the relationship between the 33% discrimination index and the 33% difficulty index based on the entire-test-total score for the items with poor discriminating power.

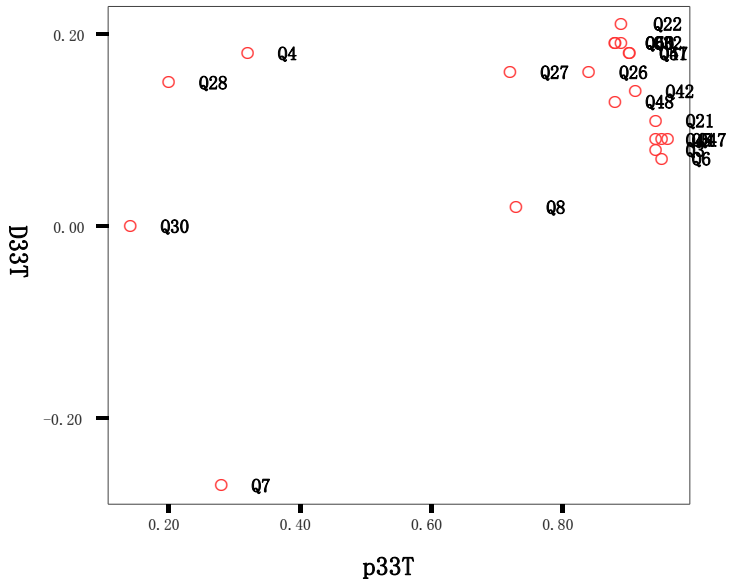


Table 9

The Discrimination Index (D) *for 33% and the Corresponding Discrimination Coefficient* (d_{ij}) *of Each Distractor for the Items with Poor Discriminating Power*

		Q3			Q4			Q5			Q6			Q7			Q8		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
	upper	.99	.00	.00	.51	.00	.49	.00	.00	1.00	.01	.99	.01	.23	.64	.12	.17	.00	.83
$D_{33\%}S$	lower	.89	.04	.07	.15	.02	.83	.08	.03	.89	.06	.90	.04	.32	.44	.24	.33	.03	.63
	d_{ij}	.10	-.04	-.07	.36	-.02	-.34	-.08	-.03	.10	-.05	.09	-.03	-.08	.20	-.12	-.16	-.03	.19
	upper	.98	.01	.01	.41	.00	.59	.01	.00	.99	.01	.98	.01	.14	.72	.13	.26	.00	.74
$D_{33\%}T$	lower	.91	.04	.06	.22	.03	.75	.08	.02	.90	.05	.91	.04	.41	.36	.23	.24	.04	.72
	d_{ij}	.08	-.03	-.05	.18	-.03	-.16	-.07	-.02	.09	-.04	.07	-.03	-.27	.36	-.10	.02	-.04	.02

Note. $D_{33\%}S$ is the $D_{33\%}$ based on the subtest-total score; $D_{33\%}T$ is the $D_{33\%}$ based on the entire-test-total score.
The option underscored is the correct answer.

Table 9 (continued)

The Discrimination Index (D) *for 33% and the Corresponding Discrimination Coefficient* (d_{ij}) *of Each Distractor for the Items with Poor Discriminating Power*

		Q21				Q22				Q26				Q27			
		A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
	upper	.01	.00	.00	.99	.00	1.00	.00	.00	.07	.00	.00	.93	.07	.08	.85	.01
$D_{33\%}S$	lower	.08	.01	.02	.89	.06	.79	.07	.09	.21	.01	.04	.73	.16	.17	.57	.10
	d_{ij}	-.07	-.01	-.02	.10	-.06	.21	-.07	-.09	-.14	-.01	-.04	.20	-.09	-.09	.27	-.09
	upper	.01	.00	.00	.99	.01	.99	.00	.00	.07	.00	.00	.93	.09	.11	.80	.01
$D_{33\%}T$	lower	.09	.01	.02	.88	.06	.79	.07	.09	.19	.01	.03	.76	.13	.13	.64	.10
	d_{ij}	-.08	-.01	-.02	.11	-.05	.21	-.07	-.09	-.12	-.01	-.03	.16	-.04	-.02	.16	-.09

Table 9 (continued)

The Discrimination Index (D) *for 33% and the Corresponding Discrimination Coefficient* (d_{ij}) *of Each Distractor for the Items with Poor Discriminating Power*

		Q28				Q30				Q32		Q41			
		A	B	C	D	A	B	C	D	A	B	A	B	C	D
	upper	.33	.15	.13	.38	.08	.74	.03	.14	.99	.01	.01	.00	.99	.00
$D_{33\%}S$	lower	.11	.23	.14	.52	.18	.66	.04	.11	.78	.22	.13	.06	.78	.03
	d_{ij}	.22	-.08	-.01	-.14	-.10	.08	-.01	.03	.21	-.21	-.12	-.06	.21	-.03
	upper	.28	.18	.15	.39	.07	.76	.03	.14	.99	.01	.01	.00	.99	.00
$D_{33\%}T$	lower	.13	.20	.15	.53	.21	.61	.04	.14	.80	.20	.09	.06	.81	.04
	d_{ij}	.15	-.02	.00	-.14	-.14	.15	-.01	.00	.19	-.19	-.08	-.06	.18	-.04

Table 9 (continued)

The Discrimination Index (D) *for 33% and the Corresponding Discrimination Coefficient* (d_{ij}) *of Each Distractor for the Items with Poor Discriminating Power*

		Q42				Q44				Q47				Q48			
		A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
	upper	.00	.00	.00	.99	.00	.00	.99	.00	1.00	.00	.00	.00	.01	.00	.97	.03
$D_{33\%}S$	lower	.06	.03	.07	.84	.02	.01	.91	.06	.91	.03	.03	.03	.04	.06	.81	.09
	d_{ij}	-.06	-.03	-.07	.15	-.02	-.01	.08	-.06	.09	-.03	-.03	-.03	-.03	-.06	.16	-.06
	upper	.00	.01	.00	.98	.00	.00	.99	.01	1.00	.00	.00	.00	.01	.01	.95	.04
$D_{33\%}T$	lower	.06	.03	.07	.84	.02	.01	.90	.07	.91	.03	.03	.03	.04	.07	.82	.08
	d_{ij}	-.06	-.02	-.07	.14	-.02	-.01	.09	-.06	.09	-.03	-.03	-.03	-.03	-.06	.13	-.04

Table 9 (continued)

The Discrimination Index (D) *for 33% and the Corresponding Discrimination Coefficient* (d_{ij}) *of Each Distractor for the Items with Poor Discriminating Power*

		Q56						Q57						Q60					
		A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
	upper	.00	.00	.00	.00	.00	.99	1.00	.00	.00	.00	.00	.00	.00	.98	.00	.00	.02	.00
$D_{33\%}S$	lower	.05	.08	.04	.01	.05	.77	.79	.10	.02	.01	.08	.00	.03	.76	.01	.01	.17	.02
	d_{ij}	-.05	-.08	-.04	-.01	-.05	.22	.21	-.10	-.02	-.01	-.08	.00	-.03	.22	-.01	-.01	-.15	-.02
	upper	.00	.01	.01	.00	.01	.98	.99	.00	.00	.00	.01	.00	.00	.97	.00	.00	.03	.00
$D_{33\%}T$	lower	.05	.07	.04	.01	.05	.79	.81	.08	.02	.01	.07	.00	.02	.79	.01	.01	.15	.02
	d_{ij}	-.05	-.06	-.03	-.01	-.04	.19	.18	-.08	-.02	-.01	-.06	.00	-.02	.19	-.01	-.01	-.12	-.02

In Table 9, all of the items have low discrimination index as well as the undesirable discrimination coefficient for each distractor. Moreover, the proportions of the students who chose each option in the upper and lower groups were also provided in Table 9. Most of the items with poor discriminating power were caused by the very closely high proportions of the students who chose the right answer in the upper and lower group such as Items 3, 5, 6, 21, 22, 26, 32, 41, 42, 44, 47, 48, 56, 57, and 60 while few people chose the distractors in either of the groups. Although most of the student who chose the distractors fell into the lower group, the difficulty index of these items was too high, so they could not differentiate students very well. In this case, the distractors in these items needed to be revised to make them more attractive to the students in the lower group, but not in the upper group. Take a similar question to Q21 as an example.

Example 1.

My husband had been suggesting that we should get a dog for almost ten years. There were some reasons why the _____ came up.

A. matter B. business C. order D. idea

Looking back to Table 9, few students in the lower groups chose Options B and C, but most of them gathered into the correct option D. Thus, it was better to revise Options B and C. One of the alternative revisions was to change '*B. business'* as '*B. thing'* and '*C. order'* as '*C. animal'*.

Another reason why some items such as Q4, Q7, Q8, Q27, Q28, and Q30 had low discrimination index was because some distractors in these items attracted a lot of students in the upper group. Especially in Q7, the correct option A were chosen by fewer students in the upper group even than the lower group while the exactly opposite case happened to the wrong Option B. It implied a serious problem in Q7.

Take a similar question to Q7 as another example. Q7 was selected from Listening section, so the listening material was also provided below.

Example 2.

What does Mary like best about her job?

A. Scoring goals B. Hearing cheers C. her income

Listening material: "I love to hear cheering when I'm playing, especially when I score a goal. And, of course, the pay is wonderful." Mary said.

Looking at the question and the listening material, it was not difficulty to detect that the stem of the question was sort of ambiguous. One of the revisions was recommended to change it as "What does Mary like when she score a goal?". In this case, the answer was also changed to Option B.

Among these 12 discrimination indices, it is open to doubt which ones are more dependable, simpler and more convenient to get. As a rule of thumb, the point-biserial correlation based on the entire-test-total score ($r_{pbis}T$) was regarded as the benchmark because it was obtained by evaluating each item through all the students and the entire test, which aimed to measure students' comprehensive English ability. However, in some situation, it may not be the most convenient and simplest one. Thus, in order to fulfil the initial objective of this study, the 11 other discrimination indices were further compared by looking at the mean of the differences between them and $r_{pbis}T$ for each subtest as a unit. Also, the standard deviations of the differences were also provided to ensure the validity of the comparison. They were shown in Table 10, Figures 15 and 16. In addition, since there is some relationship between the discrimination index and its corresponding difficulty index, the comparison between the difficulty indices were also conducted to guarantee the reasonable and comprehensive evaluation for each item. It also considered the difficulty index, the

single *p*, based on the entire-test-total score (*pT*) as the yardstick. The mean of the differences for each of the first three subtests and its standard deviation between the other 9 difficulty indices and *pT* were presented in Table 11, Figures 17 and 18.

