JOINT MAXIMUM LIKELIHOOD ESTIMATES ON ITEMS-EXAMINEES USING THE PROX METHOD: A STUDY ON THE READING SUBTEST OF TOEFL

Widiatmoko

Pusat Pengembangan Penataran Guru Bahasa (PPPGB),
Direktorat Jenderal Pendidikan Dasar dan Menengah
Departemen Pendidikan Nasional
Jakarta, Indonesia

Abstract

Item response theory (IRT) emerges as an accurate solution to the weaknesses of the classical test theory (CTT). IRT provides more advantages than CTT does. The advantages include the requirements of unidimension for items, local independence between examinees and items, and examinee-item parameter invariance. The requirements are needed in test construction. TOEFL is so far known as the test which meets the requirements in language testing. It however concerns IRT. In this case, the research deals with the reading subtest of TOEFL with regard to IRT. The research is designed to estimate examinee-item parameters. As a parameter logistic (1PL) model in IRT, the Prox method is employed to estimate the parameters jointly. This is named joint maximum likelihood estimates. The method requires dichotomous data. Therefore, TOEFL as a good test instrument is chosen. It includes 30 persons as the examinee measure θ parameter and 20 items as the item difficulty b parameter. Unlike CTT, IRT using Prox method is able to estimate the examinee-item parameters jointly. As a result, the values of θ and b prove the ranges as the model intended in IRT, which is commonly named as the item characteristic curve.

Keywords: item response theory, one parameter logistic model, parameter estimates, the Prox method.

INTRODUCTION

Educational development in the global era is put in seriously months or even years. In a micro scope, it is concerned with learning inputs, processes, and output evaluations. On the one hand, output evaluations functions as a determinant in success and failure of the implemented

curriculum. Furthermore, the evaluation takes the form of examination. On the other hand, some experts on evaluation use a different determinant, commonly in the integrated forms, known as program evaluation. The former and the latter perspectives can be quantitatively and qualitatively approached. They, however, need measurement.

Measurement as launched in 1960s—mainly campaigned E.L. Thorndike's Mental Measurement Theory (1914)—aims at measuring difficult and unobserved mental constructs (Crocker and Algina, 1986: 5). One of them is learning outcome which is the result of a measure from the instrument applied to examinees dealing with the trait, in this case, achievement. Hence, attitude, motivation, or interest, although they are mental constructs, are never quantitatively measured in some classrooms. The observation indicates that evaluation on attitude, motivation, and interest does not involve measurement. It may be due to a teacher's unawareness and inability in constructing and analyzing test items. As this proves, measurement in education tends to decrease in much more meaningful implementation.

Measurement commonly involves composite scores in its analysis. These scores then function as a tool of decision making. In the classical test theory, a score depends upon examinee measure and in the opposite way it depends upon item difficulty. It means that an examinee of remarkable measure tends to respond to items correctly and an examinee of unfavourable measure tends to respond to items incorrectly. It also means that items are considered easy when they are responded by a large number of examinees and are considered difficult when they are responsed by few examinees (Dali, 1992: 48). It is known that a correlation between the item and the examinee is implemented in terms of a correlation between probability of the correct response and the examinee's success, although it is probabilistic correlation stating that it may not be so. Consequently, the item and the examinee characteristics, to some extent, cannot be known precisely. The basic implication is that the items produced cannot be applied in a different situation to different examinees and certain examinee measure cannot be known precisely by seeing composite scores obtained.

Item response theory (IRT) tries to overcome these weaknesses, i.e., when an item pool has item difficulties felt easy or difficult by examinees, it is then impossible to measure true examinee measure and to know the different success of examinees. It is said that the value of item difficulty is invariant to examinee measure and the value of examinee measure is invariant to item difficulty (Dali, 1992: 161). Item difficulty as an item's characteristic and examinee measure as an examinee's trait are usually connected by the model forming as a function which is expressed by some parameters. Then, they are called parameters of item's characteristic and

examinee's trait. In various cases, the parameter of item's characteristic is known before the item is applied to examinees. The parameters of item's characteristics are usually kept in item banking. Other identities concerning various items have been well-recorded. Of course, it involves item calibration. Therefore, when examinees do items, they produce composite scores which are interpreted as examinee measures θ. [Note that Anderson and Helmick (1983) and Hulin *et al.*, (1983) use the symbol θ to indicate examinee measure and the degree of examinee's attitude. In this case, the symbol θ is used only to indicate parameter of examinee measure]. Empirically, examinee measure tends to be constant when he does the item of the same characteristic. Therefore, it is said that the estimate of the parameter of examinee's trait is done by knowing the examinee's trait item's characteristic.

