Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.5, No.25, 2014

An Assessment of the Effects of Item Difficulty and Examinee Ability on the Effectiveness of LZ Appropriateness Index

Dr. Daniel K. Korir (Corresponding author)
School of Education, Psychology Department. Moi University, PO Box 3900, 30100 , Eldoret - Kenya
E-mail: dkorir@yahoo.com

Abstract
This study investigated the effectiveness of LZ appropriateness index in detecting aberrant response patterns under nine combinations of item difficulty and examinee's ability distributions, type of aberrance, and level of aberrance. Data was generated in nine combinations of item difficulty and examinee ability to simulate the responses of 2000 non-aberrant examinees' response patterns to a 60-item test according to three-parameter model. Three uniform distributions of item difficulty were used. Two samples each consisting of 500 normal response vectors (one for spuriously low and one for spuriously high modifications) were also generated in each of the nine combinations and subjected to spurious treatment. An examinee with a spuriously high test score was simulated by selecting 20% or 10% of the examinee's original responses without replacement and changing incorrect answers to correct, but they were left unchanged if correct. An examinee with a spuriously low test score was simulated by first randomly selecting 20% or 10% of the examinee's original responses without replacement and changing correct responses to incorrect, but they were left unchanged if incorrect. LZ appropriateness index was then computed for the aberrant response vectors. The effectiveness of LZ index was evaluated by examining the extent to which it separated normal and aberrant response vectors solely on the basis of appropriateness index scores. The percentile estimates obtained for each index at each false positive rate were used as cutoff scores. The LZ index identified higher proportions of aberrant response patterns in the 20% spuriously low treatment samples than in the 20% spuriously high treatment samples. Ten percent spuriously low aberrant response samples were also found to be more detectable than the 10% spuriously high aberrant response patterns. The detection rates of the 20% and the 10% spuriously high aberrant response patterns by LZ index were found to be higher under high item difficulty parameters, and were found to be low under the low item difficulty parameters. This is not surprising as it is expected that more responses are changed from incorrect to correct and fewer responses are changed from correct to incorrect under high item difficulty parameters. The 20% and the 10% spuriously low aberrant response patterns were also more detectable under the low item difficulty parameters because more responses are changed from correct to incorrect and fewer are changed from incorrect to correct under the low item difficulty parameters.
Keywords: Appropriateness index, effectiveness, maximum likelihood index, validity.

1 Introduction

One of the basic characteristics of a test is its validity. While numerous techniques are used to assess validity, these techniques generally involve the validity of the test of the group of individuals. However, while a test may be valid for a group, it may not be valid for a particular individual who may have cheated, experienced illness, or lacked motivation during the test.

Various types of unique background experiences or individual differences in motivational dispositions such as test anxiety may make an item difficult for some examinees though it may be quite easy for most other examinees. Differential exposure to and emphasis of subject matter covered by an achievement test may result not only in the mean differences on total score from class to class but also in differences in typical response patterns. For example, there are 3,003 possible response patterns that yield a score of 10 on a 15-item test.

One can imagine numerous real situations where serious mistakes might be avoided if individuals with invalid test results were identified. For example, if the test was being used for selection to a training program, an invalid test score could result in a great deal of frustration on the part of any trainee who is selected or rejected inappropriately. This inappropriate decision would also result in administration problems and unnecessary cost on the part of the training institution.

Multiple-choice and other objectively scored tests offer a distinctive way of assessing whether a test is valid for an individual because the responses of persons with a given score are expected to follow a certain pattern. One would expect, for example, that an examinee of high ability would respond correctly to most of the easy items of a test and that an examinee of low ability would miss most of the difficult items of the test. When an individual does not follow an expected pattern of responses for a test, it makes one to question the interpretation of the test result for that individual. As an example, consider the test performance of a hypothetical examinee with high ability but limited experience with machine scored tests. The individual's high ability is apparent in the first half of the test; correct responses are given to all easy and moderately difficult items as well as many of the most difficult items. After responding to all items on the first half of the test, the examinee

decides to omit a particularly complex item. Upon solving the item that followed, the examinee forgets that the previous item was omitted. Thus, for the second half of the test, the examinee's response to the i^{th} item is recorded as the $(i - 1)^{th}$ answer on the answer sheet, the $(i + 1)^{th}$ answer is recorded in the i^{th} place, and so on. Due to time limits, the examinee does not reach the final items, and the mistake remains unnoticed.

