The Philippine Statistician, 2003
Vol. 52, Nos. 1-4, pp. 81-92

Some Applications of Item Response Theory to Testing

Leonardo S. Sotaridona, Jonny B. Pornel and Arnolfo Vallejo*

received April, 2003; revised September, 2003

ABSTRACT

A test score that is obtained through the number-correct score is often used as estimate of an examinee's proficiency. A problem with this approach is that it does not take into account the characteristic of the item, such as the item difficulty, when estimating the proficiency. Furthermore, when test scores are used to evaluate the performance of a school, the presence of examinees who do not respond according to their true ability (e.g., guessing and copying), could yield estimates of a school performance that is lower than its actual performance. In this study, it was shown that the problem is circumvented using an approach based on the item response theory (IRT). Simulation studies were conducted to illustrate the points.

Key words: item response theory, person-fit statistic, classical test theory

1. INTRODUCTION

Oftentimes, a psychological or educational test is used as a device for obtaining a sample of behavior. This behavior is usually quantified in some way to obtain numerical score. Such scores are tabulated and counted. For example, a multiple-choice test is used to determine the ability level of an examinee to a particular subject. For example, a student has to take a "final exam"' that is designed to measure his/her proficiency of the subject after a semester of study. The score obtained from the final exam would contribute a considerable weight in determining whether or not a student has passed or failed the subject. The proficiency (or ability) of a student is often estimated using the number of correct score to the items in the test [or simply the test score]. A test score that is equal to a cut-off score or greater than a cut-off score is considered a pass; otherwise a failure. This approach of using the test score as proficiency estimate is sometimes referred to as the classical test theory (CTT) approach.

One concern of using a test score as ability estimate is that the estimate is less sensitive to the characteristics of the items. Item characteristics such as an item difficulty and an item discrimination were seem ignored when estimating an examinee ability or proficiency. To illustrate why it is the case, take two examinees A and B who have the same test score on the same test but the correct scores were obtained from different items in the test. If for some reasons, student A got most of his/her correct answers to relatively difficult items while student B got most of his/her correct answers to relatively easy items, it is reasonable and even natural to expect that examinee A showed better performance than examinee B. It is then appropriate to consider the item difficulty in the scoring process. For example, some items should have more weight than the others items. This is not the case using the CTT approach.

A second concern relates to the CTT approach in estimating the item characteristics such as the item difficulty and item discrimination. For example, the CTT item difficulty is

* Mr. Sotaridona and Mr. Vallejo are both from the Mathematics Department, Western Visayas College of Science and Technology, Iloilo City. emails: l.s.sotaridona@edte.utwente.nl, alvallejo47@yahoo.com, while Mr. Pornel is from Special Science, Iloilo National High School, Iloilo City. email: jonnybpornel@yahoo.com

often estimated by the proportion of examinees who got an item correctly. It is obvious that this estimate is highly dependent on the sample of examinees who took the test. The estimate will necessarily be high if the sample of examinees are of high ability and the estimate will be low if the sample of examinees are of low ability. Ideally, it is desirable that the estimate of the item characteristic is independent on the sample of examinees. Furthermore, it is desirable that the ability estimate does not only takes into account the item difficulty but is also invariant on the set of items in the test. A solution to these concerns is to use an item response theory (IRT) to estimate the item characteristics and the examinee ability.

A third concern relates to the practice of using an examinee test score to measure or to evaluate the performance of a school or teacher, or to evaluate the quality of educational program. In these situations, the result of a test has no direct consequence to an examinee. For example, a student's test score will not affect his/her grade in the class. And because the test has no bearing on the student's grade, it is not surprising to find students responding to the test that is not according to his/her ability, for example, by simply guessing at random among the options in the test. If there are considerable number of students who are not responding according to their ability, an estimate of the performance of a school would be lower than the actual performance. We say that the estimate is bias against the school. Ideally, we want to identify those unmotivated examinees and exclude them when estimating the school performance. Using an IRT, it is possible to identify these examinees through person-fit analysis (PFA).

