# The Effect of Test Length and IRT Model on the Distribution and Stability of Three Appropriateness Indexes

**Brian W. Noonan, Saskatoon Catholic Schools**

**Marvin W. Boss and Marc E. Gessaroli, University of Ottawa**

The extent to which three appropriateness indexes—$Z_3$, ECIZ4, and W (a variation of Wright's person-fit statistic)—are well-standardized was investigated in a monte carlo study. To assess the effects of the item response theory (IRT) model and test length on the distribution of the indexes and their cutoff values at three false positive rates, nonaberrant response patterns were generated. ECIZ4 most closely approximated a normal distribution, showing less skewness and kurtosis than $Z_3$ and W. The ECIZ4 cutoff values were affected less by test length and the IRT model than were $Z_3$ and W. In contrast, the distribution of W was the least stable over replications, and its cutoff values varied greatly depending on the IRT model and test length. *Index terms: appropriateness measurement, caution index, item response theory (person fit), person-fit statistics, unusual response patterns.*

Valid interpretation of test scores is an important concern in psychological and educational measurement, especially when test scores are used to make important decisions about individuals, because a test score may be invalid for an individual, even though the test may have satisfactory properties for the group being measured. Recently, quantitative measures called appropriateness indexes have been developed to detect unusual examinee response patterns. A number of researchers (Drasgow, 1982; Drasgow & Levine, 1986; Drasgow, Levine, & McLaughlin, 1987a, 1987b; Drasgow, Levine, & Williams, 1985; Levine & Rubin, 1979; Rudner, 1983; Smith, 1986) have investigated detection rates of some of the appropriateness indexes.

Typically, two criteria are used to assess appropriateness or person-fit indexes—standardization and relative power (Drasgow et al., 1987a). Standardized indexes are those considered to be invariant in their distributions across trait levels. Standardized indexes possess two advantages. First, the high detection rates of well-standardized indexes are not attributable to differences in the trait or number-correct distributions between nonaberrant and aberrant response patterns. Second, a single cutting score at a given false positive rate (Type I error) may be used to classify examinees' responses as aberrant or nonaberrant.

The second criterion for evaluating appropriateness indexes is relative power. The most powerful indexes are those that have the highest detection rates of aberrant indexes at a given false positive rate. Using an index that is not standardized will cause problems in providing good detection rates over all trait levels. Drasgow et al. (1987a) and Drasgow et al. (1985) have demonstrated this; therefore, a necessary condition for a powerful index is that it be well-standardized.

Several studies have shown that standardized item response theory (IRT) model-based indexes best meet the criteria for standardized indexes. Drasgow et al. (1987b) examined 11 indexes over five groups of examinees with varying trait levels. They found that the $Z_3$ index (Drasgow et al., 1985), the ECIZ2 and ECIZ4 indexes (Tatsuoka,

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 16, No. 4, December 1992, pp. 345-352*

*0146-6216/92/040345-08$1.65*

1984; Tatsuoka & Linn, 1983; Tatsuoka & Tatsuoka, 1982), and the W3 index (Rudner, 1983) were fairly well standardized and effective in identifying aberrant response patterns across different trait levels.

Drasgow et al. (1987a) studied the distributions of $Z_3$, ECIZ4, and W3 over 1,000 nonaberrant response patterns. The indexes were distributed with the following means and variances, respectively: $Z_3$ (.09, .97), ECIZ4 (.14, .86), and W3 (.99, .12). Using the Kolmogorov-Smirnov test statistic, they concluded that $Z_3$, ECIZ4, and W3 did not depart significantly from a normal distribution. Tatsuoka (1984) showed that ECIZ2 and ECIZ4 followed an approximately normal distribution with mean = 0 and standard deviation (SD) = 1. The distribution of these indexes also was reported to be independent of trait level. Drasgow et al. (1985) found that $Z_3$ was distributed similarly over different trait levels, but departed somewhat from normality in the left tail of the distribution. Using eight datasets, Reise (1990) showed that $Z_3$ did not possess a linear relationship with trait levels. These studies, therefore, generally show that $Z_3$, ECIZ2, ECIZ4, and W3 are well-standardized.