It is apparent that the total score for this particular answer sheet substantially underestimates the examinee's ability. However, routine scoring of the answer sheet assigns a spuriously low score to the individual. Further, the inappropriate test score may be used in a college admissions or job selection decision. Note that the pattern of responses made by the hypothetical examinee is atypical. There are many correct answers to difficult items on the first half of the test and many incorrect answers to easy items on the second half. Consequently, these item responses would not be very well fitted by an item response theory (IRT) model that assumes the probabilities of correct responses are functions of a single ability of the examinee. In this example there is no bias inherent in the test, nor is this testing anomaly likely to occur systematically for a given individual or subpopulation. Because the anomaly is probably not stable and may not be related to group membership, it would go unnoticed by standard test scoring and item analysis procedures.

Again, there may be evidence of cheating in the examinee's pattern of responses. It is not likely that all cheaters, either those who occasionally glance at a neighbor's examination or those with a test preview, will achieve perfect scores on the test. The length of most aptitude and achievement tests probably prevents complete memorization of all items by an examinee. Careful vigilance by test administrator should preclude precise reproduction of a neighbor's answer sheet. It is more likely that an examinee may copy or memorize answers to blocks of items. The resulting answer sheet for a low-ability examinee may appear to have been generated by a bizarre process i.e. blocks of correct responses are intermixed with blocks of responses that are nearly all incorrect. This type of response pattern is quite different from that of the general exam-taking population and may be detectable by appropriateness measurement. Incidents of tampering with answer sheets have also been reported. In his study of bias in mental testing, Jensen (1980) states that some school teachers admitted that they completed students' tests. The teacher would simply record the correct response on the answer sheet whenever the student had skipped or failed to reach the item. Such efforts add an unknown degree of error to measurement and also are likely to produce atypical patterns of responses.

There are many other factors that can make a person's response pattern inappropriate. Among them is how clearly the instructions are understood by the examinee, familiarity with test materials and with the concepts used, previous experience with test tasks or with similar tasks and with working under pressure, and motivational factors (Van der Flier, 1982). Birenbaum (1985) notes different causes of aberrant (unexpected) response patterns; misconceptions concerning the subject matter, cultural bias, test anxiety, exceptional creativity, lack of concentration resulting in carelessly reading the questions, guessing, and occasional copying a more able neighbour's work. Wright (1977) mentions tendencies such as sleeping, fumbling, and plodding as causes of unusual response patterns. He defines sleeping as those examinees that get bored with a test and do poorly in the beginning because of confusion with test format. Examinees who never get to the latter items on the test are plodders. Unusual response patterns can also result from technical problems such as answer sheet alignment.

However, these factors jeopardize the validity of the response patterns and they are not directly reflected by a total test score. Checking the validity of the response pattern, therefore, becomes a necessity for ensuring an accurate assessment of performance. This validity check of response patterns is done with the help of appropriateness indices or person-fit indices which provide automated means for identifying response patterns where total test score may provide misleading information. Information from these indices is very useful when used with a measure of student ability. The test user would then be able to use both the total score and the index information to evaluate an examinee's performance.

Indices of appropriateness of a test for an individual could be used in a variety of ways. They could lead to the identification of subgroups of people for whom the test is inappropriate. In addition, the items that contribute most to an index indicating that the test is inappropriate for particular subgroups might also be identified. Furthermore, the content of such items could be analyzed toward the eventual end of reducing inappropriateness through test revision.

2 Review of the Literature

Several indices for detecting aberrant (unusual) response patterns have been developed. These indices describe the degree to which an individual's pattern of item responses is unusual. These indices can be classified into two groups. One group consists of indices based solely on the actual observed response patterns of the group of examinees. Examples of these indices include Sato's caution index (1975), Van der flier's U'" index (1982), Donlon and Fischer's personal biserial (1968), Tatsuoka and Tatsuoka's norm conformity index (1982), and Harnish and Linn's modified caution index (1981).

The other group consists of indices based solely on Item Response Theory (IRT models). Examples of these indices include the fit indices developed by Wright and his associates (1977), the appropriateness indices

developed by Levine and Rubin (1979), and the group of extended caution indices developed by Tatsuoka and Linn (1983). The first group of these indices is group dependent; the second group is IRT based. IRT based appropriateness indices can be sub divided into:

(1) Unstandardized and Standardized Extended Caution Indices,
(2) Maximum Likelihood Indices.

