Effect of Item Arrangement on Test Reliability Coefficients: Implications for Testing

Naibi, Louisa
Department of Educational Foundations,
University of Lagos, Lagos State, Nigeria
E-mail: l.naibi@yahoo.com

ABSTRACT

This study investigated the effects of two types of item arrangement formats (ascending and specified mixed order) via test scores, on the two types of test reliability coefficients (test-retest and Kuder Richardson 20). A repeated measures two-group within-subject design was used, with a sample of four hundred and eighty (480) secondary school students from three Local Government Areas in Bayelsa State, South-South Nigeria. A 40-item Mathematics Achievement Test was used to gather data, which was analyzed using the z-test and the t-formula for testing the significance of reliability coefficients. The study finds significant effect on test scores, with the specified mixed order format having significantly higher scores, but found no effect on either of the reliability coefficients. It is recommended that item arrangement be strictly considered when constructing tests in schools and for standardized examinations.

Keywords: *item arrangement, ascending order, specified mixed order, reliability coefficient, test-retest, Kuder-Richardson 20 (K-R 20).*

INTRODUCTION

Educational measurement is an essential and invaluable part of the teaching and learning process. The term 'psychometric' refers to the measurement of such psychological aspects of individuals as personality, intelligence, and educational achievement (Nunnally and Bernstein, 1994). For any measuring instrument to be adjudged as serving its purpose, it must possess the two most basic psychometric properties of validity and reliability. Validity is the ability of a measuring instrument to (actually) measure or predict the trait or attribute that it was designed to measure or predict (Messick, 1989; Airasian, 1997), and is considered to be the singular, most important attribute that a test should possess (Anastasi and Urbina, 1996). A test is valid only when the inferences, decisions and conclusions made from its results, such as, whether a student passed, should be promoted, or failed, are correct, accurate and justifiable.

Reliability, on the other hand, is a technical term referring to the ability of a measuring instrument to give consistent or similar scores when it is administered to the same person(s), at different times and/or locations. It is an especially important criterion because it is a necessary requirement for validity. When a test gives diversely different scores when it is re-administered to the same group of persons, it implies that scores are varying when the trait (or item) is not; such a test cannot be validly measuring what it was designed to measure. So if a test is not reliable, it cannot be valid. Hence, the absence of score consistency limits the validity of the predictions and decisions to be made with the test (Davidshoffer

ISSN: 2141 - 6753

and Murphy, 1994). Reliability reflects test score precision and generalizability, two essential pre-requisites of validity. Test scores have precision when they are consistently exact. Test scores should also be generalizable to real life situations. One of the underlying issues in testing is the degree to which observed test performance can match unobserved performance or application. It is only when what a teacher observes in the class can be translated to a real life scenario outside the class, that conclusions drawn can be seen as valid. Consistent scores also give confidence that the results are related to the attribute being measured. Inconsistent scores, on the other hand, work against students' progress by implying that a student is strong, or weak, in a particular subject, when that may not be the case, leading to among other things, wrong or unsupportable decisions concerning admissions, class placement, job placement and promotion. Such candidates may come under pressure to perform more or less than they can. Reliability is thus a first step in the validation process.

Reliability is obtained through statistical correlation, and is expressed mathematically as a numerical value called a coefficient, represented by the symbol *r, and* with values ranging between + 1.00 to 0.00. A value nearer to 0.00 indicate inconsistent scores, while values nearer to + 1.00 shows greater score similarities. Reliability deals with test score consistency, thus, reliability theories were developed to estimate the effects of score inconsistencies on the accuracy of psychological and educational measurements. One of the main theories used to explain the concept of reliability is the Classical Test Theory (CTT), whose basic tenet is that any observed score is made up of a true and an error component. It is symbolized by the simple equation: X = T + E, where X is the actual obtained score, T is the true score and E is the error component.

The true score comprises the candidate's stable characteristics on the trait of interest, and reflects the exact value of his ability [or achievement] on that trait (Thissen, 1990), while the error score (technically called measurement error) comprises the characteristics of the candidate, the test itself, and the testing situation, that have nothing to do with the trait of interest, but which nevertheless affect the score and cause it to vary (McCormick and Pressley, 1997). These include student error, fatigue, stress, illness, motivation, excitement; poor examination environment like lighting or seating; or ambiguous test instruction and/or item. Essentially, the CTT assumes that the ability on a trait (or true score) is constant, and that it is the random or measurement errors resulting from these sources that cause score variation. So, while the true score contributes to score consistency, the error score causes inconsistencies (Guy, 2000).

