Detecting Differential Item Functioning (DIF) in Standardized Multiple-Choice Test: An Application of Item Response Theory (IRT) Using Three Parameter Logistic Model

by

Ling Siew Eng
Lau Sie Hoe
Universiti Teknologi MARA Sarawak

ABSTRACT

Multiple-choice tests are the most common format used in standardized test for measuring cognitive ability. In Malaysia, standardized test score were often used to evaluate the education quality, promoting students to a higher grade, recruiting of new staff and for promotion purposes. The presence of Differential Item Functioning (DIF) item in the test may lead to test bias. Therefore, it is very important to avoid bias which may unfairly influence examinees' test score. An item will be detected as DIF item when individuals from different subgroups who have the same ability but different probability of getting the item right. Several methods for identifying DIF have been introduced in the literature. From the comparison studies on methods to identify DIF, IRT techniques are the theoretically soundest procedure. Therefore, the three-parameter logistic model was used to identify DIF items in Mathematics Paper 1 of 'Sijil Pelajaran Malaysia (SPM)' Trail Examination for Sri Aman/Betong Division for the year 2003 across urban and rural students. The study flagged only item 15 as DIF across urban and rural students. The positive area for item 15 indicated that item 15 was in favour of the urban group. The difference between the signed and unsigned area for item 15 shows that item 15 was a non-consistent DIF and was not in favour of urban group over the entire ability range.

Keywords
Differential Item Functioning, Item Response Theory, Three-Parameter Logistic Model Standardized Multiple-choice Test.

INTRODUCTION

From the age that children learn how to read and write, tests play an important role in their lives. Multiple-choice tests are the most common format used in standardized test for measuring cognitive ability (Kurz & Barber, 1999). Standardized test scores are usually used to make decisions about programs and individuals such as evaluating the education quality, promoting students to a higher grade, recruiting new staff and for promotion purposes (Brescia & Fortune, 1988). When these important decisions are made based on the test scores, it is very important to avoid bias test, which may unfairly influence examinees' scores (Hambleton & Swaminathan, 1985). A bias test may be due to the presence of irrelevant, non-target constructs which are related to gender, ethnicity, race, linguistic background, socioeconomic status or handicapping conditions (Flores, 2000; Lam, 1995), differences in upbringing environment, culture (Brescia et al., 1988; Flores, 2000) and daily life experiences (Brescia et al., 1988; Fortune, 1985).

Under Malaysia education system, students irrespective of their cultural, social economic backgrounds, learning environment and upbringing will be taking standardized test in public examination. These test score are used in promoting students to a higher grade and in awarding certificates. Therefore it is very important to ensure that these standardized tests do not contain DIF items, which might cause bias to students from different ethnic, sex, culture, religions or social economic background.

The purpose of this study are to detect items showing DIF in Mathematics Paper 1 of 'Sijil Pelajaran Malaysia' Trail Examination for Sri Aman/Betong Division for the year 2003 across urban and rural students, to identify which subgroup the DIF items are in favour of and to determine whether the DIF items are uniform or non-uniform.

Theoretical Framework

DIF was originally called item bias. According to Pine (1977), a test item is unbiased if all individuals with the same underlying ability have equal probability of getting the item correct, regardless of subgroup membership. The word "bias' is also being used simultaneously by several writers (Linn, Levine, Hastings & Wardrop, 1981; Shepard, Camilli & Averill, 1981; Ironson, 1982; Linn & Drasgow, 1987) for at least two entirely different meanings, statistical and social. Finally, Holland & Wainer (1993) proposed and used DIF to refer to the simple observation that an item display different statistical properties in different group. The definition of DIF accepted by psychometricians is:

> *'…an item shows DIF if individuals from different subgroups who have the same ability but they do not have the same probability of getting the item right.' (Hambleton & Swaminathan, 1991, p. 110)*

In a test construction, discrimination and difficulty parameters work effectively in the process of items selection (Crocker & Algina, 1986). However, the addition of item responses in the item selection can improve the test development decisions (Crocker et al., 1986). To know how an item functions in a test development is an approach of Item Response Theory (IRT).
According to Hambleton and Swaminathan (1985),

> *'Any theory of item responses supposes that, in testing situations, examinee performance on a test can be predicted (or explained) by defining examinee characteristics, referred to as traits, or abilities; estimating scores for examinees on these traits (called 'ability scores') and using the scores to predict or explain item and test performance.' (p. 9)*

Item Characteristic Curve (ICC) is the key concept of IRT. An ICC is a monotonically increasing function which relates the relationship between examinees' item performance, $P_i(\theta)$ and the traits (θ) underlying item performance (Figure 1).