The parameter of item's characteristic and examinee's trait in IRT can be estimated marginally or jointly. The estimate is not found in the classical test theory. It means that an examinee of certain measure θ can be known as he does an item of certain difficulty b_j. It also means that an item of certain difficulty b_j can be known as it is done by an examinee of certain measure θ.

There are many estimates of parameters, the procedures or methods of which can be used for the IRT models. One of the models is one parameter logistic (1PL) model or commonly named the Rasch model. The model only contains the parameter of examinee measure θ for the examinee's trait and the parameter of item difficulty b_j for the item's characteristic. The model is initially known before the estimate occured. Then, the estimate using the Prox method is employed for the 1PL model.

There are some questions that need to be addressed for future research: Can the Prox method be applied to estimate the parameters of items-examinees in a two parameter logistic model?; Can the Prox method be applied to estimate the parameters of items-examinees in a three parameter logistic model?; Can the Prox method be applied to estimate the parameters of items-examinees in a four parameter logistic model?; Can the Prox method be applied to estimate the parameters of items-examinees in polytomous scores?; Does the Prox method produce item characteristic curves monotonically?; Does the Prox method prove that the result of the estimate is the same as the item characteristic curve when the number of items and examinees are increased?; Is it true that Prox method is an appropriate method to estimate parameters of joint items-examinees in 1PL model?

Certainly, there are many other questions to be investigated. However, this article is concerned with the last of the seven questions mentioned above.

THEORETICAL REVIEW

Item-Examinee Parameters in 1PL Model

In the classical test theory and IRT, the central analysis is the item and the examinee. Basically, they are different. IRT provides some advantages such as local independence between an item and an examinee, examinee's invariance on an item and vice versa, and unidimension of an item (Hulin, *et al.,* 1983: 40-43). The classical test theory, however, does not provide such advantages. It is shown that subpopulation of items cannot be separated from subpopulation of examinees. As a result, when a subpopulation of homogenous items is undertaken by subpopulation of different examinees, the characteristics of items will change. In other words, item difficulty and item discriminating power will change due to a different examinee measure, and examinee measure will change due to his doing different items (Dali, 1992: 4-5). Therefore, the classical test theory states that examinee-item depends on each other, that examinee measure is not invariant to item difficulty, and that an item tends to be unidimension.

The advantages of IRT are explained in the following. First, local is supposed to be a point in a continuum of examinee's trait parameter θ, which can be an interval form containing homogenous subpopulation of examinees. However, independence is interpreted as all examinees in the subpopulation, which are independent with regards to the items in the subpopulation. This means that composite scores of items responded by the homogeneous subpopulation of examinees should be independent (Dali, 1992: 170-171). Local independence can also be interpreted as responses which are conditionally independent in the subpopulation where examinee's latent trait F_1, ..., F has constant value f_1, ..., f. In other words, it means that two items or more are uncorrelated in a homogenous subpopulation with a particular level of fixed latent traits θ (Hulin *et al.*, 1983: 43; McDonald, 1999: 255). [Note that in a heterogeneous population where θ varies, item scores should be correlated]. Furthermore, Lord and Novick (1968: 361) state that local independence means that within any group of examinees, all characterized by the same values θ_1, θ_2, ..., θ_k, the (conditional) distributions of the item scores are all independent of each other.

Secondly, parameter invariance is interpreted as a function of the single measure θ or the item characteristic b which does not change across subpopulation whenever the subpopulation changes. Parameter invariance is also interpreted as an examinee's trait, which does not change whenever the item chosen changes (Hulin *et al.*, 1983: 44; Dali, 1992: 173).

Thirdly, unidimension is interpreted as an item that measures one trait or characteristic over the examinees (Dali, 1992: 164). It also means

that the probability of an item response is a function of a single latent characteristic of the examinee θ (Hulin *et al.*, 1983: 40). Since every characteristic is determined by one measure, one type of measure can also be interpreted as the requirement to measure only one dimension of examinee's latent trait over subpopulation. Usually, items that meet the requirement of unidimension can be found in a certain battery. Mostly, the items are available in the homogenous test battery (Lord and Novick, 1968: 381). The items have so far been developed and kept in item banking. The advantages are then considered why they are appropriately used in test construction.

