American Review of Mathematics and Statistics
June 2015, Vol. 3, No. 1, pp. 34-51
ISSN: 2374-2348 (Print), 2374-2356 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/arms.v3n1a5
URL: http://dx.doi.org/10.15640/arms.v3n1a5

A Statistical Analysis of Testlets - A Parametric Approach

Chwan-Chin Song[1], Yan-Teng Shih[1] & Jann-Huei Jinn[2]

Abstract

Based on the basic concepts in classical test theory (CTT), Song et al. (2014) proposed a parametric method to develop the computational formulas of difficulty index and discrimination index for independent test items. In this article, modeling testlets with appropriate probability structures, we generalize their results to items in testlets. This parametric approach considers the dependence between the items within each testlet. It would also take the effect of performance of middle-scoring group on these two index values into account. In addition, we provide an efficient computing algorithm for obtaining both of them by using the probability generating function technique. Real data taken from the English Test Items of the Second Basic Competence Test for Junior High School Students in 2007 in Taiwan are used for empirical study, and the results are compared with those obtained by the traditional nonparametric method. Discrepancies between these two methods are also discussed in this study.

Keywords: Classical test theory, Difficulty index, Discrimination index, High-scoring group, Item, Low-scoring group, Testlet

1. Introduction

A test is defined as the collection of items used to measure a specific goal. A testlet (item set) in a test is used to measure different sub-goals. The purpose of analyzing test items is to improve the "quality" of a test. A good quality test would be able to classify students into different scoring groups, for example, high, middle, and low, commensurate with their abilities. We can also identify good, bad, or appropriate items in a test by analyzing the test result. Items which are too difficult or too easy may not accurately reflect students' study abilities. We usually use difficulty and discrimination indexes available in CTT to evaluate items in a test. A difficulty index is used to indicate the difficulty of an item in a test. The difficulty index used in this article is $P_i = (P_{iH} + P_{iL})/2$, where $P_{iH} = R_{iH} / N_H$, $P_{iL} = R_{iL} / N_L$, R_{iH} is the number of students who got correct answer to the i^{th} item in the high-scoring group (top 25%~33%) with a total number of N_H students and R_{iL} is the number of students who got correct answer to the i^{th} item in the low-scoring group (bottom 25%~33%) with a total number of N_L students. The larger the P_i value, the easier the item. Similarly, the smaller the P_i value, the harder the item. When the P_i value is close to 0.50, it indicates a moderate item (not too hard or too easy). If all the students were unable to correctly answer the i^{th} item then P_i=0, likewise, if all the students got correct answer to the i^{th} item then P_i=1. A discrimination index value of a test item is used to detect whether an item can distinguish a students' learning capability. The discrimination index of the i^{th} item used in this article is defined as $D_i = P_{iH} - P_{iL}$. Where $-1 \le D_i \le 1$, and the larger the D_i value, the more discriminating the item.

[1] Department of Mathematical Sciences, National Cheng-Chi University, Taipei, Taiwan, 11605, R.O.C.
[2] Department of Statistics, Grand Valley State University, Allendale, Michigan, 49401, USA.

When $D_i = 1$, it indicates all the students in the high-scoring group got correct answer and all the students in the low-scoring group answered incorrectly. When $D_i = -1$, it indicates an opposite situation. When D_i =0, it usually indicates that the item is either too hard or too easy so that all the students in the high-scoring and low-scoring groups answered it either correctly or wrong. We adopt the following standards in using the discrimination index value D_i to evaluate the i^{th} item (see Ebel and Frisbie, 1991).

Table 1: The Evaluation Standards for the Discrimination Index

D_i -value	Item Evaluation
$D_i \geq 0.40$	Very good item
$0.30 \leq D_i < 0.39$	Reasonably good but possibly subject to improvement
$0.20 \leq D_i < 0.29$	Marginal item, usually needing and being subject to improvement
$D_i \leq 0.19$	Poor item, to be rejected or improved by revision

When the difficulty index value is near 0.5, the discrimination index value would approach the extreme value. The criteria for choosing the test items are: (1) Select items with larger discrimination index values, then from which choose the ones with difficulty index values closer to 0.5. (2) Select items with difficulty index values closer to 0.5. For more details about difficulty and discrimination indexes in CTT, readers may refer to Ebel and Frisbie (1991), Crocker and Algina (2008). The traditional nonparametric method in studying the difficulty and discrimination indexes considers only the performance of high-scoring and low-scoring groups of students on the test. This method almost ignores the performance of middle-scoring group of students. Using the traditional nonparametric method to do data analysis may lead to biased results. In this article, based on the difficulty and discrimination indexes mentioned above, we propose a parametric method to calculate them for items in testlets by modeling testlets with appropriate probability distributions. This method considers the dependence between items within each testlet. It also takes into account the effect of performance of the middle-scoring group on these two index values. In addition, we provide an efficient computing algorithm for obtaining both of them by using the probability generating function technique. Real data taken from the English Test Items of the Second Basic Competence Test for Junior High School Students in 2007 in Taiwan (abbreviated to English Test in Taiwan (2007)) is given to study the difficulty and discrimination index values. The results are compared with those obtained by the traditional nonparametric method. Finally, some discrepancies between these two methods are also discussed.

2. Review of Traditional Nonparametric Method

We use an example to demonstrate how to calculate the difficulty and discrimination index values in traditional nonparametric method.

Example 1: Assume that there are six items in a test, each item has four multiple choices, and twelve students took the test. The high-scoring group is defined as those students who got at least five correct answers, and the low-scoring group is defined as those students who got at most one correct answer. The test results are arranged in descending order of total number of correct answers in Table 2.

Table 2: Test Results

	Case Number	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	# of correct answers
H-Group	1	C	B	D	A	D	B	6
	2	C	B	D	A	D	D	5
	3	A	B	D	A	D	B	5
	4	C	B	C	A	D	B	5
M-Group	5	C	B	D	A	A	D	4
	6	C	D	C	A	D	D	3
	7	C	B	C	A	A	D	3
	8	A	D	C	A	D	D	2
L-Group	9	A	B	C	B	A	D	1
	10	C	D	C	B	A	D	1
	11	A	D	C	A	A	D	1
	12	A	D	C	B	A	D	0
Correct Answer		C	B	D	A	D	B	
Correct Answer Rate		7/12	7/12	4/12	9/12	6/12	3/12	

H-Group: high-scoring group; L-Group: low-scoring group

We use the first item's result to show the calculation of the difficulty index and discrimination Index. Similar calculation can be applied to the other items.

P_{1H} = (# of students answered correctly in the high-scoring group) / (total # of students in the high-scoring group) = 0.75.

P_{1L} = (# of students answered correctly in the low-scoring group) / (total # of students in the group) = 0.25.

$P_1 = \dfrac{P_{1H} + P_{1L}}{2} = 0.50$, and $D_1 = P_{1H} - P_{1L} = 0.50$.

Table 3 shows the correct answer rates of each group, and the difficulty and discrimination index values for each item.

Table 3: Correct Answer Rates, Difficulty and Discrimination Index Values

I #	Correct Answer Rate of H-Group P_{iH}	Correct Answer Rate of L-Group P_{iL}	Difficulty Index Value P_i	Discrimination Index Value D_i
1	0.75	0.25	0.5	0.5
2	1.00	0.25	0.625	0.75
3	0.75	0.00	0.375	0.75
4	1.00	0.25	0.625	0.75
5	1.00	0.00	0.50	1.00
6	0.75	0.00	0.375	0.75

I #: Item #

The results in Table 3 will be used to compare with those of testlets examples in Sections 3.4 and 3.5 given by the parametric method.

3. Parametric Method

In this section, we will introduce the notation, the model of parametric method, the distribution of the total number of correct answers, and an efficient algorithm to compute difficulty index and discrimination index values.

Assume that in a test there are $t(t \geq 1)$ testlets, in the i^{th} testlet there are $r_i(\geq 1)$ items, and each item has $s(s \geq 2)$ multiple choices.

3.1 Notation

Let n_{ijk} denote the number of students who choose the k^{th} multiple choice for the j^{th} item in the i^{th} testlet.