According to Birnbaum (1968) the equation of ICC of the three-parameter IRT logistic model is

$$P_i(\theta) = c_i + (1 - c_i) \quad \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (i = 1, 2, 3, 4....n) \qquad [\text{Equation 1}]$$

Where

$P_i(\theta)$ is the probability that a randomly chosen examinee with ability θ answers item i correctly.
n is the number of items in the test,
e is the transcendental number with a value of 2.718 (correct to three decimals)
θ is the ability level,
D is the scaling factor 1.7 introduced to make the logistic function as close as possible to the normal ogive function.
a_i is the discrimination parameter of item i
b_i is the difficulty parameter of item i and
c_i is the pseudo-chance-level (guessing) of item i.

The four characteristics of IRT models (Hambleton et al., 1985, p.9) are:

1. *It is a model, which supposes that examinees' performance on a test can be predicted (or explained) in terms of one or more characteristics referred to as traits.*
2. *IRT model specifies the relationship between the observable examinee item performance and the traits or abilities assumed to underlie performance on the test.*
3. *A successful IRT model provides a means of estimating scores for examinees on the traits or abilities.*
4. *The traits or abilities must be estimated (or inferred) from observable examinee performance on a set of test items (p.9).*

The assumptions to IRT models (Crocker et al., 1986; Hambleton et al., 1991; Hambleton & Cook, 1977) are:

1. Unidimensionality – it is assumed that the examinee performance in a test can be accounted for by a single latent trait or ability.
2. Local independent – The examinees' responses to any pair of items are statistically independent. This means that an examinee's performance on one item does not affect his or her performance on the other item in the test.. Assumption of local independent is equivalent to assumption of undimensionality.
3. Item Characteristic Curve (ICC) - ICC explains the relationship between the probability of an examinee responding an item correctly, $P_i(\theta)$ and his/her trait or ability θ. How well the ICC explains this relationship depends on how well the chosen model account for the data. When the data fit the chosen model, the probability of a correct response to an item will not depend on the distribution of ability in the group of examinees used to estimate the item parameters.

4. Speededness- Examinee fail to answer the items correctly because of their limited ability, not because they are not given enough time to answer them.

Once the assumptions of IRT models are met and there is a close fit between the chosen response model and the test data set, the following features of IRT will be obtained.

- *Item parameter estimates are independent of the group of the examinees used from the population of examinees for whom the test was designed.*
- *Examinee ability estimate are independent of the particular choice of test items used from the population of items which was calibrated.*
- *Precision of ability estimates is known.*

(Hambleton et al., 1985, P.11)

DIF can be investigated by evaluating the area between the ICCs (Shephard et al., 1981; Shephard, Camilli, & William,1984; Raju, 1988, 1990; Rudner, Getson, & Knight, 1980). The area between the two ICCs shows the degree of DIF (Osterlind, 1983). The larger the area, the greater is the degree of DIF between these two subgroups (Figure 2). The signed areas are computed by using Equation 2 and the unsigned area are computed by using Equation 3

$$\text{Unsigned area} = \int_{-4}^{4} \left[P_{iR}(\theta) - Pi_{F}(\theta)\right] d\theta \qquad \text{[Equation 2]}$$

$$\text{Signed area} = \int_{-4}^{4} \left|P_{iR}(\theta) - P_{iF}(\theta)\right| d\theta \qquad \text{[Equation 3]}$$

where

$P_{iR}(\theta)$ is the probability that a randomly chosen examinee from reference group (R) with ability θ answers item i correctly.

$P_{iF}(\theta)$ is the probability that a randomly chosen examinee from focal (F) group with ability θ answers item i correctly.

According to Camilli and Shepard (1994), DIF can be divided into two broad categories.