In terms of power, these four indexes have been shown to be more effective than most other indexes. Rudner (1983) studied the detection rates of five indexes and found that U3, W3 (three-parameter versions of U1 and W1), and $L_g$ (Levine & Drasgow, 1982) showed higher detection rates than U1 and W1 (Wright, 1977). These rates ranged from 50% to 75% for response strings modified to be 15% and 20% spuriously high. These results were obtained using a .05 false positive rate. Using three levels of aberrance for spuriously high and spuriously low scores, Drasgow et al. (1985) found that $Z_3$ identified from 65% to 95% of aberrant responses for spuriously high scores at a false positive rate of .10. For spuriously low scores, approximately 40% were identified.

Drasgow et al. (1987a) found that $Z_3$, ECIZ2, ECIZ4, and W3 produced better detection rates than other indexes. At a .05 false positive rate, $Z_3$ correctly identified 35% to 98% of modified response patterns, ECIZ2 identified 32% to 98%, ECIZ4 identified 29% to 97%, and W3 identified 29% to 97%. Drasgow et al. (1987b), in a follow-up study, examined the detection rates for a 30-item test and a 50-item test. They found that $Z_3$, $Z_h$ (Levine & Drasgow, 1982), ECIZ2, and ECIZ4 were the most effective. In general, detection rates were lower for the 30-item test. It is important to note that in the above studies response patterns were simulated and the degree of aberrance was known.

Although the power of an index ultimately will determine its usefulness, there are a number of other concerns related to standardization that need to be investigated. Because the effectiveness of an appropriateness index is related to the use of a single cutoff score at a selected false positive rate, it is necessary to examine the stability of the cutoff value of the selected index at the selected false positive rate.

Other issues related to appropriateness indexes include the effects of test length and IRT model on the distributions of the indexes and on the cutoff scores for selected false positive rates. Studies related to the effect of test length (Drasgow et al., 1987b; Rudner, 1983) have shown that nonstandardized appropriateness indexes tend to be more effective with longer tests. However, in those studies, different types of tests were used in the comparison—which may have confounded the results. Only one researcher (Drasgow, 1982) compared the effects of different IRT models, but used only nonstandardized indexes.

This study had two main purposes. First, although it has been suggested that W, $Z_3$, and ECIZ4 approximate normal distributions under conditions of no aberrance, the distributions of these indexes have not been systematically investigated under conditions in which the IRT model and test length are varied. Second, the stability of these indexes at various false positive rates is not known. In most studies, a sample of nonaberrant examinees is simulated. Cutoff values at specified false positive rates then are

used to assess detection rates in simulated response strings with aberrance introduced. Thus, this study examined the effects of varying the IRT model and test length on the distributions of $Z_3$, ECIZ4, and W and on the stability of the cutoff values of the indexes at selected false positive rates when no aberrance is introduced.

## Method

### Person-Fit Indexes

$Z_3$ (Drasgow et al., 1985) is the standardized maximum likelihood estimate of $L_o$, the index first suggested by Levine & Rubin (1979). ECIZ4 is the standardized version of ECI4 (Tatsuoka, 1984; Tatsuoka & Linn, 1983; Tatsuoka & Tatsuoka, 1982), and is based on Sato's caution index (Sato, 1975; Tatsuoka & Linn, 1983). It is computed as the complement of the ratio of two covariances:

$$\text{ECI4} = 1 - \frac{\text{cov}(\mathbf{Y}_i\mathbf{P}_i)}{\text{cov}(\mathbf{P}_i\mathbf{G})} \quad , \tag{1}$$

where $\mathbf{Y}_i$ is the observed response vector of person $i$, $\mathbf{P}_i$ is the probability vector associated with person $i$, and the elements of $\mathbf{G}$ are

$$G_j = \frac{1}{N}\sum_{i=1}^{N} P_{ij} \quad , \tag{2}$$

where $i$ indexes persons, $j$ indexes items, $n$ is the number of items, and $N$ is the number of examinees (Tatsuoka, 1984; Tatsuoka & Linn, 1983).