Most of the previous researchers in appropriateness measurement have compared the effectiveness of appropriateness indices (Levine & Rubin, 1979; Rudner, 1983; Parsons, 1983; Birenbaum, 1985; Drasgow, Levine, & Williams, 1985, 1986, 1987; Tomsic, 1986; Noonan, 1990; Candell, 1990); others have investigated the distribution of appropriateness indices under different conditions of item and ability parameters (Molenaar & Hoijtink, 1990, Hoijtink, 1986; & Drasgow, 1985). Recent studies in apprpriateness measurement have investigated the distributions and effectiveness of IRT based indices in varying conditions of testing.

In this study, the effects of item difficulty and examinee ability on the effectiveness of the standardized maximum likelihood index (LZ) were investigated. Levine and Rubin (1979) proposed an index denoted by Lo, which is closely related to the likelihood function, L. The likelihood function of N examinees on an n-item test is given by;

$$L = \prod_{i=1}^{N} \prod_{j=i}^{n} P_{ij}(\Theta) u^{ij} (1 - P_{ij}(\Theta))^{(1-uij)}$$

If a particular examinee (denoted by Θ) does not contribute much to the maximizing likelihood function, L, then it is likely that examinee i is not an appropriate representative of the group whose abilities are measured by the test. Thus Levine and Rubin (1979) defined appropriateness index, Lo, as;

$$Lo = \Sigma [u_{ij} \text{ Ln } P_{ij}(\Theta) + (1-u_{ij}) \text{ Ln} Q_{ij} (\Theta)]$$

Where $P_{ij}(\Theta)$ is the probability of examinee i answering item j correctly, u_{ij} is the observed item responses (0 for wrong, 1 for right) and $\text{Ln} Q_{ij} (\Theta)] = 1 - P_{ij}(\Theta)$.

With Lo, aberrance of an individual's pattern of item responses is indicated by a relatively low maximum of the function Θ. An atypical examinee pattern is expected to have a relatively low likelihood function because it is not likely that high ability examinees miss easy items or low ability examinees pass hard items.

LZ is the standardized version of the likelihood index (Lo) proposed by Levine and Rubin (1979). Levine and Drasgow (1985) found the Lo index to be less effective with examinees with high omit rates. However, Drasgow, Levine and Williams (1985) showed that the 'standardization' process is useful in accounting for the item omitting as well as controlling for the confounding effect of ability and appropriateness. They transformed Lo into a standard normal distribution and denoted it as LZ. LZ is defined as follows:

$$LZ = \frac{Lo - E(Lo)}{\sqrt{Var(Lo)}}$$

Where $E(Lo) = \Sigma \{ P_{ij}(\Theta) \text{ Ln } P_{ij}(\Theta) + (1 - P_{ij}(\Theta)) \text{ Ln}(1 - P_{ij}(\Theta)) \}$

And $Var(Lo) = \Sigma P_{ij}(\Theta) (1 - P_{ij}(\Theta) \{ \text{Ln} [P_{ij}(\Theta)/(1 - P_{ij}(\Theta)] \}^2$

3 Methods

Simulated data were used in this study. Data were generated according to the three parameter model to simulate the responses of examinees to 60 multiple choice items using Datagen, a fortran computer program developed by Hambleton and Rovinelli (1977) and modified by Carlson (1985). In previous research, the three parameter logistic model has been found to be adequate for modelling the multiple choice items on the Scholastic Aptitude Test Verbal section (Drasgow, 1982; Levine and Rubin, 1979; Levine & Drasgow, 1982; Rudner, 1983; Drasgow et al., 1985, 1986, 1987), Graduate Record examination Verbal Section (Drasgow, 1982; Levine & Drasgow, 1982) and simulated data (Noonan, 1990; Candell et al., 1990). The LOGIST computer program (Wood, Wingersky, & Lord, 1976) was used to estimate item parameters.

To evaluate the effectiveness of appropriateness indices, most researchers have used the design devised by Levine and Rubin (1979). In this design, a study begins with the test norming sample that consists of N examinees' responses (either real or simulated) to n items. Item parameters for a test model are estimated using the test norming sample. These item parameter estimates are then used to estimate examinee's ability and to compute appropriateness indices. A similar design was employed in this study and a FORTRAN77 computer program written by Drasgow (1985) was used to compute LZ scores.