Table 10

The Mean and the Standard Deviation of the Differences between the Point-biserial Correlation Based on Total Score ($r_{pbis}T^{7}$) and All the Other Discriminating Indices

Subtest	$r_{pbis}S^{1}$	$r_{bis}S^{2}$	$D_{10\%}S^{3}$	$D_{27\%}S^{4}$	$D_{33\%}S^{5}$	$D_{50\%}S^{6}$	$r_{bis}T^{8}$	$D_{10\%}T^{9}$	$D_{27\%}T^{10}$	$D_{33\%}T^{11}$	$D_{50\%}T^{12}$
The Mean of the Difference between *rpbisT* and the Discriminating Indices above											
Lis1	.086	.247	.224	.133	.116	.097	.139	.127	.067	.067	.127
LanApp	.063	.070	.230	.142	.123	.088	.154	.178	.095	.085	.110
Reading	.047	.134	.193	.109	.109	.118	.193	.120	.084	.089	.145
The Standard Deviation of the Difference between *rpbisT* and the Following Discriminating Indices											
Lis1	.058	.053	.148	.085	.066	.066	.089	.078	.052	.053	.109
LanApp	.019	.068	.118	.067	.059	.104	.103	.112	.058	.066	.094
Reading	.032	.112	.128	.079	.071	.084	.107	.087	.065	.072	.076

Note. [1], [2], [3], [4], [5], [6] are based on the subtest-total score. [7], [8], [9], [10], [11], [12] are based on the entire-test-total score.

Figure 15. The mean of the differences between $r_{pbis}T$ and the other 11 discrimination indices in the first three subtests.

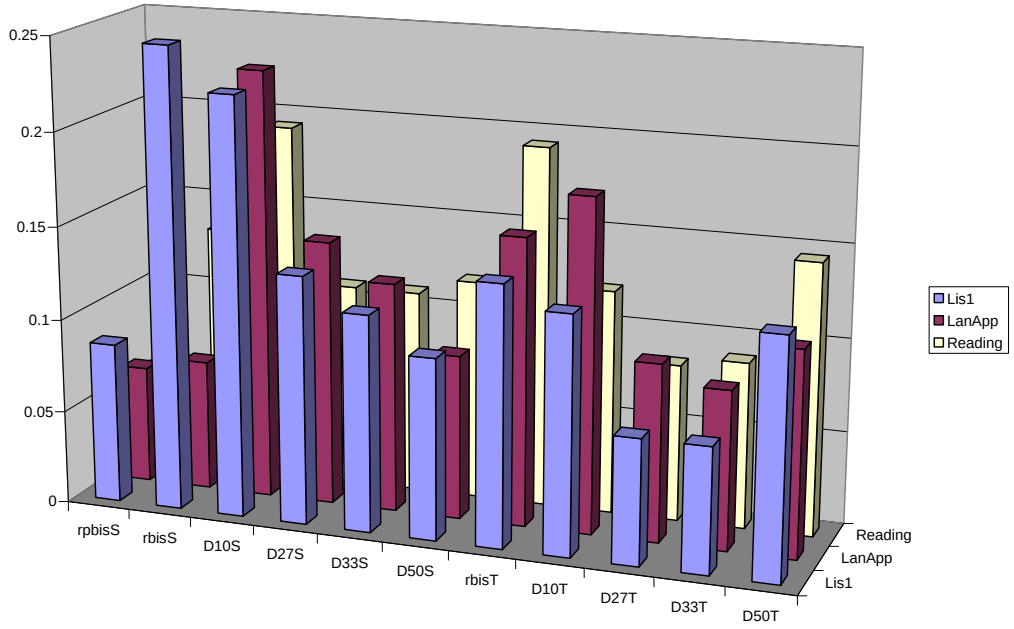


Figure 16. The standard deviation of the differences between $r_{pbis}T$ and the other 11 discrimination indices in the first three subtests.

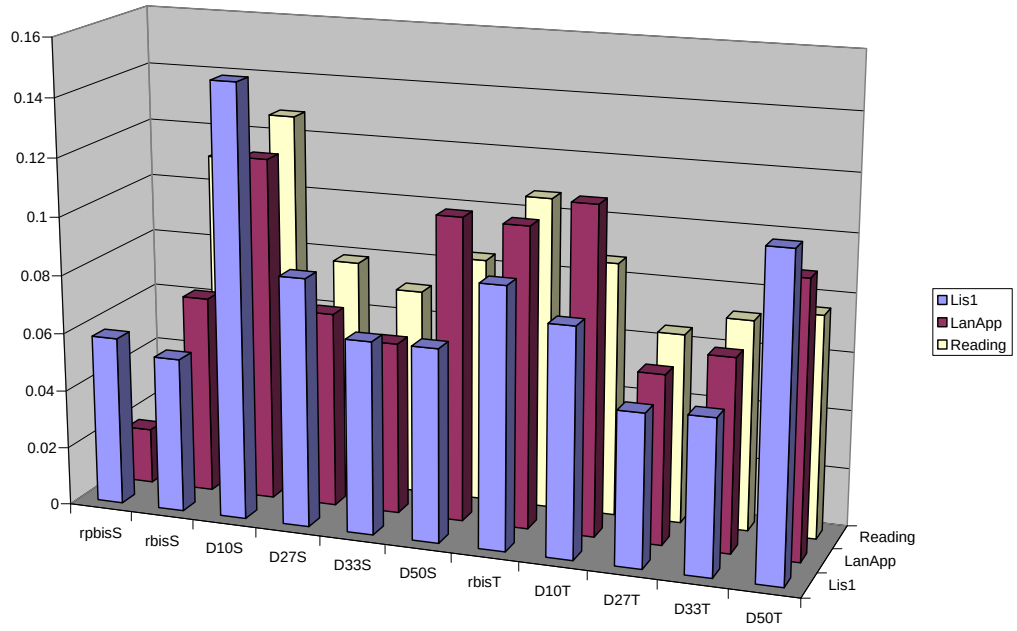


Table 11

The Mean and the Standard Deviation of the Differences between the Difficulty Index Based on Total Score (pT^6) and All the Other Difficulty Indices

Subtest	pS^1	$p_{10\%}S^2$	$p_{27\%}S^3$	$p_{33\%}S^4$	$p_{50\%}S^5$	$p_{10\%}T^7$	$p_{27\%}T^8$	$p_{33\%}T^9$	$p_{50\%}T^{10}$
The Mean of the Differences between *pT* and the Difficulty Indices above									
Lis1	.0000	.0633	.0247	.0173	.0000	.0620	.0200	.0107	.0000
LanApp	.0000	.1020	.0285	.0195	.0010	.0740	.0265	.0185	.0000
Reading	.0000	.0795	.0285	.0185	.0005	.0710	.0285	.0200	.0000
The Standard Deviation of the Differences between *pT* and the Difficulty Indices above									
Lis1	.0000	.0289	.0172	.0109	.0000	.0402	.0156	.0128	.0000
LanApp	.0000	.0812	.0208	.0119	.0031	.0607	.0195	.0123	.0000
Reading	.0000	.0812	.0208	.0119	.0031	.0607	.0195	.0123	.0000

Note. ^{1}, ^{2}, ^{3}, ^{4}, ^{5} are based on the subtest-total score. ^{6}, ^{7}, ^{8}, ^{9}, ^{10} are based on the entire-test-total score.

Figure 17. The mean of the differences between *pT* and the other 9 difficulty indices in the first three subtests.

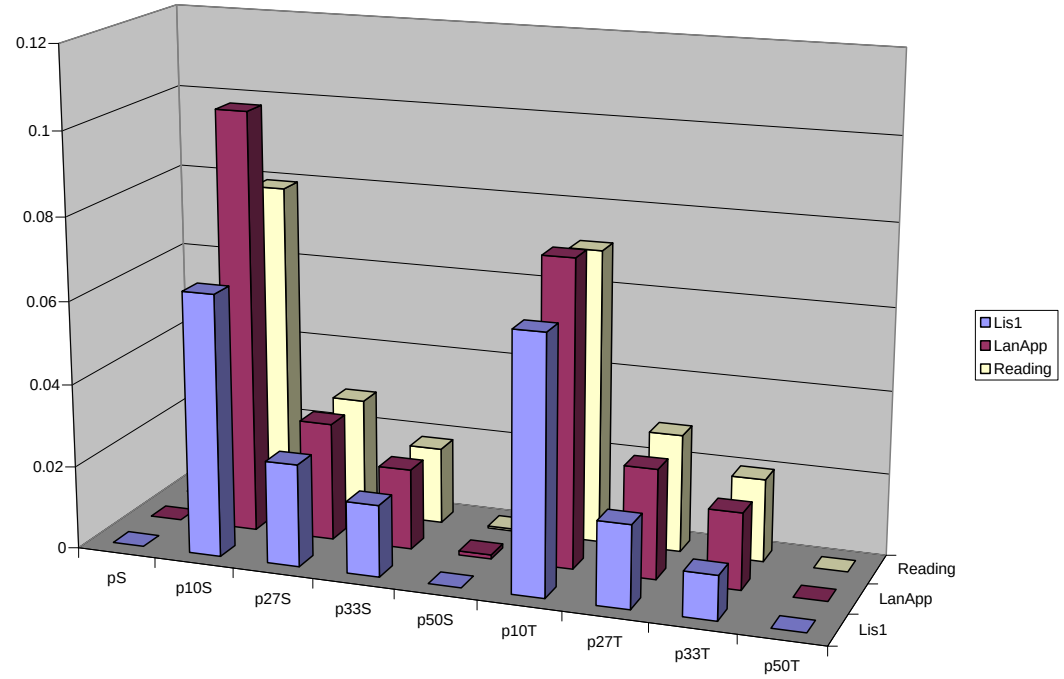
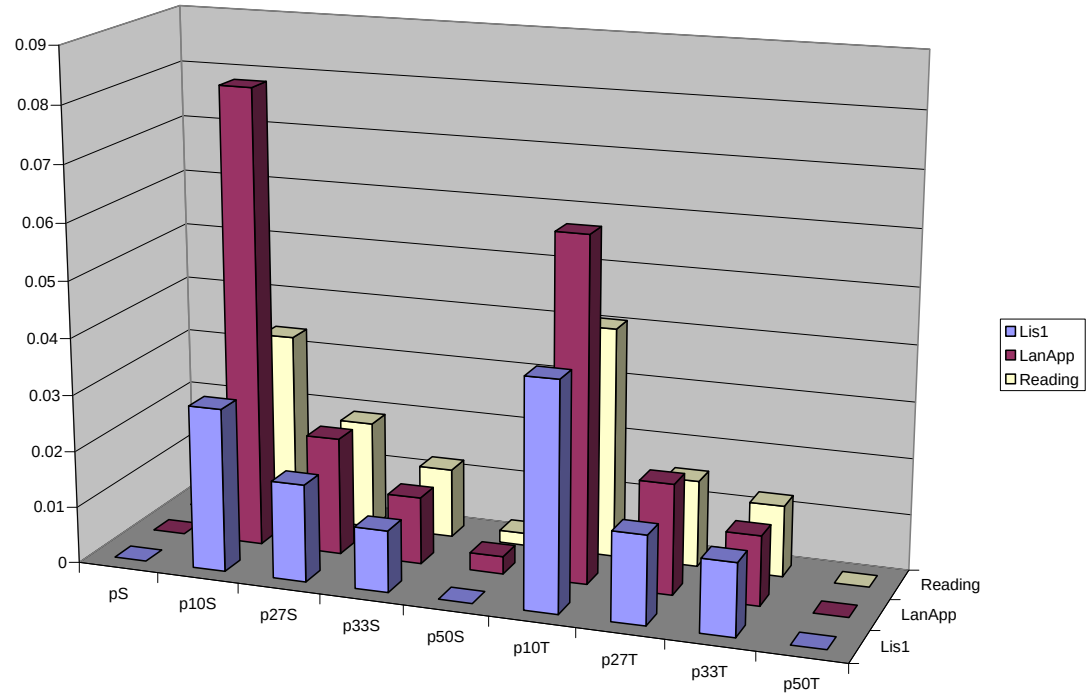


Figure 18. The standard deviation of the differences between *pT* and the other 9 difficulty indices in the first three subtests.