To standardize the index, the conditional expectation is:

$$1 - \frac{\text{Var}(\mathbf{P}_i)}{\text{Cov}(\mathbf{G},\mathbf{P}_i)} \quad , \tag{3}$$

the variance is

$$\text{varE}(\text{ECI4}\,|\,\theta) = \frac{\sum_{j=1}^{n}\sigma_{ij}^2(P_{ij} - T_i)^2}{n^2\text{cov}^2(\mathbf{G},\mathbf{P}_i)} \quad , \tag{4}$$

and $T_i$ is the average probability of correctly answering the items on the test given the person's trait level estimate.

The third index, W, is based on the person-fit statistic, W1, introduced by Wright (1977), and is defined by:

$$\text{W} = \frac{\sum_{i=1}^{n}(U_{ij} - P_{ij})^2}{\sum_{i=1}^{n}[P_{ij}(1 - P_{ij})]} \quad , \tag{5}$$

where $U_{ij}$ is the observed response,
$P_{ij}$ is the probability of a correct response, and
$n$ is the number of items.

W is an extension of W1 to the two- or three-parameter model.

### Data Generation and Analysis

To investigate the distributions of the appropriateness indexes and the stability of the indexes at various false positive rates, a monte carlo approach was used. Simulated data were used so that examinee response patterns could be generated with known item parameters, known examinee trait levels, and no aberrance. The data were generated using a modified version of the program DATAGEN (Hambleton & Rovinelli, 1977).

The distribution of item parameters was similar to those recommended by Hambleton & Swaminathan (1985) for a general achievement test; thus, uniform distributions of item difficulty (-2.00 to 2.00) and item discrimination (.40 to 1.50) were used. Two IRT models, the two-parameter logistic (2PLM) and three-parameter logistic (3PLM), and two test lengths (40 items and 80 items) were the independent variables. For the 3PLM, a uniform distribution of the pseudo-guessing parameter (.05 to .20) was used. For each combination of test length and IRT model, 2,000 response vectors were generated using examinee trait levels ($\theta$) drawn from a normal (0,1) distribution. 2,000 response vectors were used because that number has been used by other researchers (Drasgow et al., 1987a) and because it is adequate for a 3PLM analysis.

The three appropriateness indexes then were

calculated for each examinee using a computer program developed by Drasgow (1985). The mean, SD, skewness, and kurtosis also were calculated for each index and for each combination of test length and IRT model. In addition, the value of each index at false positive rates of .01, .05, and .10 was saved. To examine the stability of the distribution and the cutoffs for each false positive rate, the above procedures were replicated 50 times.

The relationship of each index to θ also was examined. For each combination of IRT model and test length, 10,000 nonaberrant response patterns were simulated. Pearson correlations were computed to determine if there were linear relationships between θ and the indexes. Also, the absolute values of θ were correlated with each of the indexes to determine whether extreme θ values (regardless of sign) were associated with high index values.

## Results and Discussion

Pearson correlations between an index and θ ranged from −.017 to .018. None of these values was significant at the .05 level. Similar results were found when absolute values of θ were correlated with the indexes. Thus, it appears that the values of the indexes were not conditional on θ. This provides support for the standardization of these indexes.

