Title: Assessing literacy on a global scale: Assessment challenges from the Literacy Assessment and Monitoring Programme.

Author: Brenda Siok-Hoon Tay-Lim

Affiliation: UNESCO Institute for Statistics

Email: b.tay-lim@unesco.org

Abstract:

Based on the main survey field observations in Mongolia, it was suspected that some cognitive items show signs of bias for some subgroups, contrary to the expectation that the developed items are neutral. Due to the LAMP two-stage assessment design, some of the traditional DIF procedures that are commonly used in other large-scale assessment surveys are not suitable here. The Raja's DIF index is applied on this dataset. Further research is needed to explore other viable methods for this two-stage test design. Of the 7 Prose and 2 Numeracy items identified in the field observation, not all items suspected to be biased against the focal group are actually statistically biased. This may be explained as the observations in the field are few (at most 50 cases) while the statistical analysis is based on the general pattern of more than 4000 cases In the next step, it may be of interest to subset the group further to look at how certain items function differently and explore why the differences exist. Coupled with their answer to the background questionnaire, we may understand if this is due to bigger environmental factors like schooling. This exploration may help us understand what extraneous factors affect literacy skills.

Key words: Literacy Assessment, Item Response Theory, Differential Item Functioning

Introduction

The importance of obtaining literacy data, using them to formulate policies, and monitoring their effects is increasingly recognized by both national and international agencies. The Literacy Assessment and Monitoring Programme (LAMP) was initiated in 2003 by the UNESCO Institute for Statistics as a pilot project to explore the implementation of large-scale literacy assessment survey in developing countries. It was built on the International Adult Literacy Survey (IALS) as a methodological basis, with items from IALS and Adult Literacy and Life Skill Survey (ALL). However LAMP has further developed and improved after field testing the instruments and operational procedures in selected number of countries. The main survey of this project has been completed in 2011 in 4 culturally diverse countries – Jordan, Mongolia, Occupied Palestinian Territories, and Paraguay.

LAMP tests literacy in three domains: reading of continuous texts (prose); reading of non-continuous texts (document); and numeracy skills (UNESCO-UIS, 2009). In LAMP design the respondents aged 15 and above are randomly selected from the household listing and are given a background questionnaire and then a Filter booklet consists of Prose, Document, Numeracy (PDN) items. Based on his/her performance in this Filter booklet, the respondent is assigned either module A or B. Module A consists of two tests, first the Locator, which also consists of PDN items like the Filter booklet, and a second Reading Component test that consists of items testing pre-reading skills (recognizing alphanumeric, visual word recognition, word meaning, short sentence processing, and passage reading). Module B consists of two booklets, Booklet B1 and Booklet B2, but the respondent is only assigned one of the two booklets. There are 19 common items between Booklet B1 and Booklet B2. The flow of the interview is presented in Figure 1.

Insert figure 1 about here

There are 77 PDN items (24 Prose, 29 Document, 24 Numeracy) distributed across the 4 booklets (Filter, Module A Locator, and Module B Booklets 1 and 2). Out of the 77 items there are 36 items developed by the pilot countries and 41 items taken from the previous international literacy surveys IALS and ALL. The distribution of items is presented in Table 1.

Insert table 1 about here

Objectives

Based on the field observation of both field tests and main survey it was suspected that some PDN items showed signs of bias for some subgroups, contrary to the expectation that the developed items are neutral. The 'potential' bias items will need to be examined and tested to assure that the items that make up the test is valid for all examinees. A list of items that may have 'face validity' issue is presented in Table 2. There are 7 Prose, 7 Document, and 2 Numeracy items that are isolated for further analyses from the Mongolia main survey[1].

[1] The list of items is identified by an ethnographer, Dr Bryan Maddox of University of East Anglis, from the main survey data collection conducted in Ulaanbataar and Gobi desert in October 2010.

Insert table 2 about here

Mean performance differences between the two groups do not necessary indicate that there is differential item functioning (DIF) (Thissen, Steinberg, & Gerrard). An item is considered to have differential item functioning (DIF) if two individuals who belong to different group membership but with the same ability perform differently on the item. In other words, DIF exists when the probability of getting the item correct is different between the two respondents belonging to different group membership. Traditionally, in order to examine DIF two groups of examinees are first matched on the construct measure by the test and then the performance of each item is compared between the two groups (Dorans and Holland, 1993). Those items that show significant statistical DIF are identified for further examination by the test developer. DIF is normally done during the field test to identify bias items to be excluded before using them for the main survey data collection. DIF is also done prior to scaling to flag bias items needed to be deleted from calibration in the Item Response Theory analysis.