Consequently, true score has been described as the score that a candidate would have gotten had the measurement been error-free (Onunkwo, 2002). Finally, the CTT sees error-free measurements as a theoretical impossibility since all measuring tools (and situations) are subject to some degree of error (Mehrens and Lehman, 1992). This is why, despite being theoretically desired, perfect score consistency, in form of +1.00 is often impracticable in test application (Salvia and Ysseldyke, 2001). The goals of this theory therefore include estimating measurement errors, suggesting ways of reducing them, and consequently improving the reliability status of measurement tools and procedures. The

ISSN: 2141 - 6753

commonest techniques for quantifying reliability may be classified into either repeated measures or internal consistency techniques (Traub, 1994). While these techniques are conceptually based on the CTT, they are practically based on the Parallel Test Model (Davidshoffer and Murphy, 1994), which suggests developing two equivalent forms of the same test, which should give identical scores; administering these tests to the same persons; and correlating the scores to assess the degree to which they agree with each other.

Repeated measures involve repeated or multiple administrations of the same test or, of parallel forms of the same test, to the same persons; and the correlation of the obtained scores to get a coefficient. The most commonly used repeated measures technique is the test-retest (Anastasi and Urbina, 1996), which involves administering the same test twice to the same persons. The retest may be done immediately, or after a specified interval, with two weeks being the widely accepted standard (Salvia and Ysseldyke, 2001). The test-retest coefficient is called stability index since it indicates the stability, and the extent to which the test scores may be generalized over different times and different occasions (Ashworth, 1982). It is therefore suitable for such stable traits as intelligence and achievement (Gronlund and Linn, 2000), and ultimately for long term or predictive stability. Unfortunately, due to its having two administrations, the use of test-retest are not feasible in many testing situations. Internal consistency techniques are more viable because they involve only a single administration.

Internal consistency deals with the interrelatedness of a set of test items. These techniques evaluate how consistently the items in the test measure the same trait, by assessing how consistently test takers respond or perform, correctly or incorrectly, across the test items. They are used for test items that are measuring a homogenous content, such as a Mathematics test. The rationale behind their use is that since the items in such tests are similar, and/or designed to measure the same factor or construct, the candidate should respond the same way, or consistently, to all the items, or at least a large number of them. Consequently, most of the responses of a good candidate, on a homogenous test should correlate or agree with each other (Cronbach, 1984).

The obtained coefficient is called an internal consistency index. By far the most reported indexes of internal consistency reliability are the Cronbach Alpha, for scales, and the Kuder Richardson 20 (K-R 20) formula for tests (Cortina, 1993). The K-R 20 is used for dichotomously-scored tests, such as multiple-choice tests and is more suitable for power, rather than speeded, tests (Onunkwo, 2002). It is not a technique per se, but a formula, used to correlate individual item scores with the total item score. The procedure involves administering the test once to the test takers, obtaining their scores, computing the variance of the total test scores, and for each item, determining the number of candidates that scored it correctly (and incorrectly), and then using this information to calculate the proportion of students that passed it (designated p) and the proportion that failed it (designated q). This information is then substituted in the formula. In a nutshell, score inconsistency is as a result of random measurement errors, and the reliability status of a test is the degree to which its scores are free from these random errors, for a particular group of test takers. Following from this, many sources of measurement error stem from

ISSN: 2141 - 6753

the procedures involved in developing and standardizing the test items (Linn, 1988; Nunnally and Bernstein, 1994; Haladyna, 1997), and one of these is the order in which the items are arranged (Traub, 1994). Item arrangement is a formatting process in test development. It is the procedure of arranging the items in a sequential pattern, in which they will appear in the final or published edition of the test. A test is usually formatted before it is administered to the normative group to obtain the reliability and validity coefficients, and other norms. As such, item placement is tied to the process of test validation and has the potential for affecting the reliability coefficient.

Items are not arranged haphazardly, but systematically, according to some particular rationale. Items are arranged according the type of item used (whether essay, multiple-choice, true-false); or the learning or instructional objectives being assessed; and within these formats, they are usually arranged according to their complexity or difficulty level (Gronlund and Linn, 2000). The degree of difficulty or easiness of an item is expressed by a numeral, called its difficulty index (or *p*-value), computed during item analyses. *p*-values range between 0.00 to +1.00; easier items have higher values nearer +1.00, such as 0.76, 0.89, and the more difficult ones have lower values, nearer 0.00, such as 0.44 or 0.23.