1. Uniform or consistent DIF- If the signed and the unsigned area calculated are the same, it shows that the ICCs do not cross each other (Croker et al., 1986). This means that this item is in favour of one subgroup for entire level of abilities (Figure 2).

2. Non-uniform or inconsistent DIF- The signed and unsigned area will only differ when two ICCs cross (Figure 3). The unsigned area will be greater than signed area. This is because the signed area A and signed area B will cancel each other to form a smaller total area. It means that this item does not favour only one subgroup for entrie level of abilities.

METHODOLOGY

The population of the study comprised of all the Form 5 students from secondary schools in Sri Aman and Betong Division for the year 2003. The main criterion used in determining the category was the location of the schools, which was based on the guideline by 'Malaysia Ministry of Education'. Rural schools are schools categorized under 'Sekolah Luar Bandar 2' and 'Sekolah Pedalaman 1' and urban schools are schools categorized under 'Sekolah Bandar' and 'Sekolah Luar Bandar 1'. The study used census.

The instrument used for this study was the Mathematics Paper 1 of the SPM 2003 Trial Examination paper. This paper was administered to all the form five students in the Sri Aman and Betong Division. A mathematics teacher from each school was appointed as research assistants. The research assistants photocopied the students' original answer sheets (excluding the personal particulars). There were 1095 data collected from urban schools and 1117 data collected from rural schools after excluding the absentees and dropout students. One thousand samples were selected randomly using Statistical Package for Social Science (SPSS) for the purpose of comparison (Table 1).

Procedure

The first procedure was to check the model assumptions. This is followed by the estimation of item parameters, scale equating and the computation of the area between two ICCs. The last procedure was to determine the DIF items.

Checking Model Assumption

Factor analysis using the tetrachoric item intercorrelation was performed on the mathematics paper to determine the degree to which the 40 items could be considered unidimensional. When the data meet the assumption of unidimensionality, the data also meet the assumption of local independent (Hambleton et al., 1985)

Chi-square goodness of fit statistics, χ^2 was generated using BILOG-MG 3 (Zimowski, Muraki, & Mislevy, 1996) to determine the fit of the data to the Three-parameter Logistic model. If the χ^2 calculated at the 0.001 level of significance was greater then χ^2 critical at the associate degrees of freedom, the item was not fitted for the model.

The percentage of examinees completing the test, percentage of examinees completing 75% of the test, and the number of items completed by 80% of the examinees was used to check the assumption of speededness. Speed was assumed to be an unimportant factor in the test performance when nearly all examinees completed nearly all of the items.

Estimation of Item Parameters

The 40-item test was calibrated using the BILOG-MG 3 (Zimowski et al., 1996) separately for the urban group, rural group, Group 1 and Group 2 (Table 1). The logistic models with the Marginal Maximum Likelihood (Harwell, Baker & Zwarts, 1998) estimation option were used in each of the calibration to estimate the examinees abilities and item parameters.

Scale Equating

The item parameters and examinees' ability were estimated separately for each group. Hence the ICCs from separate analysis cannot be compared directly. The item parameters estimated for rural group were transformed to the scale underlying the parameters estimate for the urban group. The item parameters estimated for Group 2 were also transformed to the scale underlying the parameters estimate for group 1. This was done by using the test characteristic curve equating method (Stocking & Lord, 1983) as implemented in ST program (Hanson & Zeng, 1995).

Area Measures

The signed and unsigned area between ICCs was calculated using equation 2 and 3.

DIF Indices

The area computed between two ICCs should be zero if DIF is not present. Non-zero area does not indicate the presence of DIF item. According to Camilli et al. (1994), we must be concerned with error in estimating parameters due to sampling fluctuation even when an accurate IRT model is chosen. A cut-off value for the area statistic needs to be found to determine whether DIF is present. The largest area value between ICCs in comparison 2 was used as a cut-off value to determine the presence of DIF item (Hambleton & Rogers, 1989).

FINDINGS

In the principal component analysis, the first unrotated factors yield 10 eigenvalue greater than one with the highest eigenvalue of 14.88 accounting for 37.2% of the total variance. This met the Reckase's (1979) minimum criterion of 20% needed to assure unidimensionality of the data.