Tables 10 and 11 showed the specific mean of the differences among discrimination indices and difficulty indices and its standard deviation for each of the first three subtests, compared with $r_{pbis}T$ and pT. Figures 15 through 18 presented a clear and straightforward picture. In Figure 15, $r_{pbis}S$ had the smallest average differences for each of the subtest while the relatively smaller average differences were obtained for $D_{27\%}S$, $D_{33\%}S$, $D_{50\%}S$, $D_{27\%}T$, $D_{33\%}T$, and $D_{50\%}T$. However, Figure 16 showed $D_{50\%}S$ and $D_{50\%}T$ had too large standard deviations. Thus, for the subtest level, the $r_{pbis}S$ was the best one and $D_{33\%}S$ and $D_{27\%}S$ can also be considered. For the entire test level, there is no doubt that the $r_{pbis}T$ is the best one. Additionally, $D_{27\%}T$ and $D_{33\%}T$ were also taken into account. Correspondingly, the comparison of difficulty indices was shown in Table 10 and Figures 17 and 18. As mentioned previously, the pS was completely the same as the pT. By contrast, the $p_{33\%}$ is closer to the yardstick for either of the score levels. Overall, for the subtest level, the $r_{pbis}S$ and the $D_{33\%}S$ can be considered to estimate the discriminating power of items while for the entire test level, the $D_{33\%}T$ can also be used other than the $r_{pbis}T$.

Polytomously-scored Items

Lis2 in the Listening section and the writing section were not dichotomously-scored. Students can get partial credits for the imperfect answers. The writing section was included by two parts—essay1 and essay2. Essay1 was scored from three aspects—Wording (8 points), Content (5 points), and Consistency (2 points) while essay2 was scored from two aspects—Generalization (5 points) and Thematic writing (20 points). The points in the parentheses referred to the full score in the

respective aspect. In this study, each scoring aspect was regarded as an assumed item which was supposed to measure some specific writing skill of students. As stated previously, for polytomously-scored items, the mean, the standard deviation, and the correlations between each item and the subtest-total score or the entire-test-total score were required for the item analysis. Since the full scores of different scoring aspects in Writing were different, the mean and the standard deviation between these aspects were not comparable. Thereby the mean and the standard deviation were converted to the proportional ones by dividing them by their respective full score. All the item statistics mentioned above were shown in Table 12 through Table 15. In addition, Figure 19 through Figure 22 presented the patterns of the mean and the standard deviation.

Table 12

The Mean and the Standard Deviation of Each Item in Lis2

	Q16	Q17	Q18	Q19	Q20
Mean	.79	.66	.97	.93	.47
Standard Deviation	.41	.31	.12	.20	.41

Table 13

The Product-moment Correlation Coefficient of Each Item in Lis2 versus the Sub1-total Score and the Entire-test-total Score

	Product-moment Correlation Coefficient	
Item	Lis2 vs Sub1	Lis2 vs Total
Q16	.36	.36
Q17	.48	.58
Q18	.21	.27
Q19	.24	.32
Q20	.47	.52

Table 14

The Mean and the Standard Deviation of Each Scoring Aspect in Writing (Sub4) and Their Respective Proportional Values

	Essay1			Essay2	
	Wording	Content	Consistency	Generalization	ThWriting
Mean	5.14	3.10	1.13	2.85	13.35
Standard Deviation	0.97	0.71	0.35	0.75	2.38
Proportional *M*	.64	.62	.57	.57	.67
Proportional *SD*	.12	.14	.18	.15	.12

Note. ThWriting refers to Thematic Writing.
Proportional *M* is the proportional mean.
Proportional SD is the proportional standard deviation.

Table 15

The Product-moment Correlation Coefficient of Each Scoring Aspect in Writing (Sub4) versus the Sub4-total Score and the Entire-test-total Score

	Product-moment Correlation Coefficient	
	Writing vs Sub4	Writing vs Total
Wording	.68	.49
Content	.62	.39
Consistency	.48	.31
Generalization	.69	.53
ThWriting	.90	.63

Figure 19. The mean and the standard deviation of each item in Lis2.

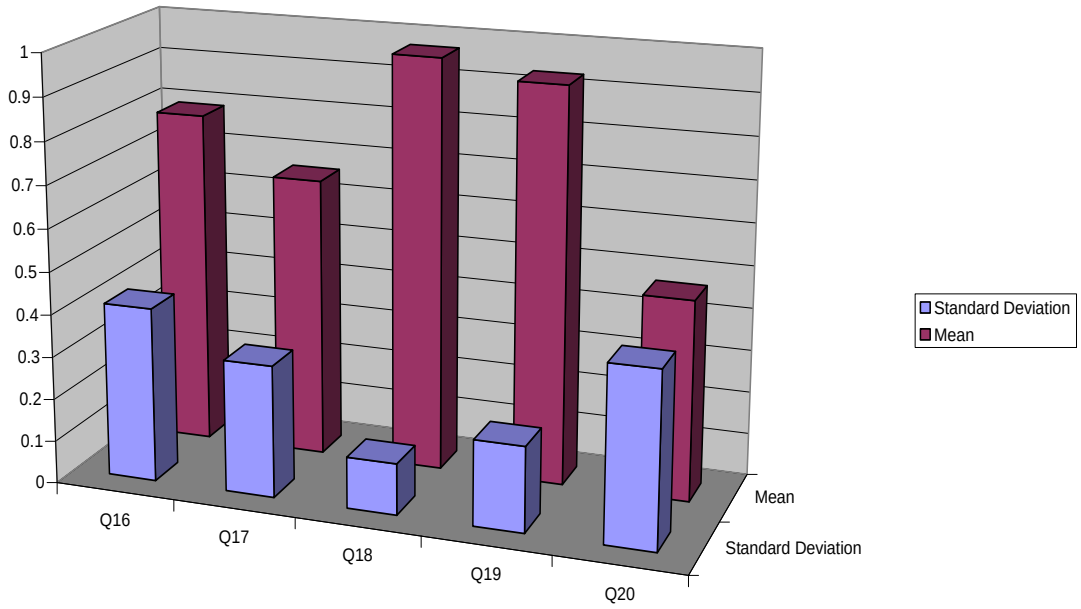


Figure 20. The Product-moment correlation coefficient of each item in Lis2 versus the Sub1-total score and the entire-test-total score.

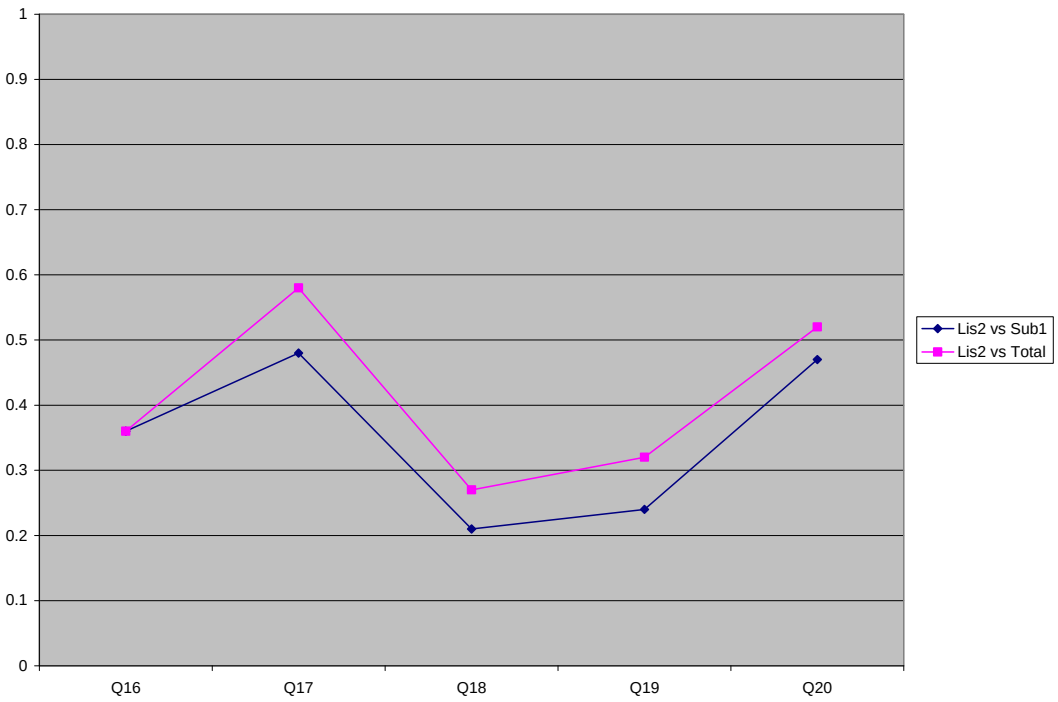


Figure 21. The proportional mean and standard deviation of each scoring aspect in Writing (Sub4).