Let p_{ijk} denote the probability of choosing the k^{th} multiple choice for the j^{th} item in the i^{th} testlet where $i = 1,2,\dots,t$; $j = 1,2,\dots,r_i$; $k = 1,2,\dots,s$.

Let $n_{i,j+1,k^*|ijk}$ denote the number of students who choose the k^{th} multiple choice for the j^{th} item but choose the k^{*th} multiple choice for the $(j+1)^{th}$ item in the i^{th} testlet.

Let $p_{i,j+1,k^*|ijk}$ denote the probability of choosing the k^{th} multiple choice for the j^{th} item but choose the k^{*th} multiple choice for the $(j+1)^{th}$ item in the i^{th} testlet.

Let $\boldsymbol{n}_{i,j+1|ijk} \equiv (n_{i,j+1,1|ijk},\dots,n_{i,j+1,k^*|ijk},\dots,n_{i,j+1,s|ijk})$ where $i = 1,2,\dots,t$; $j = 1,\dots,r_i - 1$; $k, k^* = 1,2,\dots,s$

3.2 Model of Parametric Method

Suppose that *n* students took the test and answered all the items. We assume that (1) All testlets are independent. (2) For the first item in each testlet, the number of students in s answer categories follow a multinomial distribution, that is,

$$(n_{i11},\dots,n_{i1k},\dots,n_{i1s}) \sim Mul(n, p_{i11},\dots,p_{i1k},\dots,p_{i1s}) \quad i = 1,2,\dots,t . \tag{3.1}$$

(3) In each testlet, the $(j+1)^{th}$ item's correct answer depends only on that of the j^{th} item.

(4) Given the number n_{ijk} , the number of students in s answer categories of the $(j+1)^{th}$ item follow a multinomial distribution, that is,

$$\boldsymbol{n}_{i,j+1|ijk} \big| n_{ijk} \sim Mul(n_{ijk}, p_{i,j+1,1|ijk},\dots,p_{i,j+1,s|ijk}) \quad i = 1,2,\dots,t ;\ j = 1,\dots,r_i - 1 ;\ k = 1,2,\dots,s . \tag{3.2}$$

3.3 Distribution of Total Number of Correct Answers

Assuming that getting more correct answers on the test is equivalent to getting a higher score, and, therefore, the high-scoring and the low-scoring groups can be identified by the total number of correct answers. For each randomly selected student, define random variables

$$C_{ij} = \begin{cases} 1, & \textit{correct answer to the } j^{th} \textit{ item in the } i^{th} \textit{ testlet} \\ 0, & \textit{wrong answer to the } j^{th} \textit{ item in the } i^{th} \textit{ testlet} \end{cases}$$

$$i = 1,2,\dots,t ;\ j = 1,2,\dots,r_i .$$

Let C_i denote the number of correct answers in the i^{th} testlet, $i = 1,2,\dots,t$, and let X denote the total number of correct answers in the test, that is, $C_i = \sum_{j=1}^{r_i} C_{ij}$, and $X = \sum_{i=1}^{t} C_i$. We can find the probability distribution of X through the probability generating function (PGF) of X.

The PGF of C_i is given by $G_{C_i}(t) = E(t^{C_i}) = E(t^{\sum_{j=1}^{r_i} C_{ij}}) =$

$$E(t^{C_{i1}+...+C_{i,r_i}}) = \sum_{m=0}^{r_i} \sum_{d_{i1}+...+d_{i,r_i}=m} P(C_{i1} = d_{i1},...,C_{i,r_i} = d_{i,r_i})t^m =$$

$$\sum_{m=0}^{r_i} \sum_{d_{i1}+...+d_{i,r_i}=m} P(C_{i,r_i} = d_{i,r_i} \mid C_{i,r_i-1} = d_{i,r_i-1}) \times ... \times P(C_{i2} = d_{i2} \mid C_{i1} = d_{i1}) P(C_{i1} = d_{i1}) t^m =$$

$$\sum_{m=0}^{r_i} \sum_{d_{i1}+...+d_{i,r_i}=m} P(C_{i1} = d_{i1}) \left\{ \prod_{j=1}^{r_i-1} P(C_{i,j+1} = d_{i,j+1} \mid C_{ij} = d_{ij}) \right\} t^m \qquad (3.3)$$

where $d_{ij} = 0$ or 1.

Since the test is composed of several independent testlets, the PGF of total number of correct answers is equal to the product of all PGF's of the number of correct answers in each testlet. That is,

$$G_X(t) = E(t^X) = E(t^{\sum_{i=1}^{t} C_i}) = E(t^{C_1}) \cdots E(t^{C_t}) = G_{C_1}(t) \cdots G_{C_t}(t) \qquad (3.4)$$

From the coefficients of (3.4) we can obtain the distribution of total number of correct answers, X, in the test.

3.4 Computing Difficulty Index and Discrimination Index

Based on the classical test theory, the difficulty index P_{ij} and the discrimination index D_{ij} of the j^{th} item in the i^{th} testlet used in this article are defined as follows:

$$P_{ij} = \frac{P_{ij}^{[H]} + P_{ij}^{[L]}}{2} \quad i = 1,2,...,t\,;\ j = 1,2,...,r_i \qquad (3.5)$$

$$D_{ij} = P_{ij}^{[H]} - P_{ij}^{[L]} \quad i = 1,2,...,t\,;\ j = 1,2,...,r_i \qquad (3.6)$$

In (3.5) and (3.6), $P_{ij}^{[H]}$ and $P_{ij}^{[L]}$ are the proportions of students who got correct answer to the j^{th} item in the i^{th} testlet in the high-scoring and the low-scoring groups, respectively. That is, $P_{ij}^{[H]} = P(Cij = 1 \mid H)$ and $P_{ij}^{[L]} = P(Cij = 1 \mid L)$ for $i = 1,2,...,t$; $j = 1,2,...,r_i$, where H and L respectively denote the high-scoring and low-scoring groups. Let $x^{[H]}$ and $x^{[L]}$ denote the least and the most number of correct answers in the high-scoring and the low-scoring groups, respectively, and assume both $x^{[H]}$ and $x^{[L]}$ are known. Then,

$$P_{ij}^{[H]} = \frac{P(X \ge x^{[H]}, C_{ij} = 1)}{P(X \ge x^{[H]})} = \frac{P(X \ge x^{[H]} \mid C_{ij} = 1) \times P(C_{ij} = 1)}{P(X \ge x^{[H]})}$$

$$j = 1,2,...,r_i\,; \quad i = 1,2,...,t\,. \qquad (3.7)$$

Similarly,

$$P_{ij}^{[L]} = \frac{P(X \le x^{[L]}, C_{ij} = 1)}{P(X \le x^{[L]})} = \frac{P(X \le x^{[L]} \mid C_{ij} = 1) \times P(C_{ij} = 1)}{P(X \le x^{[L]})}$$

$$j = 1,2,...,r_i\,; \quad i = 1,2,...,t\,. \qquad (3.8)$$

The probability, $P(X \ge x^{[H]})$, in (3.7) can be obtained by summing the coefficients of terms corresponding to the power at least $x^{[H]}$ in the PGF of X (Eq. (3.4)). Similarly, The probability, $P(X \le x^{[L]})$, in (3.8) can be obtained by summing the coefficients of terms corresponding to the power at most $x^{[L]}$ in the PGF of X.

