Education 2014, 4(4): 88-94
DOI: 10.5923/j.edu.20140404.02

The Effects of Item Difficulty and Examinee's Ability on the Effectiveness of ECIZ4 Appropriateness Index

Korir Daniel K.

Department of Educational Psychology, School of Education, Moi University, P O Box 3900, Eldoret – Kenya

Abstract This study examined the effectiveness of ECIZ4 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. ECIZ4 appropriateness indices were then computed for the aberrant response vectors. The effectiveness of ECIZ4 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 ECIZ4 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 than the 10% spuriously high aberrant response patterns. The detection rates of the 20% and the 10% spuriously high aberrant response patterns by ECIZ4 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, Person-fit index, Caution Index, Validity, Effectiveness

1. Introduction

A test is a systematic procedure for measuring a sample of examinee's behavior. In the strictest sense, a test measures only test taking behavior, that is, the responses a person makes to the test items. A person is not measured directly; rather a person's characteristics (traits) are inferred from his or her responses to a test. If the behaviors exhibited on the test adequately mirror the construct being measured, the test will provide useful information. If the test does not adequately reflect the underlying characteristics, inferences made from test scores are inappropriate. A test score in a multiple choice test can only be useful in estimating person ability if the person's pattern of responses to the items corresponds to his or her expected response pattern. For instance, if the test consists of k dichotomous items arranged in ascending order of difficulty (from easy to difficult), then someone who gets x of the items right is expected to have answered the first x items correctly and the last k-x items incorrectly. If it is the easy items that he or she gets wrong, his or her pattern is regarded as deviating from the expected pattern. Therefore, a score with such a response pattern is said to be inappropriate in estimating the person's ability.

There are many factors that can make a person's response 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 [1]. Birenbaum [2] notes different causes of aberrant (unexpected) response patterns; misconceptions concerning the subject matter, cultural bias,

* Corresponding author:
dkorir22@yahoo.com (Korir Daniel K.)
Published online at http://journal.sapub.org/edu
Copyright © 2014 Scientific & Academic Publishing. All Rights Reserved

test anxiety, exceptional creativity, lack of concentration resulting in carelessly reading the questions, guessing, and occasional copying a more able neighbour's work. Wright [3] 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 which provide automated means for identifying response patterns where total test score may provide misleading information.

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 [4], Van der flier's U'" index [1], Donlon and Fischer's personal biserial [5], Tatsuoka and Tatsuoka's norm conformity index [6] , and Harnish and Linn's modified caution index [7].

The other group consists of indices based solely on Item Response Theory (IRT) models. Examples of these indices are the fit indices developed by Wright and his associates [3], the appropriateness indices developed by Levine and Rubin [8], and the group of extended caution indices developed by Tatsuoka and Linn [7]. 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, and
(3) Person fit indices.

Most of the previous researchers in appropriateness measurement have compared the effectiveness of appropriateness indices [8, 9, 10, 2, 11, 12, 13, 14, and 15]; others have investigated the distribution of appropriateness indices under different conditions of item and ability parameters [16, 17, 11]. Recent studies in appropriateness 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 fourth standardized extended caution index (ECIZ4) was investigated. Extended caution indices have been developed from Sato's caution index. In the extended caution indices, the ideal response curves are replaced by examinee response curves theoretically derived from IRT. The response curve for examinee is obtained by holding a (item discrimination) constant and considering b (item difficulty) as a continuous variable in the logistic function. Intuitively, the examinee response curve at a fixed level of corresponds to a step function whose values equal one for b and zero for b. The six extended caution indices which have been developed are ECI1, ECI2, ECI3, ECI4, ECI5, and ECI6.

However, the effectiveness of unstandardized extended caution indices were found to be related to examinee's ability level. Hence, Tatsuoka & Tatsuoka [18] standardized the extended caution indices by subtracting their expected values and then dividing by their standard errors. These indices are denoted by ECIZ1, ECIZ2, ECIZ3, ECIZ4, ECIZ5, and ECIZ6. The fourth standardized extended caution index can be computed relatively easily. Let Θ_i denote the one, two or three parameter logistic maximum likelihood estimate of Θ for the i^{th} person in the test norming sample of N examinees, and let $P_{ij}(\Theta)$ be the probability of a correct response to item j by this i^{th} examinee. ECIZ4 is then defined as follows:

$$ECIZ4 = \frac{\Sigma \left\{P_{ij}(\theta) - U_i\right\} \left\{Pij\ (\theta) - P\right\}}{\left\{P_{ij}(\theta) Q_i(\theta) \left[\ P_{ij}(\theta) - P\right]^2\right\}\ 1/2}$$

Where $P = \Sigma \{P_{ij}(\Theta)/n$ i=1……..n
and $Q_i = 1 - P_{ij}(\Theta)$

Where the distribution of the fourth standardized extended caution index (ECIZ4) have been reported to approximate a normal distribution and to be least related linearly or curvilinearly to the total test score, an indication that it provides non-redundant information [11, 12, 19, 14].