In this paper, simulation studies were conducted (1) to show that the classical test theory (CTT) item difficulty estimates are dependent on the sample of examinee, (2) to compare the estimates of an examinee's proficiency based on CTT-test score and an IRT approach, (3) to investigate the effect of including the test scores of examinees who are randomly guessing when estimating the performance of a school, and (4) to investigate the usefulness of the standardized loglikelihood statistic to detect the response patterns of examinees who are simply randomly guessing in a test.

2. ITEM RESPONSE THEORY

Item response theory postulates that (a) examinee test performance can be predicted (or explained) by a set of factors called traits, latent traits, or abilities, and (b) the relationship between examinee item performance and the set of traits assumed to be influencing item performance can be described by a monotonically increasing function called an item characteristic function (Hambleton & Swaminathan, 1985). A good item response model provides a means of scoring the examinees on the abilities, where abilities must be estimated (or inferred) from the observed examinee performance on a set of test items. Some well-known unidimensional IRT models for dichotomously scored responses are the Rasch model (1PLM), the two-parameter logistic model (2PLM) and the three-parameter logistic model (3PLM). The mathematical form of these models are given as

$$P_i(\theta_j) \equiv \Pr(U_{ji} = 1) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]} \qquad (1)$$

where $P_i(\theta_j)$ is the probability of a correct response to item i=1, 2,..., I for an examinee j=1, 2,..., J with ability level θ_j. U_{ji} is a response indicator (1 if correct and 0 if not correct), and $a_i \in (0, \infty)$, $b_i \in R$, and $c_i \in [0, 1)$ are the discrimination, difficulty, and guessing parameter for item i, respectively. Equation (1) defines the 3PLM. Setting c_i =0 will yield the 2PLM and if

c_i=0 and a_i=1 yield the 1PLM or the Rasch model. Further details about IRT can be found in van der Linden and Hambleton (1997). The books of Lord and Novick (1968) and Lord (1980) are considered classic reference for IRT.

2.1. Interpretation of the Item Parameters

The intercept parameter b_i is directly related to the concept of item difficulty in classical test theory. The higher the value of b_i relative to the ability of an examinee, the lower is the value of $P_i(\theta_j)$. Note that the values of b_i could span the whole range of real number whereas that of the CTT item difficulty, the range is from zero to one. The parameter a_i functions in a similar way to an item discrimination index in classical test theory. The difference between the probabilities of a correct response at any two ability levels increases directly with the value of a_i (Hambleton and Swaminathan, 1985). The parameter c_i is a pseudo guessing parameter. It expresses the probability of a correct response to an item by an examinee who is totally ignorant of the correct answer to an item.

2.2. Parameter Estimation

To date, estimation of examinee and item parameters are reasonably well-understood. Some of these methods includes the conditional maximum likelihood (CML), marginal maximum likelihood (MML), joint maximum likelihood (JML), and Bayesian estimation method. Baker (1992), Hambleton and Swaminathan (1985), van der Linden and Hambleton (1997) are some of the excellent references on this topic. Several computer packages are available that implements these estimation procedures.

3. SIMULATION STUDIES

3.1. Study 1

The aim of this study is to show that the estimates of the item difficulty based on the classical test theory is highly dependent on the sample or group of examinees who took the test. To do this, we considered a test consisting of 40 items and 100 examinees.

3.1.1. Method

The item response pattern of each examinee was simulated using the two-parameter logistic model, assuming that the examinee and the item parameters are known. Two samples of examinees were considered, namely, high-ability and low-ability examinees. The high-ability examinees were drawn from a normal distribution with mean 2 and variance 1, $N(2, 1)$, and the low-ability examinees were drawn from a normal distribution with mean 0 and variance 1, $N(0, 1)$. The item parameters were drawn from the uniform distributions: $a \sim U(0, 2)$, and $b \sim U(-2, 2)$. Appendix A.1 shows the items parameters. The difficulty parameters were sorted in ascending order and are presented in column-wise direction.