To calculate the probability $P(X \geq x^{[H]} \mid C_{ij}=1)$ in (3.7) and $P(X \leq x^{[L]} \mid C_{ij}=1)$ in (3.8), we need to find the conditional PGF of X given that correctly answered the j^{th} item in the i^{th} testlet. This conditional PGF of X can be obtained by $E(t^X \mid C_{ij}=1) = E(t^{C_1+\ldots+C_t} \mid C_{ij}=1) = \prod_{i=1}^{t} E(t^{C_i} \mid C_{ij}=1) = \left\{\prod_{h \neq i} E(t^{C_h})\right\} \cdot E(t^{C_i} \mid C_{ij}=1)$ (3.9)

where $E(t^{C_h})$ is the PGF of number of correct answers C_h in the h^{th} testlet, and can be obtained from (3.3). On the other hand,

$$E(t^{C_i} \mid C_{ij}=1) = \sum_{m=1}^{r_i} \sum_{d_{i1}+\ldots+d_{i,r_i}=m, d_{ij}=1} [P(C_{i1}=d_{i1},\ldots,C_{ij}=d_{ij},\ldots,C_{i,r_i}=d_{i,r_i} \mid C_{ij}=1)]t^m =$$

$$\sum_{m=1}^{r_i} \sum_{d_{i1}+\ldots+d_{i,r_i}=m, d_{ij}=1} [P(C_{i1}=d_{i1},\ldots,C_{ij}=1,\ldots,C_{i,r_i}=d_{i,r_i})/P(C_{ij}=1)]t^m = \frac{1}{P(C_{ij}=1)} \times$$

$$\{\sum_{m=1}^{r_i} \sum_{d_{i1}+\ldots+d_{i,r_i}=m, d_{ij}=1} P(C_{i1}=d_{i1}) \cdot P(C_{i2}=d_{i2} \mid C_{i1}=d_{i1}) \times \cdots \times P(C_{ij}=1 \mid C_{i,j-1}=d_{i,j-1}) \times$$

$$P(C_{i,j+1}=d_{i,j+1} \mid C_{ij}=1) \times P(C_{i,j+2}=d_{i,j+2} \mid C_{i,j+1}=d_{i,j+1}) \times \cdots \times P(C_{i,r_i}=d_{i,r_i} \mid C_{i,r_i-1}=d_{i,r_i-1})\}t^m =$$

$$\sum_{m=1}^{r_i} \sum_{d_{i1}+\ldots+d_{i,r_i}=m, d_{ij}=1} [P(C_{i1}=d_{i1}) \cdot \prod_{g=1}^{r_i-1} P(C_{i,g+1}=d_{i,g+1} \mid C_{ig}=d_{ig})]t^m / P(C_{ij}=1) \qquad (3.10)$$

Hence, from (3.9) and (3.10), we obtain $E(t^X \mid C_{ij}=1) = \{\prod_{h \neq i} G_{C_h}(t)\} \cdot E(t^{C_i} \mid C_{ij}=1) =$

$$\{\prod_{h \neq i} G_{C_h}(t)\} \cdot G_{ij}^*(t) / P(C_{ij}=1)$$

where

$$G_{ij}^*(t) = \sum_{m=1}^{r_i} \sum_{d_{i1}+\ldots+d_{i,r_i}=m, d_{ij}=1} [P(C_{i1}=d_{i1}) \cdot \prod_{g=1}^{r_i-1} P(C_{i,g+1}=d_{i,g+1} \mid C_{ig}=d_{ig})]t^m$$

Let

$$F_{ij}(t) = \{\prod_{h \neq i} G_{C_h}(t)\} \cdot G_{ij}^*(t) \qquad (3.11)$$

Then

$$E(t^X \mid C_{ij}=1) = F_{ij}(t) / P_r(C_{ij}=1) \qquad (3.12)$$

From (3.12), the numerator of (3.7), $P(X \geq x^{[H]} \mid C_{ij}=1) \cdot P(C_{ij}=1)$, can be obtained by summing the coefficients of terms corresponding to the power at least $x^{[H]}$ in $F_{ij}(t)$ given by (3.11). Similarly, the numerator of (3.8), $P(X \leq x^{[L]} \mid C_{ij}=1) \cdot P(C_{ij}=1)$, can be obtained by summing the coefficients of terms corresponding to the power at most $x^{[L]}$ in $F_{ij}(t)$ given by (3.11). Under the parametric model, any two different testlets, each consisting of one single item, having the same correct answer rate, must have the same difficulty index value and discrimination index value. The following is a proof: Assume that $i \neq i^*$, and $P(C_{i1}=1) = P(C_{i^*1}=1)$. We want to prove $P_{i1}^{[H]} = P_{i^*1}^{[H]}$, and $P_{i1}^{[L]} = P_{i^*1}^{[L]}$. From (3.7), it gives

$$P_{i1}^{[H]} = \frac{P(X \ge x^{[H]} \big| C_{i1} = 1) \times P(C_{i1} = 1)}{P(X \ge x^{[H]})}$$

and

$$P_{i^*1}^{[H]} = \frac{P(X \ge x^{[H]} \big| C_{i^*1} = 1) \times P(C_{i^*1} = 1)}{P(X \ge x^{[H]})},$$

Since $P(X \ge x^{[H]} \big| C_{i1} = 1)$ and $P(X \ge x^{[H]} \big| C_{i^*1} = 1)$ are the sum of the coefficients of terms corresponding to the power at least $x^{[H]}$ in the polynomials of $E(t^X \big| C_{i1} = 1)$ and $E(t^X \big| C_{i^*1} = 1)$, respectively. Hence, to show $P_{i1}^{[H]} = P_{i^*1}^{[H]}$, it suffices to show $E(t^X \big| C_{i1} = 1) = E(t^X \big| C_{i^*1} = 1)$. By (3.9), we have

$$E(t^X \big| C_{i1} = 1) = \{\prod_{h \ne i} E(t^{C_{h1}})\} \cdot E(t^{C_{i1}} \big| C_{i1} = 1) = \{\prod_{h \ne i, i^*} E(t^{C_{h1}})\} \cdot E(t^{C_{i1}} \big| C_{i1} = 1) \cdot E(t^{C_{i^*1}})$$

$$= \{\prod_{h \ne i, i^*} E(t^{C_{h1}})\} \cdot t \cdot [P(C_{i^*1} = 0) + P(C_{i^*1} = 1) \cdot t]$$

$$= \{\prod_{h \ne i, i^*} E(t^{C_{h1}})\} \cdot [P(C_{i^*1} = 0) \cdot t + P(C_{i^*1} = 1) \cdot t^2].$$

Similarly, we can show that $E(t^X \big| C_{i^*1} = 1) = \{\prod_{l \ne i, i^*} E(t^{C_{l1}})\} \cdot [P(C_{i1} = 0) \cdot t + P(C_{i1} = 1) \cdot t^2]$

But, $P(C_{i1} = 1) = P(C_{i^*1} = 1)$ and $P(C_{i1} = 0) = P(C_{i^*1} = 0)$, hence $E(t^X \big| C_{i1} = 1) = E(t^X \big| C_{i^*1} = 1)$. By the same argument, we can show that $P_{i1}^{[L]} = P_{i^*1}^{[L]}$.

3.5 Computing Algorithm and Example

The following steps are used to compute the difficulty index P_{ij} and discrimination index D_{ij} for the j^{th} item in the i^{th} testlet:

Step1: Find the PGF $G_X(t)$ of total number of correct answers, X, in the test. That is, calculate (3.4).

Step 2: Find the polynomial $F_{ij}(t)$ in (3.11).

Step 3: Based on the PGF found in step 1, calculate the denominators $P(X \ge x^{[H]})$ and $P(X \le x^{[L]})$ in (3.7) and (3.8). The probability $P(X \ge x^{[H]})$ can be obtained by summing the coefficients of terms corresponding to the power at least $x^{[H]}$ in $G_X(t)$. Similarly, the probability $P(X \le x^{[L]})$ can be obtained by summing the coefficients of terms corresponding to the power at most $x^{[L]}$ in $G_X(t)$.

Step 4: Based on the polynomial found in step 2, calculate the numerators $P(X \ge x^{[H]}, C_{ij} = 1)$ and $P(X \le x^{[L]}, C_{ij} = 1)$ in (3.7) and (3.8), respectively. The probability $P(X \ge x^{[H]}, C_{ij} = 1)$ can be obtained by summing the coefficients of terms corresponding to the power at least $x^{[H]}$ in $F_{ij}(t)$. Similarly, the probability

$P(X \le x^{[L]}, C_{ij} = 1)$ can be obtained by summing the coefficients of terms corresponding to the power at most $x^{[L]}$ in $F_{ij}(t)$.