Test instrument calibration is always concerned with how a model intended is decided instead of unidimension, parameter invariance, and local independence requirements. The models intended include the Guttman perfect scale model, the latent distance model, the linear model, the normal ogive model, and the logistic model.

Logistic Model

The logistic model is in fact the normal ogive model, but the two are not the same. The logistic model does not require the intricate mathematic calculations, whereas, the normal ogive model requires the complicated ones.

Like the normal ogive model, the logistic model also undertakes the models of one, two, three, and four parameters. The logistic model with one parameter called the one parameter logistic (1PL) model only employs item difficulty parameter b_j. [Note that item difficulty has various symbols, like α (Anderson and Helmick, 1983), d (Henning, 1987), b_j (Hulin, *et al.,* 1983; Hambleton, 1989; Anastasi and Urbina 1997). This article uses the symbol b_j referring to item difficulty]. The two parameter logistic (2PL) model not only employs the parameter of item difficulty b_j, but it also employs the parameter of item discriminating power a. [Note that the symbol a for item discriminating power is taken from Hulin, *et al.* (1983); Hambleton (1989); Anastasi and Urbina (1997)]. The three parameter logistic (3PL) model not only employs the parameter of item difficulty b_j and the parameter of item discriminating power a, but it also employs the parameter of the examinee of low measure responding items correctly or guessing correct c. [Note that the symbol c indicating guessing correct is taken from Hulin, *et al.* (1983); Hambleton (1989); Anastasi and Urbina (1997)]. The four parameter logistic (4PL) model not only employs the parameter of item difficulty b_j, the parameter of item discriminating power a, and the parameter of guessing correct c, but it also employs the parameter of the examinee of high measure responding the items incorrectly γ. [Note that the symbol γ for the parameter

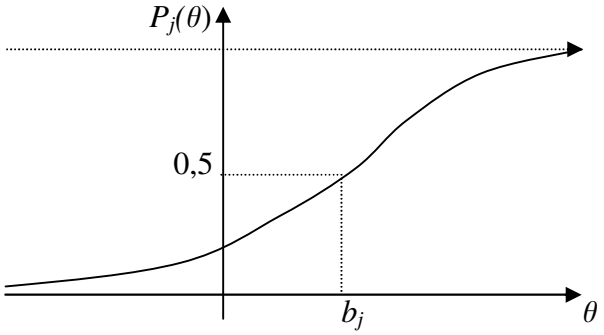
of guessing incorrect is taken from Hambleton (1989)]. The latter model as stated by McDonald (1967) and Barton and Lord (1981) as quoted by Hambleton, (1989: 157) can be considered as the difficult model to prove.

Among the four logistic models, the simplest one is the 1PL model, also called the Rasch model. The model is principally described as an item characteristic curve of the examinees' responses to the items. In other words, the item characteristic curve is a model of the probability of a 1 (correct) response on item *j*, given the examinee's parameter θ which is symbolized as the function $P_j(\theta)$. The item characteristic curve only depends upon the distance between θ and b_j, meaning that the item parameter b_j is only defined relative to the examinee's parameter θ (Andersen, 1983: 197-198). [Note that the symbol θ for examinee measure and $P_j(\theta)$ for the probability of correct response are taken from Hulin, *et al.* (1983); Hambleton (1989); Anastasi and Urbina (1997)]. If written in the mathematical form, 1PL model looks as follows:

$$Pj(\theta) = \frac{e^{D(\theta - bj)}}{1 + e^{D(\theta - bj)}}$$

(j = 1, 2, ..., n); D = a constant weighing 1,7; *e* exponential numbers (Hambleton, 1989: 154).

The 1PL model is empirically the model that proves a curve like ogive along with its low and high asymptotes. Low asymptote approximates the value of $-\infty$, and high asymptote approximates the value of $+\infty$. It means that in a certain condition the examinee of low measure will execute the probability of correct response as poorly as the groups of the homogenous subpopulation. On the contrary, in a certain condition the examinee of high measure will execute the probability of correct response as well as the groups of the homogenous subpopulation. The similar relativity in each group of subpopulations tends to approximate asymptote on each point. Theoretically, the two points describe the examinee measure ranging from $-\infty$ to $+\infty$ (Hambleton, 1989: 161; Dali, 1992: 224). Practically, it ranges from -3 to +3 (Hulin, *et al.,* 1983: 101) or from -4 to +4 (Dali, 1992: 224). [Note that range (-3, +3) or (-4, +4) rather than range $(-\infty, +\infty)$ is executed by transforming the values of some normal probability distribution into the values of standard normal probability distribution. This is done by deciding the value of the mean parameter = 0 and the value of standard deviation parameter $\sigma = 1$]. If the item characteristic curve is described, it will be as follows:



The item characteristic curve can then be the model in the study of the item-examinee parameters. One of them is used as a model of the parameter estimates. Therefore, the estimate is determined by the model of item characteristics intended. The estimates of the parameter can be done jointly or marginally, which are known as the maximum likelihood estimates. Among them is the simplest estimate of the parameter conformed in 1PL model. So far, the estimate executes the Prox method.