In this study, DIF is used in a non-traditional way. DIF is used in the post-hoc analysis. Items that are suspected to perform differently between respondents belonging to different group membership are first identified; then each item is tested to see if it has DIF. The analysis on differential item functioning will be performed on the PDN items only, as these items are usually presented in “context” which may cause differential performance. Since each item is suspected to DIF in different group composition, each item in table 2 will be grouped differently for the performance of DIF. For example, for the parking item, the comparison group is respondents by gender.

There are two types of DIF, uniform and non-uniform DIF. Uniform DIF exists when there is no interaction between ability level and group membership, i.e., the probability of answering an item correctly is greater for one group consistently across the ability level. Non-uniform DIF exists when there is an interaction between ability level and group membership, i.e., the probability of answering an item correctly is not consistent between the two groups across the ability level. The index used to identify DIF should apply to both uniform and non-uniform DIF.

There are three commonly used DIF approaches: contingency table - Mantel-Haenszel (MH), regression model - logistic regression, and Item Response Theory (IRT). MH is attractive as it is easy to implement and has an associated test of significance. However it is not designed to identify non-uniform DIF. Logistic regression method is more powerful than MH procedure in identify non-uniform DIF and is as powerful as MH procedure in identifying the uniform DIF (Swaminathan, and Rogers, 1990). In logistics regression group membership and the ability level (i.e., the total test score) are usually the predictor variables used in the model. In LAMP the respondents were assigned different booklets only respondents who are assigned same set of books will have same maximum total score. For example, a respondent who is assigned Module A will have a maximum total score of 35 since there are 35 dichotomously scored items; a respondent who is assigned Module B Booklet 1 will have a maximum total score of 47. Since maximum total score is different for different respondent the logistic regression procedure may not be a suitable for LAMP. On the other hand, IRT-DIF which “integrate out” the ability may be a more appropriate alternative since it does not rely on the ability level as an input.

Methodology

One of the major advantages of IRT over the classical test theory is that the item response function (IRF) is invariant over subgroups of respondents (Hambleton, et. al., 1991), provided that the assumptions (e.g., unidimensionality and model fit) are checked. Therefore this invariant property makes IRT a good choice for the analysis of DIF. (Oshima & Morris, 2008)

The IRT-based DIF analysis is performed separately by domain. The IRT-DIF methods use the same approach as Mantel-Haenszel (MH) and logistic regression. It first separates the respondents into two groups (focal=female and reference=male) based on the respondents' background characteristics. Each group is calibrated to obtain a set of item parameters. The two sets of item parameters give two Item Response Functions (IRFs). Before comparing the two IRFs, the item parameters are put on the same scale. The IRT approach focus on determining the area between the two IRFs based on item parameters of each of the two groups. Unlike the contingency table or regression method, the IRT approach does not have to match the groups on the observed total score. It assumes that the ability distribution has been "integrated out" before computing the area between the two IRFs across the distribution of continuum ability.

The proposed DIF statistic is developed by Nambury Raju (1995) and it is called the area-based method or simply the area difference between two IRFs. This index is based on the 2-parameter (2PL) IRT model. The 2PL is chosen as all PDN items are dichotomously scored. In addition the item biserials are not similar across items. The 2PL model is represented by the following equation:

$$P_i(x_i = 1 \mid \theta, a_i, b_i) = \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}$$

where x_i = the response to item *i*, 1 if correct and 0 if not,
a_i = the discrimination parameter of item *i*, and
b_i = the threshold parameter of item *i*.

PARSCALE (du Toit, 2003) is used to calibrate the item parameters and the item parameters are used to compute the DIF index presents below. It estimates the item threshold and item discrimination parameters separately for each of the two groups. The different groups assume different empirical posterior distributions with means, μ_g, and standard deviation, σ_g, and the posterior distributions are not necessary normal. Separate prior distribution is used for each group, and the prior distribution is updated after each estimation cycle of the posterior distribution from the previous cycle. In order to see if there is DIF in the item, the two groups were put on common scale. The program constraints the overall difficulty level of the set of common items given to both focal and reference groups to be the same, the item difficulty parameters for the focal group are then adjusted accordingly (Muraki & Engelhard, 1989). An individual item will be flagged as having DIF if the index, which is the area between the two IRFs across the distribution of continuum ability, is greater than 0.006 and the χ^2_{NF} with N_F *df* is statistical significant (Raju, et. al., 1995). The formula of the DIF index and the χ^2_{NF} are presented below:

$$DIF_i = \int_{-\infty}^{\infty} | P_{iF}(\theta) - P_{iR}(\theta) |^2 \; f_F(\theta) d\theta$$

where DIF_i = differential item functioning between two IRFs for item i,
$P_{iF}(\theta)$ = the focal group probability function for item *i*,
$P_{iR}(\theta)$ = the reference group probability function for item *i*. and
$f_F(\theta)$ = the focal group distribution function.

