# Appropriateness Fit and Criterion-Related Validity

**Neal Schmitt, José M. Cortina, and David J. Whitney**

**Michigan State University**

Unmotivated or suspicious test takers in concurrent validation studies can cause numerous problems for test users. The effects of these problems, however, have not been carefully examined. This study used item response theory-based appropriateness fit indexes to identify and remove from a validation sample those examinees whose response patterns did not match their trait levels (e.g., examinees with low trait levels who answered difficult items correctly). The person-fit index $l_z$ described in Drasgow, Levine, & Williams (1985) had little effect on validities. The multitest index $l_{zm}$ described by Drasgow & Hulin (1990) was more promising. Implications for selection research and practice are discussed. *Index terms: aberrant response patterns, appropriateness fit, concurrent validity, distorted responses, item response theory, person fit.*

Psychometrically adequate tests may provide inappropriate estimates of some examinees' trait levels. For example, examinees with low trait levels may copy the answers of examinees with high trait levels, or examinees may deliberately try to distort their answers on personality tests in a socially desirable manner or in a way that they believe will make them more employable when the test is used in an employment context. The latter issue continues to be of significant concern (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990). On ability tests, highly capable examinees may receive scores that are too low when they are poorly motivated, interpret items unconventionally, or make mistakes in using answer sheets.

In the last 15 years, researchers using item response theory (IRT) have developed methods of efficiently identifying individuals whose item response patterns are inconsistent with their trait levels as measured by the test as a whole (see Drasgow & Hulin, 1990, for a review). In personnel selection or educational contexts, aberrant response patterns can be very important both to the individual and the employing organization or academic institution. For the individual whose trait level is underestimated because of misinterpretation of some test items or some carelessness in answering computer-scored answer sheets, the outcome may be lack of access to a desirable job or academic opportunity. For the organization, examinees with low trait levels whose scores are overestimated may produce costly failures in expensive training programs or on-the-job performance.

Drasgow & Guertler (1987) outlined how the utility of various outcomes associated with test use, the base rates of those outcomes, and the accuracy of an appropriateness index all affect the importance of detecting individuals whose response patterns are unlikely. In addition to the two practical reasons described above for being concerned about identifying aberrant response patterns, researchers also may be concerned because a sizable proportion of such response patterns in a given sample may distort estimates of criterion-related validity and estimates of the interrelationships between trait levels and performance constructs.

## Purpose

Appropriateness indexes of model fit were used here to identify aberrant responders and to assess

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 17, No. 2, June 1993, pp. 143-150*

*0146-6216/93/020143-08$1.65*

the effect of such aberrant responses on estimates of concurrent criterion-related validity for a set of employment tests. Of interest was whether identification and removal of individuals with aberrant response patterns affected concurrent criterion-related test validities. In concurrent criterion-related research, job incumbents are required to take a test battery whose validity is being examined. Frequently, no personal benefit accrues to the job incumbents who take these tests, and occasionally they react with resentment and suspicion as to the company's motivation. This may result in some examinees either responding randomly to some items to thwart the company's objectives or copying from colleagues. This type of situation produced the validities and test responses studied here. Specifically, the effect on observed validities of the removal of individuals whose response patterns were unlikely given their trait levels was examined. A positive result could lead to the use of this procedure as a guard against the underestimation of validities in criterion-related research.

## Appropriateness Indexes

Donlon & Fischer's (1968) proposed use of the personal biserial coefficient may represent the earliest attempt to provide a quantitative estimate of the degree to which an examinee's trait level as estimated by a test score matches the examinee's responses to items of varying difficulty as estimated in a norm group. The personal biserial coefficient is the correlation between the dichotomously scored item responses of a particular individual and the group-determined item difficulties. A low or negative personal biserial coefficient results when an individual answers many easy items incorrectly and answers difficult items correctly.