Parameter Estimates Using the Prox Method

There are many estimate methods in IRT. One needs dichotomous data of items, and others need polytomous ones in their analysis. Parameter estimates in IRT can be applied to 1PL, 2PL, 3PL, or 4PL models. The estimate in 4PL model is more difficult than that in 3PL model; 3PL model is more difficult than that in 2PL model; 2PL model is more difficult than that in 1PL model. The estimate in 1PL model is then called the estimate using the Prox method which requires dichotomous data of items. Since the estimate is for 1PL model, then it estimates two parameters, i.e., parameters of examinee measure and item difficulty. The estimate with the Prox method is known as a joint maximum likelihood estimate. It means that when the estimate brings about, all values of parameters of examinees' traits and items' characteristics are not known (Dali, 1992: 264).

Principally, the Prox method arranges initial item difficulty b_j for the parameter of item's characteristic and initial examinee measure θ for the parameter of examinee's trait. The initial value b_j and θ are based on logit incorrect value and logit correct value. Usually, the initial value is expressed as deviation from the mean of the logit, so both values will be b_A and θ_A. Referring to variance of the logit, the Prox method forms expansion factors for two parameters, i.e., $F(b_j)$ and $F(\theta)$. By using initial value b_j, initial value θ, expansion factor $F(b_j)$, and expansion factor $F(\theta)$, the parameter of item's characteristic and the parameter of examinee's trait are estimated.

The steps of the estimates using the Prox method are as follows (Henning 1987: 118-122). First, dichotomous response of correct 1 and incorrect 0 matrix is edited. It is done in such a way that every examinee or item for which all responses are correct or all responses are incorrect are eliminated. It means that the examinee responding all items correctly and all items incorrectly is not put in the matrix. It also means that the item responded by all examinees correctly and responded by all examinees incorrectly is not put in the matrix. Therefore, the matrix only contains responses of correct-incorrect proportionally and improportionally in each column and row. Finally, the examinees' scores are put in order vertically from the smallest to the largest, and the items' proportions of correct responses are put in order horizontally from the largest to the smallest. In other words, the examinees are ordered from the lowest measure to the highest one, and the items are ordered from the easiest to the most difficult.

Second, the initial item difficulty b_j calibration is computed. This is done by using logit incorrect value for each possible number correct. The logit incorrect value for each item is computed as the natural logarithm of the ratio of the proportion incorrect to the proportion correct. This is a reference to calibrate the initial item difficulty b_j. Then, the examinees do N items, so the mean of logit incorrect values among the items is computed. Considering that the items are sample, variance of logit incorrect value is obtained. To compute variance of the logit incorrect value, it is necessary to compute the sum of logit incorrect value squared minus N items times the mean adjustment squared, all divided by the number of items minus one. The variance is required for the expansion factor computation. The initial item difficulty b_j calibration is meant to decide deviation from the mean of logit incorrect value. This is done for all items.

Third, the initial examinee measure θ is calculated. Unlike the case with the items, the calculation is done by using logit correct value instead of logit incorrect value. The logit correct value for each examinee is computed as the natural logarithm of the ratio of the proportion correct to the proportion incorrect. This is a reference to compute the initial examinee measure θ. Then, the items are done by M examinees, so the mean of logit correct values among the examinees is computed. Considering that the examinees are sample, variance of logit correct value is obtained. To compute variance of the logit correct value, it is necessary to compute the sum of logit correct value squared minus M examinees times the mean adjustment squared, all divided by the number of the examinees minus one. The variance is required for the expansion factor computation. The initial examinee measure θ is meant to decide deviation from the mean of logit correct value. This is done for all examinees.

Fourth, the expansion factor for items $F(b_j)$ and the value of b_j are calculated. The Prox method does not do the estimate cycle repeatedly. Nevertheless, it executes the statistic estimate on sample variance of logit incorrect value and logit correct value. The factor for the estimate is the expansion factor, which is then multiplied by the initial item difficulty b_j to obtain the final estimate. This is done for all items.