$$\chi^2_{N_F} = \frac{N_F(DIF_i)}{\sigma^2_{d_i}}$$

$$d_i(\theta) = P_{iF}(\theta) - P_{iR}(\theta)$$

where $\chi^2_{N_F}$ = chi-square significant test for item *i*,

N_F = focal group sample size,

DIF_i = differential item functioning between two IRFs for item *i*,

$\sigma^2_{d_i}$ = variance of the difference in the probability of a correct response for item *i*,

$d_i(\theta)$ = the difference in the probability of a correct response for item *i*,

$P_{iF}(\theta)$ = the focal group probability function for item *i*, and

$P_{iR}(\theta)$ = the reference group probability function for item *i*.

Results

Based on the initial data received from Mongolia, a preliminary analysis was performed[2]. All items have been calibrated, the DIF index and the chi-square statistics for items identified in table 2 are computed for the following reference/focal groups: male/female, and age 15-24/25+. Of the 7 Prose items identified by the ethnographer that was suspected to favor the reference groups (male or age 15-24 group) only one item (Book1 item 5 – Medical dosage) is flagged as having DIF in the age comparison. Given same ability, the probability of getting the item correct is higher for the 15-24 age group (than the 25+) in the low ability range but lower in the high ability range. In other words, the item functions non-uniformly between the 15-24 and 25+ age groups across the ability continuum. That is the item favors age 15-24 group in the low ability range but favor the age 25+ in the high ability range. The plot of two groups' IRFs is presented in Figure 2.

Insert figure 2 about here

Of the 2 Numeracy items (F010 – Gas gauge and BL17 – Parking time) identified by the ethnographer, both are flagged as functioning differently in the male versus female comparison but not in the 15-24 versus 25+ comparison. This may mean that estimation and computation items favor male than female.

Discussion

Based on the main survey field observation in Mongolia it was suspected that some cognitive items showed signs of bias for some subgroups, contrary to the expectation that the developed items are neutral. Furthermore, due to the LAMP two-stage assessment design, some

[2] A preliminary analysis was performed on the initial data. Final analysis will be performed after receiving the final weights from the country and more in-depth analyses will be conducted. The details of the analyses will be presented in the Technical Report to be published in 2012 by the UNESCO Institute for Statistics.

of the traditional DIF procedures and methodologies that are commonly used in other large-scale assessment surveys are not suitable here. The Raja's DIF index is applied on this dataset. Further research is needed to explore other viable methods that can be applied in this two-stage test design. Due to time constraints, I have only looked at the Prose and Numeracy items. Of the 7 Prose, and 2 Numeracy items identified in the field observation, not all items suspected to be biased against the focal group are actually statistically biased. This may be explained as the observations in the field are few (at most 50 cases) while the analysis is based on the general pattern of more than 4000 cases In the next step, it may be of interest to subset the group further to look at how certain items function differently, e.g., between males living in urban and males in living rural areas and explore why the differences exist. Based on their answers to the background questionnaire, we may understand if this is due to bigger environmental factors like schooling or work experience. This exploration may help us understand what extraneous factors affect literacy skills.

References

Dorans, N.J. & Holland, P.W. (1992). DIF detection and description: Mantel-Haenszel and standardization. (RR-92-10). Princeton, NJ: Educational Testing Service.

Muraki, E. & Engelhard, G. (1989). Examining differential item functioning with BIMAIN. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Oshima, T.C. & Morris, S. B. (2008). Raju's differential functioning of items and tests. *National Council on Measurement in Education: Instructional Topics on Educational Measurement Series*, 43-50.

Raju, N.S., van der Linden, W., & Fleer, P.F. (1995). IRT-based internal measure of differential functioning of items and tests, *Applied Psychological Measurement,* 19(4), 353-368.

Swaminathan, H. & Rogers, H. (1990). Detecting differential item functioning using logistic regression procedures, *Journal of Educational Measurement,* 27(4), 361-370.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Item bias. *Psychological Bulletin*, 99, 118-128.

