DIFFERENTIAL ITEM FUNCTIONING ANALYSIS, AN ISSUE ON UMPTN ITEM BANKING WITH IRT PROCEDURES

Saifuddin Azwar

National College Admission Tests (Ujian Masuk Perguruan Tinggi Negeri - UMPTN) was first administered to select candidates for five most prominent universities in Indonesia more than fifteen years ago. The system was then improved and has been widely used with the inclusion of more than 10 participating universities across the country.

High school graduates from different areas of the country are eligible to register for the exams. With the advancement of the administration system and the communication network, graduates have access to register and can take the exams in their high school region without having to go to the city where the intended university is located.

That had been good so far until several years ago when apparent differences in performance on UMPTN among high school graduates from different geographical areas were for the first time noticed.

Generally, the tendency showed that graduates from high schools that were located in Java (in-Java students) performed better than those from high schools that were located in other islands (out-Java students). This tendency resulted in smaller proportion of out-Java students admitted to the prominent universities.

If out-Java students failed UMPTN mostly because they had not been as smart as in-Java students, then there would not have been a measurement problem but rather an educational problem. To think of out-Java students as less capable than in-Java students so they do not deserve places in good universities would not be justifiable. It is far more likely that out-Java students have been deprived from environmental and academic conditions conducive to teaching-learning process.

It is realized that there exist academically potential out-Java students. Among small number of out-Java students that are currently attending universities, many of them have been achieving excellently and outperforming in-Java students.

Why then the proportion of out-Java students passing UMPTN has been so small over years is getting attention from government officials, publics, high school teachers, and is becoming concerns of education and measurement specialists.

To many, this problem can be attributed to the conditions of out-Java schools which are believed to be much less satisfactory than in-Java schools are. Among the shortcomings are that national curricullum and syllabi were not properly followed, environmental-related lack of motivation for learning among students, deprivation from modern information media, unstimulating teaching-learning situation, et cetera. Whatever the condition is, seemingly unfairness of UMPTN becomes an intriguing issue. Parents and teachers, especially of out-Java students, are most concerned about UMPTN favoring in-Java students.

Efforts have been done to ensure fair opportunity for out-Java students when they are taking UMPTN. Item banking procedures has been improved, researches have been conducted for calibrating UMPTN items through equating procedures. Trainings for item writers were held intensively. Still another way of improvement needs to be applied, i.e. analysis of differential item functioning.

This analysis will give information on potentially bias items that need to be deleted from the exams. Such information will be very useful for test compilers that they can better select for the test only items that can detect "true" ability of the students regardless of what school group they are from. Unbiased items will lead to more valid test scores interpretation. Valid interpretation will lead to fair decisions. It is very crucial because fairness of the test is the one characteristic we can not afford to lose. Ideally, potentially bias items should be identified first before equating and banking procedures are carried out.

THEORETICAL BASES

IRT Frame Work

Item Response Theory (IRT) assumes that an examinee's probability of answering a given item correctly depends on the examinee's ability or abilities and the characteristics of the item (Hambleton, Swaminathan, and Rogers, 1991).

One of the advantages of IRT model over Classical Test Theory (CTT) is that item characteristics (item parameters) in IRT are not group dependent, i.e. parameters of item are invariant across groups of subjects. This makes way of comparing group ability on a set of items comprising the test.

Estimates of item parameters can be obtained by administering the item to many examinees whose ability levels are known.

Parameters of item are

b = item difficulty index,
a = item discrimination index, and
c = pseudo-guessing probability parameter.

Once estimates of item parameter are obtained, the relationship between ability and probability of answering item correctly can be depicted in a diagram called item characteristics curve (ICC). Because the shape of ICC is determined by item parameters, two items will have identical ICCs if they both have the same parameters.

The appropriateness of ICC is dependent on the appropriateness of the mathematical model for the item of interest. If the model being used fits the data, ICC will give good information on item parameters, probability of correct response at certain ability level, and can be used to detect items that function differently for different group at the same ability level.

It is very important to assess model-data fit in applying analysis of item based on IRT approach. The invariance of item and ability parameter estimates can not be assured if the model does not satisfactorily fit test data set.

Definition of DIF

When different groups of subject with the same ability level do not have the same probability of answering an item correctly, then we have item bias problem.

To distinguish item bias from test bias researchers usually use the term differential item functioning (DIF) to replace the term item bias (Scheuneman & Bleistein, 1989).

Hambleton et.al. stated that an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right (Hambleton, Swaminathan, and Rogers, 1991).