Fifth, the expansion factor for examinees $F(\theta)$ and the value of θ are calculated. The Prox method does not do the estimate cycle repeatedly. It executes the statistic estimate on sample variance of logit correct value and logit incorrect value. The factor for the estimate is also the expansion factor, which is then multiplied by the initial examinee measure θ to obtain the final estimate. This is done for all examinees.

METHODOLOGY

This study is a survey which undertakes examinees-items of TOEFL's reading subtest. Subpopulation was taken by using random purposive sampling technique which involves some steps. First, population (examinees in Jakarta) was determined. Second, target population (the examinees doing TOEFL in 2004-2005) was determined [Note that subpopulation is the term used to substitute for sample]. Due to the cost and time, 30 persons as a subpopulation of the examinees were taken in a simple random way. Third, since the population is mostly concerned with this subtest, one subtest of reading in TOEFL is determined purposively. Fourth, from the subtest, in a simple random way, 20 items as a subpopulation of the items were taken. Therefore, the research analysis units are 30 examinees and 20 items of the subtest of reading in TOEFL.

ANALYSIS

The Prox method employs some steps to estimate item difficulty b_j and examinee measure θ. First, dichotomous responses of correct 1 and incorrect 0 matrix are edited. It is done in such a way that every examinee or item for which all responses are correct or all responses are incorrect is eliminated (see Table 1).

The examinees' scores are put in order vertically from the smallest to the largest, and the items' proportions of correct responses are put in order horizontally from the largest to the smallest. In other words, the examinees are ordered from the lowest measure to the highest one, and the items are ordered from the easiest to the most difficult (see Table 2).

The calibration of initial item difficulty b_j is computed. It uses logit incorrect value (LG_i) as a reference to calibrate initial value of parameters of item difficulty b_j.

$$LGi = \ln \frac{Q(\theta)}{P(\theta)}$$; where $Q(\theta)$ = the probability of incorrect response, $P(\theta)$ = the probability of correct response.

Then, the examinees do N items so that the mean of logit incorrect values among the items can be computed. Variance of logit incorrect value (S^2_{LG}) is obtained.

$$S^2_{LG} = \frac{1}{N-1}\left[\sum_{j=1}^{N}(LGi)^2 - N\mu^2_{LG}\right],$$ where N = a number of items, $_{LG}$ = the mean of logit incorrect value.

The initial item difficulty b_j is meant to decide the deviation from the mean of logit incorrect value. This is done for all items (see Table 3).

Then, the initial examinee measure θ is calculated. This is done by using logit correct value (LS_j) as a reference.

$$LSj = \ln \frac{P(\theta)}{Q(\theta)}$$

The items responded by M examinees are necessary to compute the mean of logit correct values among the examinees. Then, the variance of logit correct value is obtained.

$$S^2_{LS} = \frac{1}{M-1}\left[\sum_{i=1}^{M}(LSj)^2 - M\mu^2_{LS}\right],$$ where M = a number of examinees, $_{LS}$ = the mean of logit correct value.

The value of the parameter of the examinee measure θ is determined as deviation from the mean of logit correct. This is done for all examinees (see Table 4).

Then, the expansion factor for item characteristic $F(b_j)$ and the value of b_j is computed.

$$F(b_j) = \sqrt{\frac{1+\frac{S^2_{LS}}{2{,}89}}{1-\frac{S^2_{LS}S^2_{LG}}{8{,}35}}}$$

The estimate of the parameter of item difficulty b_j is obtained by multiplying the initial value of the parameter of item difficulty by the value of the expansion factor for item difficulty b_j gained. This is done for all items (see Table 5).

The expansion factor of examinee's trait $F(\theta)$ and the value of θ are computed.

$$F(\theta) = \sqrt{\frac{1+\frac{S^2{}_{LG}}{2,89}}{1-\frac{S^2{}_{LS}\,S^2{}_{LG}}{8,35}}}$$

Finally, the estimate of parameter of examinee measure θ is obtained by multiplying the initial value of examinee measure by the value of the expansion factor for examinee measure θ gained. This is done for all examinees (see Table 6).

From the values of the estimates of θ and b, it is proved that the item characteristic curve conforms to the 1PL model.

CONCLUSION

From the data analysis, it can be concluded that the examinee measure θ, as examinee's trait, and item difficulty $b_{,j}$ as item's characteristic, can be estimated jointly by using the Prox method. The estimate using the Prox method factually provides the accurate result. It so happens that the estimate forms the item characteristic curve of the 1PL model. Therefore, the joint estimate of item-examinee parameters can be used as a proof of intended model accuracy, i.e., the examinee measure θ is as similar as the item difficulty b_j by condition of $P_j(\theta) = 0,5$.