The conventional strategy for positioning educational (achievement) test items according to their difficulty is in ascending order (Anastasi and Urbina, 1996), starting with easy items and ending with the most difficult ones. The rationale behind this is that when candidates encounter the easiest items first, and can successfully solve them, it will increase their confidence and gives them a mental boost which will motivate and encourage more successful solutions of the subsequently less easy items (Mehrens and Lehman, 1992). A test candidate coming across the harder items first, especially in a timed test, might spend a lot of time there and not eventually get to the easier items that were placed at the latter part of the test.

Items may also be placed in a random or specified mixed order. A particular variant of this method involves placing difficult items (defined as items with *p*-values of less than 0.50), throughout the test at specified intervals, each followed by subsequently easier ones. The rationale behind this method is that the typical ascending order technique frustrates the candidate when they encounter and attempt too many difficult items in a row. Consequently, they end up not attempting these items at all, guessing, or worse, cheating on them, and this neither reflects the candidates' ability on that trait, nor bodes well for testing. The specified mixed order is particularly noteworthy because it is the technique used by the Stanford Achievement Test, one of the oldest and foremost standardized achievement tests in the United States (Davidson, 1985). It can also be argued that a candidate, who can perform well on a test in spite of the organisation of items, has mastered the material in question. Research on item arrangement spans the last sixty years. These studies have involved comparison of the standard ascending order format, against either of the descending order or the random or mixed order formats, and have revealed significant effects on test scores, and through this, on other test features. A majority studies on item arrangement found significantly higher scores with the ascending order format (Soureshjani, 2011; Carlson and Ostrosky, 1992; Amadioha, 1991; Gohman and Spector, 1989;

ISSN: 2141 - 6753

Hambleton and Traub, 1974; Flaugher, Melton and Myers, 1968; Sax and Cromach, 1966). Some found declining scores with the ascending order (Leary and Dorans, 1985), and higher scores with random order format (Chidomere, 1989). Some found the differences specific to item types or subject matter: Kingston and Dorans (1984), using the Graduate Record Examination (GRE), found effects on the Analysis of Explanation and Logical Diagram items, two item types that were consequently removed from the GRE tests: while Mollenkopf (1950), using the College Entrance Examination Board (CEEB) tests, found higher scores with the ascending order format, with the Verbal Analysis, but not with the Complex Mathematical items.

On the other hand, a few studies have found item positioning to have immaterial effects (Plake, Thomson and Lowry, 1981; Aiken, 1964; MacNichol, 1956), while some found results that were inconclusive, for instance, Newman, Kundert, Law and Bull (1988) found non-significant differences, even though their control group (ascending order), had a higher mean; while Monks and Stalling (1970) carried out t-tests analyses on eleven pairs of equivalent tests, with their items arranged in different positions, and found significant effects in only two of the analyses. Studies that have focused on psychometric properties have been few and far between. While MacFarland, Ryan and Ellis (2002) found enhanced effects on the psychometric properties, favouring the random placement of items; Amadioha (1991), Monks and Stalling (1970) and Flaugher, Melton and Myers (1968) found results that favoured ascending order. Sanders, Erikson and Dawis, (1988) found no significant effects, and labelled their results inconclusive. However, Marso (1970) and Kande-Bawa (2003) have argued that more empirical evidence is needed.

The preceding review has shown results that have sometimes been conflicting, depending on the situations or variables involved. All the variables have been tied to the test scores, and several of the studies have shown that item arrangement affect test scores. Since test scores determine the psychometric criteria, which are essential for any measurement strategy, any factor that impinges on the reliability of test scores is a threat to precision, accuracy, objectivity, external validity and overall validity. Item arrangement is hypothesized as a source of measurement error. Therefore ways to assess, control and minimize its effects should be studied. The studies that have specified reliability coefficients have been inconclusive and most of the reviewed studies have been done outside Nigeria. These were some of the gaps that this study sought to fill. The main thrust of this study was to analyze and compare the effects of two item arrangement formats (ascending order and the specified mixed order), via test scores, on two test reliability coefficients (test-retest and K-R 20). The study tested the hypotheses of no significant difference between, the following variables (the mean test scores, the test-retest reliability coefficient, and the K-R 20 reliability coefficient) of a Mathematics achievement test, when its items were arranged in ascending order format and when the items were arranged in a specified mixed order format.