Since that time, several researchers (e.g., Drasgow & Levine, 1986; Drasgow, Levine, & McLaughlin, 1987; Rudner, 1983) have investigated the use of various indexes of fit. This literature has indicated that the $l_z$ index (Drasgow & Levine, 1986) provides the most accurate identification of aberrant response patterns. $l_z$ is the standardized estimate of $l_o$ (Birnbaum, 1968). To compute $l_o$, first obtain the trait level estimate for each examinee using an appropriate IRT model. $l_o$ is the logarithm of the compound probability of the correct and incorrect responses given by an examinee with a given trait level as estimated by the model. Formally,

$$l_o = \sum_{i=1}^{n} \{u_i \ln P_i(\hat{\theta}) + (1 - u_i) \ln[1 - P_i(\hat{\theta})]\} , \quad (1)$$

where

$n$ refers to the number of items in the test,

$u_i$ is the response of the individual to the $i$th item (1 = correct, 0 = incorrect), and

$P_i(\hat{\theta})$ is the probability of the response to item $i$ given the estimate of the examinee's trait level.

$l_o$ is then standardized using the following formula:

$$l_z = \frac{l_o - \mathrm{E}(l_o)}{[\mathrm{Var}(l_o)]^{1/2}} , \quad (2)$$

where $\mathrm{E}(l_o)$ is the expected value of $l_o$ and $\mathrm{Var}(l_o)$ is the variance of $l_o$, which are computed as follows:

$$\mathrm{E}(l_o) = \sum_{i=1}^{n} \{[P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + [1 - P_i(\hat{\theta})] \ln[1 - P_i(\hat{\theta})]\} \quad (3)$$

and

$$\mathrm{Var}(l_o) = \sum_{i=1}^{n} P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\{\ln[P_i(\hat{\theta})/[1 - P_i(\hat{\theta})]]\}^2 . \quad (4)$$

These equations were derived in Drasgow, Levine, & Williams (1985) and were also presented in Drasgow & Guertler (1987).

Previous research (Drasgow et al., 1985) has indicated that the distribution of $l_z$ is close to standard normal at all θ levels. This is important because some indexes easily detect aberrant responses of examinees with high or low θ levels, but are not particularly effective for examinees whose θ levels are relatively average. Other research indicates that $l_z$ consistently outperforms other indexes of fit; however, the range of accuracy (65% to 90% accurate) varies with conditions, particularly the false positive rates associated with a normal response pattern (e.g., Drasgow & Levine, 1986; Drasgow et al., 1987). Finally, accuracy in detection of aberrant response patterns also deteriorates substantially as the test becomes shorter.

Drasgow, Levine, & McLaughlin (1987), for example, showed that for low θ simulated examinees who were given the answers to 30% of the items in a 30-item test, the most powerful appropriateness index identified only 45% of the "cheaters," as opposed to 93% accuracy when the test was comprised of 85 items (Drasgow et al., 1987). For this reason, Drasgow, Levine, & McLaughlin (1991) developed a multitest extension of $l_z$ (see also Drasgow & Hulin, 1990). This multitest extension can be used when the data from several tests or scales are combined, as they might be in a selection test battery. This multitest appropriateness index proved to be much more powerful in detecting aberrant response patterns—accuracy rates for combinations of tests were approximately equivalent to single tests with an equal number of items (Drasgow et al., 1991).

$l_{zm}$ is defined by

$$l_{zm} = \frac{\sum_{j=1}^{J} \{l_o^{(j)} - \mathrm{E}[l_o^{(j)}]\}}{\left\{\sum_{j=1}^{J} \mathrm{Var}[l_o^{(j)}]\right\}^{1/2}}, \tag{5}$$

where $l_{zm}$ is the multitest extension of $l_z$, and $j$ refers to the individual tests.

## Method

### Sample

201 maintenance mechanics in a small manufacturing firm were asked to take a battery of tests that were being considered for use in selecting new employees. All but two of the employees had worked more than a year for the company; three-fourths had worked more than three years. 176 had completed a high school education; 52 reported some post-high school education. 20 were female; 179 were white; and their average age was approximately 32.

### Tests and Criteria

*Tests.* The four tests administered included (1) a 47-item knowledge test (KN) constructed by the authors consisting of basic questions about plumbing, electricity, hydraulics/pneumatics, and safe work practices; (2) a 24-item graphic arithmetic (GA) test (Personnel Designs, Inc., 1990a); (3) a 60-item mechanical comprehension (MC) test (Personnel Designs, Inc., 1990b); and (4) the 64-item Space Relations (SR) test of the Differential Aptitude Battery (Bennett, Seashore, & Wesman, 1973). These tests differed with respect to speededness. They ranged in speededness from unspeeded (the KN test had no time limit and everyone finished) to highly speeded (the SR test had a time limit and few employees finished). All tests were multiple-choice, and all were administered to small groups of employees.