It is suggested that those who are concerned with measurement, evaluation, test, and assessment pay much attention to examinee's trait and item's characteristic. However, it deals with the decision on the part of the examinees including pass-fail, accepted-rejected, and so forth, as well as the decision on the items including good-bad items, valid-not valid items, and so on.

The research implication is in line with the recommendation of the requirement of test construction consisting of good items kept in the item banking, which tells us about item and examinee identity.

REFERENCES

Anastasi, A. & S. Urbina. 1997. *Psychological testing.* New Jersey: Prentice-Hall.

Anderson, E.B. Helmick. 1983. Analyzing data using the Rasch model. In Scarvia B.A. & John S.H. (Eds.). *On educational testing.* San Francisco, California: Jossey-Bass Inc.

Crocker, L.& J. Algina. 1986. *Introduction to classical and modern test theory.* Florida: Holt, Rinehart and Winston.

Dali, S.N. 1992. *Pengantar teori sekor pada pengukuran pendidikan.* Jakarta: Penerbit Gunadarma.

Hambleton, R.K. 1989. Principles and selected applications of item response theory. In Robert L. Linn (Ed.). *Educational measurement.* New York: American Council on Education and Macmillan Publishing.

Henning, G. 1987. *A guide to language testing: development, evaluation, research.* Cambridge, Massachusetts: Newbury House Publishers.

Hulin, C.L., F. Drasgow, & C.K. Parsons. 1983. *Item response theory: application to psychological measurement.* USA: Dow Jones-Irwin.

Lord, F.M. & M.R. Novick. 1968. *Statistical theories of mental test scores.* Canada: Addison-Wesley Publishing Company.

McDonald, R.P. 1999. *Test theory: a unified treatment.* New Jersey: Lawrence Erlbaum Associates.

APPENDIX

Examinees	Items of Reading Subtest of TOEFL																				Scores
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
1	0	0	1	0	0	0	0	0	0	1	0	0	1	1	0	1	1	0	1	1	8
2	0	0	0	1	0	1	1	0	1	1	1	1	0	1	0	1	0	0	1	0	10
3	1	0	1	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0	1	0	9
4	1	0	1	0	0	0	0	0	1	1	1	1	1	0	0	1	1	1	1	0	11
5	0	1	1	1	0	1	1	0	0	0	1	1	1	0	1	1	1	0	1	1	13
6	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	1	0	1	1	0	9
7	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1	1	0	10
8	1	1	1	1	0	0	0	1	0	1	1	1	1	0	1	1	1	1	1	0	14
9	0	0	0	1	0	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	11
10	1	0	1	1	0	1	0	1	1	1	1	1	0	1	0	1	0	0	0	0	11
11	0	0	0	0	1	1	0	1	1	0	1	1	0	1	1	1	0	0	1	0	10
12	0	0	0	1	1	0	1	0	1	0	0	0	0	0	1	1	1	1	0	0	8
13	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	0	0	0	13
14	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	15
15	0	0	0	0	0	0	0	1	1	1	1	0	0	1	0	1	1	1	1	0	9
16	1	1	1	1	0	0	0	1	1	1	0	0	1	0	1	0	1	0	1	0	11
17	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	0	14
18	0	0	1	1	0	1	1	1	0	0	0	0	1	1	1	1	1	1	0	1	12
19	1	1	1	0	1	0	1	0	1	1	0	1	1	0	1	1	1	1	1	0	14
20	0	0	0	0	0	1	0	1	0	1	0	0	1	1	1	1	0	0	0	0	7
21	0	0	0	1	1	1	1	0	0	0	0	1	1	0	0	1	0	1	0	1	9
22	1	1	1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1	0	0	9
23	0	1	1	0	1	0	0	1	1	1	1	0	1	1	0	0	1	0	1	1	12
24	0	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	1	0	1	0	6
25	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1	0	11
26	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	1	0	1	0	0	5
27	1	1	1	1	0	0	1	1	0	0	1	1	1	1	1	1	1	1	1	1	16
28	0	1	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	0	13
29	0	0	0	1	1	1	1	1	1	0	0	0	0	1	1	1	0	0	0	0	9
30	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	17
No. of Correct	12	13	16	18	9	14	12	18	15	19	15	18	20	19	17	27	19	16	21	8	