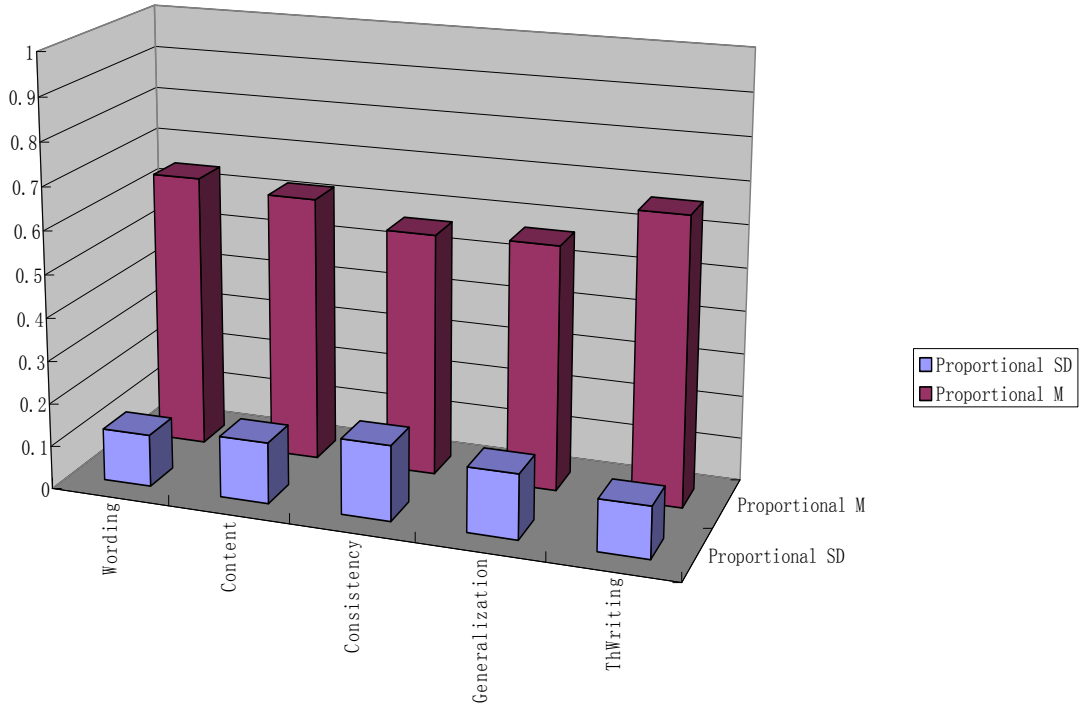
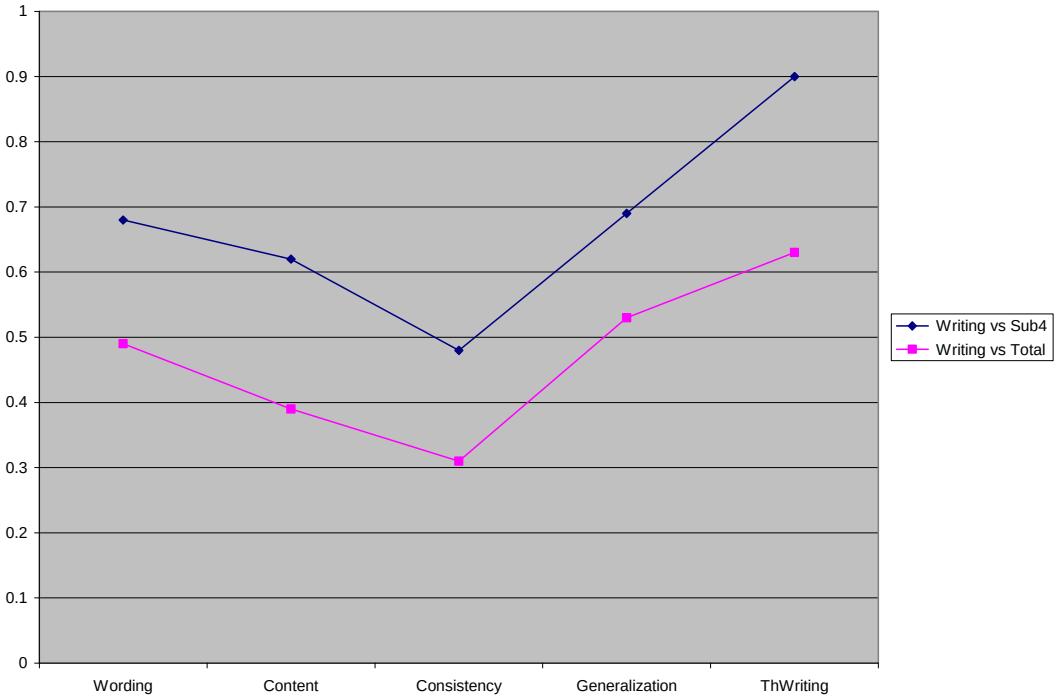


Figure 22. The Product-moment correlation coefficient of each scoring aspect in Writing (Sub4) versus the Sub4-total score and the entire-test-total score.



In Figures 20 and 22, the trend of the correlations tended to be similar for either of the subtests no matter which internal criteria were used, the subtest-total score or the entire-test-total score. It apparently showed that the correlation between the item and the subtest-total score was all higher than the correlation between the item and the entire-test-total score. That was because the item score covered more proportion in the subtest-total score than in the entire-test-total score. Figures 19 and 20 showed Q18 and Q19 had very high average scores but small standard deviations and correlations with either of the internal criteria. It implied that these two items had low discriminating power. However, Q20 had a low average item score but a high standard deviation and correlation with both of the internal criteria. Thus, Q20 was an ideal item to differentiate students with different ability levels. In addition, Q17 could also be considered as a good one. In Figure 21, the average proportional scores and their respective proportional standard deviation of each scoring aspect were very close to each other. By contrast, Consistency seemed to have a higher proportional standard deviation, so it was supposed to have a high discriminating power. However, it had the relatively lowest correlations with both of the internal criteria in Figure 22. It implied that the high-scored students in Consistency might not get the high score in the subtest of writing and the overall test. Generalization and Thematic Writing can, therefore, be considered to have a strong discriminating power because they had a relative high proportional standard deviation and correlations. In addition, the students in the third grade seemed to be good at the thematic writing while the students who were scored highly in the subtest and the overall test should get more training in terms of the aspect of consistency in writing.

CHAPTER 5

SUMMARY AND DISCUSSION

Summary

The objective of the study was to find out the efficient and simple discrimination indices for the classroom item analysis by comparing some popularly used ones based on different internal criteria. In this study, the subtest-total score and the entire-test-total score were employed as the two different internal criteria. The comparison was conducted by means of the modern computer technology. This study also aimed to apply the selected discrimination indices to the classroom or school item analysis, especially for the high school teachers. Additionally, difficulty indices and distractor statistics were also combined with the selected discrimination indices to analyze the items as a whole.

As mentioned previously, there have been many studies related to the comparisons between different discrimination indices, especially in 1930s through 1960s. However, most of the comparisons were fulfilled by using the sample data set collected from the national or state-wide official exams rather than the regular classroom or school-sized exams. However, the classroom item analysis is also in great demand among the high school teachers, especially in the society where the intense competition exists among schools in pursuit of the high acceptance by colleges. The classroom item analysis not only improves the testing quality but also helps to orientate the teachers' instructions corresponding to the student's learning status. Furthermore, no related comparison was priorly conducted in the other cultural and educational setting. Decades ago, the limited computer technology and

accessibility also restricted the development of the related studies. Therefore, this study had some superior characteristics including that (a) the comparison were carried out based on the data set provided by a regular high school; (b) the data set were collected from China and the items analyzed in the study were constructed by a common classroom teacher; (c) the updated computer programs (Lertap5) and SPSS were employed in estimating the item statistics and made the analysis more efficient and precise; (d) the exam used in the study involved four different subtests, which measured students' four different English skills, and also involved different item types including multiple choices, fill-in blanks, matching, and essays; and (e) the study also used the subtest-total score as the internal criterion other than the entire-test-total score. Finally, the study attempted to make the complicated and sophisticated item analysis become simple and straightforward for the specific targeted population—high school teachers.

Normally, a student with a high total score was supposed to have a high score in each of the subtests. Thus, it is open to doubt if the subtest score can provide the similar information as the entire-test-total score in estimating the item statistics. In virtue of the modern computer programs—Lertap5 and SPSS, the item statistics were obtained in terms of the discrimination indices, the difficulty indices, and the distractor statistics based on these two internal criteria. The entire-test-total score was always used as the internal criteria in the classical item analysis, so the $r_{pbis}T$ was generally regarded as the dependable discrimination index for the dichotomously-scored items. Thus, this study compared the $r_{pbis}T$ as the benchmark with the other discrimination indices including $r_{pbis}S$, $r_{bis}S$, $D_{10\%}S$, $D_{27\%}S$, $D_{33\%}S$, $D_{50\%}S$, $r_{bis}T$, $D_{10\%}T$, $D_{27\%}T$, $D_{33\%}T$, and $D_{50\%}T$. As a result, the

other discrimination indices yielded the similar trend in measuring the discriminating power of the items. Therein, the $r_{bis}S$, $D_{10\%}S$, $r_{bis}T$, and $D_{10\%}T$ produced the relatively larger gap with the $r_{pbis}S$ in the first three subtests. Theoretically, the $r_{bis}S$ and $r_{bis}T$ were caused by the recognized computing difference yielded by the formula between the r_{pbis} and r_{bis} . The $D_{10\%}S$ and $D_{10\%}T$ resulted from the fact that 10% was not representative for a large sample size. In addition, although the $D_{50\%}$ yielded the similar value to the r_{pbis} , it also yielded a very large standard deviation in the differences from the r_{pbis} , so the $D_{50\%}$ is not optimal. By contrast, the ones very close to the $r_{pbis}T$ were the $r_{pbis}S$, $D_{27\%}S$, $D_{27\%}T$, $D_{33\%}S$, and $D_{33\%}T$. However, the values of the $D_{27\%}S$ and $D_{27\%}T$ for each item fluctuated more than the values of $D_{33\%}S$ and $D_{33\%}T$. Thus, this study recommended to use the $r_{pbis}S$, $D_{33\%}S$, and $D_{33\%}T$. In the subtest level, $r_{pbis}S$ is the most appropriate one if the related computer program is available. As known, obtaining the $r_{pbis}S$ needs to analyze the original responses of all the students to the items in the related subtest. It is time-consuming if doing it by hand or by some rudimentary computer programs. Thus, if no direct values of the $r_{pbis}S$ are provided by the computer program, the $D_{33\%}S$ can be considered because it only analyzes the original responses of part of the students. It saves time and is easy to calculate through the other computer program such as Excel and SPSS. Likewise, in the entire test level, the $r_{pbis}T$ is a priority unless no direct outputs are provided by the computer programs. Otherwise, $D_{33\%}$ is a good choice to measure the item discriminating power in some other situations.

The difficulty indices are also very important in judging the quality of the items. Since it has a close relationship with the discrimination indices, the study also conducted the comparisons between the respective difficulty indices. In theory, there are no difference between pT, pS, $p_{50\%}S$, and $p_{50\%}T$. Also, as expected, $p_{33\%}S$ and $p_{33\%}T$ had the smallest difference with the pT. They were separately used in the similar situations as mentioned in terms of the corresponding discrimination indices. Moreover, the study also verified the recognized relationship between the difficulty indices and the discrimination indices. That is, the items with extreme difficulty level, too easy or too difficulty, always have the poor discriminating power. After checking the difficulty index and the discrimination index of the item, some problem items were sorted out for further revision or even elimination. In this regard, the distractor statistics like d_{ij}, which is supposed to be negative, can provide more reliable information in revising these items. It is not desirable that the d_{ij} is close to or equal to 0, especially larger than 0. Two examples were given in this study to show how to utilize the d_{ij} in the process of the item analysis.

As for the polytomously-scored items, the correlations between the item score and the subtest-total score as well as between the item score and the entire-total score yielded the similar pattern. That is, the subtests contributed the similar information as the overall test to evaluating the quality of items. Moreover, the related item analysis can be achieved by comparing the mean, the standard deviation, and the correlation between the item and the internal criteria. It is always preferable that a polytomously-scored item has a moderate mean value compared to the full score, a large standard deviation and a strong correlation with the internal criteria.

Discussion

Nowadays, since most of the researchers concentrated on the studies related to IRT, CTT seemed to be outdated and impractical. CTT is, however, still widely used by many professors, testing constructor, and developers. Especially CTT has some unnegligible advantages over IRT in some situations. The previous studies have also demonstrated that CTT can always yield the similar results as IRT. This study has applied the classical item-analyzed technique to one of the situation. That is, the prospective users are simply the classroom teachers in the high school, who have never or seldom been exposed to this technique before. Also, most of them have to focus lots of time and energy on the teaching and instructions to their students. Thus, an efficient and simple item analysis system is more applicable to the specific group. Compared to IRT, CTT is relatively easy to understand and calculate, especially with the help of the modern computer programs. As is known, there are many correlated indices in the classical item analysis. In the classroom and school setting, which one is more appropriate is in question. This study not only placed the classical technique to the regular high school test in China, but also made an extensive comparison between these indices based on two different internal criteria. The test applied in the study was representative of the regular classroom test in the high school of China. Considering the career characteristics of high school teachers, the study made great efforts to investigate the feasible and efficient discrimination indices and other related item statistics to fulfil the item analysis in the classroom. For example, if the specific analysis computer programs are available and operable such as Lertap5 and Iteman, the teacher can easily analyze the items by the direct outputs regarding to $r_{pbis}S$, $r_{pbis}T$, pT, pS, and d_{ij} provided by these programs. If the teacher can only collect the original responses of students regarding some certain subtest or he is only

interested in the items in one of the subtests, he can use $r_{pbis}S$, pS, and d_{ij} to check the quality of the subtest items because $r_{pbis}S$ and pS can lead us to get the similar results as $r_{pbis}T$ and pT do. On the other hand, it can save lots of time. Besides, if the teacher has no knowledge on running the eligible computer programs, the item analysis can also be fulfilled by using Excel, which is almost installed in each of the computers. Most of the classroom teachers are proficient in using Excel, so $D_{33\%}S$, $D_{33\%}T$, $p_{33\%}S$, $p_{33\%}T$, and d_{ij} can bring the item analysis into effect. The indices stated above are derived from some simple arithmetic computation. In virtue of the function commands in Excel, these indices can easily obtained. $D_{33\%}S$ and $p_{33\%}S$ serve for the subtest level. Otherwise, $D_{33\%}T$ and $p_{33\%}T$ can be used.

