Herald Journal of Education and General Studies
Volume 3 (1), Pp. 001- 012 March, 2017

ISSN Online: (ISSN. 2350 – 2177)

Full Length Research Paper

The comparison of item and person item response theory (IRT) parameter estimates for the anchor-items and common-persons designs

Monamodi Kesamang Ph.D.

Corresponding author Email address: kesamangm@yahoo.com

Received 03rd October 2016 **Accepted 8th January 2017**

Abstract

An assessment system should be able to identify the potential of each learner and the quality of education is dependent on providing valid score and grades to every examinee. This empirical study investigated the use of Item Response Theory (IRT) test-linking techniques (anchor-items and common-person designs) as a method of maintaining equivalent standards across years. IRT parameter estimates are assumed to be invariant. The accuracy with which each method was able to estimate item parameter estimates was also investigated. Below are the study questions, How do the item parameter estimates for the two methods compare?. How do the Item Response Functions (IRFs) and the Test Response Functions (TRFs) compare for the two designs? How do the Item Information Functions (IIFs) and the Test Information Functions (TIFs) compare for the two designs? How does the reliability and Standard Error of Measurement (SEM) compare for the two designs? The study made the following findings: The Pearson correlation coefficients for the item parameters are significantly different. IRFs and TRFs are significantly different for the two linking designs with the anchor-item test IRFs/TRFs approximating the theoretical Item characteristic curve (ICC). IIFs and TIFs for the anchor-item test provide more information. The anchor-item test is more reliable.

Keywords: Equating, linking, anchor items, common persons, IRT

INTRODUCTION

In IRT the examinee's score is not only dependent on the total examinee's score, but also on the statistical characteristics of items scored correctly or incorrectly (Weiss and Yoes, 1991). Classical Test Theory (CTT) relies on reliability based on the true score. The true score cannot be directly measured, but can only be estimated from the observed score. The coefficient of reliability is also examinee dependent. Standard error of measurement (SEM) is calculated from this reliability coefficient. These problems with CTT make it extremely difficult for standard setter to maintain equivalent standards across the years. If a national examination cannot ascertain with certainty, that the grades awarded to candidates' year on year are equivalent or have elicited the same mental processes to reach the target grade, then its reliability and validity are questionable and compromised. CTT scoring method is dependent on the test takers (examinees) or the type of items in the test. How would one therefore know with certainty that a grade A awarded in one year is of equivalent standard with a grade A awarded the next year in the same subjects. This is because the forms used in the two years may not have been parallel or equivalent. How would the standard setting group know that a cut-off score in one year was set such that it separate students of different ability levels and can be comfortably be compared to any cut-off score in that examinations for all the other years. These are really challenging questions that need to be addressed if one can speak of maintaining equivalent standards year-on-year. Fairness in psychological measurement has always been a challenge to test developers and psychometricians for decades. Michaelides (2006), on the comparability of standards states that;

"Testing programs repeatedly administer different versions of the same test, and it is a matter of fairness to individual examinees to preserve comparability of the different test forms. Equating methods provide the statistical adjustments for placing scores on a common scale. Common scales are also useful for tracking trends in group performance over time, indicating how examinee cohorts perform compared to their counterparts. The significance of measuring trends accurately has recently received much attention. Under the No Child Left Behind Act of 2001 in America, educational institutions are expected to demonstrate adequate yearly progress in reading and mathematics. There are rewards and sanctions attached to their progress. Equating is an essential part of linking achievement scores across years and maintaining a common longitudinal scale. To implement the legislation, and measure the magnitude of gains or declines from one year to the next properly, the equating must be accurate" (p 61)

On setting cut-scores, Haertel (2004) states that 'students are admitted to colleges based on their test scores, and the meaning of a given scale score in one year should be the same as for the previous year. Agencies set scale-score cut points defining passing levels for professional certification, and fairness requires that these standards be held constant over time'. There is no doubt that this emphasis on the maintenance of standards by setting equivalent passing scores is a very important assessment issue and national assessment or examination Boards are bound by the spirit of fairness that equivalent standards are maintained.

There have been some dissenting voices in Botswana about the state of the education system. These include politicians, educators, the media, industry, trade unions and some governments departments. A Mmegi editorial on the state on the education and assessment standards of the Botswana education system, state that;

'An examination usually ascribes symbols to a child's performance. This does not however tell us much other than just to rank students in performance categories. We do not know what the student's competencies and capabilities are' (Mmegi, 2007).

The author argues that, this symbol ascribed, for example, Grade A at Junior Certificate (JC), are actually meaningless as they do not tell whether a child is capable of doing something. Is the child able to write a communicative letter, or sew a skirt or construct a wooden candle stand? All what the parent is concerned about is the child finds a place in the next level. What the Ministry of Education is concerned about is that, it releases results on time, not necessarily what these results mean. Parents do not have any idea what the results mean. All what everybody does is to make assumptions. In fact, we do not even make assumptions, all we do really is engage in the expanse journey of emotions, joy and tribulations. We are either excited that our child passed with Grade A or go berserk because our boy could only get a Grade D when the playful boy next door excelled with a Grade B. Faint voices of some 'malcontent' mutter their dissatisfaction over the declining standards but nobody seems to care. The minister cannot even dignify these murmurings with a response. This has been the culture and it shall be. But we beg to differ and we are asking you to hear us loud and clear. What is the responsibility of the Minister of Education if not to improve the quality of education? We are not asking you to revolt but to demand answers from you, a public servant. It is no longer a matter of speculation that JC results are falling mainly in quality. The minister should tell the nation what he intends to do to improve the already deplorable education standards. Some are even demanding a national symposium on the viability of automatic promotion from primary to secondary education. To inject more quality into our education system, we must be talking about the size of our classes. This is all we ask for. This certainly cannot be an arm and a leg (Mmegi, 2007).