DIF is not the same with test bias which implies "unfairness" in interpretation and use of test results. Jensen (1980) and Reynold (1982) distinguished between fairness in testing as opposed to unbiased tests. According to Shepard, item bias and test bias methodologies are superficially different but should not imply different conceptualizations of bias (Shepard, 1980). This statement does not seem in accordance with the terminology used by Scheuneman and Bleistein (1989).

In this paper, we are going to use the term DIF in the sense that proposed by Scheuneman and Bleistein (1989) pertaining to what actually detected by statistical procedures.

Methods of Detecting DIF

Among methods of detecting DIF that are based on Classical Test Theory are Transformed Item Difficulty (TID) Methods, Item Discrimination Procedures, Contingency-table Approach, Partial Correlation Method, Mantel-Haenszel Procedure, Standardization procedure, and Distractor Analysis. Item Response Theory approaches in detecting DIF among others are 3-Parameter Methods and the use of the Rasch Model.

A popular method used by Angoff(1972,1975) uses transformation of p-values to delta-values which is normal deviates with a mean of 13 and a standard deviation of 4, i.e. delta = 13 + 4x.

Plots of delta-values from different groups of the same ability will form an ellipse along the 45 degree line cross the origin, which represents equal difficulty of the items. If the points of delta plots from equal-ability groups are away from the 45 degree line then the ellipse will be rotated from the line and that can mean DIF presents. Precisely, indices of presence of DIF include (a)the distance of each item from the major axis of the ellipse (b)the standard deviation of this distance, and (c)the difference between the delta-values for the two groups.

Advantages of delta-plot method are that it is simple, inexpensive, easily explained, and does not require large sample. A principle disadvantage is that when two groups differ in their mean ability, an item that is unusually discriminating will result in larger item difficulty differences, whereas an item having particularly low discrimination will show smaller differences than other items on the test, even when the item are not functioning differentially (Scheuneman & Bleistein, 1989).

Scheuneman (1975,1979) suggested contingency table method for analyzing DIF. Based on a definition that if DIF is not present then persons of equal ability have equal probability of a correct response regardless of their group membership, a two-way table of contingency is then established. Once a group Membership x Ability table is established, a C2 index is computed based on correct response in each group being compared as well as the total number of responses at each score-level interval.

A modification of C2 index using full chi-square that includes both the correct and the incorrect responses in the contingency table was proposed by Veale (1977) and then Camilli (1979). This modification was intended to overcome critiques saying that the C2 was not distributed as a chi-square.

Mantel and Haenszel (1959) developed a procedure that had been widely used in biomedical researches which is very closely related to log-linear procedure. This procedur was then adopted and popularized by Holland and Thayer (1986) for DIF analysis. The Mantel-Haenszel (MH) statistic may be interpreted as the average factor by which the likelihood that a member of one group (either focal or reference) answers an item correctly exceeds the corresponding likelihood for a member of the other group. An MH value of 1.00 indicates that a correct response is equally likely for both groups. If

reference group members are more likely to respond correctly then the MH value exceeds 1.00 and if focal group members are more likely to respond correctly, the MH value will be less than 1.00.

Dorans and Kulick (1983) developed an approach that compares empirical item-test regression. This approach was called Standardization method which is primarily a descriptive approach and so provides no significance test. In standardization approach, estimates of the conditional probability of success at each score level are developed on the base group. The base group is usually the larger sample. Two indices of DIF (one signed and one unsigned) use a weighting function supplied by a standardization group. The signed item-discrepancy index is the standardized p difference between focal group and base (reference) group numbers for each item.

A method of detecting DIF using IRT frame work is the Three-Parameter Method. This method uses the item parameters to relate the probability of a correct response to ability. The basic idea in this method is if DIF presents then ICC of an item for different group will not the same. In order to make the two ICCs of both groups comparable, item parameters for each group is estimated separately and then transformed onto a common metric.

Rudner (1977) proposed calibrating items separately for each groups being compared and transformed onto a common scale. The area between the two curve is then approximated by summing the difference between the respective probability of a correct response at small ability increments.

Another IRT based method is the Rasch model (Rasch, 1960) which assumes the discrimination of the item to be constant and the lower asymptote of the ICC to be zero. The difference in the difficulty parameter then becomes indication of DIF.

IMPLEMENTING DIF ANALYSIS ON UMPTN ITEMS

Every year, groups of UMPTN item writers are summoned to discuss domains of content of UMPTN, to review CTT-oriented item analysis results for last year exams and to discuss possible improvements on technical aspects of item writing.