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 [20]. In previous research, the three parameter logistic model has been found to be adequate for modeling the multiple choice items on the Scholastic Aptitude Test Verbal section [11, 8, 19, 9], Graduate Record examination Verbal Section [11, 19] and simulated data [14, 15]. The LOGIST computer program [21, 22] was used to estimate item parameters.

To evaluate the effectiveness of appropriateness indices, most researchers have used the design devised by Levine and Rubin [8]. 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 program written by Drasgow [11] was used to compute ECIZ4 scores.

In this study, the effects of item difficulty and examinee's ability distributions on the effectiveness of ECIZ4 appropriateness index was investigated. Hoijtink [17] using the Rasch model, reported that examinees ability and item difficulty distributions affect the effectiveness of appropriateness indices. Molenaar and Hoijtink [23], using the Rasch model, reported that examinees ability distributions affect the effectiveness of ECIZ4 index.

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 [1]. 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. In summary, 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 to +1.50 and to have uniform distributions. The guessing parameters were generated in such a manner that 0.05 to 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, a normal distribution of examinees' ability with a standard deviation of 0.6 but with different means was used. Molenaar et al. [16] 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 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 effectiveness of ECIZ4 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 [21] was used to estimate item parameters. ECIZ4 appropriateness indices were computed for each examinee in each of the nine combinations of item difficulty and examinee ability distributions.

The values of ECIZ4 at the 99^{th}, 95^{th}, 90^{th}, and 75^{th} percentile points were also computed. A total of 50 replications were used in each combination. The means and standard deviations were computed over the 50 replications for the four percentiles of each index. These statistics were used to determine the cutoff scores of ECIZ4 index under the varying conditions of item difficulty and examinee ability distributions.

To examine the effectiveness of ECIZ4 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 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. ECIZ4 appropriateness indices were then computed for the aberrant response vectors. The effectiveness of ECIZ4 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.

4. Results

Table 1 present the 99^{th}, 95^{th}, 90^{th}, and 75^{th} percentile estimates of ECIZ4 over 50 replications. As shown in table 1, the four percentile estimates of ECIZ4 were found to be different from the expected values. Except for very few cases, the 95^{th}, 90^{th} and 75^{th} percentile estimates of ECIZ4 were less

than the expected values 0f 1.65, 1.29 and 0.68 respectively. The 99th percentile estimates did not show any pattern in terms of its magnitude. The results also showed that the percentile estimates of ECIZ4 deviated most when the item difficulty parameters did not match the ability distributions. For example, very low percentile estimates were observed under combinations of low item difficulty and high ability distributions and under combinations of high item difficulty and low ability distributions, suggesting that item difficulty and ability distributions have some impact on the percentile estimates.

Table 1. The 99th, 95th, 90th, and 75th Percentile Estimates of ECIZ4 Over 50

Item		Ability Distributions		
Diff.	FP.	Low	Medium	High
Low.	0.01	2.365	2.388	2.079
	0.05	1.616	1.620	2.079
	0.10	1.227	1.230	1.185
	0.25	0.599	0.611	0.763
Med.	0.01	2.114	2.428	2.315
	0.05	1.412	1.637	1.566
	0.10	1.064	1.240	1.183
	0.25	0.515	0.605	0.578
High.	0.01	1.817	2.261	2.442
	0.05	1.201	1.508	1.652
	0.10	0.894	1.136	1.257
	0.25	0.425	0.554	0.619

However, the percentile estimates ECIZ4 were found to be very close to the expected values when the ability estimates matched the item difficulty parameters. The 99th, 95th, 90th, and the 75th percentile estimates of ECIZ4 significantly differed among all the three ability groups under the medium and under the high item difficulty. The 99th, 95th and 75th percentile estimates of ECIZ4 significantly differed between the low and the high and between the medium and the high ability groups under the low item difficulty. This suggests that the percentile estimates of ECIZ4 are more stable under the low item difficulty than they are under the medium and under high item difficulty.

However, the four percentile estimates for ECIZ4 index were found to be different from combination to combination. With respect to the cutoff points, the results of this study showed that the percentile estimates ECIZ4 were affected by item difficulty parameters and ability distributions. It is difficult to have exact cutoff scores for determining the detection rates of the two indices. The results of the present study indicated that different cutoff points should be used for the nine combinations of item difficulty and ability distributions.