Botswana political parties such as the Botswana Congress Party also lament the state of the education system when it states that "The standard of education in our country has deteriorated at all levels. The education system has failed to produce skilled individuals able to develop and harness new technologies for our country. As testimony to the failure of the education system, the BDP (Botswana Democratic Party) leadership escapes the declining standards by sending their children to study for both primary and secondary abroad. The elite have long taken flight of the moribund public education into private education. Thus, the poor education system that the BDP leadership is presiding over is not for them but for others – the poor", (BCP 2009). The Botswana Federation of Trade Unions (BFTU) and Botswana Institute of Development and Policy Analysis (BIDPA), are also concerned about the education system in Botswana which they view as academically oriented and does not encourage diversity of career opportunity but tend to emphasise academic achievement. This type of education which is not market based has been blamed for increased unemployment, [Ministry of Labour and Home Affairs,(2004), BIDPA (2006), BFTU (2006)].

Item Response Theory

IRT yields invariant item and latent trait estimates and offers a common trait scale for measurement, this makes it highly useful in facilitating detection of item bias or Differential Item Functioning (DIF), adaptive testing and test equating Nenty (2004). Nenty (2001) stresses that assessment information plays a critical role in formulation and implementation of education policy through informed decision making. These decisions are on equity and access as espoused by Jomtein UNESCO Conference of 1990, Education For All (EFA). The validity of this assessment information influences the quality of decisions made by governments (Nenty 2001). Challenges to the validity of assessment information in Africa hinges on the following factors;

- The validity of what is to be measured
- The faulty basis for developing and administering assessment instruments
- The validity in scoring
- The lack of objectivity in setting of standards, (Nenty, 2001).

An item elicits some form of behaviour from an examinee, the examinee brings in some cognitive ability (θ) to overcome an item. On the other hand, an item has some inherent cognitive resistance (b). The probability of an examinee overcoming an item is dependent on these two factors (θ - b_i). If the examinee cognitive ability is higher than the item's cognitive resistance, then the examinee's probability to overcome the item is higher. It therefore follows that a good measurement requires that;

- a person with high ability has a better chance of overcoming an item than a person with lower ability.
- any person has a better chance of overcoming a less demanding item than a more demanding one.
- these conditions can only be the consequence of the person's and the item's position on the trait under measurement (Wright, Mead and Draba,1976).

Purpose of Study

The purpose of this comparative study is to investigate the linking of test forms by the use of two designs, that is, the use of anchor-items and common-persons. The item/person parameter estimates for the two linking designs are compared to determine (i) how different are their item parameter estimates, (ii) how different are their item and test response functions, (iii) how different are their item and test information functions and (iv) how reliable are their item parameter estimates- or how much error does each linking design introduce.

Research Questions

The study intends to answer some of the questions posed below,

1. How do the item parameter estimates for the anchor-items method of testing linking and the common-persons method compare?
2. How does the Item Response Functions (IRFs) and the Test Response Functions (TRFs) for the two linking methods compare across the θ-scale?
3. How does the Item Information Functions (IIFs) and Test Information Functions (TIFs) for the two linking methods compare across the θ-scale?
4. How does the reliability and Standard Error of Measurement (SEM) compare for the two linking designs across the θ-scale?

Test Equating by the use of Anchor-Items and Common-Persons

Anchor-Items

Yu and Osborn Popp (2005) offered a practical technique of equating test by the use of anchor-items and common-subjects. Although this was not an empirical study, it offers some insights on how one has to deal with different approaches to test equating. Two IRT programs for test equating are discussed at length, namely BILOG-MG (multiple group) and WINSTEPS. BILOG offers options of calibrating items with the use of the 1-PL, 2-PL and the 3-PL models, while WINSTEPS is a Rasch model. Preparation of data for analysis, input and output files are discussed. The paper offers a practical guide to test equating using (i) alternate form equating (where anchor-items are analysed at the same time with the rest of the items), (ii) across sample equating (where sets of test items are separately based on fixed item parameters of calibrated anchor-items) and (iii) common-subject equating, (same examinees taking different tests assessing the same construct). The present study will adopt the alternate form equating method. Real data sets were used in the two equating methods.

Data from "Mathematical competence test" was used for this analysis. This dataset was collected in a large metropolitan school district. Figure 1 shows the setup for the alternate form equating. There are eighteen common items for all the forms. Each form had 46 new items. The common items are placed in the data set such that they come before the new data set for all the three tests. Form ID column identifies each of the three tests which come one after the other.

Test Equating- Common-Persons

The separate test could be analysed separately to estimate examinee scores and then the scores with their 95% confidence band based on standard errors