### Effects of IRT Model and Test Length on Distributions

Table 1 shows the mean, SD, skewness, and kurtosis calculated for the indexes across the 2,000 examinees, then averaged across the 50 replications of each condition. As expected, when no aberrance was present, $Z_3$ and ECIZ4 had means and SDs that approximated a (0,1) distribution (Drasgow et al., 1987a; Tatsuoka, 1984), regardless of test length or IRT model. The fit statistic, W, had a mean of approximately 1.00. However, its SD had a mean over replications that varied from .232 to .144 depending on test length and IRT model. Both the longer test and the 3PLM resulted in a lower SD for W. Over the 50 replications, the mean of the W distribution was more stable than that for $Z_3$ and ECIZ4. This was

**Table 1**
Mean (M) and SD Over Fifty Replications for Mean, Standard Deviation, Skewness, and Kurtosis Across 2,000 Simulated Examinees for $Z_3$, ECIZ4, and W by Experimental Condition

| Statistic and Index | 40 Items | | | | 80 Items | | | |
|---|---|---|---|---|---|---|---|---|
| | 2PLM | | 3PLM | | 2PLM | | 3PLM | |
| | M | SD | M | SD | M | SD | M | SD |
| Mean | | | | | | | | |
| $Z_3$ | −.007 | .022 | −.009 | .021 | −.004 | .022 | −.005 | .022 |
| ECIZ4 | .004 | .020 | .004 | .018 | .003 | .024 | .004 | .023 |
| W | 1.002 | .005 | 1.002 | .004 | 1.000 | .003 | 1.000 | .003 |
| Standard Deviation | | | | | | | | |
| $Z_3$ | 1.003 | .016 | 1.003 | .015 | .999 | .016 | .999 | .017 |
| ECIZ4 | .999 | .015 | .999 | .015 | 1.002 | .014 | 1.001 | .015 |
| W | .232 | .011 | .205 | .011 | .162 | .007 | .144 | .007 |
| Skewness | | | | | | | | |
| $Z_3$ | −.621 | .081 | −.513 | .069 | −.432 | .063 | −.355 | .063 |
| ECIZ4 | .277 | .065 | .232 | .066 | .205 | .048 | .177 | .051 |
| W | .624 | .250 | .580 | .306 | .413 | .241 | .328 | .228 |
| Kurtosis | | | | | | | | |
| $Z_3$ | .509 | .297 | .340 | .196 | .222 | .223 | .139 | .209 |
| ECIZ4 | .014 | .160 | −.022 | .115 | −.008 | .114 | −.036 | .113 |
| W | 2.599 | 2.342 | 2.990 | 3.212 | 2.211 | 1.998 | 1.957 | 1.691 |

expected because the SD of W was so much smaller than the SDs of $Z_3$ and ECIZ4.

The mean skewness over the 50 replications was more than twice as large for $Z_3$ and W, as compared to ECIZ4. In addition, ECIZ4 showed less variability in skewness over replications. In fact, the variability of the skewness for W suggests that it was quite unstable. The 80-item test and the 3PLM resulted in less skewness for all indexes. It is interesting to note that for each index the mean skewness reflected non-normality in the tail of the distribution that is used to identify aberrance.

The W distributions were quite leptokurtic and had much variability across replications. The mean kurtosis over 50 replications was extremely large (1.957 or greater). In addition, the SD of kurtosis over 50 replications ranged from 1.691 to 3.212. The maximum kurtosis value for each combination of test length and IRT model was 9.29 or greater. $Z_3$ was somewhat leptokurtic and was less stable over replications than ECIZ4. For $Z_3$, increased test length and the 3PLM reduced kurtosis. Overall, ECIZ4 appeared to be neither platykurtic nor leptokurtic; however, it still lacked stability—the SDs over replications ranged from .113 for the 80-item-3PLM condition to .160 for the 40-item-2PLM condition.

Thus, under conditions of no aberrance, ECIZ4 most closely approximated a normal distribution. ECIZ4 also showed the most stability over replications and over various combinations of test length and IRT model. The W distributions showed the greatest departures from normality and the least stability over replications. Although the means of the W distributions were quite stable, the SDs were quite variable—they tended to have large skewness, and extremely large and variable kurtosis.