In addition, the indices recommended by this study can also benefit the classroom teachers and school administrators in some other regards. Through these indices, the item analysis becomes more scientific and informative. The school administrator can give a more reasonable and objective evaluation to the teacher's capability in test construction. They can also investigate the students' learning status and the teachers' teaching efficiency for the different classes in the grade. As for some pretests before the national or state-wide exam, they can roughly predict the students' future performance in these official large-scale exams. The indices can provide more specific information for the classroom teachers. Take the English test in this study as an example. In the subtest of the language application, the first ten multiple choices tested the students' knowledge on part of speech, which included prepositions, conjunctions, verbs, adjectives, adverbs and nouns. If the discrimination index and difficulty index of one item appeared relatively inappropriate compared to the other items, one reason was that the item required further modification. Another reason may

be that the students were kind of weak in such type of items. The further investigation was needed for the teacher. As such, if the students were randomly assigned into the classes of the grade, the indices of some items resulted from a class were largely deviated from the other classes or the overall grade. The teachers in that class may have to adjust their instructions in the corresponding respect. In a word, the indices made possible to compare the classes, teachers, and schools in general or even in very detail.

However, the conclusion drawn by the study was only based on one school data set. Although it verified some recognized relationships between the indices, further similar studies are still needed to guarantee the practicality and accuracy of the conclusion. Based on the prior studies, $D_{27\%}$ was always recommended to use. Thus, more investigations on $D_{27\%}$ in competition with $D_{33\%}$ are suggested in future. As stated previously, the English test in the study employed many types of items (i.e., multiple choices, fill-in blanks, matching, and essays). It determined that the other analyzing methods may be also considered in evaluating the polytomously-scored items. For example, one section of the fill-in blanks was treated as the multiple-choice questions with two options in this study. It is possible to analyze them as the original form (fill-in blanks) by some specific computer program. Moreover, the study applied the classical item analysis to the English test, so the conclusion is still in question if it is used in the tests of other subjects. The other subjects may involve some different item types. In order to apply the technique to a large range of high school, more analyzing methods need to be explored. Additionally, another study is expected to extend the application of the classical item analysis to China's school setting. That is, the students are normally assigned to different classes according to their interests when they are in the second year of the high school. The interests include physics,

chemistry, geography, biology, history, politics, and art. Many high-ability students may gather in one class, but the students in the art class may be weak in the regular school subjects. Also, the students with different interests may tend to be advantageous in some aspect. These special situations will greatly challenge the conclusion of this study. If the test is taken in some specific class with lots of extreme high-ability or low-ability students, it is doubtful that the teacher in that class can still confidently use the indices recommended by this study. Moreover, an English teacher normally teaches two classes, which may focus on two different interests. Any different combination of the interests may lead to different conclusions related to the item statistics. Based on the assumptions mentioned above, the analysis seemingly appears to be complicated. However, above all, the indices first need to be carefully checked on the basis of a small sample size such as around 60 students in a regular class. The study also made some partial investigation regarding this topic, which was shown in the Appendix.

REFERENCES

Adams, J. F. (1960). The effect of non-normally distributed criterion scores on item analysis techniques. *Educational and Psychological Measurement, 20,* 317-320.

Aleamoni, L. M., & Spencer, R. E. (1969). A comparison of the biserial discrimination, point biserial discrimination, and difficulty indices in item analysis data. *Educational and Psychological Measurement, 29,* 353-358.

Attali, Y., & Fraenkel, T. (2000). The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement, 37,* 77-86.

Bowers, J. A. (1972). A note on comparing *r* biserial and *r* point biserial. *Educational and Psychological Measurement, 32,* 771-775.

Beuchert. A. K., & Mendoza. J. L. (1979). A Monte Carlo comparison of ten item discrimination indices. *Journal of Educational Measurement, 16,* 109-117.

Brennan, R. L. (1972). A generalized U-L item discrimination index. *Educational and Psychological Measurement, 29,* 353-358.

Brogden, H. E. (1949). A new coefficient: Application to biserial correlation and to estimation of selective efficiency. *Psychometrika, 14,* 169-182.

Camp, B. H. (1931). *The mathematical part of elementary statistics*. New York: D. C. Heath.

Carroll, J. B. (1987). Correcting point-biserial and biserial correlation coefficients for chance success. *Educational and Psychological Measurement, 47,* 359-360.

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Statistics, 25,* 31-45.

Crock, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* Belmont, CA: Wadsworth.

Cureton, E. E. (1959). Note on phi/phi max. *Psychometrika, 24,* 89-91.

Davis, F. B. (1949). Item-analysis data: Their computation, interpretation, and use in test construction. *Harvard Educational Papers, 2,* 1-42.

Dixon, W. J., Brown, M. B., Engelman, L., Frane, J. W., Hill, M. A., Jennrich, R. I., & Toporek, J. D. (1981). *BMDP statistical software.* Berkeley: University of California Press.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Englehart, M. D. (1965). A comparison of several item discrimination indices. *Educational and Psychological Measurement, 2,* 69-76.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381.

Findley, W. G. (1956). A rationale for the evaluation of item discrimination statistics. *Educational and Psychological Measurement, 16,* 175-180.

Flanagan, J. C. (1939). General consideration in the selection of test items and a short method of estimating the product-moment coefficient from data at the tails of the distribution. *Journal of Educational Psychology, 30,* 674-680.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). Educational research: An introduction. White Plains, NY Longman.

García-Pérez, M. A. (1987). A finite state theory of performance in multiple-choice tests. In E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (Vol. 1, pp. 455-464). Amsterdam: Elsevier.

García-Pérez, M. A. (1993). In defence of none of the above. *British Journal of Mathematical and Statistical Psychology, 46,* 213-229.

García-Pérez, M. A., & Frary, R. B. (1989). Psychometric properties of finite state scores versus number-correct and formula scores: A simulation study. *Applied Psychological Measurement, 13,* 403-417.

García-Pérez, M. A., & Frary, R. B. (1991). Finite state polynomic item characteristic curves. *British Journal of Mathematical and Statistical Psychology, 44,* 45-75.

Glass, G. V. (1965). A ranking variable analogue of biserial correlation: Implications for short-cut item analysis. *Journal of Educational Measurement, 2,* 91-95.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education* (4th ed.). New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests.* New York: John Wiley and Sons, Inc..

Hales, L. W. (1972). Method of obtaining the index of discrimination for item selection and selected test characteristics: A comparative study. *Educational and Psychological Measurement, 32,* 929-937.

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), 38-47.

Harris, C. W., & Subkoviak, M. J. (1986). Item analysis: A short-cut statistic for mastery tests. *Edcuational and psychological measurement, 46,* 495-507.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika, 28,* 211-218.

Henrysson, S. (1971). Gathering, analyzing and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130-159). Washington, DC: American Council on Education.

Hills, J. R. (1981). *Measurement and evaluation in the classroom* (2nd ed.). Columbus, OH: Charles E. Merrill.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1986). *Educational and psychological measurement and evaluation* (7th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Ivens, S. H. (1971). Non-parametric item evaluation index. *Educational and Psychological Measurement, 31,* 843-849.

Kelley, T. L. (1939). Selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17-24.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological development* (Vol. 1, pp. 159-168). Greenwich, CT: JAI.

Lord, F. M. (1952). The relationship of the reliability of multiple choice items to the distribution of item difficulties. *Psychometrika, 18,* 181-194.

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

Lykken, D. T. (1968). Statistical significance of psychological research. *Psychological Bulletin, 70,* 155-159.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62,* 921-943.

Magnusson, D. (1967). *Test theory*. Boston: Addison-Wesley.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

Ndalichako, J. L., & Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement, 57,* 580-589.

Nelson, L. R. (2001). *Item analysis for tests and surveys using Lertap5.* West Australia: Curtin University of Technology.

Nitko, A. J., & Hsu, T. (1984, April). *Item analysis appropriate for domain-referenced classroom testing.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Oosterhof, A. A. (1976). Similarity of various item discrimination indices. *Journal of Educational Measurement, 13,* 145-149.

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika, 1,* 33-49.

Shannon, G. A., & Cliver, B. A. (1987). An application of item response theory in the comparison of four conventional item discrimination indices for criterion-referenced tests. *Journal of Educational Measurement, 24,* 347-356.

Thorndike, R. L. (1982). Educational measurement: Theory and practice. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology: Contributions of latent trait theory* (pp. 3-13). Princeton, NJ: ERIC Clearinghouse of Tests, Measurements, and Evaluations. (ERIC Document Reproduction Service No. ED 222 545)

Tinsley, H. E., & Dawis, R. V. (1977). Test-free person measurement with the Rasch simple logistic model. *Applied Psychological Measurement, 1,* 483-487.

APPENDIX

The study above mainly investigated the efficient and simple discrimination indices based on two internal criteria and also discussed the results combined with the difficulty indices and distractor analyses. The conclusion was based on the data set which included the original responses from 1059 students of the third grade in a high school of China. If the later more investigations carry out and their results can also verify the recommendations made by this study, it will greatly help the classroom teachers and the school administrators to evaluate the items and adjust their instructions according to the students' learning status. However, this study can most likely give some suggestions on the school or grade-scale tests. As for the individual teachers, they might more care about the item quality in their own tests which is constructed for the one or two-class students. In the case, it is doubted that the recommendations made by this study on the discrimination indices are still feasible and appropriate. Thus, in order to give more efficient recommendations to the single classroom teacher, another tentative study were also conducted based on the same data set. The data set was composed of twenty classes which focused on seven learning interests, so these twenty classes with different characteristics were totally not the same including the number of the students in one class. Yet, the present study is simply to find out if it can make the similar recommendation as the previous one based on a small sample size. Also, in China a class normally has sixty students. Thus, Class 4 orientated in Physics was selected from these twenty classes because it has the regular class size—fifty-nine students. The similar analysis procedures were made for Class 4. Some results were shown in Figures 1 through 12.