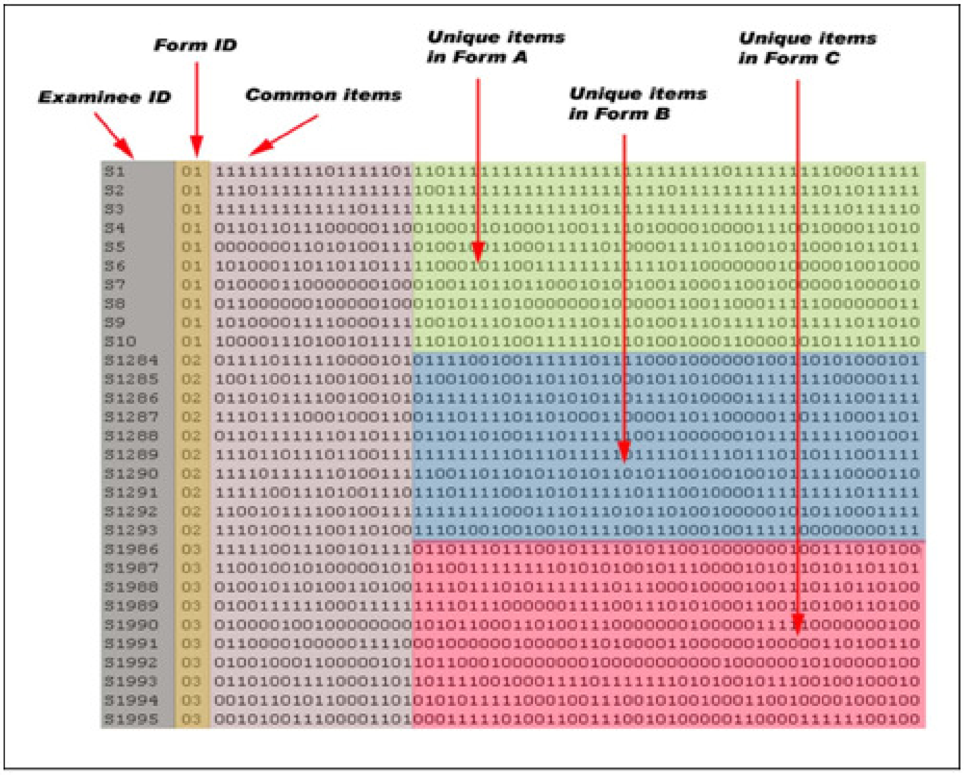

Figure 1: Data setup for alternate form equating in BILOG-MG. (Yu & Osborn Popp, 2005)

are plotted to evaluate to what extent the two tests the same construct within reasonable degree of measurement errors. Yu and Osborn Popp (2005) conclude by admitting that there is no '..single best test equating methodology' and that different contexts call for different approaches', p16. Anchor-item equating may be useful in that fewer resources are used as the two tests are administered once.

METHODOLOGY

The study intends to compare two linking designs, these being, the anchor-item (anchor-item design) and the common-person (single-group design) designs. The accuracy with which each method is able to estimate item/person parameter estimates is also investigated. The study employs a quantitative analysis in the estimation of IRT item/person parameter estimates for the two linked tests. The item statistics and parameter estimates for all the three tests are compared to establish whether there exists a significant difference (correlation) among them. IRT 1-PL, 2-PL and 3-PL IRT models were employed to calibrate item parameter estimates. The analysis is intended to establish the comparability between all the three test forms. The programs that were used are PARSCALE (1-PL, 2-PL and 3-PL models) and XCALIBRE (2-PL and 3-PL models).

Sampling and Sampling Procedure

The population of the study is all the Form Three examinees that were about to sit for their final JCE examinations in 2008. A representative sample of Form Three JCE students is given pilot forms (tests) by the Botswana Exanimations Council (BEC) in the preparation for subsequent examinations. The study instruments were administered at this time to the non-pilot schools. The sampling for the study was stratified according to school size. Four schools were randomly sampled from each region for this study according to the strata. At least two sampled schools from the four schools from each region were large schools. The data collectors were instructed to administer the study instruments to at least three of these four sampled schools, but fortunately the data collectors managed to administer to all the four sampled schools.

The study sample consisted of about 3000 Form Three students. There were three instruments and each of the sub-samples of about 1000 students were

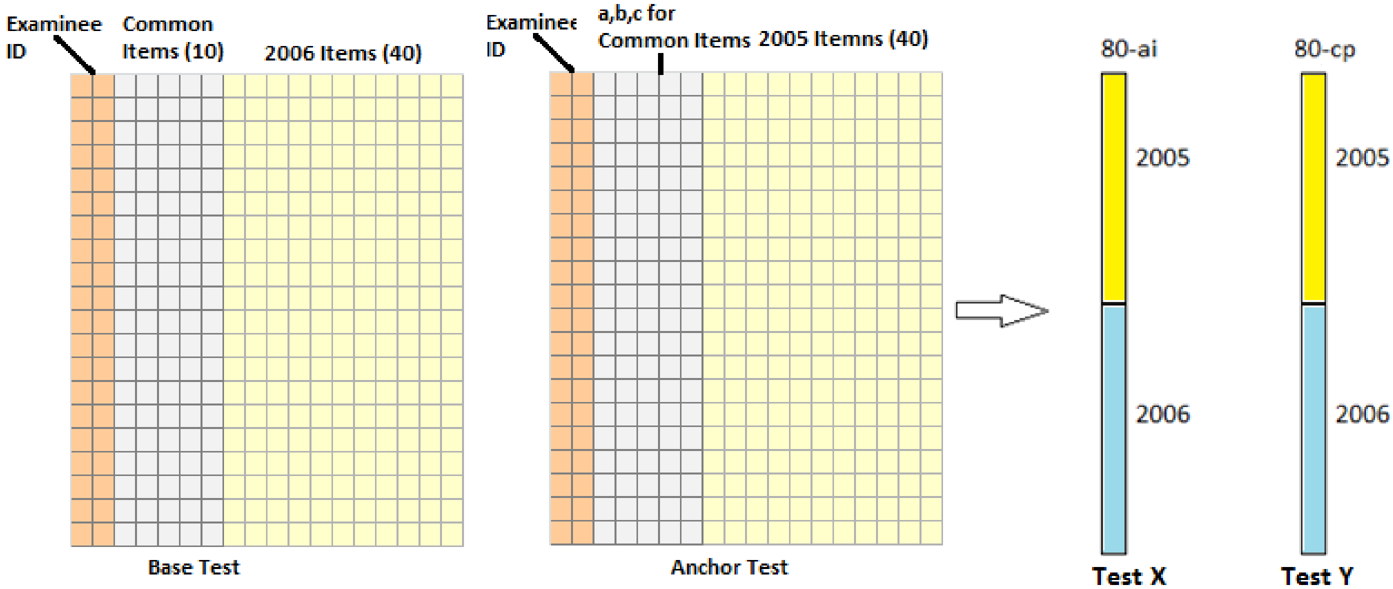

Figure 2: The combined 80-item test (X) for the anchor-item and the common-person test (Y).

administered one of the three instruments. The three sub-samples sat for one of these instruments; (a)

Instrument 1, 2005 Science paper with ten (10) anchor-items, (b) Instrument 2, 2006 Science paper with ten (10) anchor-items, and (c) Instrument 3, an 80 item test consisting of the 2005 and 2006 papers (common-persons). A sample size of 1000 examinees for each instrument is deemed sufficiently large for most IRT models. Most researchers on IRT such as Lord (1969), Warm (1978) and Fan (1998) are agreeable that this sample size gives stable IRT estimates.