The percentile estimates of ECIZ4 index were found to be sensitive to the variations of item difficulty and ability distributions. The four percentile estimates obtained for ECIZ4 index were found to be different from the expected values in all the nine combinations. The marginals of the percentile estimates of ECIZ4 were found to be 2.244, 1.520, 1.157 and 0.585 respectively.

Detection rates ECIZ4 in Aberrant Response Patterns

The strengths and weaknesses of the 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.

The total percentage of aberrant response patterns correctly classified by ECIZ4 at the 0.01, 0.05, 0.10, and 0.25 false positive rates in the spuriously high treatment samples decreased as a function of ability distributions and increased as a function of item difficulty parameters. They ranged from 0% to 23% in the 20% spuriously high aberrant response patterns and they ranged from 0% to 6% in the 10% spuriously high aberrant response patterns. But they increased as a function of ability distributions and decreased as a function of item difficulty parameters in the spuriously low treatment samples. They ranged from 1% to 36% in the 20% spuriously low aberrant response patterns and they ranged from 0% to 13% in the 10% spuriously low aberrant response patterns. Spuriously low aberrant response patterns were also found to be more detectable by ECIZ4 than the spuriously high aberrant response patterns at 0.01 false positive rates.

Spuriously high treatment samples were less detectable by ECIZ4 under high item difficulty and they were more detectable by ECIZ4 under the low and under the medium item difficulty parameters. Spuriously low treatment samples were less detectable by ECIZ4 under the low and under the medium item difficulty parameters and they were more detectable by ECIZ4 under high item difficulty parameters.

In the spuriously high treatment samples, the detection rates of ECIZ4 decreased as a function of ability under high item difficulty. Under high item difficulty, the detection rates of ECIZ4 ranged from 9% to 43% for the 20% spuriously high aberrant response patterns and they ranged from 8% to 18% for the 10% spuriously high aberrant response patterns. In the spuriously low treatment samples, the detection rates of ECIZ4 were high under the low and under the medium item difficulty, and they increased as a function of ability. Under the medium item difficulty, the detection rates of ECIZ4 ranged from 19% to 57% in the 20% spuriously low aberrant response patterns and they ranged from 5% to 30% for the 10% spuriously low aberrant response patterns under the low item difficulty.

Table 2. The percentage of the 10% and 20% spuriously high and spuriously low aberrant response patterns correctly identified by ECIZ4 at selected false positive rates

		Ability Distributions					
		SPURIOUSLY HIGH					
		10%			20%		
Item Diff	FP.	Low	Med.	High	Low	Med.	High
Low	0.01	2	2	0	0	0	0
	0.05	9	6	2	10	3	1
	0.10	17	12	6	16	8	4
	0.25	34	28	16	33	21	13
Med							
	0.01	3	1	2	5	0	0
	0.05	15	8	7	13	3	0
	0.10	22	1 6	12	21	26	2
	0.25	44	33	27	37	22	13
High							
	0.01	6	4	2	23	12	3
	0.05	18	11	8	43	31	9
	0.10	28	18	13	54	43	17
	0.25	50	39	29	71	63	37
		SPURIOUSLY LOW					
		10%			20%		
Item diff.	FP.	Low	Med.	High	Low	Med.	High
Low							
	0.01	1	5	5	5	14	36
	0.05	5	15	30	12	33	56
	0.10	9	23	39	21	47	64
	0.25	24	42	57	41	65	78
Med.							
	0.01	0	5	9	8	15	32
	0.05	4	12	18	19	37	57
	0.10	8	20	26	26	50	69
	0.25	22	37	46	43	22	84
High							
	0.01	0	2	3	2	1	2
	0.05	6	5	11	8	4	12
	0.10	12	10	17	13	10	19
	0.25	27	27	35	28	24	38

At 0.05 false positive rates, spuriously low aberrant response patterns were more detectable by ECIZ4 index than the spuriously high aberrant response patterns. It was also observed that high proportions of the spuriously high aberrant response patterns were more detectable under the high item difficulty parameters whereas spuriously low aberrant response patterns were more detectable under the low item difficulty parameters. This could be attributed to the fact that more responses are changed from incorrect to correct and fewer are changed from correct to incorrect under the high item difficulty parameters. But more responses are changed from correct to incorrect and fewer are changed from incorrect to correct under the low item difficulty parameters.

At 0.10 false positive rates, the detection rates of ECIZ4 decreased as a function of ability distributions in the spuriously high aberrant response patterns and increased as a function of ability distributions in the spuriously low aberrant response patterns. Spuriously low aberrant response patterns were more detectable by ECIZ4 than the spuriously high aberrant response patterns.