Step 5: Based on the results obtained in step 3 and 4, calculate $P_{ij}^{[H]}$ and $P_{ij}^{[L]}$ in (3.7) and (3.8).

Step 6: Based on the results obtained in step 5, calculate the difficulty index

$P_{ij} = (P_{ij}^{[H]} + P_{ij}^{[L]})/2$, and the discrimination index $D_{ij} = P_{ij}^{[H]} - P_{ij}^{[L]}$ for the

j^{th} item in the i^{th} testlet, respectively.

Next, we use a simple example to demonstrate the execution of the algorithm in computing difficulty and discrimination indexes by parametric method. Note that, we substitute the maximum likelihood estimates for the unknown parameters when it needs.

Example 2 (Example 1 modified): Let testlet #1consist solely of item #1, let testlet #2 consist of items #2 and #3, and let testlet #3 consist of items #4, #5, and #6. Rewrite Table 3 as Table 4.

Table 4: Test Results

		Testlet #1	Testlet #2		Testlet #3			
	Case Number	1	1	2	1	2	3	# of correct answers
H- Group	1	C	B	D	A	D	B	6
	2	C	B	D	A	D	D	5
	3	A	B	D	A	D	B	5
	4	C	B	C	A	D	B	5
M- Group	5	C	B	D	A	A	D	4
	6	C	D	C	A	D	D	3
	7	C	B	C	A	A	D	3
	8	A	D	C	A	D	D	2
L- Group	9	A	B	C	B	A	D	1
	10	C	D	C	B	A	D	1
	11	A	D	C	A	A	D	1
	12	A	D	C	B	A	D	0
Correct Answer		C	B	D	A	D	B	
Correct Answer Rate		7/12	7/12	4/12	9/12	6/12	3/12	

We use the three items in the third testlet to show the calculation of difficulty and discrimination Indexes.

Step1: Find the PGF $G_X(t)$ of total number of correct answers, X, in the test.

In the first testlet, the PGF of C_1 is $\frac{7}{12}t+\frac{5}{12}$. In the second testlet, the probability of getting two correct answers is $P(C_{21}=1,C_{22}=1)=P(C_{22}=1|C_{21}=1)\cdot P(C_{21}=1)=\frac{4}{7}\times\frac{7}{12}=\frac{4}{12}$. The probability of getting one correct answer is $P(C_{21}=1,C_{22}=0)+P(C_{21}=0,C_{22}=1)$

$=P(C_{22}=0|C_{21}=1)\cdot P(C_{21}=1)+P(C_{22}=1|C_{21}=0)\cdot P(C_{21}=0)=\frac{3}{7}\times\frac{7}{12}+0=\frac{3}{12}$. The probability of no correct answers is $P(C_{21}=0,C_{22}=0)=P(C_{22}=0|C_{21}=0)\cdot P(C_{21}=0)$

$=\frac{5}{5}\times\frac{5}{12}=\frac{5}{12}$. Therefore, the PGF of C_2 is $\frac{4}{12}t^2+\frac{3}{12}t+\frac{5}{12}$.

In the third testlet, the probability of getting three correct answers is

$P(C_{31}=1,C_{32}=1,C_{33}=1)=P(C_{33}=1|C_{32}=1)\cdot P(C_{32}=1|C_{31}=1)\cdot P(C_{31}=1)$

$=\frac{3}{6}\times\frac{6}{9}\times\frac{9}{12}=\frac{3}{12}$. The probability of getting two correct answer is

$P(C_{31}=1,C_{32}=1,C_{33}=0)+P(C_{31}=1,C_{32}=0,C_{33}=1)+P(C_{31}=0,C_{32}=1,C_{33}=1)=$

$P(C_{33}=0|C_{32}=1)\cdot P(C_{32}=1|C_{31}=1)\cdot P(C_{31}=1)+P(C_{33}=1|C_{32}=0)\cdot P(C_{32}=0|C_{31}=1)\cdot P(C_{31}=1)$

$+P(C_{33}=1|C_{32}=1)\cdot P(C_{32}=1|C_{31}=0)\cdot P(C_{31}=0)=\frac{3}{6}\times\frac{6}{9}\times\frac{9}{12}+0\times\frac{3}{9}\times\frac{9}{12}+0\times\frac{0}{3}\times\frac{3}{12}=\frac{3}{12}$.

The probability of getting one correct answer is

$P(C_{31}=1,C_{32}=0,C_{33}=0)+P(C_{31}=0,C_{32}=1,C_{33}=0)+P(C_{31}=0,C_{32}=0,C_{33}=1)=$

$P(C_{33}=0|C_{32}=0)\cdot P(C_{32}=0|C_{31}=1)\cdot P(C_{31}=1)+P(C_{33}=0|C_{32}=1)\cdot P(C_{32}=1|C_{31}=0)\cdot P(C_{31}=0)$

$+P(C_{33}=1|C_{32}=0)\cdot P(C_{32}=0|C_{31}=0)\cdot P(C_{31}=0)=\frac{3}{3}\times\frac{3}{9}\times\frac{9}{12}+0\times\frac{0}{3}\times\frac{3}{12}+\frac{0}{3}\times\frac{3}{3}\times\frac{3}{12}=\frac{3}{12}$.

The probability of no correct answers is

$P(C_{31}=0,C_{32}=0,C_{33}=0)=P(C_{33}=0|C_{32}=0)\cdot P(C_{32}=0|C_{31}=0)\cdot P(C_{31}=0)$

$=\frac{3}{3}\times\frac{3}{3}\times\frac{3}{12}=\frac{3}{12}$. Hence, the PGF of C_3 is $\frac{3}{12}t^3+\frac{3}{12}t^2+\frac{3}{12}t+\frac{3}{12}$.

Therefore, $G_X(t)=(\frac{7}{12}t+\frac{5}{12})(\frac{4}{12}t^2+\frac{3}{12}t+\frac{5}{12})(\frac{3}{12}t^3+\frac{3}{12}t^2+\frac{3}{12}t+\frac{3}{12})$

$$=\frac{7}{144}t^6+\frac{23}{192}t^5+\frac{119}{576}t^4+\frac{1}{4}t^3+\frac{29}{144}t^2+\frac{25}{192}t+\frac{25}{576}$$

Step 2: Find the polynomials $F_{31}(t)$, $F_{32}(t)$, and $F_{33}(t)$.

$F_{31}(t)=G_{C_1}(t)\cdot G_{C_2}(t)\cdot G_{31}^*(t)$ where $G_{C_1}(t)=\frac{7}{12}t+\frac{5}{12}$, $G_{C_2}(t)=\frac{4}{12}t^2+\frac{3}{12}t+\frac{5}{12}$,

and $G_{31}^*(t)=[P(C_{31}=1,C_{32}=0,C_{33}=0)]\cdot t+[P(C_{31}=1,C_{32}=1,C_{33}=0)+P(C_{31}=1,C_{32}=0,C_{33}=1)]\cdot t^2$
$+[P(C_{31}=1,C_{32}=1,C_{33}=1)]\cdot t^3=$

$[P(C_{31}=1)\cdot P(C_{32}=0|C_{31}=1)\cdot P(C_{33}=0|C_{32}=0)]\cdot t+$

$[P(C_{31}=1)\cdot P(C_{32}=1|C_{31}=1)\cdot P(C_{33}=0|C_{32}=1)+$

$[P(C_{31}=1)\cdot P(C_{32}=0|C_{31}=1)\cdot P(C_{33}=1|C_{32}=0)]\cdot t^2+$

$[P(C_{31}=1)\cdot P(C_{32}=1|C_{31}=1)\cdot P(C_{33}=1|C_{32}=1)]\cdot t^3=\frac{3}{12}t^3+\frac{3}{12}t^2+\frac{3}{12}t$

Hence, $F_{31}(t)=(\frac{7}{12}t+\frac{5}{12})(\frac{4}{12}t^2+\frac{3}{12}t+\frac{5}{12})(\frac{3}{12}t^3+\frac{3}{12}t^2+\frac{3}{12}t)$

$$=\frac{7}{144}t^6+\frac{23}{192}t^5+\frac{119}{576}t^4+\frac{29}{144}t^3+\frac{25}{192}t^2+\frac{25}{576}t.$$