METHOD

The research was quasi experimental, using participants from intact classrooms, with a repeated measures two-group within-subject design, where each participant received each

ISSN: 2141 - 6753

level of the independent variable (each of them wrote the test twice, each with the different item arrangement format). This eliminated the need for random assignment to experimental groups. The ABBA counterbalancing design was built-in to control for order effects (which along with practice effects, are inevitable with the use of repeated measures. Counterbalancing involved presenting the experimental conditions to the participants in an order that controls for any outside effects caused by the particular order of presentation. The study was carried out in in three local government areas (Odi, Sagbama and Yenagoa) of Bayelsa State, Nigeria. A population of Senior Secondary School Class One (SSS1) students was used, from which a sample of four hundred and eighty (480) students (male = 216; female = 264) was randomly selected.

The main study instrument was a Mathematics Achievement Test (MAT) developed and validated by the researcher, consisting of forty (40) five-optioned multiple choice items. It was content and face validated by two psychometry experts and four practicing Mathematics teachers, who critically evaluated the test format and items, in terms of congruence to the test blueprint, suitability of vocabulary, clarity of figures and graphs, font type and size, test length, test difficulty level and time limit given. The items were arranged under the two item placement formats and designated MAT Form A (Ascending order) and MAT Form B (Specified Mixed order). Four testing sessions were carried out, two for each test format, with participants being tested and retested with each format. Testing spanned a period of five days, while retesting spanned fifteen days.

RESULTS AND DISCUSSION

The first null hypothesis of no significant difference between the mean scores of the MAT when items are arranged in ascending order and when they are arranged in specified mixed order, was rejected when the results revealed a significant difference, as shown on Table 1. The second null hypothesis of no significant difference between the obtained test-retest *r* coefficients of the MAT when items are arranged in ascending order and when they are arranged in specified mixed order was not rejected, when the results, as shown on table 2, indicate that there was indeed no difference. The last null hypothesis of no significant difference between the obtained K-R 20 *r* coefficient of the MAT when items are arranged in ascending order and when they are arranged in specified mixed order was also rejected, as shown on table 3, when no difference was found.

This study focused on the effect of item arrangement on test scores and test reliability coefficients. Overall, the findings revealed the impact of item arrangement on test scores (with the mixed specified format having higher scores than the ascending order format), but not on test reliability. Hypothesis one rejection was in line with Chidomere (1989), who found higher scores with the random specified order when compared against the ascending order format. Leary and Dorans (1985) argue that scores of item arranged in ascending order declined because the difficulty level built up to a point where the items in the latter part of the test became too difficult and had to be guessed at or answered haphazardly. Flaugher Flaugher, Melton and Myers (1968) agree, adding also that the scores are affected because the test is timed, and the candidates do not get to the difficult

ISSN: 2141 - 6753

items in the allotted time. Sax and Cromach (1966) assert that under more generous timing, the difference between the formats would be non-existent. Deaton, Glasnapp and Poggio (1980) and Wu, Douglas and Monseur (2001) attribute the differential responses brought on by different item arrangement, to the differential positioning, as well as to the length of the test, and to the items terms being positively or negatively stated. Hypothesis two, which did not reject the null hypothesis, was in line with Brenner (1964) and Sanders, Erikson and Dawis (1988), though it contradicted MacFarland, Ryan and Ellis (2002) who found that item arrangement affected psychometric properties. Overall, this study found results similar to Carlson and Ostrosky (1992), Gohman and Spector (1989) and Brenner (1964), all who, while finding significant changes in test scores, found none with reliability (and validity).