*The criterion.* 10 behaviorally anchored rating scales (Smith & Kendall, 1963) and a single-item overall performance scale (anchored with "superior," "average," and "clearly deficient") were constructed, and ratings by supervisors using these scales served as the criterion. Ratings on all 11 dimensions were made on seven-point scales. The 11 criterion scales were Inspection (INSP), Communication (COMM), Maintenance (MAIN), Diagnosis (DIAG), Calibration (CALI), Housekeeping (HOUS), Safety (SAFE), Equipment Transportation (EQUI), Interpersonal Work Relationships (RELA), Mechanic (MECH), and Overall Performance (OVER).

Collection of these ratings was preceded by a short training session in which types of rating errors were described, along with suggestions on how to avoid these errors and an explanation of the nature and importance of the study and the need to collect accurate performance indexes. There was no opportunity to evaluate the interrater reliability of the ratings because only one supervisor was available to rate any given employee. Intercorrelations among the 11 rating dimensions ranged from .34 to .77.

### Calculation of Appropriateness Indexes

Appropriateness indexes were computed and were used to remove from the sample those employees whose fit indexes suggested that their answers to some items did not match their ability levels. Theoretically, this would include both employees of low ability who answered difficult items correctly and those of high ability who answered easy questions incorrectly. Criterion-related validities then were recalculated. Because the analyses of the adequacy of the test scores as estimates of their ability levels were internal to the test, this procedure is not the same as removing outliers before reassessing validity. The latter is circular and increases the size of validity coefficients. This method of removing persons whose test scores are "suspicious" relies only on estimates of ability levels derived from the test, not from the test-criterion relationship.

To use $l_z$ as a measure of appropriateness, the difficulty, discrimination, and guessing parameters were estimated by BILOG (Mislevy & Bock, 1990) for each item in the four tests. IRT assumes that test items are unidimensional; therefore, the items were factor analyzed (using matrices of tetrachoric correlations) and eigenvalues were examined to determine if the items in each test were indeed unidimensional. The first factor eigenvalues and the percentage of variance they accounted for were 13.09 (27.9%), 8.77 (36.5%), 16.11 (25.2%), and 14.8 (24.7%) for the KN, GA, MC, and SR tests, respectively. Because these all accounted for large percentages of variance, and because they were all at least three times as large as their respective second factor eigenvalues, it was concluded that the tests were adequately unidimensional.

The IRT parameters were used to calculate $l_z$ for each of the examinees for each of the tests, as well as the combination of the four tests using $l_{zm}$. Because $l_z$ is distributed approximately normally with a mean of 0 and a standard deviation of 1 (Drasgow et al., 1985), an $l_z$ value less than –2 was taken as an indication that the response pattern was aberrant in some way. As indicated above, these employees were removed from the sample and the validities were recalculated.

## Results

### Removal of Examinees With Inappropriate Response Patterns

Descriptive statistics for the tests are presented in Table 1, which also shows the number of persons removed from the analysis for each test. The cutoff value of –2 for $l_z$ led to the removal of 36 examinees from the KN test, 39 from the GA test, 18 from the MC test, and 70 from the SR test. This left 162, 159, 180, and 128 examinees, respectively, for the recalculation of validities. Incomplete data on one or more tests precluded use of data for three additional examinees.