At the time, sets of new items on particular subjects are handed in by the appointed item writers. These items are to be reviewed by a team of item reviewers. Items that are judged to have flaw will either be modified or discarded depending on how serious the flaw is. Items passing the reviewing stages are then collected in a pool of items from which sets of item are drawn according to UMPTN test specifications. Eventually, these items are compiled to be administered in coming years either as scored test items or as field-tested items.

As good items are accumulating and domains of knowledge are getting better defined, the needs of a good item bank is inevitable. An item bank is a collection of good items with certain criteria and specifications. There is no reason for calling any large collection of test questions an item bank if it includes items that don't meet the previously defined criteria and specifications.

Wright and Bell (in Bollwark, 1988) stated that an item bank is a composition of coordinated questions that develop, define, and quantify a common theme and thus provide an operational definition of a variable. This definition implies that not every item can be stored in an item bank.

Items qualified for banking should have undergone some evaluation procedures scrutinizing practical and psychometric characteristics of item. That is they have to have been field tested, empirically examined, and fulfilled certain requirements.

The importance of evaluation of items prior to banking can not be overempashized, because the main purpose of item banking among others are to provide access to items of high quality and to reduce test construction time (Bollwark, 1988).

For tests like UMPTN, which are released to students and every year sets of new tests have to be prepared, an item bank will facilitate the tests compilation.

In relation to the issue of fairness of the UMPTN exams, analysis of DIF should be included as part of item banking procedure to avoid selecting candidates based on irrelevant variables, that is wasting highly potential students due to improper characteristics of items being used in the tests.

In the case of UMPTN, DIF analysis should be conducted for groups of in-Java students (reference group) and out-Java students (focal group). Prior to calibrating items for banking, data of the newly administered exams are collected and tabulated accordingly.

Because UMPTN are scored using guessing formula, i.e. applying penalty for wrong answers, a three-parameter IRT model might not fit the data. Tendency to guess decreases when examinees are told that there will be punishment for incorrect responses. So Rasch model would seem to be more appropriate.

Rasch model (one parameter logistic model) assumes no guessing factor involved, discriminating index be constant, and characteristic of ICC is determined solely by difficulty index of the item.

Mathematical function for the model takes a form of:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1,2, \ldots, n$$

where

$P_i(\theta)$ = the probability that a randomly chosen examinee with ability θ answers item i correctly
θ = ability level
bi = item i difficulty parameter
n = number of item
e = a transcendental number whose value is 2,718

Application of the Rasch model to DIF analysis requires item difficulty parameter estimates for out-Java student and in-Java student groups. If DIF is present there will be differences in item difficulty parameter between the two groups (difficulty shift). This difference can only correctly judged if estimates of the b-parameter are placed on the same scale. A t-satistic is then used to test hypothesis of no DIF.

There is still a possibility that three-parameter model would fit the UMPTN test data. For three-parameter model, the mathematical function takes a form of:

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1,2, \ldots, n$$

where

c_i = pseudo-chance level parameter
D = scaling factor introduced to make the logistic function as close as possible to the normal ogive function
a_i = discrimination parameter which is proportional to the slope of the ICC at the point b_i on the ability scale

If this is the case, then ICC area method can be used for detecting the presence of DIF. ICC area method has advantage over Mantel-Haenszel procedure for it can detect non uniform item bias (Hambleton & Swaminathan, 1985). Whereas with Rasch model the shape of ICC is determined by b-parameter only, in three-parameter model the shape of ICC is determined by b-parameter, a-parameter, and c-parameter.

Comparison between the ICCs of out-Java students and in-Java students is made by calculating the area between the ICCs obtained for each group separately. The area between two ICCs is directly related to the differences in probability of success for the two groups at every ability level and hence is a natural index of bias (Hambleton & Rogers, 1989). A large area indicates that a DIF is present.

Procedure for analysing DIF of UMPTN items would follow Hambleton & Rogers (1989) study. Intervals of ability would be between lower group mean -3 SD and upper group mean +3 SD. Because there is no significance test for null hypothesis of no DIF available, a cutoff values will be obtained by carrying out analysis on two randomly chosen samples of in-Java students. The largest area statistic obtained between ICCs of these equivalent groups is considered to be due to chance factor and so will serve as a cut-off point.

Any item resulting area of ICC difference between in-Java and out-Java students greater than the cut-off point will be flagged as potentially biased. Items not indicating any potential bias will proceed through equating procedure and eventually will be stored in item bank according to domain specification where the items are supposed to be. Those items indicating DIF will be put aside and be examined further to identify characteristics causing DIF.