Tables and figures

Figure 1. Administration procedure of the two-stage test

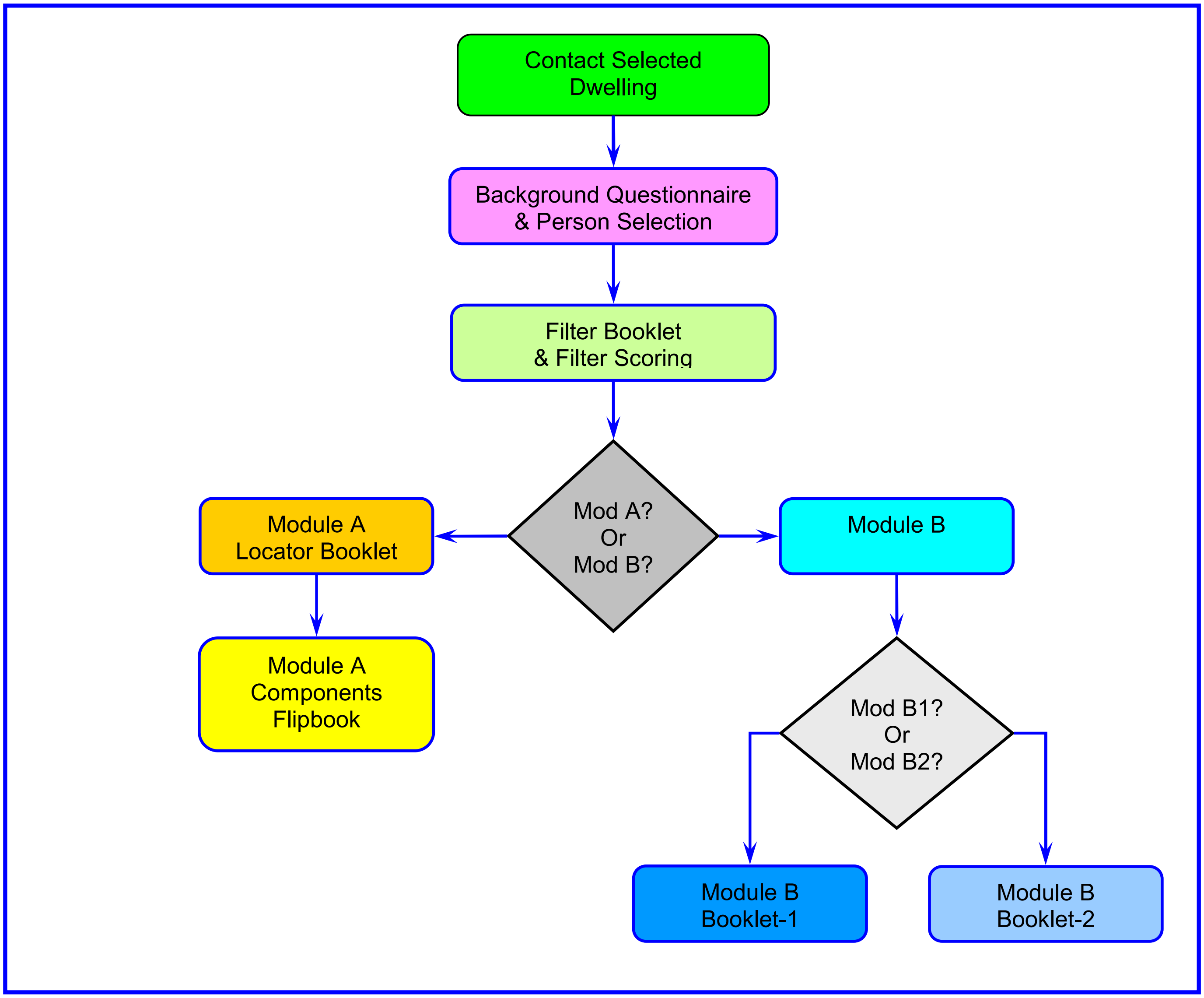
Contact Selected Dwelling
Background Questionnaire & Person Selection
Filter Booklet & Filter Scoring
Mod A? Or Mod B?
Module A Locator Booklet
Module B
Module A Components Flipbook
Mod B1? Or Mod B2?
Module B Booklet-1
Module B Booklet-2

Figure 2. Prose item B105 - Medical dosage
Reference group=15-24, Focal group=25+