### Effects of Test Length and IRT Model on Cutoff Values

The summary statistics by IRT model and test length over 50 replications for cutoff values at three false positive rates are shown in Table 2. Several differences among the indexes are apparent. For the W index, the cutoff values at each of the three false positive rates were not much larger than the mean of the total distribution. These differences from the mean (which was approximately 1.00) ranged from .181 for the 80-item-3PLM condition at the .10 false positive

**Table 2**
Mean (M) and SD Over Fifty Replications of Three Indexes at Selected False Positive Rates

| Index, Test Length, and IRT Model | False Positive Rate | | | | | |
|---|---|---|---|---|---|---|
| | .01 | | .05 | | .10 | |
| | M | SD | M | SD | M | SD |
| $Z_3$ | | | | | | |
| 40, 2PLM | -2.791 | .139 | -1.826 | .062 | -1.348 | .043 |
| 40, 3PLM | -2.724 | .120 | -1.799 | .053 | -1.347 | .041 |
| 80, 2PLM | -2.652 | .120 | -1.771 | .054 | -1.328 | .051 |
| 80, 3PLM | -2.593 | .109 | -1.750 | .054 | -1.328 | .041 |
| ECIZ4 | | | | | | |
| 40, 2PLM | 2.522 | .098 | 1.715 | .046 | 1.315 | .043 |
| 40, 3PLM | 2.466 | .089 | 1.715 | .047 | 1.322 | .042 |
| 80, 2PLM | 2.492 | .091 | 1.714 | .047 | 1.312 | .040 |
| 80, 3PLM | 2.454 | .095 | 1.698 | .046 | 1.309 | .034 |
| W | | | | | | |
| 40, 2PLM | 1.629 | .038 | 1.398 | .022 | 1.294 | .015 |
| 40, 3PLM | 1.555 | .047 | 1.347 | .019 | 1.259 | .013 |
| 80, 2PLM | 1.423 | .024 | 1.273 | .013 | 1.205 | .010 |
| 80, 3PLM | 1.372 | .029 | 1.240 | .012 | 1.181 | .009 |

rate to .629 for the 40-item-2PLM condition at the .01 false positive rate.

The cutoff values for $Z_3$ departed more from the mean of its distribution (approximately 0) than did the cutoff values for ECIZ4. This is likely a reflection of the greater skewness and kurtosis noted above. Over the 50 replications, the cutoff values for the false positive values showed greater variability for $Z_3$ than for ECIZ4. However, W showed considerably less variability than either $Z_3$ or ECIZ4. This was expected because the mean cutoff values differed so little from the distribution mean for W.

A multivariate analysis of variance was performed to examine the effect of test length and IRT model on the three indexes at the three false positive rates. Test length, IRT model, and their interaction were significant at the .05 level. In the follow-up univariate analyses of variance (.01 level of significance), considerable differences were found on the indexes for the effects of test length and IRT model. These results, shown in Table 3, indicated that the interaction affected only W and that it was most highly affected by test length and IRT model.

The results in Table 3 indicate that the W index was the most highly affected by test length and IRT model. Cutoff values for establishing false positive rates, therefore, are highly influenced by test length and IRT model. Although $Z_3$ was not as highly affected as W, both the IRT model and test length had a significant effect on cutoff values. Although the IRT model did significantly affect ECIZ4 at the .01 false positive rate, the performance of ECIZ4 was certainly better than either $Z_3$ or W. The data in Table 3 also show that test length influenced cutoff values more than IRT model did, especially for W and $Z_3$. Significant differences were found for ECIZ4 only for IRT model at the .01 false positive rate.