Similarly, we can obtain $F_{32}(t)=\frac{7}{144}t^6+\frac{23}{192}t^5+\frac{91}{576}t^4+\frac{25}{192}t^3+\frac{25}{576}t^2$ and

$$F_{33}(t)=\frac{7}{144}t^6+\frac{41}{576}t^5+\frac{25}{288}t^4+\frac{25}{576}t^3.$$

Step 3: Calculate $P(X\geq 5)$ and $P(X\leq 1)$. Based on the PGF $G_X(t)$ found in step 1, the probability $P(X\geq 5)$ can be obtained by summing the coefficients of terms corresponding to the power at least 5 in $G_X(t)$, i.e., $P(X\geq 5)=\frac{7}{144}+\frac{23}{192}=\frac{97}{576}$. Similarly, the probability $P(X\leq 1)$ can be obtained by summing the coefficients of terms corresponding to the power at most one in

$G_X(t)$, i.e., $P(X \le 1) = \frac{25}{192} + \frac{25}{576} = \frac{25}{144}$.

Step 4: Calculate $P(X \ge 5, C_{31} = 1)$ and $P(X \le 1, C_{31} = 1)$. Based on the polynomial $F_{31}(t)$ found in step 2, the probability $P(X \ge 5, C_{31} = 1)$ can be obtained by summing the coefficients of terms corresponding to the power at least 5 in $F_{31}(t)$, i.e., $P(X \ge 5, C_{31} = 1) = \frac{7}{144} + \frac{23}{192} = \frac{97}{576}$. The probability $P(X \le 1, C_{31} = 1)$ can be obtained by summing the coefficients of terms corresponding to the power at most one in $F_{31}(t)$, i.e., $P(X \le 1, C_{31} = 1) = \frac{25}{576}$. Similarly, we can obtain $P(X \ge 5, C_{32} = 1) = \frac{97}{576}$; $P(X \ge 5, C_{33} = 1) = \frac{23}{192}$; $P(X \le 1, C_{32} = 1)$ =0; $P(X \le 1, C_{33} = 1)$ =0.

Step 5: Based on the results obtained in step 3 and 4, calculate $P_{31}^{[H]}, P_{31}^{[L]}, P_{32}^{[H]}, P_{32}^{[L]}$,

$P_{33}^{[H]}$, and $P_{33}^{[L]}$. For example, $P_{31}^{[H]} = \frac{P(X \ge 5, C_{31} = 1)}{P(X \ge 5)} = \frac{(97/576)}{(97/576)} = 1.00$ and

$P_{31}^{[L]} = \frac{P(X \le 1, C_{31} = 1)}{P(X \le 1)} = \frac{(25/576)}{(25/144)} = 0.25$. Similarly, we can obtain

$P_{32}^{[H]}$ =1.00, $P_{32}^{[L]}$ =0.00, $P_{33}^{[H]}$ =0.7113 , and $P_{33}^{[L]}$ =0.00

Step 6: Based on the results obtained in step 5, calculate the difficulty index values P_{31}, P_{32}, and P_{33} ,and the discrimination index values D_{31}, D_{32}, and D_{33}. For example,

$$P_{31} = \frac{p_{31}^{[H]} + P_{31}^{[L]}}{2} = \frac{1}{2}(1 + \frac{1}{4}) = 0.625;\ D_{31} = P_{31}^{[H]} - P_{31}^{[L]} = 1 - \frac{1}{4} = 0.75.$$

Similarly, we can obtain P_{32} =0.50, D_{32} =1.00; P_{33} =0.8557 and D_{33} =0.7113.

Table 5 shows the values of difficulty and discrimination indexes for each item in the Example 2.

Table 5: The Values of Difficulty and Discrimination Indexes in Example 2

T #	I #	Correct Answer Rate of H-Group $P_{ij}^{[H]}$	Correct Answer Rate of L-Group $P_{ij}^{[L]}$	Difficulty Index Value P_{ij}	Discrimination Index Value D_{ij}
1	1	0.7936	0.3500	0.5719	0.4438
2	1	1.0000	0.4500	0.7250	0.5500
	2	0.7835	0.0000	0.3918	0.7835
3	1	1.0000	0.2500	0.6250	0.7500
	2	1.0000	0.0000	0.5000	1.0000
	3	0.7113	0.0000	0.3557	0.7113

T #: Testlet #

4. Real Data Analysis and Comparison between Methods

In this section, we use real data taken from the English Test in Taiwan (2007) to study the difficulty and discrimination index values. The explanation of the difference between the difficulty (or discrimination) index values obtained by the classical nonparametric method and the parametric method is also given.

In the end, we use examples to show that the performance of the middle-scoring group is a possible important factor which accounts for differences in results from the parametric method.

4.1 Analysis of English Test in Taiwan (2007)

There were 60,225 students in the three major test districts in the northern and middle part of Taiwan taking the English Test in Taiwan (2007), a proportional stratified sample of 82 students was chosen from these three test districts. The English test contains 18 independent items, 11 testlets. Each testlet contains either two or three items. Each item has four multiple choices. There are 45 items in this English test. The high-scoring group (top 25%) consists of students who got at least 34 correct answers, and the low-scoring group (bottom 25%) consists of students who got at most 11 correct answers, i.e., $x^{[H]}$=34, $x^{[L]}$=11. The algorithm demonstrated in the previous section is used to compute the difficulty index and discrimination index values for parametric method. The TESTER for Windows is used to run the two index values for nonparametric method. The results of the difficulty index and discrimination index values for the first eighteen items are provided in Table 6, and the results for the other eleven testlets are provided in Table 7. Some important values used either to compute these two index values or to explore the data are included in Tables 8, 9 and 10.

Table 6: Difficulty Index and Discrimination Index Values of the First Eighteen Items in English Test in Taiwan (2007)

	Difficulty Index Values			Discrimination Index Values		
I #	Parametric Method (1)	Nonpara-metric Method (2)	Difference (1) – (2)	Parametric Method (3)	Nonpara-metric Method (4)	Difference (3) – (4)
1	0.9397	0.8500	0.0897	0.1003	0.3000	-0.1997
2	0.9202	0.8000	0.1202	0.1317	0.4000	-0.2683
3	0.9105	0.8000	0.1105	0.1471	0.4000	-0.2529
4	0.8820	0.7500	0.1320	0.1916	0.5000	-0.3084
5	0.9202	0.8250	0.0952	0.1317	0.3500	-0.2183
6	0.8634	0.6750	0.1884	0.2200	0.6500	-0.4300
7	0.9397	0.8500	0.0897	0.1003	0.3000	-0.1997
8	0.9105	0.8000	0.1105	0.1471	0.4000	-0.2529
9	0.9397	0.8750	0.0647	0.1003	0.2500	-0.1497
10	0.9897	0.9750	0.0147	0.0173	0.0500	-0.0327
11	0.9010	0.7500	0.1510	0.1622	0.5000	-0.3378
12	0.8727	0.7250	0.1477	0.2059	0.5500	-0.3441
13	0.7999	0.7000	0.0999	0.3112	0.6000	-0.2888
14	0.7822	0.6500	0.1322	0.3348	0.5000	-0.1652
15	0.8088	0.6500	0.1588	0.2989	0.7000	-0.4011
16	0.8727	0.6750	0.1977	0.2059	0.6500	-0.4441
17	0.9010	0.7750	0.1260	0.1622	0.4500	-0.2878
18	0.9202	0.8250	0.0952	0.1317	0.3500	-0.2183

Table 7: Difficulty Index and Discrimination Index Values of Items in Testlets in English Test in Taiwan (2007)