Table 1: Correlated z-test summary table showing the analysis of the mean test scores of the MAT Test Forms A and B

Test Forms	N	Mean	S.D.	Df	z-calc	z-crit	Result
A (Ascending Order)		21.34	7.89				
	480			478	± 6.65	± 1.96	P > .05
B (Spec. Mixed Order)		24.91	8.93				

Source: Quasi experimentation, 2013

Table 2: t-analysis of test-retest r obtained with the MAT Test Forms A and B

Test Forms	N	r-values	t-calc	z-crit	Result
A (Ascending Order)		T-R $r^A = 0.84$			
	480		± 0.33	± 1.96	P > .05
B (Spec. Mixed Order)		T-R $r^B = 0.85$			

Source: Quasi experimentation, 2013

Table 3: t-analysis of K-R 20 r obtained with the MAT Test Forms A and B

Test Forms	N	r-values	t-calc	z-crit	Result
A (Ascending Order)		K-R 20 $r^A = 0.84$			
	480		± 0.33	± 1.96	P < .05
B (Spec. Mixed Order)		K-R 20 $r^B = 0.85$			

Source: Quasi experimentation, 2013

CONCLUSION AND RECOMMENDATIONS

Based on the findings of this study, it was concluded that item sequencing affects test scores (causing test-takers to perform better when Mathematical items, based on their difficulty levels, are scattered round the test, rather than placed in ascending order) but do not affect them enough to affect the reliability coefficients. Item arrangement is used by examiners and examination bodies to control for examination malpractice and to create equivalent forms of tests. Kande-Bawa (2003) states that re-arrangement can minimize cheating and examination malpractice. The fact that a particular placement format can differentially affect scores and other related test features is relevant for both test construction and for the practice of using alternate forms of a test to control for cheating. Since it affects test scores, it may follow that such re-arranged items may not be completely equivalent. Various researchers have agreed that item arrangement should be taken into consideration when constructing tests. Hambleton and Traub (1974) do not want item arrangement to

ISSN: 2141 - 6753

be used to develop equivalent tests, nor to compute test-retest coefficient; Deaton, Glasnapp and Poggio (1980) want open-ended and difficult items to be scattered round the test and not placed at the end; while Harris and Pommerich (2003) suggest that pretesting the items in different positions before they are selected, would mitigate the effects of item arrangement. Item placement should thus be done strictly and with proper understanding. If alternate tests are used for a particular examination, the specified order should be used in a way that gives each test paper (or combination) similar average difficulty level.

REFERENCES

Aiken, L. R. (1964). Item context and position effects on multiple-choice tests. *Journal of Psychology*, 58, 369-373.

Airasian, P. W. (1997). *Classroom assessment* (3rd ed.) New York: McGraw-Hill.

Amadioha, A. (1991). *Effects of item placement on test scores*. Unpublished M. Ed thesis, University of Port Harcourt, Nigeria.

Anastasi, A. and **Urbina, S.** (1996). *Psychological testing* (7th ed.). Engelwood Cliffs, NJ: Prentice-Hall.

Ashworth, A. E. (1982). *Testing for continuous assessment*. Ibadan: Evans Brothers.

Brenner, M H. (1964). Test difficulty, reliability, and discrimination as a function of item difficulty order. *Journal of Applied Psychology*, 48, 98-100.

Carlson, J. L. and **Ostrosky, A. L.** (1992). Item sequence and student performance in multiple-choice examinations. *Journal of Economics Education*, 23, 232-235.

Chidomere, R. C. (1989). Test item arrangement and student performance in Principles of Marketing examinations: A replication study. *Journal of Marketing Education*, (Fall), 36- 40.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98-104.

Cronbach, L. J. (1984). *Essentials of psychological testing (4th ed.)*. New York: Harpers and Row.

Davidshoffer, C. O. and **Murphy, K. R.** (1994). *Psychological testing: Principles and applications*. Engelwood Cliffs: Prentice-Hall.

Davidson, M. I. (1985). Review of the Stanford Achievement Test. In J. V. Mitchell, (Ed.) *9th Mental Measurement Yearbook*. Lincoln: Buros Institute of Mental Measurement. Retrieved November 13 2013 from *http:/www.unl.edu/buros.html.*

Deaton, W. L ., Glasnapp, D. R. and **Poggio, J. P.** (1980). Effects of item characteristics on psychometric properties of forced choice scales. *Educational and Psychological Measurement*, 40, 599-610.

Flaugher, D. A., Melton, R. S. and **Myers, C. W.** (1968). Item arrangement under typical test conditions. *Educational and Psychological Measurements*, 28, 813-824.

Gohman, S. F. and **Spector, L. C.** (1989). Test scrambling and student performance. *Journal of Economics Education*, 20 (3), 235-238. *doi: 10.2307/1182298*

Gronlund, N. E. and **Linn, R. L.** (2000). *Measurement and evaluation in teaching (8th ed.)*. New York: MacMillan.