Figure 1. The distribution of discrimination indices for each item in Lis1 based on the Sub1-total score for Class 4

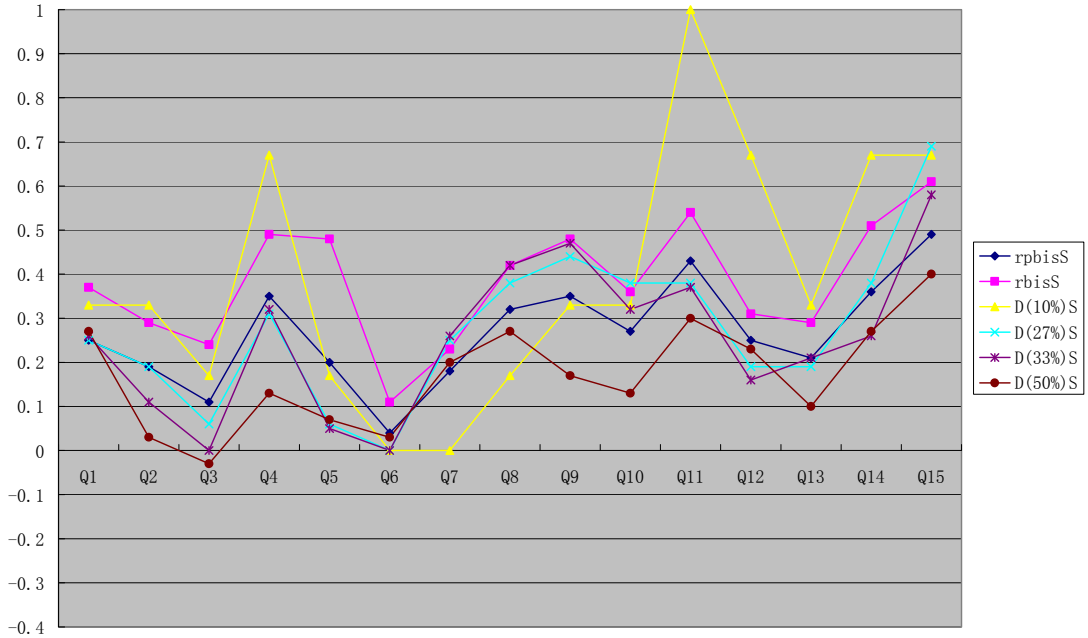


Figure 2. The distribution of discrimination indices for each item in Lis1 based on the entire-test-total score for Class 4.

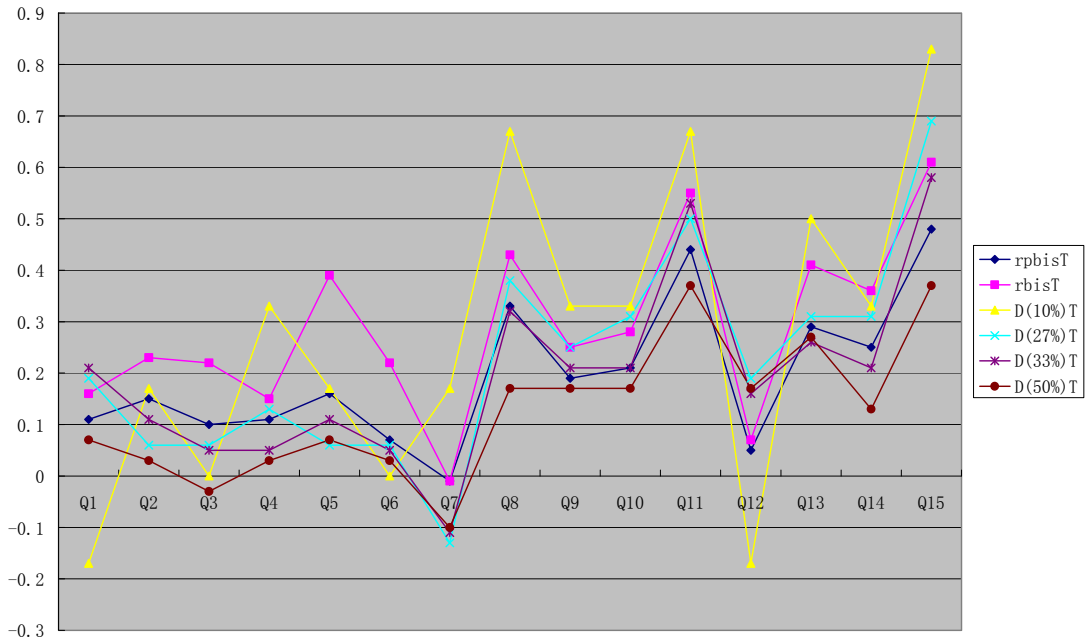


Figure 3. The distribution of difficulty indices for each item in Lis1 based on the Sub1-total score for Class 4.

Figure 4. The distribution of difficulty indices for each item in Lis1 based on the entire-test-total score for Class 4.

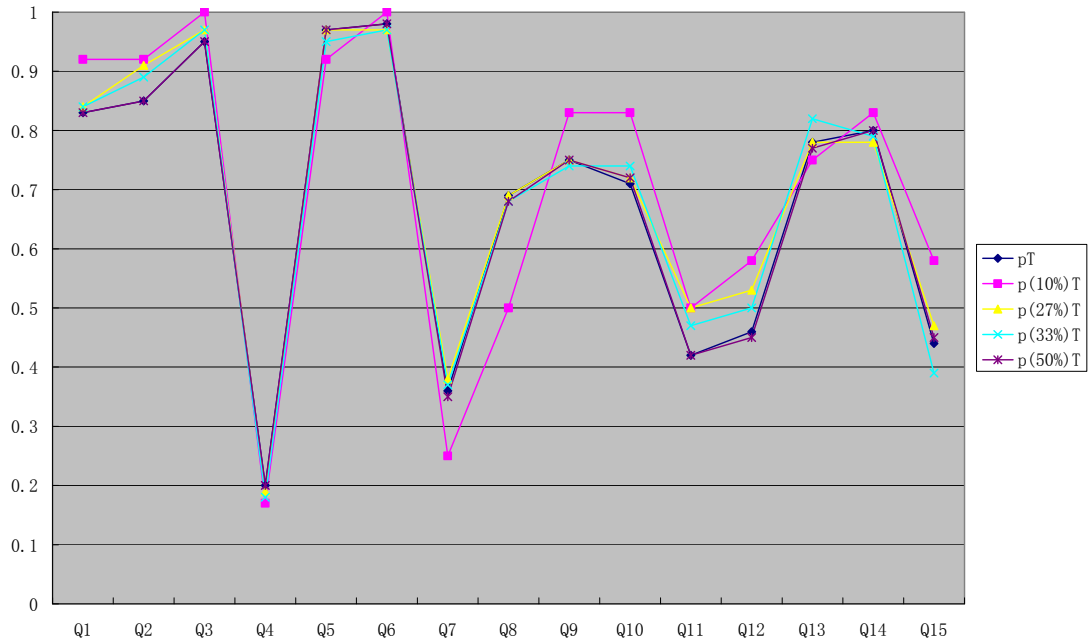


Figure 5. The distribution of discrimination indices for each item in LanApp based on the Sub2-total score for Class 4.

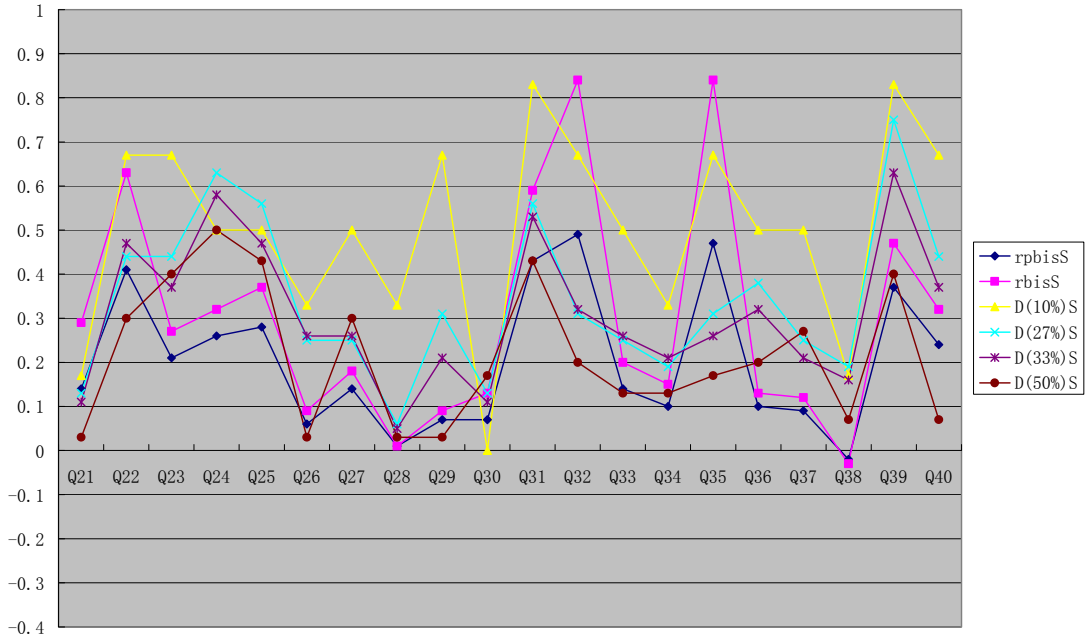


Figure 6. The distribution of discrimination indices for each item in LanApp based on the entire-test-total score for Class 4.

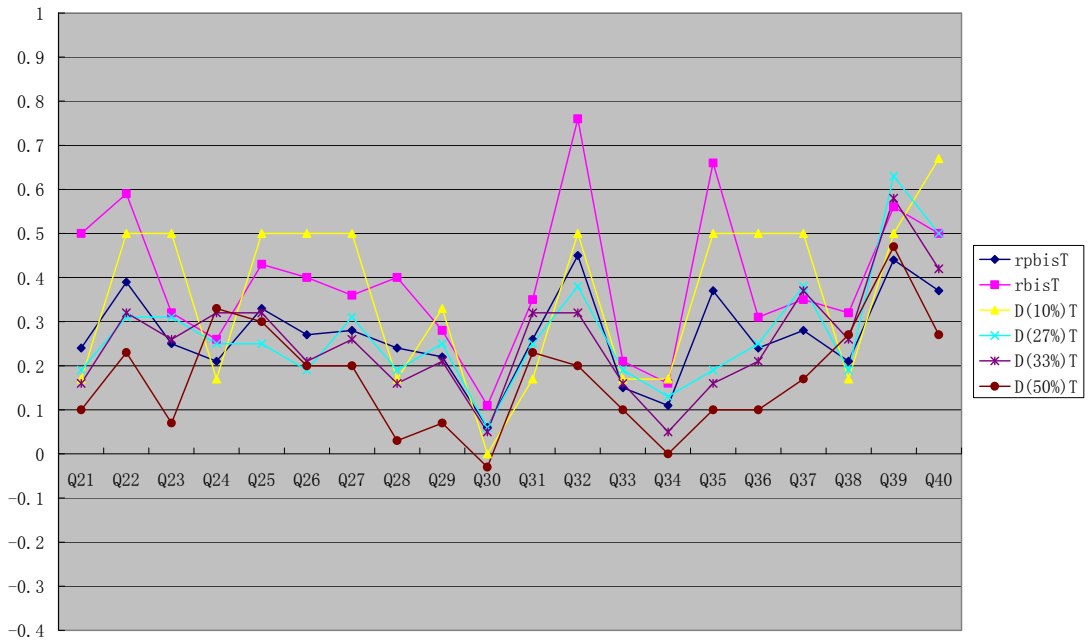


Figure 7. The distribution of difficulty indices for each item in LanApp based on the Sub2-total score for Class 4

Figure 8. The distribution of difficulty indices for each item in LanApp based on the entire-test-total score for Class 4.