The IRT area method seems ideal to be implemented for UMPTN item analysis of DIF. The analysis could use data of all the examinees which are more than 10,000 students every year, so the problem of requiring large sample size (Scheuneman & Bleistein, 1989) will not be a concern as long as computer program permits.

The main concern for using three-parameter models is the assumption of model-data fit might not be met, in which case analysis of DIF should be conducted based on Rasch model.

There are at least two factors that are not conducive to implementing IRT-based DIF analysis of UMPTN items. First, the analysis requires complex computer analysis, estimation of parameters (particularly c-parameter) is difficult, and secondly, computer program LOGIST is expensive to run and is not available for the time being in Indonesia.

--ooOoo--

REFERENCES

Angoff, W.H. (1972) A Technique for the Investigation of Cultural Differences. In Scheuneman, J.D. & Bleistein, C.A. (1989) A Consumer's Guide to Statistics for Identifying Differential Item Functioning. *Applied Measurement in Education*, 2(3), 255-275.

Angoff, W.H. (1975) The Investigation of Test Bias in the Absence of an Outside Criterion. In Scheuneman, J.D. & Bleistein, C.A. (1989) A Consumer's Guide to Statistics for Identifying Differential Item Functioning. *Applied Measurement in Education*, 2(3), 255-275.

Bollwark, J. (1988) *Recent Developments and Issues in Item Banking*. (Research Re. No. 185). Amherst, MA: UMASS, Laboratory of Psychometric and Evaluative Research.

Camilli, G. (1979) A Critique of the Chi-Square Method for Assessing Item Bias. In Sceuneman, J.D. & Bleistein, C.A (1989) A Consumer's Guide to Statistics for Identifying Differential Item Functioning. *Applied Measurement in Education*, 2(3), 255-275.

Dorans, N.J. & Kulick, E. (1983) *Assessing Unexpected Differential Item Difficulty of Female Candidates on SAT and STWE Forms Administered in December 1977: An Application of the Standardization Approach*. (Research Re. No. 83-9). Princeton, N.J.: ETS.

Hambleton, R.K. & Rogers, H.J. (1989) Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education*, 2(4), 315-334.

Hambleton, R.K. & Swaminathan, H. (1985) *Item Response Theory: Principles and Application*. Boston: Kluwer.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991) *Fundamentals of Item Response Theory.* Newbury Park: Sage.

Holland, P.W. & Thayer, D.T (1986) Differential Item Performance and the Mantel Haenszel Procedure. In H. Wainer & H.I. Braun (eds.), *Test Validity* (pp.129-145), Hillsdale, N.J.: Lawren Erlbaum Associate.

Jensen, A.R. (1980) Bias in Mental Testing. In Shepard, L.A. (1982) Definition of Bias. in R.A. Berk (ed.) *Handbook of Methods for Detecting Test Bias* (pp. 9-30), Baltimore: John Hopkins University Press.

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Nielson & Hydiche.

Rudner, C.M. (1977) An Evaluation of Select Approaches for Bias Item Identification. In Scheuneman, J.D. & Bleistein, C.A. (1989) A Consumer's Guide to Statistics for Identifying Differential Item Functioning. *Applied Measurement in Education*, 2(3), 255-275.

Scheuneman, J.D. (1975) *A New Method of Assessing Bias in Test Items*. Paper presented at the Meeting of the American Educational Research Association, Washington, D.C.

Scheuneman, J.D. (1979) A Method of Assessing Bias in Test Items. *Journal of Educational Measurement*, 16, 143-152.

Scheuneman, J.D. & Bleistein, C.A. (1989) A Consumer's Guide to Statistics for Identifying Differential Item Functioning. *Applied Measurement in Education*, 2(3), 255-275.

Shepard, L.A. (1982) Definition of Bias. In R.A. Berk (ed.), *Handbook of Methods for Detecting Test Bias* (pp. 9-30), Baltimore: John Hopkins University Press.

Veale, J.R. (1977) A Note on the Use of Chi-Square with "Correct/Incorrect" Data to Detect Culturally Biased Items. In Scheuneman, J.D. & Bleistein, C.A. (1989) A Consumer's Guide to Statistics for Identifying Differential Item Functioning. *Applied Measurement in Education*, 2(3), 255-275.

Saifuddin Azwar
The University of Massachusetts
Amherst, Fall 1991