		Difficulty Index Values			Discrimination Index Values		
T #	I #	Parametric Method (1)	Nonpara-metric Method (2)	Difference (1) – (2)	Parametric Method (3)	Nonpara-metric Method (4)	Difference (3) – (4)
1	19	0.7941	0.6500	0.1441	0.3647	0.6000	-0.2353
	20	0.7081	0.6250	0.0831	0.5436	0.6500	-0.1064
	21	0.6094	0.5500	0.0594	0.6196	0.9000	-0.2804
2	22	0.9293	0.8750	0.0543	0.1183	0.2500	-0.1317
	23	0.7706	0.7500	0.0206	0.3520	0.5000	-0.1480
	24	0.9113	0.8000	0.1113	0.1438	0.4000	-0.2562
3	25	0.9501	0.8750	0.0751	0.0823	0.2500	-0.1677
	26	0.9701	0.9250	0.0451	0.0494	0.1500	-0.1006
4	27	0.8683	0.7750	0.0933	0.2429	0.4500	-0.2071
	28	0.8499	0.7500	0.0999	0.2715	0.5000	-0.2285
5	29	0.8245	0.7000	0.1245	0.2832	0.6000	-0.3168
	30	0.9469	0.8750	0.0719	0.0935	0.2500	-0.1565
6	31	0.9096	0.8250	0.0846	0.1644	0.3500	-0.1856
	32	0.8368	0.8000	0.0368	0.2944	0.4000	-0.1056
	33	0.8977	0.8000	0.0977	0.1882	0.4000	-0.2118
7	34	0.8418	0.7000	0.1418	0.2852	0.6000	-0.3148
	35	0.8238	0.6500	0.1738	0.3122	0.7000	-0.3878
8	36	0.8905	0.8000	0.0905	0.2049	0.4000	-0.1951
	37	0.7648	0.7000	0.0648	0.3904	0.6000	-0.2096
9	38	0.8796	0.7500	0.1296	0.2370	0.5000	-0.2630
	39	0.8427	0.6750	0.1677	0.2952	0.6500	-0.3548
10	40	0.8846	0.8000	0.0846	0.2272	0.4000	-0.1728
	41	0.8607	0.7750	0.0857	0.2743	0.4500	-0.1757
	42	0.8969	0.8250	0.0719	0.2025	0.3500	-0.1475
11	43	0.7828	0.6750	0.1078	0.3763	0.6500	-0.2737
	44	0.6861	0.6500	0.0361	0.5199	0.7000	-0.1801
	45	0.9194	0.8500	0.0694	0.1558	0.2000	-0.0442

Table 8: Correct Answer Rates of High-Scoring and Low-Scoring Groups and Their Difference for Item #1 Through #18 in English Test in Taiwan (2007)

	High-Scoring Group			Low-Scoring Group		
I #	Parametric (1)	Nonpara-metric (2)	Difference (1) – (2)	Parametric (3)	Nonpara-metric (4)	Difference (3) – (4)
1	0.9898	1.0000	-0.0102	0.8895	0.7000	0.1895
2	0.9860	1.0000	-0.0140	0.8543	0.6000	0.2543
3	0.9841	1.0000	-0.0159	0.8370	0.6000	0.2370
4	0.9778	1.0000	-0.0222	0.7863	0.5000	0.2863
5	0.9860	1.0000	-0.0140	0.8534	0.6500	0.2043
6	0.9734	1.0000	-0.0266	0.7534	0.3500	0.4034
7	0.9898	1.0000	-0.0102	0.8895	0.7000	0.1895
8	0.9841	1.0000	-0.0159	0.8370	0.6000	0.2370
9	0.9898	1.0000	-0.0102	0.8895	0.7500	0.1395
10	0.9984	1.0000	-0.0016	0.9811	0.9500	0.0311
11	0.9820	1.0000	-0.0180	0.8199	0.5000	0.3199
12	0.9756	1.0000	-0.0244	0.7697	0.4500	0.3197
13	0.9555	1.0000	-0.0445	0.6443	0.4000	0.2443
14	0.9496	0.9000	0.0496	0.6148	0.4000	0.2148
15	0.9583	1.0000	-0.0417	0.6593	0.3000	0.3593
16	0.9756	1.0000	-0.0244	0.7697	0.3500	0.4197
17	0.9820	1.0000	-0.0180	0.8199	0.5500	0.2699
18	0.9860	1.0000	-0.0140	0.8543	0.6500	0.2043

Table 9: Correct Answer Rates of High-Scoring and Low-Scoring Groups and Their Difference for Item #19 Through #45(Items in Testlets) in English Test in Taiwan (2007)

		High-Scoring Group			Low-Scoring Group		
T #	I #	Parametric (1)	Nonpara-metric (2)	Difference (1) – (2)	Parametric (3)	Nonpara-metric (4)	Difference (3) – (4)
1	19	0.9765	0.9500	0.0265	0.6118	0.3500	0.2618
	20	0.9799	0.9500	0.0299	0.4363	0.3000	0.1363
	21	0.9192	1.0000	-0.0808	0.2996	0.1000	0.1996
2	22	0.9885	1.0000	-0.0115	0.8702	0.7500	0.1202
	23	0.9466	1.0000	-0.0543	0.5946	0.5000	0.0946
	24	0.9832	1.0000	-0.0168	0.8394	0.6000	0.2394
3	25	0.9913	1.0000	-0.0087	0.9089	0.7500	0.1589
	26	0.9948	1.0000	-0.0052	0.9454	0.8500	0.0954
4	27	0.9897	1.0000	-0.0103	0.7469	0.5500	0.1969
	28	0.9856	1.0000	-0.0144	0.7142	0.5000	0.2142
5	29	0.9661	1.0000	-0.0339	0.6829	0.4000	0.2829
	30	0.9936	1.0000	-0.0064	0.9001	0.7500	0.1501
6	31	0.9918	1.0000	-0.0082	0.8274	0.6500	0.1774
	32	0.9840	1.0000	-0.0160	0.6896	0.6000	0.0896
	33	0.9918	1.0000	-0.0082	0.8036	0.6000	0.2036
7	34	0.9844	1.0000	-0.0156	0.6992	0.4000	0.2992
	35	0.9799	1.0000	-0.0201	0.6677	0.3000	0.3677
8	36	0.9929	1.0000	-0.0071	0.7880	0.6000	0.1880
	37	0.9600	1.0000	-0.0400	0.5697	0.4000	0.1697
9	38	0.9981	1.0000	-0.0019	0.7611	0.5000	0.2611
	39	0.9904	1.0000	-0.0096	0.6951	0.3500	0.3451
10	40	0.9982	1.0000	-0.0018	0.7710	0.6000	0.1710
	41	0.9978	1.0000	-0.0022	0.7235	0.5500	0.1735
	42	0.9982	1.0000	-0.0018	0.7957	0.6500	0.1457
11	43	0.9709	1.0000	-0.0291	0.5946	0.3500	0.2446
	44	0.9460	1.0000	-0.0540	0.4261	0.3000	0.1261
	45	0.9973	0.9500	0.0473	0.8415	0.7500	0.0915

Table 10: Correct Answer Rate of Each Item in English Test in Taiwan (2007)

I #	Correct Answer Rate	I #	Correct Answer Rate	I #	Correct Answer Rate
1	0.9268	16	0.8415	31	0.9024
2	0.9024	17	0.8780	32	0.8172
3	0.8902	18	0.9024	33	0.8902
4	0.8537	19	0.7561	34	0.8171
5	0.9024	20	0.6341	35	0.7927
6	0.8293	21	0.4756	36	0.8780
7	0.9268	22	0.9146	37	0.7073
8	0.8902	23	0.7439	38	0.8780
9	0.9268	24	0.8902	39	0.8293
10	0.9878	25	0.9390	40	0.9024
11	0.8780	26	0.9634	41	0.8780
12	0.8415	27	0.8537	42	0.9146
13	0.7439	28	0.8293	43	0.7317
14	0.7195	29	0.7805	44	0.5976
15	0.7561	30	0.9390	45	0.9146