Guy, F. Y. (2000). *Measurement and evaluation*. Bunton, Illinois: Apton.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn and Bacon.

Hambleton, R. K. and **Traub, R. E.** (1974). Effects of item order on test performance and stress. *Journal of Experimental Education*, 43, 40-46.

Harris, D. and **Pommerich, M.** (2003). Effects of context of pre-testing on item statistics and examinee scores. *Paper presented at the AERA Annual Meeting Chicago, Illinois*. Abstract retrieved October 3rd 2013 from *http:/www.home.att.net~pommie/AERA2003-All.pdf*

Kande-Bawa, A. (2003). *Multiple perspectives and potential problems with item ordering*. Pittsburgh: Koenig.

ISSN: 2141 - 6753

Kingston, N. M. and **Dorans, N. J.** (1984). Item location effects and their implication for IRT equating and adaptive testing. *Applied Psychological Measurements*, 2, 147-154.

Leary, L. F. and **Dorans, N. J.** (1985). Implication for altering the context in which items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55(3), 387-413.

Linn, R. L. (1988). *Educational measurement (3rd ed.)*. New York: MacMillan.

MacFarland, L., Ryan, A. and **Ellis, A.** (2002). Effects of item placement on faking behavior and test measurement properties in personality testing. *Journal of Personality Assessment*, 78 (2), 348-369.

MacNichol, K. (1956). *Effects of varying order of difficulty on an unspeeded verbal test.* An unpublished manuscript. Princeton, New Jersey: Educational Testing Services.

Marso, R. N. (1970). Test item arrangement, testing time, and performance. *Journal of Educational Measurement*, 7, 113-118.

McCormick, C. B. and **Pressley, M.** (1997). *Educational psychology: Learning, instruction, assessment.* New York: Addison Wesley Longman.

Mehrens, W. A. and **Lehman, I. J.** (1992). *Measurement and evaluation in education and psychology (4th ed.).* Forth Worth: Harcourt Brace College Publishers.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed.), pp 13- 103. New York: American Council of Education / MacMillan.

Mollenkopf, W. G. (1950). An experimental study on the effects of item analyses data on changing item placement and test time limits. *Psychometrika*, 1 (5), 291- 465.

Monks, J. J. and **Stalling, W. M.** (1970). Effects of item order on test scores. *Journal of Educational Research*, 63, 436-465.

Newman, D. L., Kundert, D. K., Law, D. S. and **Bull, K. S.** (1988). Effects of varying item order on multiple- choice test score: Importance of statistical and cognitive difficulty. *Applied Measurement in Education*, 1, 89-97.

Nunnally, J. C. and **Bernstein, I. H.** (1994). *Psychometric theory.* (3rd ed.). New York: McGraw Hill.

Onunkwo, G. I. N. (2002). *Fundamentals of educational measurement and evaluation*. Owerri: Cape Publishers.

Plake, B. S., Thomson, P. A. and **Lowry, S.** (1981). Effects of item arrangement, knowledge of arrangement, and test anxiety on two scoring methods. *Journal of Experimental Education*, 49, 214-219.

Salvia, J. and **Ysseldyke, J.** (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin.

Sanders, A. M., Erikson, C. A. and **Dawis, S. F.** (1988). Scrambled order- scrambled brains: The effects of presenting items in sequential vs. random order. Paper presented at the Annual Meeting of the South-Western Psychological Association, Tulsa, Okl.

Sax, G. and **Cromach, T.** (1966). The effects of various forms of item arrangement on test performance. *Journal of Educational Measurement*, 3, 309-311.

Soureshjani, H. K. (2011). Item sequencing on test performance: Easy items first? *Language Testing in Asia* 2011, **1**:46-59

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.). *Computerized and adaptive testing : A primer.* (pp. 161-185). Hillsdale, NJ.: Lawrence Erlbaum.

Traub, R. E. (1994). *Measurement methods for the Social Sciences - Reliability for the Social Sciences: Theory and applications, Vol. 3*. Thousand Oaks, CA.: Sage

Wu, M. L., Douglas, A. R. and **Monseur, C.** (2001). The PISA (Program for International Students Assessment) Project: SAI (Student Assessment Instrument) design study. *American Council for Educational Research.* Retrieved October 4th 2013 from *http://www. Leonline.com/ dui/pdf*

ISSN: 2141 - 6753