Instrumentation

The study intended to compare the 2005 and the 2006 objective science papers after they have been linked by (i) anchor-items and (ii) common-persons. These are (i) a 50 item test (2005), (ii) a 50 item test (2006), for both of these, 10 of the items are linking items, and (iii) an 80 item test, (2005 and 2006 tests). The ten (10) linking items were from the 2004 examinations. These items were selected for their good CTT item statistics. The last instrument was an 80-item test, these was made of the two tests combined sat by the same examinees (common persons)

Data

The three tests were answered on the optical mark reader forms (OMRs). The data was analysed by IRT program, XCALIBRE (Assessment Systems Corporation, 1995) and PARSCALE (Scientific Software International, 1992). These programs give both the CTT and IRT item statistics and person parameter estimates. The ten linking items in this file had to be formatted such that the system recognizes that they are common to the two tests. The 2006 test was used as a base test (fixed based procedure), and the 2005 test was anchored on it by the fixed parameters from the ten anchor-items. The common-person 80 item test was already linked by common examinees. The data was prepared as shown in Figure 2. The internal anchor test procedure was used in which the anchor items and the unique items were administered together as a single test to each group.

The two tests/examinations were then combined into a one 80-item test known as the anchor-item (ai) test (Test X). The a, b and c item parameters for Test X were then were then compared to the 80-item common-person (cp-Test Y) test a, b and c item parameters.

Data analysis and interpretation of results in this study compared the 2005 and 2006 examinations without linkage and when the two examinations had been linked.

Cohen and Kim (1998) found that for small number of anchor items, separate calibration could be preferred, but for a large number of anchor items, both separate and concordant calibrations give similar results. Beguin and Hanson (2002) found that when compared to the fixed parameter (for anchor items) and separate calibration, concurrent calibration is favoured by a larger sample size for the parameter estimation for the anchor items.

Table 1: The Summary Results of the Three Instruments

NO# of Items	Paper	Sample size
50	2005 + 10 anchor-items	**1386**
50	2006 + 10 anchor-items	**1287**
80	2005 + 2006 (common-person)	**1803**

Analysis

The three tests were equated on a common scale by the use of IRT item or test equating programmes. The item parameters from the three tests were linked either by anchor-items or common-persons, compared to check whether there exists a significant difference or correlation between their statistics. XCALIBRE and PARSCALE statistics were used to estimate item/person parameters. These programs are iterative, that is, the initial estimates output are used as the input into the second cycle and the second estimates become the input into the third cycle and so on. The residuals or error of estimation after every iterative cycle decreases until the summed absolute parameter change is less than 0.05 or until a certain number of iterations set have run. IRT estimation programs runs through EM (Expectation-Maximisation) loops (cycles) until either all items have converged to within a specified tolerance level or when the specified number of loops is reached, XCALIBRE manual (1996). If the parameter estimation of a given item fails to converge, the program will indicate that such an item be deleted from the analysis. Such an item should be excluded for the next run. The IRT item parameter estimates were carried out using the Marginal Maximum-Likelihood Estimation (MMLE) program. Person parameter are estimated by the Maximum-Likelihood Estimation (MLE) program. This option is a default in the XCALIBRE program for Windows Version 1.10. PARSCALE comes with the Bayesian MAP and EAP (*Maximum a posteriori*) and (*Expectation a posteriori*) scoring options.

RESULTS

A comparison of the anchor-item and the common-person linking designs was conducted. Item/test response functions and item/test information functions for the two linking designs were investigated. Pearson correlation analysis and group means would be used for interpretation of results.

IRT Model Assumptions

Item response theory imposes some restrictions on the type of data used for analysis. The data used for IRT models should (i) fit the type of IRT model used, (ii) data must adhere to the assumption of unidimetionality, that is, the data set must be a measure of only a single trait, (iii) items in a test must be uncorrelated and (iv) satisfies sampling adequacy and sphericity. The instruments for the study have been summarised in Table 1.

The test of Unidimentionality

(a) Principal Component Analysis (PCA)
Confirmatory factor analysis and principal component analysis (PCA) method were used to test for unidimentionality. These methods would determine whether the data set measures one or more constructs. For all the three instruments, over 43% and above of the variance was explained. These analyses indicate that the instruments were measuring a single construct, therefore satisfying unidimentionality test.
(b) Scree Plots
Scree plots provide a convenient way of visualising a dominant factor in principal component analysis. A steep slope for the first factor is an indication of unidimentionality. All the three scree plots showed this characteristic steep slope.

KMO and Bartlett's Tests of Sampling Adequacy and Sphericity

The data sets meet all the assumptions of PCA method. Kaiser-Meyer-Olkin (KMO) measured sampling adequacy with a value of over 0.90 (which is greater than 0.6) for all the three tests. Bartlett's test of sphericity gives a significant value in all the tests. These two tests confirm factorability of the correlation matrix. The results of these two tests are shown in Table 2.