4.2 Comparison between Parametric Method and Nonparametric Method

In this section, the two methods are compared through the calculating results of real data (English Test in Taiwan (2007)) given in Table 6 and Table 7. The difficulty index values given by the classical nonparametric method are all between 0.5500 and 0.9750. There are no “hard” items. Items #6, #13, #14, #15, #16, #19, #20, #21, #29, #34, #35, #37, #39, #43 and #44 are with difficulty index values closer to “0.5”, and they could be identified as “moderate” items. The rest of items are identified as “easy” ones. The difficulty index values given by the parametric method are all between 0.6094 and 0.9897. There are no “hard” items, either. Only items #20, #21 and #44 could be identified as “moderate” items, and the rest of items are identified as “easy” ones. Both methods identify items #20, #21 and #44 as “moderate” ones. The discrimination index values in the traditional nonparametric method are all between 0.0500 and 0.9000. Based on the evaluation standards in Table 1, items #10 and #26 are “poor”, items #9, #22, #25, #30 and #45 are “marginal”, items #1, #5, #7, #18, #31 and #42 are “reasonably good”, and the rest of items are “very good”. The discrimination index values in the parametric method are all between 0.0173 and 0.6196. Items #1, #2, #3, #4, #5, #7, #8, #9, #10, #11, #17, #18, #22, #24, #25, #26, #30, #31, #33 and #45 are “poor”, items #6, #12, #15, #16, #27, #28, #29, #32, #34, #36, #38, #39, #40, #41 and #42 are “marginal”, items #13, #14, #19, #23, #35, #37 and #43 are “reasonably good”, and items #20, #21 and #44 are “very good”. Both methods simultaneously identify items #10 and #26 as “poor” items, and identify items #20, #21 and #44 as “very good” items. Table 11summarizes our discussions as follows:

Table 11: Categories of English Test Items

	Nonparametric Method		Parametric Method	
	Category	Item #	Category	Item #
Difficulty	Hard	None	Hard	None
	Moderate	6,13,14,15,16,19,**20**, **21**,29,34,35,37,39,43, **44**	Moderate	**20,21,44**
	Easy	The rest of items	Easy	The rest of items
Discrimination	Poor	**10, 26**	Poor	1,2,3,4,5,7,8,9,**10**,11,17,18, 22,24,25,**26**,30,31,33,45
	Marginal	9,22,25,30,45	Marginal	6,12,15,16,27,28,29,32,34, 36,38,39,40,41,42
	Reasonably Good	1,5,7,18,31,42	Reasonably Good	13,14,19,23,35,37,43
	Very Good	2,3,4,6,8,11,12,13,14, 15,16,17,19,**20**,**21**,23, 24,27,28,29,32,33,34, 35,36,37,38,39,40,41, 43,**44**	Very Good	**20**,**21**,**44**

Real data analysis (see Tables 6 and 7) shows that, for almost all items, the parametric method gives the difficulty index values slightly larger than the nonparametric method, however, the parametric method gives the discrimination index values much smaller than the nonparametric method. How do we explain these phenomena? Since the parametric method, in essence, considers the performance of the middle-scoring group, the correct answer rate of the high-scoring group calculated by the parametric method, denoted as CAR(H), is usually smaller than that calculated by the nonparametric method, denoted as CAR*(H), for each item. Similarly, the correct answer rate of the low-scoring group calculated by the parametric method, denoted as CAR(L), is usually larger than that calculated by the nonparametric method, denoted as CAR*(L), for each item. Readers may refer to Tables 8 and 9. That is, we usually have $CAR(H) \leq CAR^*(H)$ and $CAR(L) \geq CAR^*(L)$. It then yields that $(CAR(H) + CAR(L))/2$, the difficulty index value given by the parametric method, would tend to be close to $(CAR^*(H) + CAR^*(L))/2$, the difficulty index value given by the nonparametric method, and $CAR(H) - CAR(L)$, the discrimination index value given by the parametric method, would be smaller than $CAR^*(H) - CAR^*(L)$, the discrimination index value given by the nonparametric method.

4.3 How Does the Performance of Middle-Scoring Group Affect Both Index Values?

In nonparametric method, the value of $P_{ij}^{[H]}$ depends only on the performance of high-scoring group, and the value of $P_{ij}^{[L]}$ depends only on the performance of low-scoring group. Therefore, both the difficulty index value given by (3.5) and the discrimination index value given by (3.6) would not be affected by the performance of middle-scoring group. In parametric method, based on (3.7) and (3.8), we know that the values of $P_{ij}^{[H]}$ and $P_{ij}^{[L]}$ involve probabilities which are related to the performance of all the students. Therefore, each student might contribute to the values of $P_{ij}^{[H]}$ and $P_{ij}^{[L]}$. That is, the performance of middle-scoring group may affect both index values. We provide concrete examples below to do the demonstration.

Example 3 (Example 2 modified): We use the data in Example 2 (see Table 4 in Section 3.5) but change the number of correct answers for students #6, #7 and #8. All students now in the middle-scoring group got four correct answers, and they belong to the same group before and after making changes. In this example, the performance of middle-scoring group is close to that of high-scoring group. The following table shows the new data:

Table 12: Test Results of Example 3

		Testlet #1	Testlet #2		Testlet #3			
Case Number		1	1	2	1	2	3	# of correct answers
	1	C	B	D	A	D	B	6
	2	C	B	D	A	D	D	5
H-Group	3	A	B	D	A	D	B	5
	4	C	B	C	A	D	B	5
	5	C	B	D	A	A	D	4
	6*	C	D	C	A	D	B	4
M-Group	7*	C	B	C	A	D	D	4
	8*	A	B	C	A	D	B	4
	9	A	B	C	B	A	D	1
	10	C	D	C	B	A	D	1
L-Group	11	A	D	C	A	A	D	1
	12	A	D	C	B	A	D	0
Correct Answer		C	B	D	A	D	B	

The values of difficulty and discrimination indexes under the parametric method are given in Table 13.

Table 13: The Values of Difficulty and Discrimination Indexes in Example 3

T #	I #	Correct Answer Rate of H-Group $P_{ij}^{[H]}$	Correct Answer Rate of L-Group $P_{ij}^{[L]}$	Difficulty Index Value P_{ij}	Discrimination Index Value D_{ij}
1	1	0.7938	0.3443	0.5690	0.4496
2	1	1.0000	0.2459	0.6230	0.7541
	2	0.7113	0.0000	0.3557	0.7113
3	1	1.0000	0.1639	0.5820	0.8631
	2	1.0000	0.0000	0.5000	1.0000
	3	0.7835	0.0000	0.3918	0.7835

Example 4 (Example 2 modified): We still use the data in Example 2 (see Table 4 in Section 3.5) but change the number of correct answers for students #5 and #8. All students now in the middle-scoring group got three correct answers, and they belong to the same group before and after making changes. In this example, the performance of middle-scoring group is in between the performances of high-scoring and low-scoring groups. The following table shows the new data:

Table 14: Test Results of Example 4

	Case Number	Testlet #1 1	Testlet #2 1	Testlet #2 2	Testlet #3 1	Testlet #3 2	Testlet #3 3	# of correct answers
H-Group	1	C	B	D	A	D	B	6
	2	C	B	D	A	D	D	5
	3	A	B	D	A	D	B	5
	4	C	B	C	A	D	B	5
M-Group	5*	C	B	D	B	A	D	3
	6	A	D	C	A	D	B	3
	7	C	B	C	A	A	D	3
	8*	A	B	C	A	D	D	3
L-Group	9	A	B	C	B	A	D	1
	10	C	D	C	B	A	D	1
	11	A	D	C	A	A	D	1
	12	A	D	C	B	A	D	0
Correct Answer		C	B	D	A	D	B	

The values of difficulty and discrimination indexes under the parametric method are given in Table 15.