Examinees	Items of Reading Subtest of TOEFL																				Scores
	16	19	13	17	14	10	12	8	4	15	18	3	11	9	6	2	7	1	5	20	
26	1	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0	5
24	0	1	1	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	6
20	1	0	1	0	1	1	0	1	0	1	0	0	0	0	1	0	0	0	0	0	7
1	1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	8
12	1	0	0	1	0	0	0	0	1	1	1	0	0	1	0	0	1	0	1	0	8
15	1	1	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	0	0	0	9
21	1	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	1	0	1	1	9
29	1	0	0	0	1	0	0	1	1	1	0	0	0	1	1	0	1	0	1	0	9
3	1	1	0	1	0	1	0	0	0	1	0	1	1	1	0	0	0	1	0	0	9
6	1	1	1	0	0	0	1	0	1	0	1	1	0	0	0	1	0	1	0	0	9
22	1	0	0	0	1	1	1	1	0	0	1	1	0	0	0	1	0	1	0	0	9
2	1	1	0	0	1	1	1	0	1	0	0	0	1	1	1	0	1	0	0	0	10
7	1	1	1	1	1	1	0	0	0	1	1	0	1	1	0	0	0	0	0	0	10
11	1	1	0	0	1	0	1	1	0	1	0	0	1	1	1	0	0	0	1	0	10
9	1	1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	1	0	0	1	11
4	1	1	1	1	0	1	1	0	0	0	1	1	1	1	0	0	0	1	0	0	11
10	1	0	0	0	1	1	1	1	1	0	0	1	1	1	1	0	0	1	0	0	11
16	0	1	1	1	0	1	0	1	1	1	0	1	0	1	0	1	0	1	0	0	11
25	1	1	0	0	1	1	1	1	1	0	1	0	0	0	1	1	0	1	0	0	11
18	1	0	1	1	1	0	0	1	1	1	1	1	0	0	1	0	1	0	0	1	12
23	0	1	1	1	1	1	0	1	0	0	0	1	1	1	0	1	0	0	1	1	12
5	1	1	1	1	0	0	1	0	1	1	0	1	1	0	1	1	1	0	0	1	13
28	1	1	1	1	1	0	1	0	1	1	1	0	0	0	1	1	1	0	1	0	13
13	1	0	1	1	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	0	13
17	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	0	0	0	14
8	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	0	1	0	0	14
19	1	1	1	1	0	1	1	0	0	1	1	1	0	1	0	1	1	1	1	0	14
14	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	15
27	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	0	1	16
30	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	17
No. of Correct	27	21	20	19	19	19	18	18	18	17	16	16	15	15	14	13	12	12	9	8	

Table 3. Initial Value of Item Difficulty

Items	No of Correct	P(θ)	Q(θ)	LG_i	$(LG_i)^2$	b_{Bi}
16	27	0.90	0.10	-2.197225	4.827796	-1.994355
19	21	0.70	0.30	-0.847298	0.717914	-0.644428
13	20	0.67	0.33	-0.693147	0.480453	-0.490277
17	19	0.63	0.37	-0.546544	0.298710	-0.343674
14	19	0.63	0.37	-0.546544	0.298710	-0.343674
10	19	0.63	0.37	-0.546544	0.298710	-0.343674
12	18	0.60	0.40	-0.405465	0.164402	-0.202595
8	18	0.60	0.40	-0.405465	0.164402	-0.202595
4	18	0.60	0.40	-0.405465	0.164402	-0.202595
15	17	0.57	0.43	-0.268264	0.071966	-0.065394
18	16	0.53	0.47	-0.133531	0.017831	0.069339
3	16	0.53	0.47	-0.133531	0.017831	0.069339
11	15	0.50	0.50	0.000000	0.000000	0.202870
9	15	0.50	0.50	0.000000	0.000000	0.202870
6	14	0.47	0.53	0.133531	0.017831	0.336401
2	13	0.43	0.57	0.268264	0.071966	0.471134
7	12	0.40	0.60	0.405465	0.164402	0.608335
1	12	0.40	0.60	0.405465	0.164402	0.608335
5	9	0.30	0.70	0.847298	0.717914	1.050168
20	8	0.27	0.73	1.011601	1.023336	1.214471