## Relationships Among Indexes

Relationships among the indexes were examined using a Pearson product-moment correlation. $Z_3$ and W correlated highly (–.944 to –.951) over combinations of test length and IRT model. $Z_3$ and ECIZ4 produced lower correlations (–.652 to –.682) across the combinations. ECIZ4 and W showed the lowest correlations (.571 to .594) over combinations of test length and IRT model. The extremely high relationship between W and $Z_3$ was somewhat surprising. W is based on the squared residuals between observed item scores and probability of success given $\theta$. $Z_3$ is a standardized

**Table 3**
Results of Univariate Analysis of Variance for Effects of Test Length and IRT Model on Three Indexes at Three False Positive Rates ($df$ = 1,196)

| Index and False Positive Rate | Value of $F$ | | |
|---|---|---|---|
| | Test Length | IRT Model | Interaction |
| $Z_3$ | | | |
| .01 | 60.63* | 13.03* | .04 |
| .05 | 43.76* | 8.76* | .15 |
| .10 | 10.14* | .02 | .01 |
| ECIZ4 | | | |
| .01 | 2.60 | 12.80* | .50 |
| .05 | 1.88 | 1.41 | 1.52 |
| .10 | 1.83 | .14 | .70 |
| W | | | |
| .01 | 1,464.17* | 150.06* | 4.85 |
| .05 | 2,325.92* | 299.94* | 13.93* |
| .10 | 2,474.95* | 303.61* | 9.61* |

*Significant at the .01 level

version of $L_o$ that is the maximum of the logarithm of the dichotomous model likelihood function.

## Conclusions

The results show that the ECIZ4 index most closely approximated a normal distribution. Its distribution was the least skewed and the least subject to extreme kurtosis. Over replications, these values (skewness and kurtosis) were much more stable than for W and $Z_3$. The values of ECIZ4 at each false positive rate were also less affected by test length and IRT model. These estimates were consistently less variable over replications than those of $Z_3$.

The W statistic did not perform well, because it was limited in the range of its distribution, it was very leptokurtic, and it was skewed. In addition, its distribution was quite unstable over replications. The manner in which the false positive cutoff values were affected by IRT model and test length calls its use into question.

Some caution should be exercised in eliminating the use of $Z_3$. At this point, it is not known how the indexes compare in identifying unusual response patterns. Drasgow et al. (1987a) reported on the effectiveness of $Z_3$ using the 3PLM for specific item parameters. Insofar that choice of IRT model and test length are shown here to affect the distribution of the indexes, it will be of interest to know if those effects exist when the relative power of the indexes is examined. Information obtained from the present study does not provide answers to this question. Certainly, the evidence suggests that the stability of the cutoff values for false positive rates is of concern. It is definitely a concern for researchers when assessing the standardization of indexes. It would seem that stability over replications would be desirable for a well-standardized index.

Three expectations for a standardized index can be identified. First, the index must not be related to trait levels—results supported this for each of the three indexes. Second, the index should fit a known distribution, in this case a normal distribution. The W index violated this requirement. ECIZ4 most closely characterized a normal distribution. Third, an index value for identifying aberrant response patterns should be stable over occasions and for different kinds of tests. Again, ECIZ4 most closely met this criterion; W departed the most from this criterion. However, based on the results, the question of whether any of the indexes fully meets this last expectation remains.

## References

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* [Unpublished computer program].

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10,* 59–67.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987a). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 57–59.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987b). *Appropriateness measurement* (AFHRL-TP-87-6). Texas: Manpower and Personnel Division, Brooks Air Force Base.

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

Hambleton, R. K., & Rovinelli, R. J. (1973). A Fortran IV program for generating examinee response data from logistic test models. *Behavioral Science, 17,* 73–74.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 22–56.

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

Reise, S. P. (1990). A comparison of item and person-fit methods of assessing model data-fit in IRT. *Applied Psychological Measurement, 14,* 127–137.

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 (in Japanese).

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46,* 359–372.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

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,* 81–96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). *Standardized extended caution indices and comparison of their rule detection rates* (Research Report 82-4-ONR). Urbana: University of Illinois, Computer-Based Educational Research Laboratory.

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

## Author's Address

Send requests for reprints or further information to Brian Noonan, Saskatoon Catholic Schools, St. Paul's R.C.S.S.D. #20, 420 22nd Street East, Saskatoon, Saskatchewan, S7K 1X3, Canada.