Model Data Fit-Chi-Square Statistic

A Chi-square test was run on the data sets to establish

Table 2: The Summary Results of the KMO and Bartlett's Test for the Five Tests

Kaiser-Meyer-Olkin Measure of Sampling Adequacy	**2005** .921	**2006** .931	**2005 + 10** .940	**2006 + 10** .944	**80 (common)** .952
Bartlett's Test of Sphericity Approx. Chi-Square	6132.299	6847.069	9727.024	10144.034	23012.359
df	780	780	1225	1225	3160
Sig.	.000	.000	.000	.000	.000

Table 3: *The Pearson Correlation Coefficient Statistics for Test X and Test Y for the 2-PL Parscale Model*

Variable		a_{ai}	b_{ai}
a_{80cp}	Pearson Correlation	0.715*	
	Sig. (2-tailed)	.000	
	N	80	
b_{80cp}	Pearson Correlation		0.908*
	Sig. (2-tailed)		.000
	N		80

**: correlation is significant at 0.01 level (2-tailed), cp: commo-person,ai: anchor-item*

whether the test data/items fit the 1 PL, 2-PL and the 3-PL IRT Parscale Model. Chi-square test for Model-data Fit for the two instruments indicates that the instruments fit the 1-PL, 2-PL and the 3-PL Parscale IRT model. Only six items out of the forty items do not fit this IRT model for the 2005 data set and five items for the 2006 data set do not fit this model. All the other instruments had a poor fit. For the 2005 and 2006 + 10 items, this could be that some items had been fixed and for the 80 item test, it could be due to a large sample size. Chi-square test is sensitive to sample size.

Comparability of IRT Item and Person Parameter Estimates for the Anchor-item and the Common-person Linking Designs

H_1:1 There is significant correlation between the item parameter estimates for the anchor-item linking design with the common-person linking design.

H_1:2 There is significant correlation between the person parameter estimates for the common-person linking design and the anchor-item linking design.

The linking designs used are as per Figure 1.7. The 2006 test + 10 anchor-items were calibrated together to obtain parameter estimates for the 40 test items and the 10 anchor-items in 1-PL, 2-PL and 3-PL Parscale IRT model. The 2005 test was then calibrated next now with the item parameter estimates for 10 items fixed in the command syntax as anchor-items. The resulting parameter estimates for the two tests, Test X were then correlated to the parameter estimates from Test Y. The parameter estimates are the item slope parameter, (a-parameter), the item location parameter, (b-parameter) and the pseudo-guessing parameter, (c-parameter).

Comparison of the a-parameter: 2-PL Model

Table 3 shows the Pearson Correlation coefficient statistics for the anchor-item test and the common-person tests for the 2-PL Parscale model. A correlation analysis ran on the two data set yielded a Pearson correlation coefficient of r_{XY} = 0.715 for the a-parameter. This is high correlation coefficient and is significant at the 0.05 significant level. This indicates that the two data sets are positivity related.

Comparison of the b-parameter

A correlation analysis ran on the two data set yielded a Pearson correlation coefficient of r_{XY} = 0.908 for the b-parameter. This is high correlation coefficient and is significant at the 0.05 significant level.

Comparison of the a-parameter: 3-PL Model

A correlation analysis run on the two data set yielded a

Table 4: *The Pearson Correlation Coefficient Statistics, 3-PL model*

Variable		a_{ai}	b_{ai}	c_{ai}
a_{80cp}	Pearson Correlation	-0.103		
	Sig. (2-tailed)	.366		
	N	80		
b_{80cp}	Pearson Correlation		0.798*	
	Sig. (2-tailed)		.000	
	N		80	
c_{80cp}	Pearson Correlation			0.624*
	Sig. (2-tailed)			.000
	N			80

**: correlation is significant at 0.01 level (2-tailed), cp: common-person, ai: anchor-item*

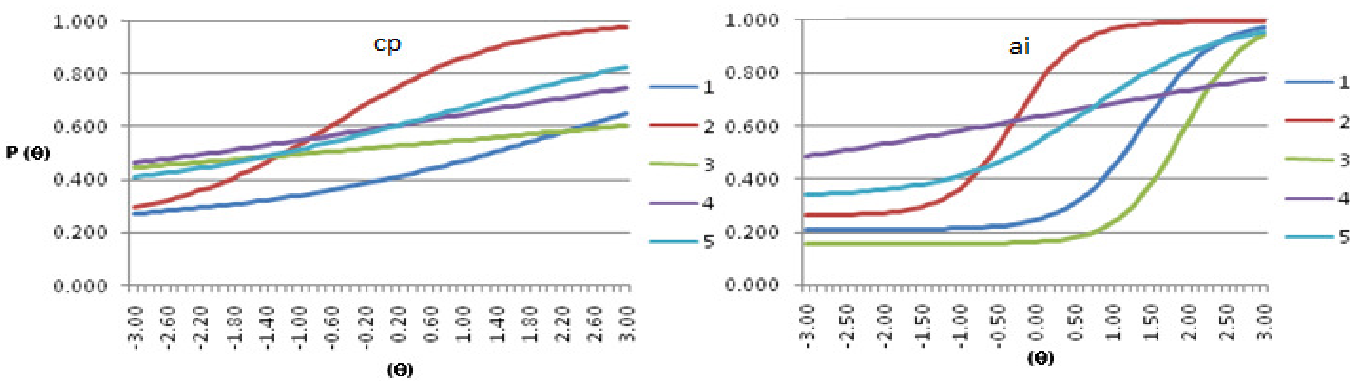

Figure 3: Item Response Functions for the two Linking Designs-Common-Person (cp) and Anchor-Item (ai).

Pearson correlation coefficient of r_{XY} = -0.103 for the slope parameter. This is a very low correlation coefficient, indicating that the two data sets are uncorrelated. This correlation is not significant. Table 4 shows the Pearson correlation coefficient statistics for the anchor-Item and the common-person item test.

Comparison of the b-parameter

A correlation analysis was run on the two data sets and yielded a Pearson correlation coefficient of r_{XY} = 0.798 for the b-parameter. This is a relatively high correlation coefficient and is significant at the 0.05 significant level. This indicates that the two data sets are positivity related, a high b-parameter for the common-person data set is associated with a high value for the anchor-item data set.

Comparison of the c-parameter

A correlation analysis run on the two data set yielded a Pearson correlation coefficient of r_{XY} = 0.624 for the c-parameter. This is high correlation coefficient and is significant at the 0.05 significant level. This indicates that the two data sets are positivity related, a high c-parameter for the common-person data set is associated with a high value for the anchor-item data set.

Comparability of the IRF/TRF for the Anchor-item and the Common-person Linking Designs

H_0:3 There is no significant correlation between the Item Response Functions (IRFs) of the two tests across the θ-scale.