**Table 1**
Test Means and Standard Deviations (SD), Number of Examinees Before and After Removal Based on $l_z$ Values, and Test Intercorrelations Before Removal (Below the Diagonal) and After Removal (Above the Diagonal) Based on $l_{zm}$

| Test | Mean | SD | Number of Persons Removed | Number of Persons After Removal | Test | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | KN | GA | MC | SR |
| KN | 32.30 | 7.30 | 36 | 162 | | .20 | .41 | .10 |
| GA | 13.38 | 5.11 | 39 | 159 | .43 | | .48 | .54 |
| MC | 49.61 | 7.73 | 18 | 180 | .62 | .62 | | .53 |
| SR | 35.45 | 12.41 | 70 | 128 | .39 | .68 | .68 | |

Very few of the employees eliminated were of high ability. That is, employees who obtained low $l_z$ scores for the most part had low or average ability scores, but answered some of the more difficult items correctly. The tendency of $l_z$ to remove only low ability examinees also applied to the SR test, but on a larger scale. Examination of the data suggested that the large number of aberrant $l_z$s may have been caused by the speededness of the SR test. In other words, the only employees who responded to all the items in the SR test were employees who apparently guessed at some or all of the items. Items at the end of the test that were not reached by low ability employees were scored as incorrect. This included a large number of low ability employees. These employees, by chance, answered some of the later items in the test correctly—items that most employees did not attempt and therefore answered incorrectly.

## Validities

Correlations between the four tests and the supervisory ratings constituted the estimates of concurrent criterion-related validities; these are shown in the BF (before) columns in Table 2. Most of the validities were relatively low given results reported in the literature for similar tests (Hunter & Hunter, 1984; Schmitt, Gooding, Noe, & Kirsch, 1984).

**Table 2**
Validities Before (BF) and After (AF) Removal of Persons with Aberrant Response Patterns Based on $l_z$ for Single Tests and Adjusted Multiple Correlations Before and After Removal of Persons with Aberrant Response Patterns Based on $l_{zm}$ for Four Tests Combined

| | Test | | | | | | | | 4 Tests Combined | |
|---|---|---|---|---|---|---|---|---|---|---|
| | KN | | GA | | MC | | SR | | | |
| Criterion | BF | AF | BF | AF | BF | AF | BF | AF | BF | AF |
| INSP | .090 | .122 | .040 | .024 | .014 | .047 | .070 | .200 | 0.00 | .09 |
| COMM | .070 | .098 | .030 | -.015 | -.027 | -.088 | .010 | 0.000 | 0.00 | .18 |
| MAIN | .170* | .152* | .070 | .039 | .162* | .140 | .060 | .160* | .14 | .12 |
| DIAG | .090 | .058 | .170* | .112 | .192* | .091 | .100 | .150* | .09 | .02 |
| CALI | .060 | .051 | .040 | -.018 | -.014 | -.021 | .030 | .040 | 0.00 | 0.00 |
| HOUS | .010 | .020 | -.030 | -.084 | -.078 | -.112 | .040 | .060 | .12 | .11 |
| SAFE | .120 | .166* | .070 | .010 | .005 | -.050 | -.010 | -.080 | .09 | .23* |
| EQUI | .150 | .233* | -.010 | -.065 | .006 | .060 | -.020 | .010 | .09 | .19* |
| RELA | -.020 | .011 | .020 | -.028 | .029 | .063 | .070 | .240* | 0.00 | 0.00 |
| MECH | .110 | .114 | .090 | .059 | .112 | .137 | .120 | .160* | 0.00 | .05 |
| OVER | .080 | .103 | .110 | .080 | .089 | .130 | .190* | .180* | 0.00 | .08 |

*Significantly different from 0.0 at $p < .05$.

Table 2 also shows in the AF (after) columns validities for each of the four tests without the employees whose responses were aberrant. As can be seen, the removal of these examinees had little consistent impact on validities. Of the 44 pairs of validities (11 for each of the four tests), 19 show small increases in validities after the removal of employees who responded "inappropriately," and 25 show small decreases, with little consistency across criteria or tests.

### Multitest $l_z$

Calculation of $l_{zm}$ led to the removal of 55 examinees from the sample. Table 1 also shows the intercorrelations of the predictor tests both before removing employees with inappropriate response patterns (lower triangle) and after their removal (upper triangle), based on $l_{zm}$. Test intercorrelations were consistently lower after the inappropriate response vectors were eliminated.

The multiple correlations, adjusted for shrinkage (Wherry, 1931), between the tests and each of the criteria before and after removal then were compared. These results are displayed in the last two columns in Table 2 and suggest that removal of examinees using $l_{zm}$ generally increased validities, especially with respect to EQUI, SAFE, and COMM. The multiple correlations ($R$s) associated with these criteria increased by .10, .14, and .18, respectively. Of the 11 $R$s computed after removal of the aberrant responders, only one (DIAG) decreased by more than .02, and this change was nonsignificant.