Three distributions of item difficulty were used. These distributions were those which are usually found in real life situations and they were generated to simulate the distributions of items typical of diagnostics tests (items used to identify students who need remedial courses), power (placement) tests, and certification and licensing tests. Items typical of diagnostic tests were generated to have uniform distributions in the interval -3.0 to + 1.2. These test items were expected to provide maximum information (differentiate) at the low ability range.

Item difficulties typical of those found with power tests were generated to have a uniform distribution in the interval -3.0 to +3.0. These items were expected to provide equal information (differentiate) over the ability range (Van der Flier, 1982). Item difficulties typical of those found with certification and licensing examinations were generated in such a manner that they would provide maximum information at the high ability range. They were generated to have a uniform distribution in the interval +1.2 to +3.0. All the three distributions of item difficulties were generated to have uniform distributions. Uniform distribution of item difficulties is what to be expected for most tests.

Since the objective of this study was to investigate the effects of item difficulty and ability distributions and not item discrimination or the guessing parameters, the same distributions of item discrimination and guessing parameters and within the same interval were used for each replication. In all the applications, the discrimination parameters were generated in such a manner that $0.60 \leq a \leq 1.50$ and to have uniform distributions. The guessing parameters were generated in such a manner that $0.05 \leq c \leq 0.20$ and to have uniform distributions. Such distributions of guessing parameters are typical of five option multiple choice tests.

Three distributions of ability were considered. In each replication, normal distributions of examinees' ability with a standard deviation of 0.6 but with different means were used. Molenaar et al. (1990) in a simulation study found that the distributions of appropriateness indices were affected by the position of the mean and the standard deviation of the examinees' ability distribution even when the examinees' ability distribution remained normal. The ability distributions used were those typical of low, medium, and high ability examinees. Thetas (ability measure) typical of low ability examinees were generated to have normal distributions with a mean of -1.2 with a standard deviation of 0.6. Medium ability thetas typical of medium ability examinees were generated to have a normal distribution with a mean of zero and a standard deviation of 0.6. High ability thetas typical of high ability examinees were generated to have a normal distribution with a mean of +1.2 and a standard deviation of 0.6.

To examine the effects of item difficulty and examinee ability distributions on the percentile estimates of LZ appropriateness index, data were generated in nine combinations of item difficulty and examinee ability distributions. In each replication, data were generated to simulate the responses of 2000 examinees to 60 test items according to the three-parameter model. LOGIST (Wood, Wingersky, & Lord, 1976) was used to estimate item parameters. LZ appropriateness index values were computed for each examinee in each of the nine combinations of item difficulty and examinee ability distributions. The values of LZ at the 1^{st}, 5^{th}, 10^{th}, and 25^{th} percentile points were computed. A total of 50 replications were used in each combination.

To examine the effectiveness of LZ appropriateness index in detecting aberrant response patterns under different combinations of item difficulty and examinee ability distributions, type of aberrance, and level of aberrance, response vectors were generated using Datagen. Two samples each consisting of 500 response vectors (one for spuriously low and one for spuriously high modifications) were also generated in each of the nine combinations and subjected to spurious treatment. An examinee with a spuriously high test score was simulated by selecting 20% or 10% of the examinee's original responses without replacement and changing incorrect answers to correct, but they were left unchanged if correct. An examinee with a spuriously low test score was simulated by first randomly selecting 20% or 10% of the examinee's original responses without replacement and changing correct responses to incorrect, but they were left unchanged if incorrect. LZ appropriateness indices were then computed for the aberrant response vectors. The effectiveness of LZ index was evaluated by examining the extent to which it separated normal and aberrant response vectors solely on the basis of appropriateness index scores. The percentile estimates obtained at each false positive rate were used as cutoff scores.