H_0:4 There is no significant correlation between the Test Response Functions (TRFs) of the two tests across the θ-scale

Figure 3 shows the Item Response Functions from the common-person (cp) and anchor-item (ai) linking methods for the first five items. The IRFs are very different with the anchor-item method giving more reliable IRFs compared to the common-person method. IRFs for the common-person method changes slightly monotonically across the theta scale except item 2 which seems to be the best item of the five items. Item 4 seems not to be affected by the linking as it behaves

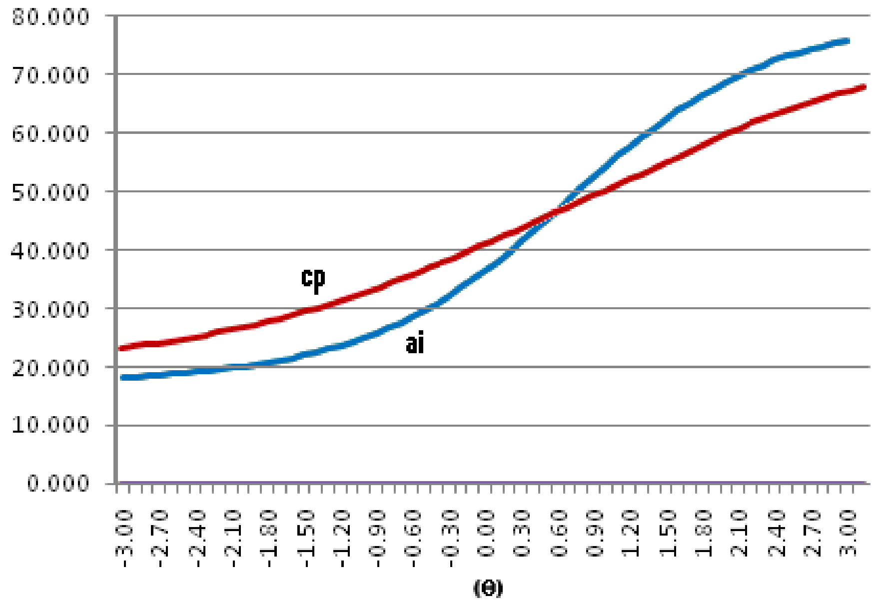

Figure 4: Test Response Functions for the two Linking Designs

the same across the two linking designs.

Figure 4 shows the Test Response Functions for the two tests. These are the probability of an examinee with a given theta responding correctly to items in the test or the number of items out of the 80 that an examinee at any theta level is expected to get correct. The TRFs for the two tests indicate that for low levels of theta, more items will be gotten correct from the common-person linking design and for high levels of theta examinees will get more items correct from the anchor-item linking design. The two tests behave inversely at each extreme values of theta. These could be an indication that cp-test has more SEM compared to ai-test at the extreme theta values. The ai-test is more discriminating.

Comparability of the IIF/TIF for the Anchor-item and the Common-person Linking Designs

H_0:5 There is no significant correlation between the Item Information Functions (IIFs) of the two tests across the θ-scale.

H_0:6 There is no significant correlation between the Test Information Functions (TIFs) of the two tests across the θ-scale.

The Item Information Functions-IIFs for the first five items from the Common-Person (cp) and Anchor-Item (ai) linking methods are shown in Figure 5. The items from the cp method provide far less information compared to the same items when they have been linked by the use of anchor-items. Item 2 just like for the IRF looks a better item as it provides sufficient information in the two linking designs. Item 4 provide no information in both linking designs, therefore this item should not have been in the test as it has poor measurement characteristics. The two graphs were placed side-side for easier comparison. Items 1 and 3 behave quite differently between the two linking designs.

The Test Information Functions-TIFs for the entire test (80 items) from the common-person (cp) and anchor-item (ai) linking methods are shown in Figure 6. The two graphs show that the anchor-item design provides significantly more information than the common-person design. Most of the information is also provided where the two graphs are steepest, and this is in the middle of the graphs.

Comparison of the Test Reliability and Standard Error of Measurement (SEM) for the Anchor-item and the Common-person Linking Designs.

H_0:7 There is no significant correlation between the Standard Error of Measurement (SEM) for the anchor-item and the common-person linking designs.

The Standard Error of Measurement (SEM) indicates the precision of a measuring instrument or tool. In this case the tool is the individual items or the test. A larger SEM is observed where there in less information provided by an item or test. An item that introduces less SEM contributes positively to the total test reliability. Examinees with lower theta level tend to provide a large SEM than examinees with high theta