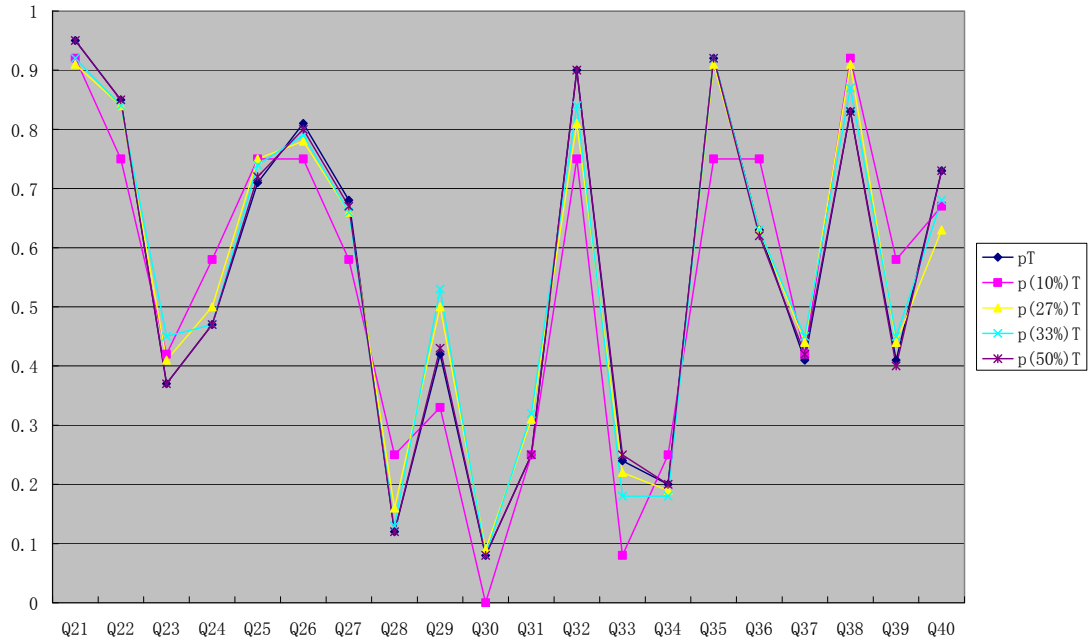


Figure 9. The distribution of discrimination indices for each item in Reading based on the Sub3-total score for Class 4.

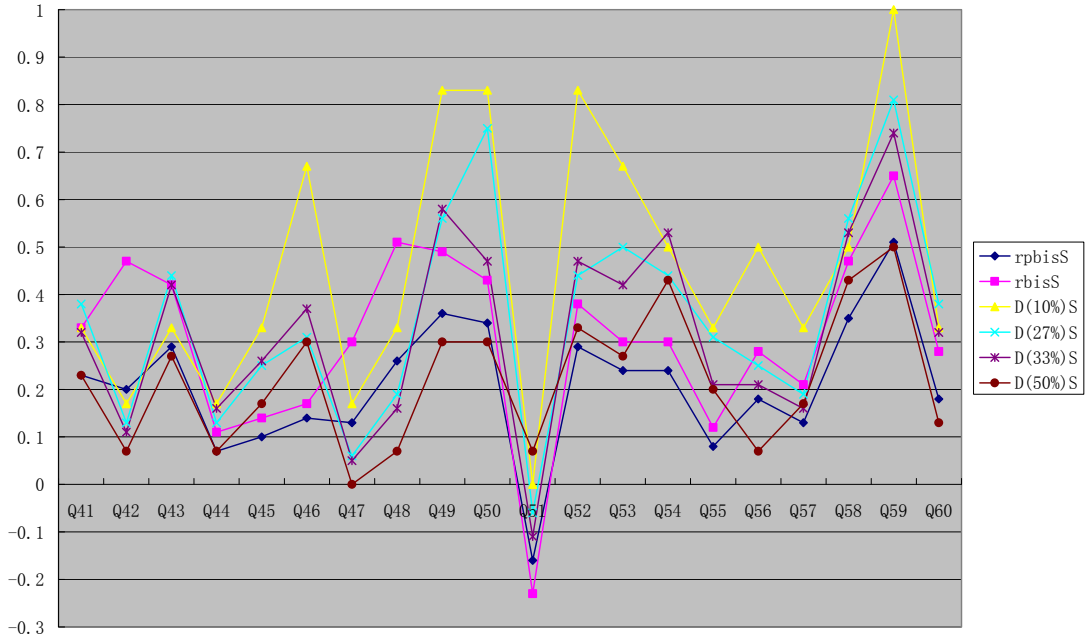


Figure 10. The distribution of discrimination indices for each item in Reading based on the entire-test-total score for Class 4

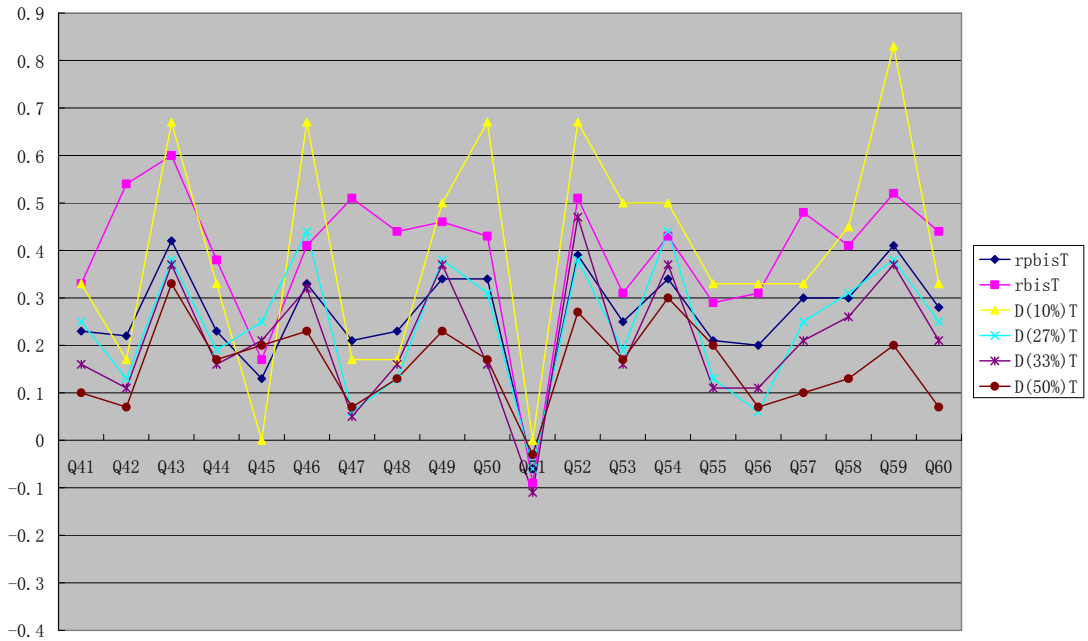


Figure 11. The distribution of difficulty indices for each item in Reading based on the Sub3-total score for Class 4.

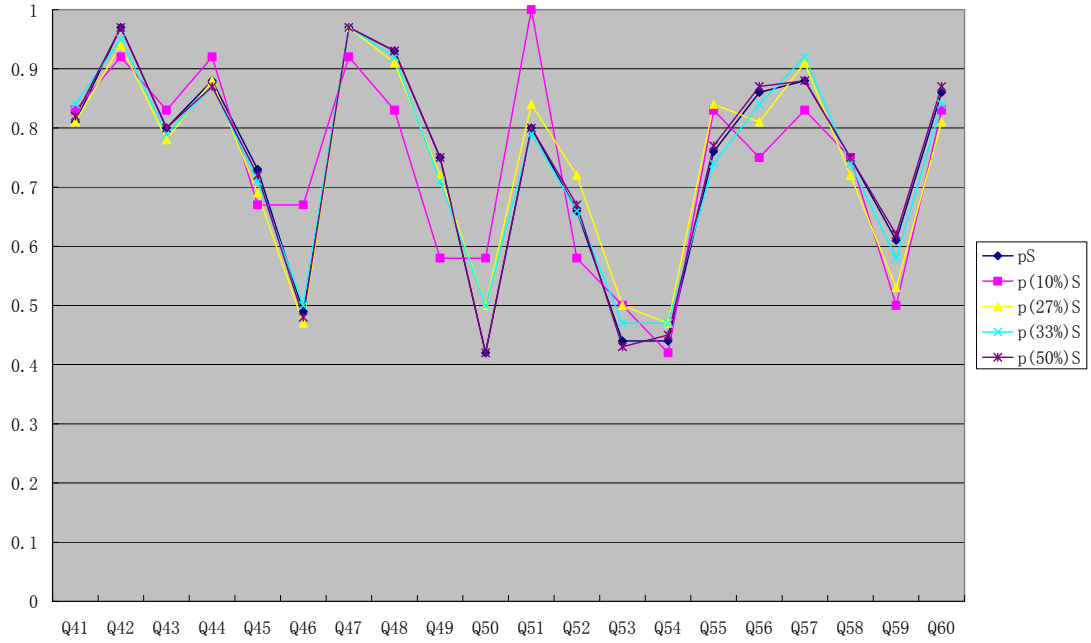
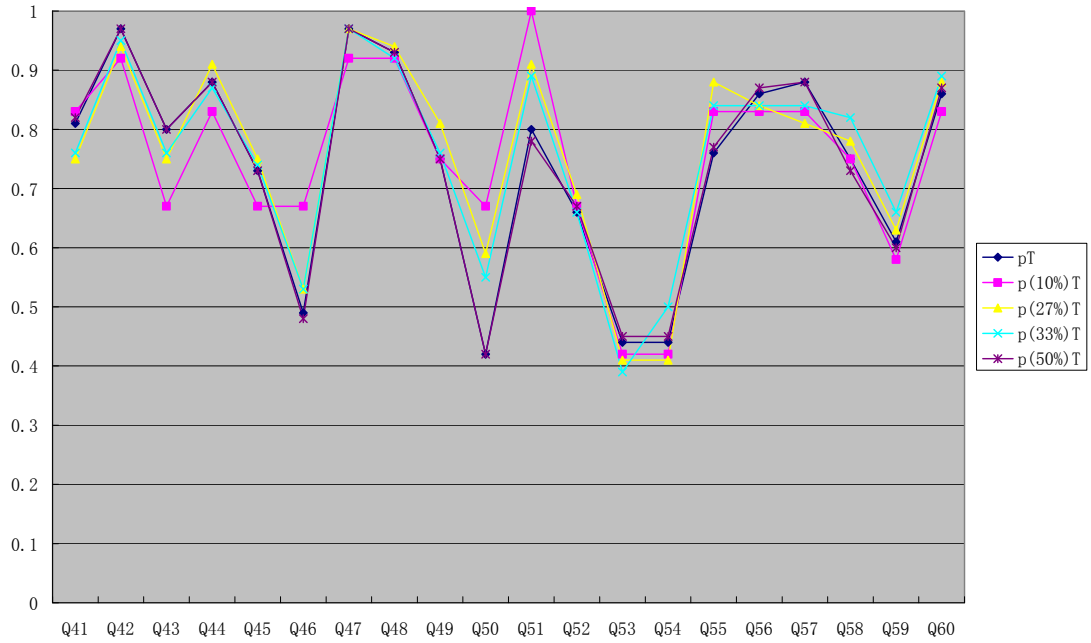


Figure 12. The distribution of difficulty indices for each item in Reading based on the entire-test-total score for Class 4.