From the Chi-square goodness of fit statistics, it was found that item 1, 2, 3 and 6 did not fit the model for urban group while item 1, 2 and 6 did not fit the model for rural group. Item 1, 2, 6 and 10 did not fit the three-parameter logistic model for group 1 while item 1, 2, 3, 6 and 27 did not fit the model for group 2. Hence, items 1, 2 and 6 were excluded in the subsequent analysis for DIF in comparison 1 and 2.

100% of the examinees completed 75% of the test and all the items were completed by 99.3% of the examinees. This indicated that the test was non-speeded.

The largest area between the two ICCs calculated in comparison 2 was 0.5481. This value was used as a cut-off value in determining DIF in comparison 1. Only item 15 had calculated area greater than this cut-off value. The unsigned area for item 15 is 0.6432 while the signed area was 0.4092. This is a non-uniform DIF as the signed and unsigned areas were different (Figure 4).

DISCUSSION AND CONCLUSION

Item 15 was identified as DIF item. Item 15 in the Mathematics Paper 1 of the SPM Trial Examination for Sri Aman/Betong Division for the year 2003 is as follows;

> *P(43^{o}S, 65^{o}T) dan Q ialah dua titik di permukaan bumi dengan keadaan PQ ialah diameter bumi. Cari longitud bagi Q.*
>
> *A.* *43^{o}U*

B. *65^oB*
C. *115^oB*
D. *137^oU*

There are several possible explanations for this. However, further research needs to be done to determine the actual causes of these DIF. Urban students have better access to reference books and tuition classes. They might become more test-wise after being exposed to more tests or exercises while attending tuition classes. Hopkins and Stanley (1981) define testwiseness as “an examinee’s ability to use the characteristics and formats of the test and /or the test-taking situation to increase his/her score (p.141). Brescia et al. (1988) say that a person becomes test-wise if he or she is exposed to more tests of the same kind.

The difficulty parameter b for the urban and the rural group are 1.5257 and 2.3672 respectively. Since the b parameters for the rural group are greater than the urban group, this indicates that item 15 is more difficult for the rural group. According to Lam (1995), inadequate or undue assistance from parents, peers and teachers and lack of resources inside and outside of schools are among the possible sources of bias. Urban students might have more advantages in learning this topic through technology-aided facilities or through attending tuition classes. Therefore, urban students will have the extra edge when they are asked to apply their knowledge on item 15.

A difficult item may trigger guesswork. According to Camilli et al., (1994), multiple-choice tests with fewer options per item have a larger c parameter. The guessing parameters (c) for the urban and the rural group are 0.3165 and 0.3423 respectively. High guessing also contributes to the area between two ICCs which will determine the presence of DIF in this item.

CONCLUSION

The results show that DIF item exist in mathematics examination paper. This finding is consistent with that of Teresi (2000), Collins, Raju and Edwards (2000), Takala and Kaftandjieva (2000), Maller (2001), and Jones & Gallo (2002) where some of the items were detected as DIF. The presence of DIF may unfairly influence examinee’s score (Hambleton et al, 1985). Therefore, the ministry of education should take effort to ensure that test items in standardized tests are free of DIF across factors such as rural-urban, gender, ethnic, culture and religion background. This is to ensure that no segment in our society will be unfairly panelized when taking standardized tests.

BIBLIOGRAFI

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp.404-405). Reading, MA: Addison-Wesley

Brescia, W., Fortune, J.C., (1988) Standardized Testing of American Indian Students. Eric Clearinghouse on Rural Education and Small Schools, Las Cruces, N. Mex. Retrieved July 31, 2004, from http://www.enc.org/topics/equity/articles/document.shtm?=ACQ-111498-1498

Camili, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items.* Nesbury. CA: Sage.

Collins, W.C., Raju, N.S., & Edward, J.E.(2000). Assessing Differential Functioning in a satisfaction Scale. *Journal of Applied Psychology*. Vol.85, Issue 3, P.451.

Crocker, L., & Algina, J. (1986). An investigation of item bias. *Educational and Psychological Measurement*, 28, 61-75.