The response pattern of examinee j to item i was obtained by drawing a sample from the set $v=\{0,1\}$, where $v=1$ has a probability of being drawn equal to $P_i(\theta_j)$ and $v=0$ has a probability of being drawn equal to $Q_i(\theta_j) = 1 - P_i(\theta_j)$. The CTT item difficulty estimate is usually define as the proportion of examinees who responded correctly to an item. With this

definition, a high value of the estimate implies an easy item whereas a low value implies a difficult item.

3.1.2. Result

Appendix A.1 showed the estimates of CTT item difficulty. The data consistently showed that the estimates differs between the two ability groups. An item looks easy when administered to high-ability examinees while it looks difficult when administered to low-ability examinees. Clearly, the estimates are dependent on the group of examinees who took the test.

3.2. Study 2

The aim of this study is to show that the difficulty of an item seems to have been ignored in the estimation of the examinee ability based on the CTT. It will be shown that this is not the case for the IRT-based ability estimate.

3.2.1. Method

We used a 40-item test with the same item parameters as in Study 1. Note that the item difficulty parameters were sorted from low to high as shown in Appendix A.1. Let the 0/1 (0=correct and 1=incorrect) response patterns of two examinees A and B be as follows:

A = [1110000100100000011001011010111011101110]

B = [1111111111111111111000010000000000000000].

Notice that A and B have the same number of correct response, that is, both got 20 correct responses out of 40 items. Examinee B, however, got the correct answers on easy items whereas examinee A got almost half of his/her correct answers on difficult items.

Using the 2PLM, we estimated the ability of examinees A and B using the maximum likelihood estimate (MLE). The ability estimation using MLE when item parameters are considered known are discussed in van der Linden and Hambleton (1997), Hambleton and Swaminathan (1985), and Baker (1992). Appendix A.2 shows a function in S-Plus (S-PLUS 2000, MathSoft, Inc.) that can compute the MLE of the ability parameter of the 2PLM given the item parameters.

3.2.2. Result

The estimate of the ability of examinee A is .38 and that of examinee B is .007. Clearly, examinee A have higher ability estimate than examinee B as expected.

3.3. Study 3

The aim of this study is to investigate the effect of *guessing* examinees when test scores are used to evaluate school performance and a test score has no direct consequence on the examinee. It is also shown that examinees who are simply guessing randomly in a test can be detected using person-fit analysis. In the following discussion, the guessing examinees

will be sometimes referred to as aberrant examinees and their response patterns are referred to as aberrant response patterns.

3.3.1. Method

In this study, we again considered a 5-option mutiple-choice test consisting of 40 items and 100 examinees. The item parameters were the same as in Study 1. The examinee parameters were drawn from standard normal distribution. The response pattern was simulated similarly as in Study 1.

The effect of guessing on the estimate of a school performance is evaluated by comparing the mean score a group of examinees without the guessers (group X) to another group of examinees with the guessers included (group Y). The statistical test for comparing the means of the two groups will be based on Welch's modified two-sample t-test. In this case, the statistic is

$$t = \frac{(\mu_x - \mu_y) - \mu_0}{\sigma} \tag{2}$$

where,

$$\sigma = \sqrt{\frac{\text{var}(x)}{n_x} + \frac{\text{var}(y)}{n_y}} \tag{3}$$

n_x and n_y are the number of observations in group X and group Y respectively. Assume that X and Y follows a normal distribution. The distribution of t under the null hypothesis, $\mu=0$, can be approximated by a t-distribution with (non-integral) degrees of freedom

$$df = \frac{1}{\frac{c^2}{n_x - 1} + \frac{(1-c)^2}{n_y - 1}} \tag{4}$$

where,

$$c = \frac{\text{var}(x)}{n_x \sigma^2}. \tag{5}$$

Some of the relevant literatures are Hogg and Craig (1970), Mood, Graybill, and Boes (1974), and Snedecor and Cochran (1980). The function *t.test* in S-Plus implements the Welch's modified two-sample t-test.