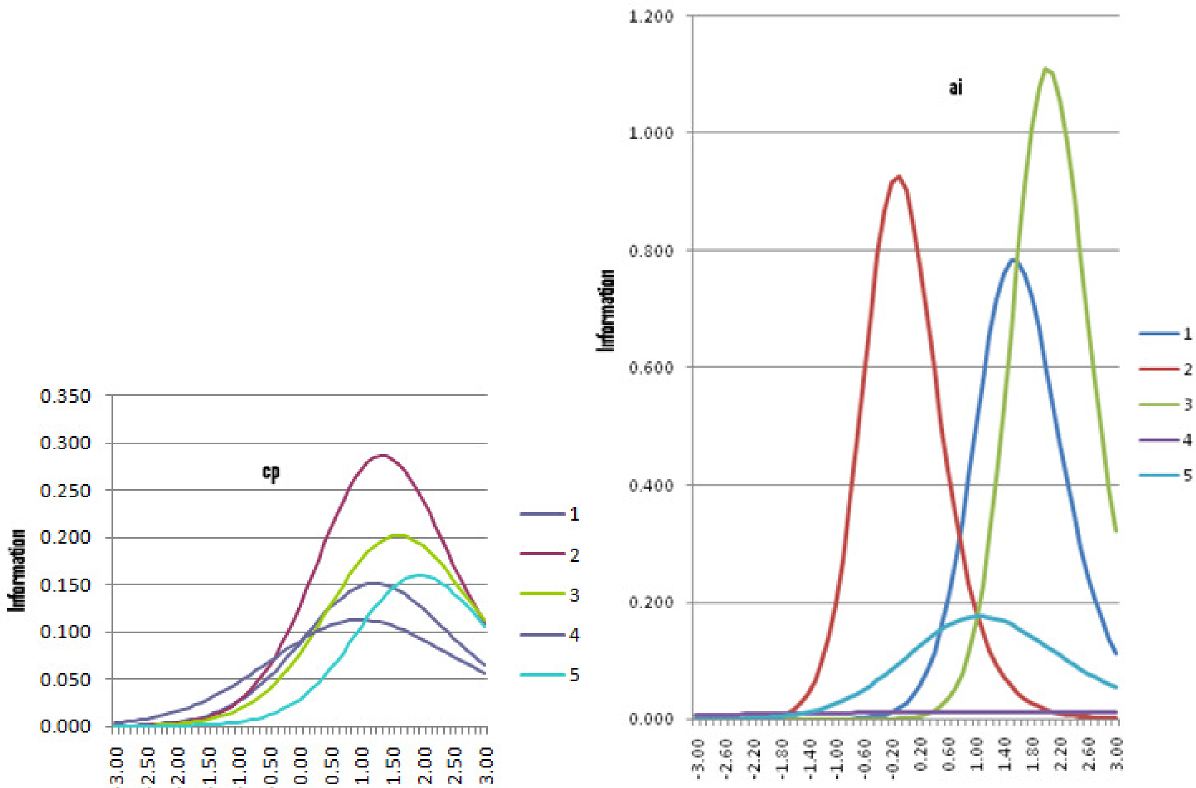

Figure 5: Item information functions

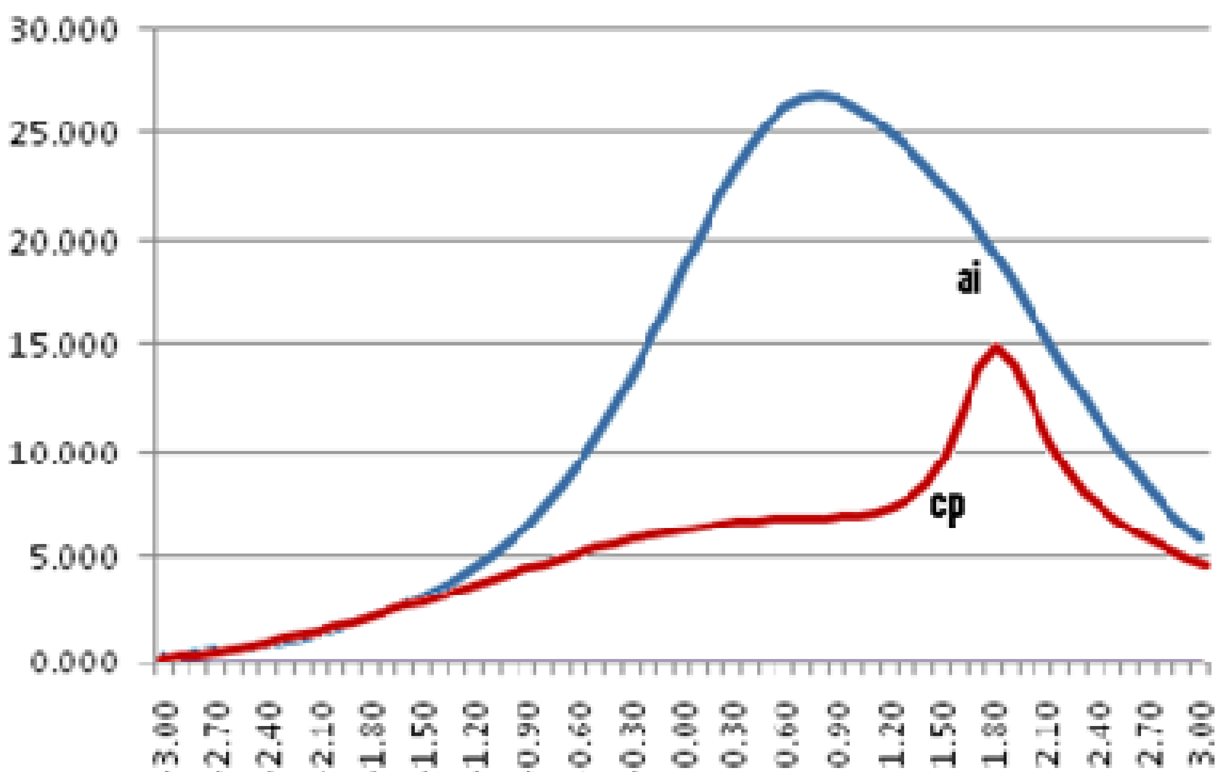

Figure 6: Test information functions for the two linking designs.

levels. Figure 7 shows the two SEM graphs for the two linking designs, more SEM is observed with the common-person design and at the low levels of theta. But generally, the two graphs are not so different.

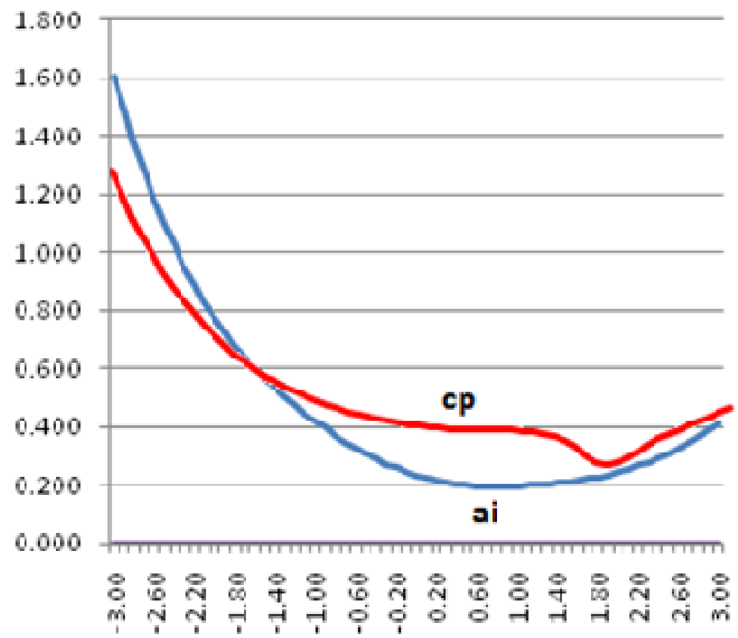

Figure 7: The two SEM graphs for the two linking designs.