Table 4. Initial Value of Examinee Measure

Examinees	No of Correct	P(θ)	Q(θ)	LS_j	$(LS_j)^2$	θ_{Aj}
26	5	0.25	0.75	-1.098612	1.206949	-1.293709
24	6	0.30	0.70	-0.847298	0.717914	-1.042394
20	7	0.35	0.65	-0.619039	0.383210	-0.814136
1	8	0.40	0.60	-0.405465	0.164402	-0.600562
12	8	0.40	0.60	-0.405465	0.164402	-0.600562
15	9	0.45	0.55	-0.200671	0.040269	-0.395767
21	9	0.45	0.55	-0.200671	0.040269	-0.395767
29	9	0.45	0.55	-0.200671	0.040269	-0.395767
3	9	0.45	0.55	-0.200671	0.040269	-0.395767
6	9	0.45	0.55	-0.200671	0.040269	-0.395767
22	9	0.45	0.55	-0.200671	0.040269	-0.395767
2	10	0.50	0.50	0.000000	0.000000	-0.195097
7	10	0.50	0.50	0.000000	0.000000	-0.195097
11	10	0.50	0.50	0.000000	0.000000	-0.195097
9	11	0.55	0.45	0.200671	0.040269	0.005574
4	11	0.55	0.45	0.200671	0.040269	0.005574
10	11	0.55	0.45	0.200671	0.040269	0.005574
16	11	0.55	0.45	0.200671	0.040269	0.005574
25	11	0.55	0.45	0.200671	0.040269	0.005574
18	12	0.60	0.40	0.405465	0.164402	0.210368
23	12	0.60	0.40	0.405465	0.164402	0.210368
5	13	0.65	0.35	0.619039	0.383210	0.423943
28	13	0.65	0.35	0.619039	0.383210	0.423943
13	13	0.65	0.35	0.619039	0.383210	0.423943
17	14	0.70	0.30	0.847298	0.717914	0.652201
8	14	0.70	0.30	0.847298	0.717914	0.652201
19	14	0.70	0.30	0.847298	0.717914	0.652201
14	15	0.75	0.25	1.098612	1.206949	0.903516
27	16	0.80	0.20	1.386294	1.921812	1.191198
30	17	0.85	0.15	1.734601	3.008841	1.539504

Table 5. Value Estimate of *b*

Items	b_{Bi}	F(b)	b
16	-1.994355	1.079806	-2.153516
19	-0.644428	1.079806	-0.695857
13	-0.490277	1.079806	-0.529404
17	-0.343674	1.079806	-0.371101
14	-0.343674	1.079806	-0.371101
10	-0.343674	1.079806	-0.371101
12	-0.202595	1.079806	-0.218764
8	-0.202595	1.079806	-0.218764
4	-0.202595	1.079806	-0.218764
15	-0.065394	1.079806	-0.070613
18	0.069339	1.079806	0.074872
3	0.069339	1.079806	0.074872
11	0.202870	1.079806	0.219060
9	0.202870	1.079806	0.219060
6	0.336401	1.079806	0.363248
2	0.471134	1.079806	0.508733
7	0.608335	1.079806	0.656884
1	0.608335	1.079806	0.656884
5	1.050168	1.079806	1.133978
20	1.214471	1.079806	1.311393

Table 6. Value Estimate of *θ*

Examinees	θ_{Aj}	F(θ)	θ
26	-1.293709	1.090018	-1.410166
24	-1.042394	1.090018	-1.136229
20	-0.814136	1.090018	-0.887423
1	-0.600562	1.090018	-0.654623
12	-0.600562	1.090018	-0.654623
15	-0.395767	1.090018	-0.431394
21	-0.395767	1.090018	-0.431394
29	-0.395767	1.090018	-0.431394
3	-0.395767	1.090018	-0.431394
6	-0.395767	1.090018	-0.431394
22	-0.395767	1.090018	-0.431394
2	-0.195097	1.090018	-0.212659
7	-0.195097	1.090018	-0.212659
11	-0.195097	1.090018	-0.212659
9	0.005574	1.090018	0.006076
4	0.005574	1.090018	0.006076
10	0.005574	1.090018	0.006076
16	0.005574	1.090018	0.006076
25	0.005574	1.090018	0.006076
18	0.210368	1.090018	0.229305
23	0.210368	1.090018	0.229305
5	0.423943	1.090018	0.462105
28	0.423943	1.090018	0.462105
13	0.423943	1.090018	0.462105
17	0.652201	1.090018	0.710911
8	0.652201	1.090018	0.710911
19	0.652201	1.090018	0.710911
14	0.903516	1.090018	0.984848
27	1.191198	1.090018	1.298427
30	1.539504	1.090018	1.678088