Simulation of Aberrant Response Pattern. Let κ denotes the number of aberrant examinees and λ denotes the number of items responded by an examinee in an aberrant way. We considered a test with 10% and 20% examinees who are not responding according to their ability. We used random guessing as a response process for the aberrant behavior. For each of the aberrant examinees, 50% and 100% of the total items were generated by guessing. For 100-examinee and 40-item test, κ=(10,20) and λ=(20,40).

An aberrant response was generated by drawing a sample from the set v={0,1}, where v=1 has a probability of being drawn equal to 0.2 and v=0 has a probability of being drawn

equal to 0.8. for all the items. This setup simulates a test situation consisting of a multiple-choice test with 5 options in every item. The probability of a correct response by random guessing (assuming no partial knowledge) is 1 out of 5 or 0.2 and the probability of an incorrect response is 0.8.

Detection of Aberrant Response Patterns. We used the standardized loglikelihood statistic l_z (Drasgow, Levine, & Williams, 1985) to identify aberrant response pattern. The mathematical form of l_z is given as

$$l_z = \frac{\mathrm{W}}{\sigma\sqrt{I}} \tag{6}$$

where,

$$W = \sum_{i=1}^{I}\left[\left(U_{ij} - P_i(\theta_j)\right)^2\left(\log\frac{P_i(\theta_j)}{1-P_i(\theta_j)}\right)\right] \tag{7}$$

$$\sigma^2 = \frac{1}{I}\sum_{i=1}^{I}\left[P_i(\theta_j)\left(1-P_i(\theta_j)\right)\left(\log\frac{P_i(\theta_j)}{1-P_i(\theta_j)}\right)^2\right]. \tag{8}$$

The distribution of l_z is asymptotically standard normal. For aberrant examinees, the value of l_z is expected to deviate from zero. Large deviation from zero can be used to classify an examinee as aberrant or not aberrant. In this study, guessing is expected to reduce the score of an examinee. Hence, the null hypothesis that $l_z = 0$ is tested against the alternative hypothesis that $l_z < 0$. The statistical test is a one-sided test. The critical value of the test is the largest value of z such that $\Pr(l_z \le z) \le \alpha$, and this value is equal to -1.645 for $\alpha = .01$. An examinee is classify as aberrant if the value of l_z is less than -1.645. Equivalently, an examinee is classify as aberrant examinee if the probability of observing a normal variate at most as large as l_z is less than $\alpha = .01$.

Appendix A.3 shows an S-Plus function that computes the standardized loglikelihood person-fit statistics.

van Krimpen-Stoop and Meijer (1999) discussed the null distribution of person-fit statistics for conventional and adaptive test. Detection of misfits in computerized adaptive tests is discussed in van Krimpen-Stoop and Meijer (2000). For detection of misfit due to cheating, see Sotaridona and Meijer (2002, 2003), van der Linden and Sotaridona (2002).

3.3.2. Results

Mean Score Difference. The mean score of non-aberrant examinees was found to be 20.1. This mean score is compared to the mean score of groups of examinees with aberrant examinees as shown in Table 1. For example, when 10% of the examinees ($\kappa = 10$) are guessing on 50% of the items ($\lambda = 20$), their mean score reduced from the baseline mean score of 20.1 to 19.0. Generally, as shown in Table 1, the mean scores of aberrant-group of examinees reduced as the number of guessers increased or the number of items guessed increased. Results also showed that the reduction in the mean scores of aberrant-groups are significant ($\alpha = .01$) except for the group with $\kappa = 10$ and $\lambda = 20$.