Flores,G.S. (2000); *What is Cultural Validity in Assessment*? Retrieved July 13, 2003, from http://www.edgateway.net/cs/cvap/print/docs/cvap/news.htm.

Fortune, J.C. (1985) *Choctaw Comprehensive School Study*. Philadelphia, MS: Choctaw Heritage Press,

Hambleton, R.K., & Cook, L.L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 14, 75-96.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principle and applications.* Boston, MA: Kluwer. Nijhoff.

Hambleton, R..K., & Swaminathan, H. (1991). *Fundamentals of item response theory*. SAGE Publications, Inc. Newbury Park, California.

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), 313-334.

Hanson, B., & Zeng, L. (1995). *A computer program for IRT scale Transformation*. Iowa; American college Testing.

Harwell, M.R., Bakar, F., & Zwarts,M. (1998). Item parameter estimation via marginal maximum likelihood and EM algorithm: A didactic. *Journal of Educational Statistics*, 13, 243-271

Holland, P.W., & Wainer, H. (1993). *Differential Item Functioning* (ed.) Hillsdale, N.J: Lawrance Erbaum Associate.

Hopkin, K.D., & Stanley, J.C. (1981) Extraneous factors that influence performance on cognitive tests. In G.V. Glass (Ed.), Educational and Psychological measurement and evaluation (6th edition. pp141-157) New Jersey: Prentiss-Hall.

Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), *Handbook of Methods for Detecting Test Bias*. Baltimore: Johns Hopkins University Press.

Jones, R.N., & Gallo, J.J. (2002). Education and Sex Differences in the Mini-Mental State Examination: Effects of Differential Item Functioning. Journal of Gerontology Series B: Psychological Sciences & Social Sciences. Vol. 57B, Issue 6, p.548.

Kurz. & Barber, T., (1999) A review of Scoring Alogorithms for Multiple Choice Tests. Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, Tx, Jan 21-23, 1999)

Lam,T.C.M ,(1995) Fairness in Performance Assessment, Eric Clearinghouse on Counseling and Student Service Greensboro NC. Retrieved July 31, 2004, from http://www.ericfacility.net/ericdigest/ed391982.html

Linn, R. L., Levine, Hasting & Wardrop.(1981) An investigation of item bias in a test of reading comprehension. *Applied Psychological measurement*, 5, 159-173,

Linn, R.L., & Drasgow, F. (1987) Implication of the Gorden Rule settlement for test construction. *Educational Measurement : Issue and Practice*, 6, 13-17.

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

Maller, S.J. (2001). Differential Item Functioning in the WISC-III: Item Parameters for Boys and Girls in the National Standardization Sample. *Educational & Psychological Measurement.* Vol.61, Issue 5, P.793

Osterlind, S. J.(1983). *Test item bias*. Newbury Park , CA: Sage.

Pine, S.M. (1977) .Applications of item response theory to the problem of test bias, in D.J. Weiss (ed.) *Applications of Computerized Adaptive Testing*. Research Report 77-1. Minneapolis: University of Minnesota, Psychometric Methods Program, Department of Psychology.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, 14(2),197-207,

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Ststistics*, 4, 207-230.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement*, 17, 1-10.

Rudner, L. M., Getson, P. R., & Knight, D.L. (1980). *Journal of Educational Statistics*, 5, 213-233.

Shepard, L. A., Camili, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Measurement*, 6, 317-375

Shepard, L. A., Camili, G., & williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Measurement*, 9, 93-128.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6, 317-375.

Stocking, M. L.,& Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210

Takala, S. & Kaftandjieva, F. (2000). Test fairness : a DIF analysis of an L2 Vocabulary test. *Language Testing*, Vol.17. Issue 3, p.323.

Teresi, J.A. (2000). Application of Item Response Theory to the Examination of Psychometric Properties and Differential Item Functioning of the Comprehensive Assessment and Referral Evaluation Dementia Diagnostic. *Research on Aging*; Vol.22 Issue 6, p.738

Zimowski, M.F., Muraki, E., & Mislevy, R.J. (1996).BILOG-MG: *Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International, Inc.

APPENDIX

Table 1: Study samples for the four comparisons.