Table 15: The Values of Difficulty and Discrimination Indexes in Example 4

T #	I #	Correct Answer Rate of H-Group $P_{ij}^{[H]}$	Correct Answer Rate of L-Group $P_{ij}^{[L]}$	Difficulty Index Value P_{ij}	Discrimination Index Value D_{ij}
1	1	0.7143	0.2857	0.5000	0.4285
2	1	1.0000	0.2857	0.6429	0.7143
	2	0.7143	0.0000	0.3571	0.7143
3	1	1.0000	0.1429	0.5714	0.8571
	2	1.0000	0.0000	0.5000	1.0000
	3	0.8571	0.0000	0.4285	0.8571

Example 5 (Example 2 modified): We still use the data in Example 2 (see Table 4 in Section 3.5) but change the number of correct answers for students #5, #6 and #7. All students now in the middle-scoring group got two correct answers, and they belong to the same group before and after making changes. In this example, the performance of middle-scoring group is close to that of low-scoring group. The following table shows the new data:

Table 16: Test Results of Example 5

	Case Number	Testet #1 1	Testlet #2 1	Testlet #2 2	Testlet #3 1	Testlet #3 2	Testlet #3 3	# of correct answers
H-Group	1	C	B	D	A	D	B	6
	2	C	B	D	A	D	D	5
	3	A	B	D	A	D	B	5
	4	C	B	C	A	D	B	5
M-Group	5*	A	B	D	B	A	D	2
	6*	A	D	C	A	D	D	2
	7*	C	B	C	B	A	D	2
	8	A	D	C	A	D	D	2
L-Group	9	A	B	C	B	A	D	1
	10	C	D	C	B	A	D	1
	11	A	D	C	A	A	D	1
	12	A	D	C	B	A	D	0
Correct Answer		C	B	D	A	D	B	

The values of difficulty and discrimination indexes under the parametric method are given in Table 17.

Table 17: The Values of Difficulty and Discrimination Indexes in Example 5

T #	I #	Correct Answer Rate of H-Group $P_{ij}^{[H]}$	Correct Answer Rate of L-Group $P_{ij}^{[L]}$	Difficulty Index Value P_{ij}	Discrimination Index Value D_{ij}
1	1	0.6627	0.0550	0.3588	0.6077
2	1	1.0000	0.2386	0.6193	0.7614
	2	0.7590	0.0000	0.3795	0.7590
3	1	1.0000	0.0795	0.5398	0.9205
	2	1.0000	0.0000	0.5000	1.0000
	3	0.7590	0.0000	0.3795	0.7590

Since the performance of the middle-scoring group does not affect both index values under the nonparametric method, we only summarize the correct answer rates, the difficulty index and discrimination index values of Examples 2, 3, 4 and 5 given by the parametric method for comparison in the following tables:

Table 18: Comparison of Difficulty Index Values of Four Examples

T #	I #	Example 2	Example 3	Example 4	Example 5
1	1	0.5719	0.5690	0.5000	0.3588
2	1	0.7250	0.6230	0.6429	0.6193
	2	0.3918	0.3557	0.3571	0.3795
3	1	0.6250	0.5820	0.5714	0.5398
	2	0.5000	0.5000	0.5000	0.5000
	3	0.3557	0.3918	0.4285	0.3795

Table 19: Comparison of Discrimination Index Values of Four Examples

T #	I #	Example 2	Example 3	Example 4	Example 5
1	1	0.4438	0.4996	0.4285	0.6077
2	1	0.5500	0.7541	0.7143	0.7614
	2	0.7835	0.7113	0.7143	0.7590
3	1	0.7500	0.8631	0.8571	0.9205
	2	1.0000	1.0000	1.0000	1.0000
	3	0.7113	0.7835	0.8571	0.7590

Table 20: Comparison of Correct Answer Rates of Four Examples

T #	I #	Example 2	Example 3	Example 4	Example 5
1	1	0.7936(H) 0.3500(L)	0.7938(H) 0.3443(L)	0.7143(H) 0.2857(L)	0.6627(H) 0.0550(L)
2	1	1.0000(H) 0.4500(L)	1.0000(H) 0.2459(L)	1.0000(H) 0.2857(L)	1.0000(H) 0.2386(L)
	2	0.7835(H) 0.0000(L)	0.7113(H) 0.0000(L)	0.7143(H) 0.0000(L)	0.7590(H) 0.0000(L)
3	1	1.0000(H) 0.2500(L)	1.0000(H) 0.1639(L)	1.0000(H) 0.1429(L)	1.0000(H) 0.0795(L)
	2	1.0000(H) 0.0000(L)	1.0000(H) 0.0000(L)	1.0000(H) 0.0000(L)	1.0000(H) 0.0000(L)
	3	0.7113(H) 0.0000(L)	0.7835(H) 0.0000(L)	0.8571(H) 0.0000(L)	0.7590(H) 0.0000(L)

H: high-scoring group; L: low-scoring group

Based on the results in Tables 18 and 19, we see that the difficulty index values tend to become smaller but the discrimination index values tend to become larger when the performance of middle-scoring group is changed. In particular, the correct answer rates, and the two index values are more affected by the middle-scoring group in Example 5 (the performance of the middle-scoring group is close to that of low-scoring group).

5. Conclusions

According to the computing formulas of difficulty and discrimination indexes and the results of data analyses, we can conclude the followings:

(1) To apply the parametric method to compute the difficulty index and discrimination index values of items in testlets, we must have complete information about each student's response to all items.
(2) Though the parametric method is more complicated than the traditional nonparametric method in manipulation, it considers the dependence between items within each testlet. In addition, it takes the performance of middle-scoring group into account. In this regard, the parametric method may provide more information about the difficulty and discrimination indexes.
(3) The values of difficulty and discrimination indexes given by the nonparametric method are not affected by the performance of the middle-scoring group. However, those given by the parametric method are.
(4) For almost all items, the parametric method gives the difficulty index values slightly larger than the nonparametric method, however, the parametric method gives the discrimination index values much smaller than the nonparametric method.
(5) We have shown that, different isolated items having the same correct answer rate must have the same difficulty index and discrimination index values, respectively, in using the parametric method. This does not hold for nonparametric method. For example, the second and the fifth items in English Test in Taiwan (2007) have the same correct answer rate 0.9024, but they have different difficulty index values 0.8000 and 0.8250, and different discrimination index values 0.4000 and 0.3500, respectively. (see Tables 6 and 10).
(6) Two items, belonging to different testlets and having the same correct answer rate, do not necessarily have the same difficulty index value and discrimination index value by using the parametric method. For example, the first items in testlets #1 and #2 have the same correct answer rate 7/12 (see Table 4), but they have different difficulty index values 0.5719 and 0.7250, and different discrimination index values 0.4438 and 0.5500, respectively (see Table 5).
(7) Different performances of middle-scoring group may cause different impacts on the correct answer rate, difficulty index value and discrimination index value of each item in using the parametric method (see Tables 18, 19 and 20).

We consider the dependence between the items within each testlet, and model testlets with appropriate probability structures. Based on the difficulty and discrimination indexes in classical test theory, a parametric method is successfully developed for deriving the formulas to calculate both of them. In addition, an efficient algorithm for computing both index values is also provided by using the probability generating function technique.

References

Ahmanan, J. S. and Glock, M. D. (1981), Evaluating student process: Principles of tests and measurement (6th Edition), Boston, MA: Allyn and Bacon

Crocker, L. and Algina, J. (2008), Introduction to classical and modern test theory, New York: Holt, Rinehart and Winston.

Ebel, R. L. (1967), The relation of item discrimination to test reliability, Journal of Educational Measurement, 4, 125-128

Ebel, R. L. and Frisbie, D. A. (1991), Essentials of educational measurement (5th Edition), Englewood Cliffs, NJ: Prentice-Hall.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47,149-174.

Mehrens, W. A. and Lehmann,I. J. (1991), Measurement and evaluation in education and psychology (4th Edition), New York: Holt, Rinehart and Winston.

Song, C. C., Lu, C. C. and Jinn, J. (2014). A Statistical Analysis of Independent Test Items: A Parametric Approach. Journal of Fundamental and Applied Statistics. Vol 6 (1 and 2), pp. 1-22. Ishaan Publishing House.

Verhelst, N. D. and Verstralen, H. H. F. M. (2008). Some Considerations on the Partial Credit Model. Psicologica, 29, 229-254.

Wang, W. C. and Wilson, M. (2005b). The Rasch testlet model. Applied Psychological Measurement, 29(2), 126-149.