Figure 6 and Figure 7 are inverse of each other as information and SEM are inversely proportional.

DISCUSSION AND CONCLUSIONS

This comparative study investigated the linking of test forms by the use of two designs, that is, the use of anchor-items and common-persons. This area of educational measurement has scanty literature indicative that not many researchers are investigating on it. Linking designs by nature are tedious, especially the use of common subjects. This is because subjects may be used only once as they may have moved to a higher level or having left the school system. In this investigation though, linking tests through common items proves a more reliable method than through the use of common persons. This reliability may stem from the fact that it is easier to control variables when dealing with items than when dealing with persons. The item/person parameter estimates for the two linking designs were correlated to determine how far they are related to each other. The error estimation in the item/person parameter estimation in the two linking methods was also investigated. The following findings were made: for the a-parameter for the two

(a) The Pearson correlation coefficient for the a-parameter for the two linking designs (2-PL model) is significant at the 0.05 level.

(b) The Pearson correlation coefficient for the a-parameters for the two linking designs (3-PL model) is not significant at the 0.05 level.

(c) The Pearson correlation coefficient for the b-parameter for the two linking designs (1-PL, 2-PL and 3-PL models) is significant at the 0.05 level.

(d) The Pearson correlation coefficient for the c-parameter for the two linking designs (3-PL model) is significant at the 0.05 level.

(e) The anchor-item linking method gives more reliable IRFs and TRFs compared the common-person linking method. This curve nearly approximates the theoretical ICC.

(f) The anchor-item linking method provides more information both at the item and test level compared the common-person linking method.

(g) Examinees with lower theta level provide more SEM than examinees with high theta levels.

REFERENCES

Assessment Systems Corporation (1995). XCALIBRE programme. Minnesota: USA.

Beguin AA, Hanson B A (2002). Obtaining Common Scale from Item response Theory Item Parameter Using separate versus Concurrent estimation in the Common-item equating Design. *Applied Psychological Measurement.* 26 (1): 3-24.

Botswana Congress Party (2009). *Education.* Retrieved on 03/08/2009 from http://www.bcp.org.bw/PGContent.php?UID=576.

Botswana Federation of Trade Unions (BFTU) (2006). Position Paper

on Privatisation in Botswana. Gaborone.
Botswana Institute of Development and Policy Analysis (BIDPA) (2006). Overview of the Economy during the National Development Plan 9 (NDP 9): Paper presented by Monnane Monnane (Research Fellow) for the BFTU, Workshop on National Economic Linkage on 23rd Feb 2006. Gaborone.
Cohen AS, Kim SH (1998). A Comparison of Linking and Concurrent Calibration Under Item Response Theory. *Applied Psychological Measurement.* 22 (2): 131-143.
Fan X (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement.* 58 (3): 357-381.
Haertel EH (2004). Cresst report: *The behaviour of linking items in test equating*. Stanford University.
Lord FM (1969). Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). *Psychometrika.* 34(3): 259-299.
Michaelides MP (2006). Effects of *misbehaving common items on aggregate scores and an application of the Mantel-Haenszel statistic in test equating*. Centre for the Study of Evaluation Report. Los Angeles. CA.
Ministry of Labour and Home Affairs (2004). Review of the National Youth Policy (Unpublished).
Mmeg (2007). *Time ripe to jack up education standards, Editorial*. Retrieved on 26/07/2009 from http://www.mmegi.bw/2007/January/Friday5/15022259576.html.
Nenty HJ (2001). *Research in educational assessment in Africa: Challenge for the 21st century*. In Lefoka JP, Matsoso LM (EDs). Educational Research for Development: Challenge for the 21st century: (pp.39-54). Roma, Lesotho: Institute of Education, National University of Lesotho.
Nenty HJ (2004). From classical test theory (CTT) to item response theory (IRT): An introduction to a desirable transition. In O. A. Afemikhe and J. G. Adewale (Eds.), Issues in educational measurement and evaluation in Nigeria (in honour of 'Wole Falayajo) (pp.371 – 384).Nigeria: Institute of Education, University of Ibadan.
OA, Adewale JG (Eds). Issues in educational measurement and evaluation in Nigeria
PARSCALE (Scientific Software International, 1992).
UNESCO (1990).World Declaration on Education for All. Jomtien, Thailand.
Warm TA (1978). *A primer of Item response theory*. Oklahoma City: U.S. Coast Guard Institute.
Weiss DJ, Yoes ME (1991). Advances in educational and psychological testing. Hambleton RK, Zaal J (Eds). Boston: Kluwer Academic Publishers.
Wright BD, Mead R, Draba R (1976). Detecting and correcting test item bias with a logistic response model. MESA Research memorandum No# 22. Chicago: MESA psychometric laboratory. Retrieved on 08/03/2006 from http//www.rasch.org/memo22.htm.
XCALIBRE (Assessment Systems Corporation, 1995)
Yu CH, Osborn PSE (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment Research & Evaluation,* 10,(4).

Citation: Kesamang M. (2017) The comparison of item and person item response theory (IRT) parameter estimates for the anchor-items and common-persons designs .Herald J. Edu. Gen. StudVol. 4(1), pp. 001 – 012