At 0.25 false positive rate, the percentage of the spuriously high aberrant response patterns correctly classified by LZ at 0.25 false positive rate were slightly higher than the percentage of the spuriously high aberrant response patterns classified by ECIZ4 under high item difficulty. Under high item difficulty, ECIZ4 correctly classified 71%, 63%, and 37% of the 20% spuriously high aberrant response patterns when the ability distributions were low, medium and high respectively. ECIZ4 had higher detection rates of the spuriously high treatment samples under the low and under the medium item difficulty in the spuriously low aberrant response patterns.

In summary, 20% spuriously low aberrant response patterns were found to be more detectable by ECIZ4 than the 20% spuriously high aberrant response patterns. Further, the 10% spuriously low aberrant response patterns were 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 found to be more detectable than the detectability of aberrant response patterns in the 10% spurious treatment samples. This implies that the detection rates of aberrant response patterns by ECIZ4 increases with the level of aberrance. At low false positive rates (0.01 & 0.05), the detection rates of ECIZ4 were higher than the detection rates of ECIZ4 under the high item difficulty and under the low and under the medium 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. This could be attributed to the fact that more responses are changed from incorrect to correct and few responses are changed from correct to incorrect under high item difficulty parameters and more responses are changed from correct to incorrect and few responses are changed from incorrect to correct under low item difficulty parameters.

The 20% spuriously low aberrant response patterns were found to be more detectable by ECIZ4 than the 20% spuriously high aberrant response patterns. Higher proportions of the 10% spuriously low aberrant response patterns were also found to be more detectable than the 10% spuriously high aberrant response patterns. Aberrant response patterns in spurious high treatment samples were more detectable under high item difficulty parameters and they were more detectable under the low item difficulty parameters in the spuriously low treatment samples. The detection rates of ECIZ4 were also found to increase with the level of aberrance.

5. Discussion

The results of the detection rates of ECIZ4 appropriateness indices in this study are consistent with the results reported by researchers such as Drasgow et al. [11], Rudner [9], Noonan [14], and Candell and Levine [15]. In particular, the high detection rates of ECIZ4 confirm the findings of Noonan [14]. The power of the ECIZ4 index and the tendency to identify larger proportions of aberrant response patterns with spuriously high scores is also consistent with the findings of Rudner [9]), Birenbaum [2], and Drasgow et al. [11].

The ECIZ4 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 than the 10% spuriously high aberrant response patterns. The detection rates of the 20% and the 10% spuriously high aberrant response patterns by ECIZ4 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.

The detection rates of ECIZ4 index were found to increase as a function of both ability distributions and item difficulty parameters. Under combination of low ability distributions and high item difficulty parameters, ECIZ4 could detect 23%, 43%, 54%, and 71% at the corresponding false alarm rates of 1%, 5%, 10% and 25% respectively. Under the same ECIZ4 could detect 6%, 18%, 28% and 50% of the 10% spuriously high response patterns at the corresponding false alarm rates of 1%, 5%, 10% and 25% respectively. This suggests that ECIZ4 index performs better in detecting aberrant response patterns in the spuriously high treatment samples.

For the case of the 20% spuriously low aberrant response patterns, ECIZ4 had high detection rates of 36%, 56%, 64% and 78% 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 10% spuriously low aberrant response patterns were 13%, 30%, 39% and 57% for. The ECIZ4 index had very low detection rates of spuriously low treatment samples under the combination of high item difficulty and low ability distributions.

It is quite difficult to compare the detection rates obtained in this study with those obtained by other researchers because of a number of reasons. First, previous researchers used different experimental conditions from the ones considered in this study. Majority of them assumed that the examinee ability distributions are always normal (0, 1), a situation which is not always true. They also assumed that tests are always constructed to cover a wide range of item difficulty parameters. But different needs demand different tests. Hence, in this study, item difficulty and ability distributions were manipulated.

Secondly, most previous reseachers used different levels of aberrance. Noonan [14] used 30%, 15% and 10%; Drasgow et al. [11] used 10%, 20% and 30%. In this study only two levels of aberrance (10% and 20%) were considered. However, the high detection rates found in this study are consistent with the previous reported results.

6. Recommendations to Practitioners

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

1. ECIZ4 index could be used to detect spuriously low and spuriously high aberrant response patterns if a test consists of items with low and moderate 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.

7. 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 analysed 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 distributions and effectiveness of LZ and ECIZ4 should be investigated [2]. Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns.

REFERENCES

[1] 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.

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] Levine, M.V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269-290.

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

[10] 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.

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

[12] Levine, M.V., & Drasgow, F. (1982). Appropriateness measurement: review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology*, 35, 42-56.

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

[14] 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.

[15] 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.

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

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

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

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

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

[21] 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.