4 Results

An analysis was also conducted on how the four percentile estimates for each index differed over ability groups in each item difficulty level and how they differed over item difficulty categories in each ability distribution. These percentile estimates were used as the cut-off scores to separate aberrant from nonaberrant response patterns. Table 1 shows the 1^{st}, 5^{th}, 10^{th}, and 25^{th} percentile estimates of LZ at 0.01, 0.05, 0.10, and 0.25 false positive rates (FP). The mean values of the 1^{st} and the 5^{th} percentile estimates of LZ were smaller than their expected values of -2.33 and -1.65 respectively for most combinations of item difficulty and ability distributions and they increased as a function of ability distributions under the low item difficulty parameters. It is also evident that the 1^{st} and the 5^{th} percentile estimates of LZ tended to be underestimated while the 10^{th} and the 25^{th} percentile estimates tended to be overestimated. However, the percentile estimates of LZ were found to be very close to the expected values when the ability estimates matched the item difficulty parameters

The percentile estimates of LZ index were found to be very sensitive to the variations of item difficulty and ability distributions. The four percentile estimates obtained for LZ index were found to be very different from the expected values in all the nine conditions. The marginal mean values of LZ of the 1^{st}, 5^{th}, 10^{th}, and 25^{th} percentile estimates were -2.396, -1.660, -1.350 and -0.614 respectively. The four percentile estimates for LZ

index were found to be very different from combination to combination.

The Scheffe post hoc results further showed that the 1st, 5th, 10th and 25th percentile estimates of LZ significantly differed among all the three ability groups under the low item difficulty and they significantly differed between the low and the high and between the medium and the high ability groups under the medium item difficulty. They also significantly differed between the low and the medium and between the low and the high ability groups under high item difficulty; an indication that the percentile estimates of LZ are more stable under the medium and under high item difficulty than they are under the low item difficulty.

Table 1: The 1st, 5th, 10th, and 25th Percentile Estimates of LZ Over 50.

Item		Ability Distributions		
diff.	**FP**	**Low**	**Medium**	**High**
Low	0.01	-2.497	-2.422	-2.137
	0.05	-1.716	-1.632	-1.362
	0.10	-1.300	-1.225	-0.989
	0.25	-0.640	-0.587	-0.428
Med.	0.01	-2.556	-2.542	-2.734
	0.05	-1.726	-1.700	-1.767
	0.10	-1.308	-1.291	-1.139
	0.25	-0.640	-0.629	-0.610
High	0.01	-2.442	-2.580	-2.556
	0.05	-1.666	-1.741	-1.722
	0.10	-1.275	-1.323	-1.300
	0.25	-0.624	-0.652	-0.620

3.1Detection rates for LZ in aberrant response patterns

The strengths and weaknesses of appropriateness indices can be assessed via their detection rates of aberrant response patterns. The detection rates of indices are determined by examining the proportion of correct classifications of aberrant response patterns at given false alarm rates. An efficient appropriateness index should identify a large proportion of aberrant response patterns at very low false alarm rates. Its distribution should also be independent of ability level of non-aberrant response patterns.

Table 2 presents the percentage of the 10% and 20% spuriously high and spuriously low aberrant response patterns correctly identified by LZ at selected false positive rates (FP). The four percentile values of LZ at 0.01, 0.05, 0.10, 0.25 false positive rates were used as the cut-off points. However, the four percentile estimates for each index were found to be different from combination to combination. The detection rates of LZ index under the combination of low ability distributions and high item difficulty parameter were 26%, 47%, 57%, and 77% at the corresponding false alarm rates of 1%, 5%, 10% and 25% respectively for the 20% spuriously high aberrant response patterns. Under the same combinations LZ could detect 7%, 21%, 30% and 55% of the 10% spuriously high response patterns at the corresponding false alarm rates of 1%, 5%, 10% and 25% respectively.

Table 2: The percentage of the 10% and 20% spuriously high and spuriously low aberrant response patterns correctly identified by LZ at selected false positive rates

		Ability distributions											
		SPURIOUSLY HIGH						SPURIOUSLY LOW					
		10%			20%			10%			20%		
Item Diff	FP.	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High
Low													
	0.01	4	1	0	2	0	0	3	7	20	3	17	33
	0.05	11	7	3	6	2	4	8	18	40	11	35	54
	0.10	18	12	6	11	4	1	11	26	52	19	48	65
	0.25	37	24	17	27	14	2	27	45	66	38	68	83
Med													
	0.01	4	1	1	3	0	0	0	7	20	6	15	30
	0.05	16	5	7	8	2	0	5	14	31	15	32	55
	0.10	26	13	13	13	3	1	8	23	40	22	47	74
	0.25	44	29	22	31	9	3	21	39	50	40	67	87
High													
	0.01	7	4	2	26	16	2	0	2	4	4	1	4
	0.05	21	12	6	47	36	9	4	5	13	3	6	14
	0.10	30	21	12	57	48	15	7	9	19	6	10	24
	0.25	55	42	25	77	67	34	17	25	38	14	24	38

For the case of the 20% spuriously low aberrant response patterns, LZ had high detection rates of 33%, 54%, 65% and 83% under the combination of low item difficulty parameters and high ability distributions at the corresponding false alarm rates of 1%, 5%, 10% and 25% respectively. The corresponding detection rates of the 10% spuriously low aberrant response patterns were 20%, 40%, 52% and 66% for the same side conditions.