Looking through Figures 1, 2, 5, 6, 9, and 10, most of the corresponding discrimination indices kept stable in the subtest and the entire test. rather than r_{bis} and $D_{10\%}$, which has a large difference by comparing the trends in the subtest and the entire test. It implied the subtest can contribute the similar information as the entire test in estimating the discrimination indices in most of cases, but it should be careful if the r_{bis} and $D_{10\%}$ are used to evaluate the item quality. As for the difficulty indices, all of them were very close and the patterns produced by them were also very similar. Thus, the subtest is also as efficient as the entire test in estimating the difficulty indices. Besides, Figures 13 and 14 showed the mean and the standard deviation of the differences between the $r_{pbis}T$ and the other discrimination indices. Figures 15 and 16 showed the mean and the standard deviation of the differences between pT and the other difficulty indices.

Figure 13. The mean of the differences between $r_{pbis}T$ and the other 11 discrimination indices in the first three subtests for Class 4.

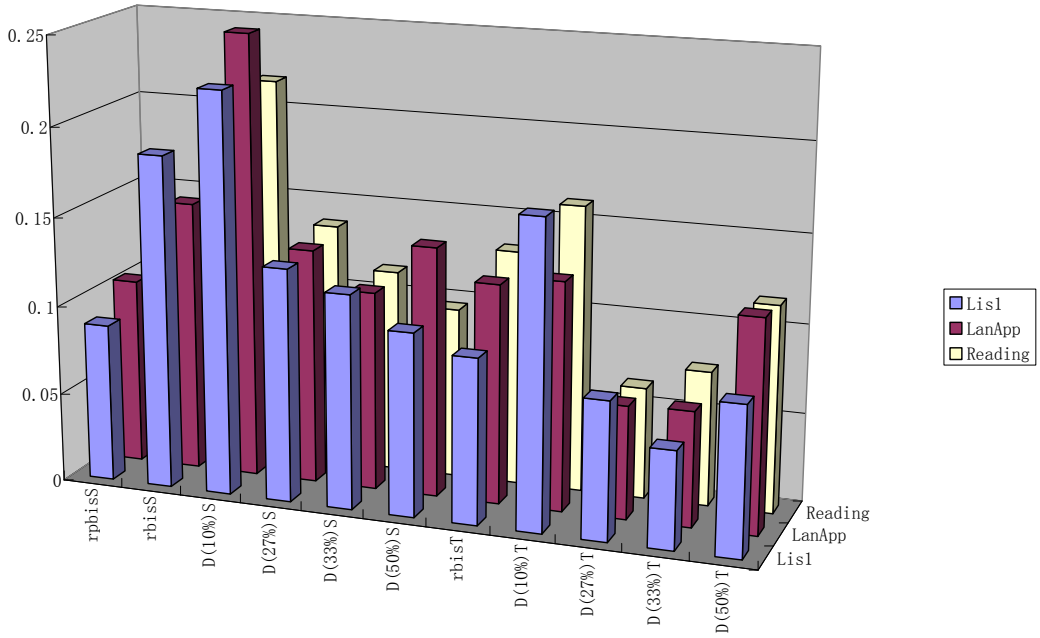


Figure 14. The standard deviation of the differences between $r_{pbis}T$ and the other 11 discrimination indices in the first three subtests for Class 4.

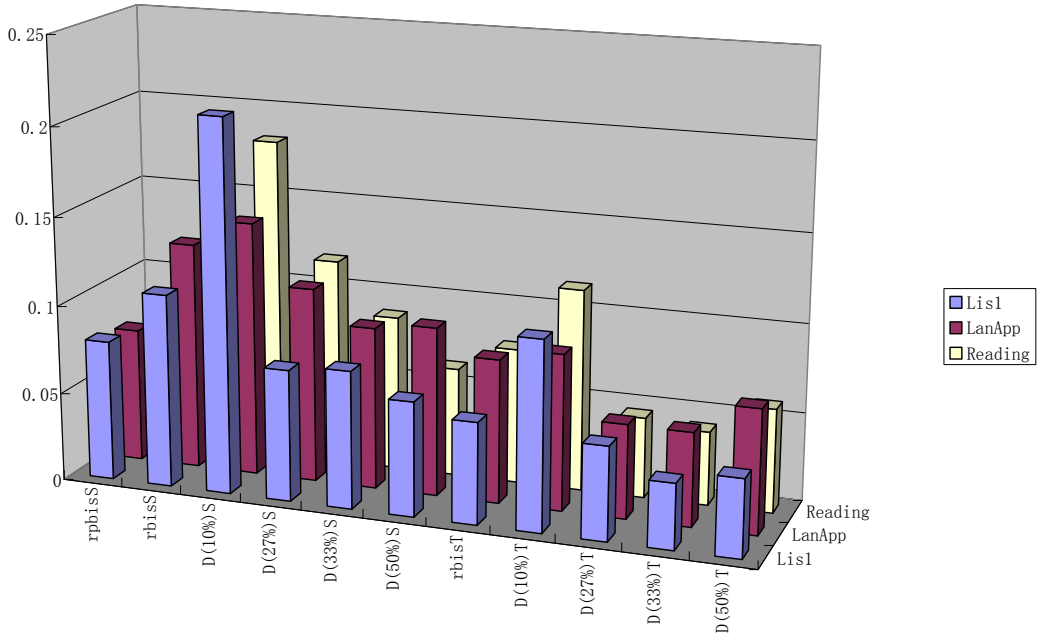


Figure 15. The mean of the differences between *pT* and the other 9 difficulty indices in the first three subtests for Class 4.

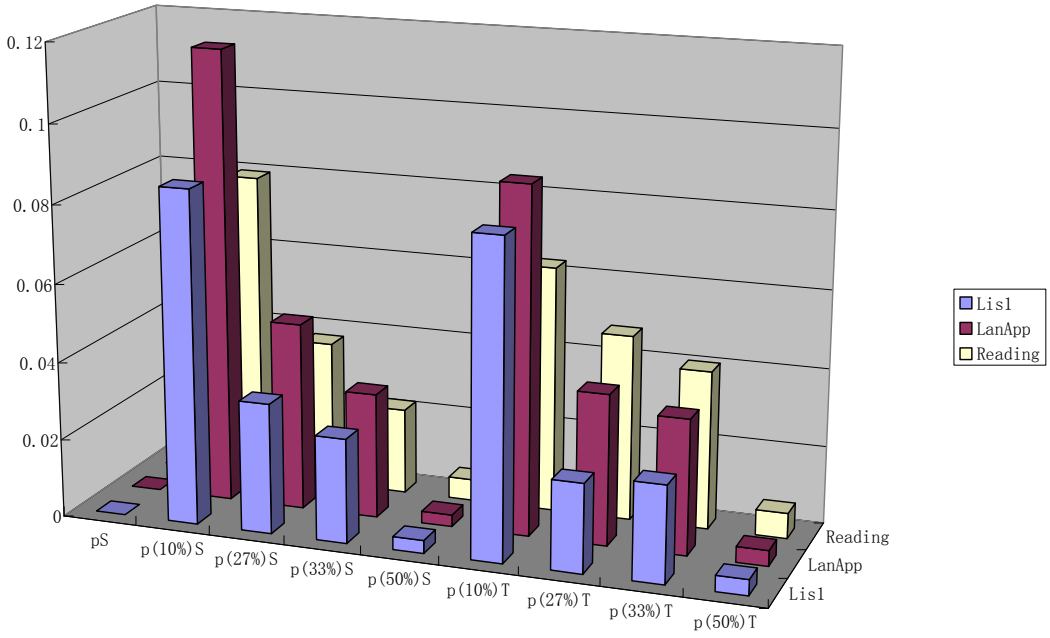
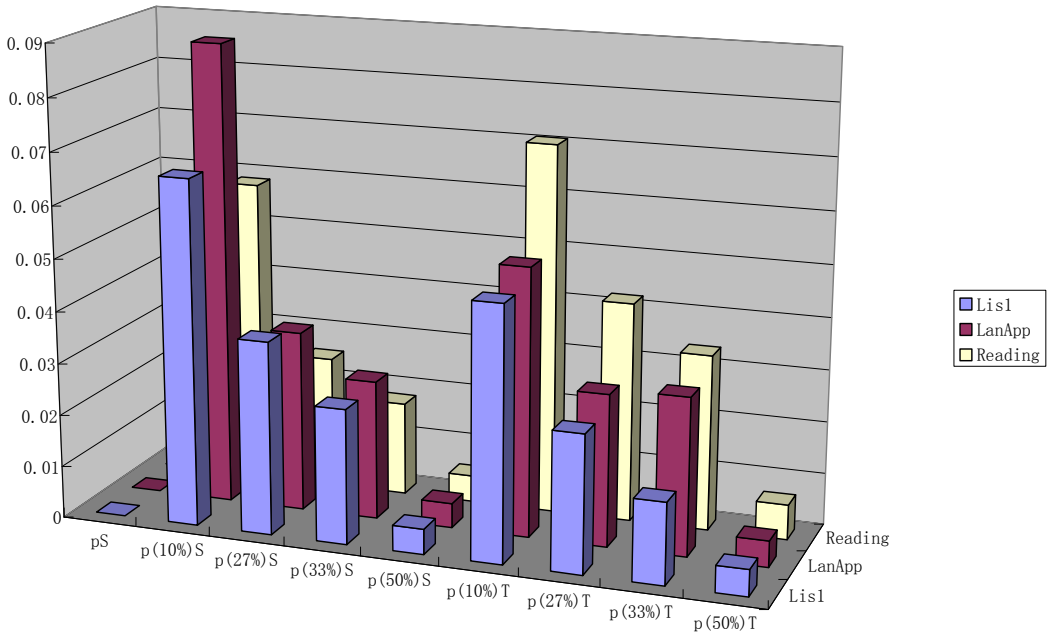


Figure 16. The standard deviation of the differences between *pT* and the other 9 difficulty indices in the first three subtests for Class 4.



In Figures 13 and 14, still $r_{pbis}S$, $D_{33\%}S$, and $D_{33\%}T$ stood out in contrast with the other discrimination indices considering the mean and the standard deviation of the differences. $D_{50\%}S$ seemed to be closer to $r_{pbis}T$ than $D_{33\%}S$. But, since $D_{50\%}T$ has a larger gap than $D_{33\%}T$, $D_{50\%}$ had to be ignored. Yet, it may provide us a new item evaluating method that $D_{50\%}$ is used in the subtest level while $D_{33\%}$ is used in the entire test level. This point needs further investigation. The corresponding $p_{33\%}$ appeared closer values with *pT*. Thus, in the unit of one class, the similar results related to the discrimination indices and the difficulty indices were obtained as in the overall grade. That is, r_{pbis} can be used if the appropriate computer programs are available. Otherwise, $D_{33\%}$ can be considered to fulfil the item analysis by means of Excel.