To further investigate the impact of respondent removal based on $l_{zm}$, a hierarchical multiple regression was conducted in which the overall measure of performance was regressed onto a composite of the scores for the four tests (Step 1), the $l_{zm}$ index score (Step 2), and the product of the two (Step 3). Table 3 shows that the interaction term was significant ($p < .02$), suggesting that the validity for the composite of the test scores was increased by the level of $l_{zm}$. These results provide additional support for the use of this index to remove respondents before computing validities.

**Table 3**
Beta Weights and Change in Adjusted $R^2$ from Moderated Hierarchical Regression of Overall Performance Rating Onto Test Score Composite, $l_{zm}$, and Their Product

| Step | Beta Weight | Change in Adjusted $R^2$ |
|---|---|---|
| 1. Composite | .11 | .007 |
| 2. $l_{zm}$ | .047 | 0.000 |
| 3. Composite by $l_{zm}$ | .698* | .024 |

*Significantly different from 0.0 at $p < .05$.

## Discussion

### The $l_z$ Index

There were four important findings with respect to the use of $l_z$. First, validities did not consistently increase after the removal of aberrant response patterns. Second, the prediction of the "Relationships" criterion benefitted most by the use of $l_z$, but this and the other significant increase observed (see Table 2) might be chance deviations. Third, a speeded test will likely produce many aberrant response patterns due to inflated item difficulties for items at the end of a speeded test, possibly due to examinees' guessing behavior.

Finally, test intercorrelations were lower after removal of examinees with high $l_z$ scores than they were when these correlations were based on the total. It is not clear why this occurred. Lowering

correlations between predictors would usually be desirable in an applied context, but this result should be replicated and explained.

### Multitest $l_z$

The general finding with respect to $l_{zm}$ was that the removal of examinees based on $l_{zm}$ had a stronger effect on observed validities than did the removal of examinees based on any combination of the single test indexes. It may be that fit indexes can lead to higher validities if enough test items are used. $l_{zm}$ did increase several of the *R*s. The *R* for the SAFE rating, for example, increased from .09 to .23. In many situations, this would be a practically significant difference. These findings agree with those of Drasgow et al. (1987). Removal of examinees based on $l_{zm}$ (or $l_z$) might result in a substantial increase in observed validities if enough items are available.

### Future Directions

This study had limitations and should be replicated. The sample size was small; three of the tests were speeded, one severely so; and the tests were not selected originally to fit an IRT model. Future research should more carefully examine the effects of variables such as speededness, the size of initial validities, and the extent of response aberrance on the impact of appropriateness on observed validities. Specifically, what types of aberrant responses exhibited by what proportion of examinees would actually affect estimates of validity substantially? Of related interest is the relationship between the "Lie" scales of personality inventories such as the MMPI and fit indexes such as $l_z$ and $l_{zm}$. If lie scales are highly correlated with statistical fit indexes such as $l_z$, then the problems of honesty and accuracy associated with lie scales could be avoided by the use of an index such as $l_z$. Finally, it might be useful to relate appropriateness fit values to various demographic or experience variables in an effort to understand the kinds of persons whose response patterns appear to be aberrant. Alternative selection procedures then might be used to obtain more appropriate estimates of these individuals' trait levels.

## References

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1973). *Differential aptitude tests: Space relations.* New York: Psychological Corporation.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

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

Drasgow, 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,* 10–18.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 577–636). Palo Alto CA: Consulting Psychological Press.

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. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 58–79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171–191.

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

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Mislevy, R. J., & Bock, R. D. (1990). *Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Personnel Designs, Inc. (1990a). *Graphic arithmetic test.* Grosse Pointe MI: Author.

Personnel Designs, Inc. (1990b). *Mechanical comprehension test.* Grosse Pointe MI: Author.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20,* 207–220.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. P. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149–155.

Wherry, R. J., Sr. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2,* 440–457.

## Author's Address

Send requests for reprints or further information to Neal Schmitt, Department of Psychology, Michigan State University, East Lansing MI 48824, U.S.A. E-mail: 10259nws@msu.