Table 1. Mean Score of Aberrant-Group of Examinees and the p-value of the *t*-test

Situation		Mean Sore	p-value
κ=10	λ=20	19.0	.086
	λ=40	17.8	.005
κ=20	λ=20	18.2	.008
	λ=40	16.2	.000

Detection of Aberrant Examinees. The values of l_z statistics for the aberrant groups are shown in Table 2. Under the null, the distribution of l_z is standard normal. The l_z statistics in Table 2 are unusually small (e.g., less than the critical value of −1.645); p-values less than .01, for all the aberrant examinees. This indicates that the l_z statistic have very good power to detect the aberrant examinees.

Table 2. Values of the Standardized Loglikelihood Person-Fit Statistics

10 examinees guessed 20 items (out of 40 items)
-7.5 -6.0 -4.8 -8.1 -8.2 -5.4 -5.7 -4.9 -4.5 -9.9

10 examinees guessed all the items (40 items)
-12.4 -12.5 -11.0 -15.8 -19.0 -17.6 -17.2 -22.3 -22.3 -23.9

20 examinees guessed 20 items (40 items)
-3.6 -5.5 -5.9 -3.7 -5.1 -7.5 -8.5 -8.2 -5.5 -6.3
-4.1 -8.7 -5.9 -8.8 -3.4 -7.8 -5.9 -7.7 -9.3 -11.1

20 examinees guessed all the items (40 items)
-9.0 -12.6 -13.9 -13.2 -15.9 -14.8 -19.2 -21.4 -18.9 -15.3
-8.8 -14.3 -17.6 -14.2 -16.6 -14.8 -14.6 -19.3 -23.7 -24.8

4. DISCUSSION

Advances in psychometric theories lead to development of new test models. The new test models open new possibilities for improving the testing practices. The approach of estimating the examinee ability using the IRT approach yield many advantages as compared to the classical approach. Some of the advantages of using IRT approach in testing have been shown in this paper. For example, if a test is used to evaluate a school and not a student, it is often the case that examinees are not motivated to answer the test seriously. Hence, it is important to identify these examinees and their scores excluded in the analysis. Failure to do

this will often lead to under-estimation of the school performance and as a consequence would yield a misleading conclusion. IRT based person-fit statistic showed some promise in detecting aberrant response pattern. Note that the decision whether or not to exclude an examinee's score in the analysis is a decision that requires some serious thought from the analyst. Person-fit analysis should not be used as the only basis for such decision. Instead, other relevant information should be considered.

It was also shown that the estimate of item characteristic, such as the item difficulty, is less appealing in the CTT framework, for example, the estimate is not invariant to the sample of examinees who took the test. Although we have not shown in this study how the estimate of the item parameters from IRT framework are affected by the sample of examinees, several studies have shown that the item parameter estimates using the IRT are invariant of the examinee population and estimates of examinee parameters are invariant of the test items. See for example, Hambleton and Swaminathan (1985), van der Linden and Hambleton (1997).

ACKNOWLEDGMENT

The authors thank the reviewer for the comments and suggestions which are very helpful in improving the manuscript.

References

BAKER, F. B. (1992). *Item Response Theory: Parameter Estimation Techniques.* NY: Marcel Dekker, Inc.

DRASGOW, F., LEVINE, M. V., & WILLIAMS, E. A. (1985). Appropriateness Measurement with Polychotomous Item Response Models and Standardized Indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.

HAMBLETON, R. K, & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston, MA: Kluwer Academic Publishers.

HOGG, R. V. and CRAIG, A. T. (1970). *Introduction to Mathematical Statistics*, 3rd ed. Toronto, Canada: Macmillan.

LORD, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

LORD, F. M., & NOVICK, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

MOOD, A. M., GRAYBILL, F. A. and BOES, D. C. (1974). *Introduction to the Theory of Statistics*, 3rd ed. New York: McGraw-Hill.

SNEDECOR, G. W. and COCHRAN, W. G. (1980). *Statistical Methods*, 7th ed. Ames, Iowa: Iowa State University Press.