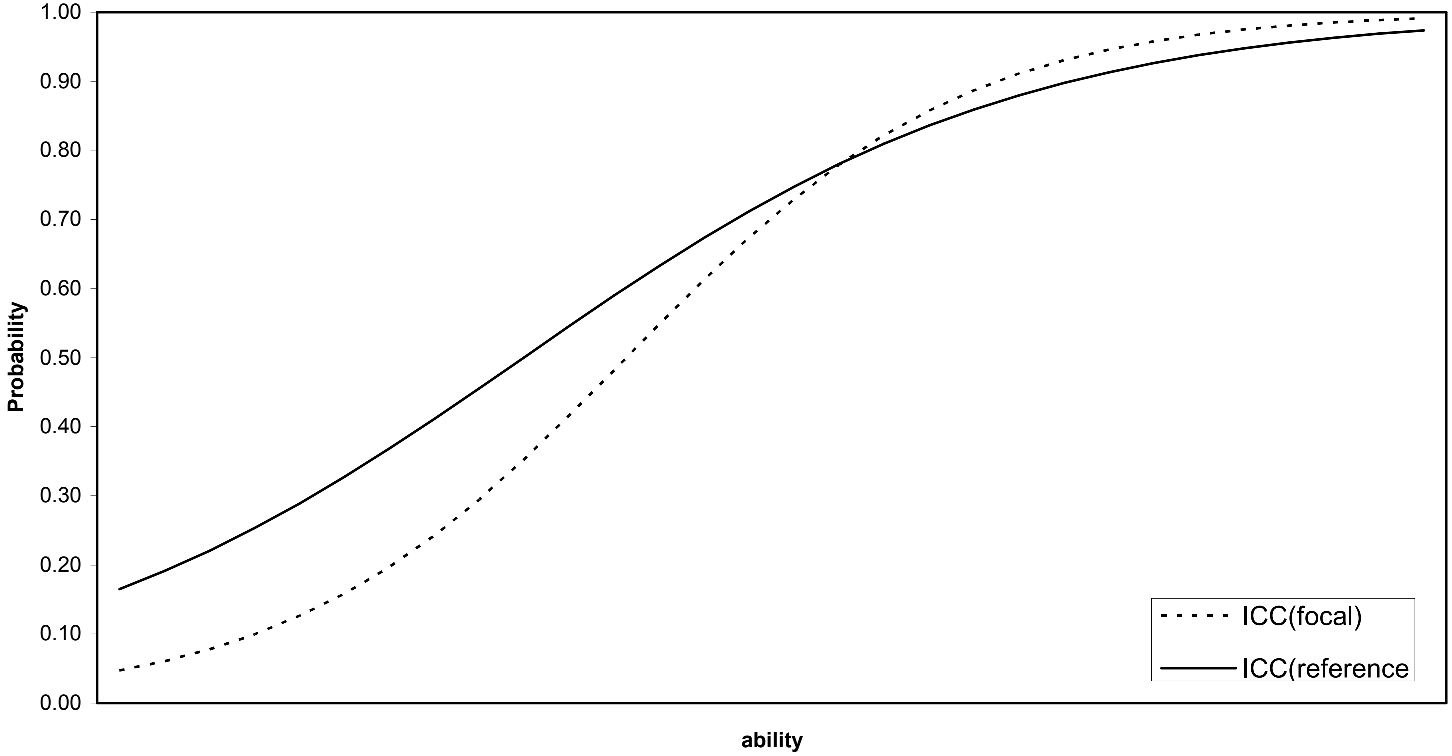

Table 1. Distribution of item

	LAMP Common	IALS/ALL	Book Total
Filter	5	12	17
Module A Locator	6	12	18
Module B Booklet 1	13	17*	30
Module B Booklet 2	12	17*	29
Total by group	36	41	77

Note:
For Module B Booklets 1 and 2 there are 19 common IALS/ALL items.

Table 2. Selected items for further analyses

	Domain	Item ID	Item description
1	Prose	F016	Mozart letter
2	Prose	B105	Medical instruction – Dosage
3	Prose	B106	Medical instruction – Side effects
4	Prose	B107	Medical instruction – Manufacturer
5	Prose	B203	Camel – Drink litres
6	Prose	B204	Camel – Countries
7	Prose	B205	Camel – Winter
8	Document	AL03	Hotel menu – Least expensive
9	Document	AL04	Hotel menu – Rice with chicken
10	Document	AL05	Hotel menu – All beverages
11	Document	BL03	Employment – Hours
12	Document	BL04	Employment – Days
13	Document	BL05	Employment – Heard about job
14	Document	BL06	Employment – Distance
15	Numeracy	F010	Gas gauge
16	Numeracy	BL17	Parking time