Comparison	Number of random samples	
	Reference Group, R (urban)	Focal Group, F (rural)
1: Urban-rural	1000	1000
2: Group1-group2	1000	1000

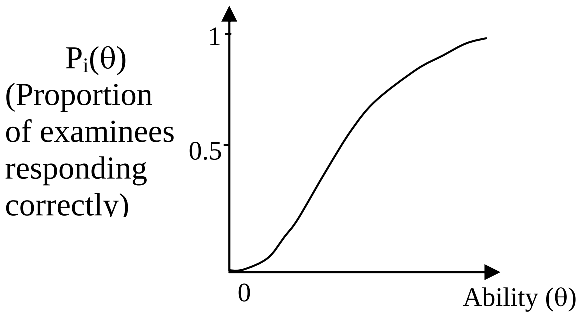

Figure 1: Item characteristic curve

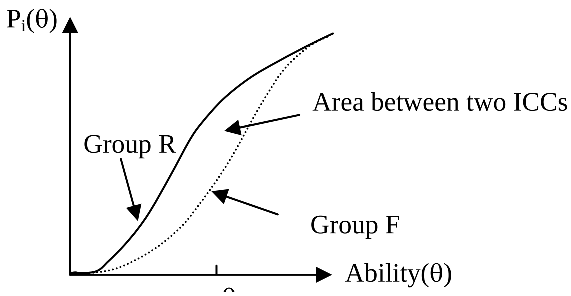

Figure 2: An area exist between two ICCs

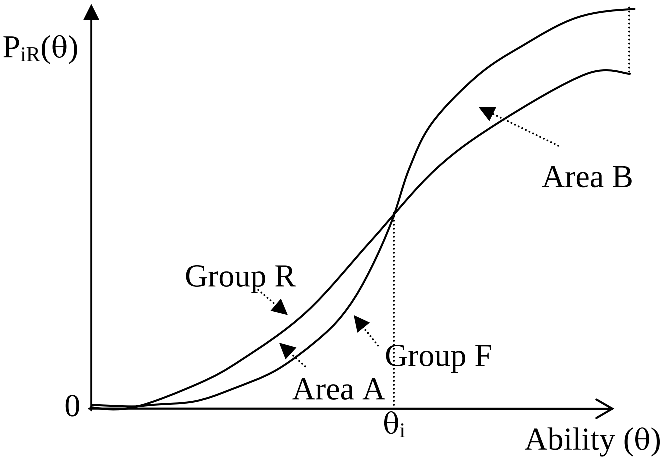

Figure 3: Two ICCs cross

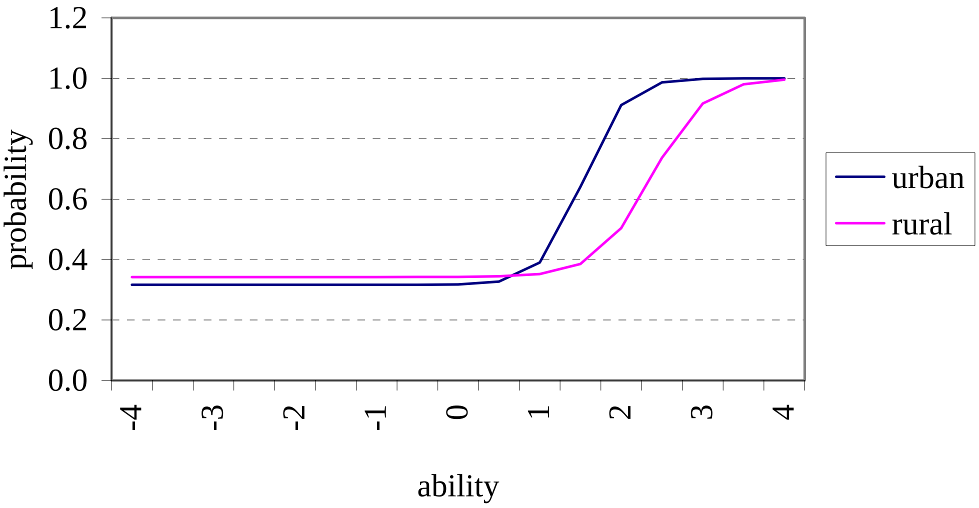

Figure 4: ICCs for Item 15