The results also showed the 20% spuriously low aberrant response patterns to be more detectable by LZ than the 20% spuriously high aberrant response patterns. Further, the 10% spuriously low aberrant response patterns were also found to be more detectable than the 10% spuriously high aberrant response patterns. Given a particular type of aberrance, aberrant response patterns in the 20% spurious samples were more detectable than the aberrant response patterns in the 10% spurious samples. This implies that the detection rates of aberrant response patterns by LZ increases with the level of aberrance. At low false positive rates (0.01 & 0.05), the detection rates of LZ were high under the high item difficulty. Spuriously high aberrant response patterns were more detectable under high item difficulty parameters and spuriously low aberrant response patterns were more detectable under the low item difficulty parameters. Noonan(1990), Drasgow et al. (1985), and Birenbaum (1985) reported similar results.

3.2 Discussion

With respect to the cutoff points, the results of this study showed that the percentile estimates of LZ were affected by item difficulty parameters and ability distributions. However, the percentile estimates did not exhibit any particular pattern with respect to their magnitude when item difficulty parameters and ability distributions were varied. The four percentile estimates of LZ index were found to be different from the expected values and these percentile estimates were used as the cut-off points to separate aberrant from nonaberrant response patterns. The detection rates of LZ in this study were found to be consistent with the results reported by researchers such as Drasgow et al. (1985, 1987), Rudner (1983), Noonan (1990) and Candell et al. (1990). In particular, the high detection rates of LZ confirm the findings of Noonan (1990). The power of the LZ index and the tendency to identify larger proportions of aberrant response patterns with spuriously low scores is also consistent with the findings of Rudner (1983), Birenbaum (1985), and Drasgow et al. (1985).

The LZ index identified higher proportions of aberrant response patterns in the 20% spuriously low treatment samples than in the 20% spuriously high treatment samples. Ten percent spuriously low aberrant response samples were also found to be more detectable by LZ index than the 10% spuriously high aberrant response patterns. The detection rates of the 20% and the 10% spuriously high aberrant response patterns were found to be higher under high item difficulty parameters, and were found to be low under the low item difficulty parameters. This is not surprising as it is expected that more responses are changed from incorrect to correct and fewer responses are changed from correct to incorrect under high item difficulty parameters and vice versa.

4. Recommendations to Practitioners

Considering the results of the present study, the following recommendations can be made:

1. LZ could be used to detect spuriously low or spuriously high aberrant response patterns if a test consists of items with high item difficulties.
2. Cutoff scores should be established using a large population. However, this study has shown that cutoff scores can vary according to the side conditions of testing. Therefore, test users should see to it that cutoff scores are reviewed regularly.

5. Limitations of the Study

The first limitation of this study is that simulated data were used. Future researchers can replicate the study using real data. The second limitation is that it was assumed in this study that all the examinees reached and attempted all the questions. However, this doesn't usually happen in real life. Future researchers can use data matrix containing omits.

Thirdly is that only one distribution of item difficulties (uniform) with varying intervals was considered. Future researchers could use skewed or normal distributions of item difficulties. Fourth is that examinee ability distributions were restricted to normal distributions with different means but with the same standard deviation. However, it is possible to have other types of ability distributions in real life situations.

The fifth limitation is that data were generated according to the three parameter model. One and two parameter models could also be used for future research. Further, spuriously high and spuriously low scores were analyzed separately. In real life, a sample may have some examinees with spuriously high scores and others with spuriously low scores. This would presumably affect the detection rates. Finally, the combined effects of test length, item difficulty and examinee ability on the effectiveness of LZ should be investigated

REFERENCES

Birenbaum, M.(1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement*, 45, 523-534.

Candell, L.G.., & Levine, M.V. (1990). Detecting aberrant responses to the initial items on computerized Adaptive Tests. *An application of appropriateness measurement*. A paper presented at the annual eeting of the American Educational Research Association. Boston.

Carlson, J. (1985). IBM Version of Datagen. [Computer program]. University of ttawa. Carroll, J.B., Meade, A., & Johnson, E.S. (1986). Test analysis with the person characteristic function. Manuscript submitted for publication.

Donlon, T.F., & Fischer, F.E. (1968). An index of individual's agreement with grouped determined item difficulties. *Educational and Psychological Measurement* 28, 105-113.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, 6, 297-308.

Drasgow, F. (1985). *A computer program to compute three appropriateness indices*.

Dragow, F., & Guertler, E. (1987). A decision theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied psychology* 72(1),10-18.

Drasgow, F., & Levine, M.V. (1986). 0ptimal detection of certain forms of inappropriateness test scores. *Applied Psychological Measurement*10(1), 59-67.

Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement* Vol. 72, No.1, p.59-79.

Drasgow, F., Levine, M.V., & Williams E.A. (1985). Appropriateness measurement and polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.

Fischer, F.E. (1970). Some properties of the personal biserial index. *Journal of Educational Measurement*, 7, 275-277.

Gafni, N. (1987). *Detection of systematic and unsystematic aberrance as a function of ability estimate by several person-fit indices*. Unpublished doctoral dissertation, University of Minnesota.

Hambleton, R.K., & Rovinelli, R. A FORTRAN7 IV program for generating examinee response data from logistic test models. *Behavioral Science*, 1973, 18-74.

Harnisch, D.L, & Linn, R.L. (1981). Analysis of item response patterns: Questional test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133-146.

Harnisch, D.L., & Tatsuoka, K.K. (1983). *A comparison of appropriateness indices based on item response theory.* In R.K. Hambleton (Ed.), Applications of item response theory. In Vancouver, B.C.: Educational Research Institute of British Columbia.

Hoijtink, H. (1986). *Detecting aberrant response patterns in the unidimensional scaling model of Rasch*.Unpublished manuscript. University of Groningen, Netherlands.

Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology*,35, 42-56. Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269-290.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum associates, inc.

Molenaar, I.W., & Hoijtink, H (1990). The many Null distributions of person fit indices. *Psychometrika*, vol. 55, 1, 75-106.

Nering M.L. (1997). The distribution of Indexes of Person Fit within the Computerized Adaptive Testing evironment. *Applied Psychological Measurement*, Vol.21 No.2, 115-127.

Nerring, M.L., & Meijer, R.R. (1998). A comparison of the person response function and the LZ person-fit statistic. *Applied Psychological Measurement*, 22, 1, 53-69.

Noonan, B.W. (1990). *The effects of test length, IRT model, type of aberrance, and level of aberrance on the distribution of three appropriateness indices*. Unpublished dissertation, University of Ottawa.

Parsons,C.K., & Hulin, C.L. (1982). An empirical comparison of item response theory and hierarchical factor analysis in applications to the measurement of job satisfaction. *Journal of Applied Biochemistry,* 67, 826-834.

Rudner, L.M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, 20, 207-219.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo:Meiji Tosho.

Tatsuoka, K.K (1984). Caution indices based on item response theory. *Psychometrika*, 49(1), 95.

Tatsuoka, K.K., & Linn, R.L. (1983). Indices for detecting unusual response patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement*. 7(1), 81-96.

Tatsuoka, K.K., & Tatsuoka, M.M. (1982a). Detection of aberrant response patterns and their effects on dimensionality. *Journal of Educational Statistics*, 7, 215-231.

Tomsic, M.L. 1986). *Stability of extended caution indices for standardized public School testing: longitudinal study*. Unpublished doctoral dissertation, University of Oregon, Eugene.

Van der Flier, H. (1982). Deviant Response patterns and comparability of test scores: *Journal of Cross cultural Psychology*, Vol.13, No.3, september 1982, 267-298.

Wood, R.L, Wingersky, M.S, & Lord, F.M. (1976). LOGIST. A computer program for estimating examinee ability and item characteristics curves (Research Memorandum No. 76 ,6). Pinceton, NJ: Educational testing.

Wright, B.D (1977). Solving measurement problems with the Rasch Model. *Journal of Educational Measurement*, 14, 97-115.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar