A literature review of detecting aberrant response patterns by using IRT-based fit indices

Hyunsoo Seol
(Chung-Ang University)

I. The Problem

1. Importance of Response Aberrance Detection

Over the years, many studies have been conducted to investigate aberrant response patterns of examinees(e.g.,Thurstone & Chave, 1929; Mosier, 1940; Glaser, 1949, 1952; Wright, 1977; Wright & Stone, 1979; Drasgow, 1982; Levine & Drasgow, 1982; Drasgow & Levine, 1986; Tatsuoka & Tatsuoka, 1982a; Tatsuoka & Linn, 1983; Tatsuoka, 1984; Wright & Masters, 1982; Smith, 1982, 1985, 1991b, 1994a). Aberrant responses are inconsistent responses, or responses that do not fit the overall pattern of an examinees responses. Recently, Meijer (1996), in a special issue of "Applied Measurement in Education", explored a number of issues concerning person-fit response analysis. Relatively less attention has been directed to aberrance associated with items constituting a test.

One of the reasons for the enthusiasm for investigating aberrant response patterns is the realization that an examinees total test score often fails to provide a meaningful measure of his or her ability level(Sato, 1975). For instance, examinee behaviors such as guessing, carelessness and plodding(see Wright, 1977, p.110-112; Wright & Stone, 1979, p.170-190; Smith, 1985, 1991b, p.146-149)can cause aberrant response patterns. These behaviors are reflected in the total score, thus making it difficult to infer an examinees ability level. Similar aberrance, as noted here with respect to persons, can occur with items. Just as person-related aberrant responses can affect the estimation of person ability through the total test score, item-related aberrance can affect the estimation of item difficulty.

To deal with this problem, persons or items that have aberrant response patterns should be detected and treated separately to get better person or item parameter estimates. One remedy following detection of aberrance is the separation and removal from the total sample, or total test, the specific persons or items that are found to be aberrant. Lee and Suen(1994) discussed a theoretical justification for the removal of subjects in item calibration.

Another reason for detecting aberrant response patterns is that this can provide diagnostic information about examinees, such as test anxiety, cultural bias(van der Flier,

1977, 1982), group differences with regard to sex or race(Donlon & Rindler, 1979; Harnisch & Linn, 1981b), misconceptions with respect to subject matter, or sporadic study habits (Blixt & Dinero, 1985). Item related diag nostic information can also be useful in exam development. van der Flier(1982, p.267) indi cated that various factors can influence the test scores of different groups, such as clarity of test construction, familiarity with specific test tasks motivation/interest, and reaction to time pressure. Drasgow, Levine, and McLaughlin(1987) described some institutional situations calling for the detection of inappropriate test scores:

"In academic admissions testing, spuriously high scores can lead to the enrollment of unqualified individuals in undergraduate, graduate, and professional programs. . . cause a more deserving student to be denied admission. . . In employment testing. . . terminating unsuccessful employees is difficult and can be very expensive if contested in the courts. Individuals who fail to complete a management or vocational training program conducted by an organization(e.g., a military training school) may cost the organization many thousands of dollars"(p.59-60).

Smith(1983) recommended a person response analysis as cost effective as follows:

"it is possible to perform a person analysis for less than 10 cents per person. This is a small price to pay to insure that the score reported accurately reflects the examinees true ability"(p.24).

Harnisch and Linn(1981a)provide additional support for Smith's recommendation:

"Indices measuring the degree to which the response pattern for an individual is unusual could be used in a variety of ways. They could identify individuals for whom the standard inter pretation of the test score is misleading, or identify groups with atypical instructional and/or experiential histories that alter the relative difficulty ordering of the items. In addition, the items that contribute most to high values on an index for particular subgroups could be identified and judgments made regarding the appropriateness of the item content for those subgroups"(p.133).

<Table 1>

	Dichotomous Response Patterns to a 10 Item Test Item										
Examinee	1	2	3	4	5	6	7	8	9	10	Diagnosis
1	1	1	1	0	1	0	1	0	0	0	Usual response pattern
2	1	1	1	0	0	0	0	1	0	1	Guessing
3	0	0	1	1	1	0	1	1	0	0	Start up/carelessness
4	0	0	0	0	0	1	1	1	1	1	Miscoding

Identifying and diagnosing aberrant response patterns can provide valuable information for understanding individual or subgroup performance on tests; and can lead to better testing practices or to the improvement of the instructional process; for example, by improving items or by identifying students in need of special help or with exceptional creativity.

2. What are the Aberrant Response Patterns ?

A response pattern that has low probability, given an IRT model, is classified as aberrant. It is assumed that examinees taking a test will generally answer easy items correctly more frequently, and hard items correctly less frequently. Any situation which contradicts this pattern interferes with a valid interpretation concerning the examinees ability. Consider the response patterns to a 10 item test below(see <Table 1>).

In the response patterns shown in <Table 1>, items are ordered from easiest(Item# 1) to hardest(Item# 10) and all of the examinees have a total score of 5. Examinee 1 shows the usual response pattern in which he/she answered easy items correctly and hard items incorrectly. Examinee 2 showed that he/she answered hard items(Item# 8 and Item# 10) correctly even though he/she failed medium difficulty items. These correct answers for hard items are highly improbable and suggest possible guessing. In contrast to Examinee 2, Examinee 3 answered easiest items(Item# 1 and Item# 2) incorrectly. One might conclude that this Examinee is unfamiliar with the test format or is careless. The response pattern of Examinee 4 is extremely unusual because he/she failed all the easy items and passed all the hard items. With this response pattern, one might investigate the answer

<Table 2>

Measurement Disturbances associated with the Person(Smith, 1982, p.127)

1. **Startup-test anxiety**. Unexpected incorrect responses at the beginning of the test which will result in an under estimate of the persons ability.
2. **Plodding/Excessive Cautiousness**. Unexpected omitted responses at the end of test. If scored as incorrect the persons ability will be under-estimated.
3. **Copying from another person**. This will result in groups of unexpected results throughout the test unless the entire test is copied. Usually this will result in an over estimate of the persons ability.
4. **Illness**. If the onset of the illness occurs during the test and impairs the persons ability to perform the persons ability will be under-estimated.
5. **External Distractions**. External distractions, e.g., mowing the lawn outside the window, can cause a person to perform below his ability on a sub-section of the test. This will result in an under estimation of the persons ability.
6. **Guessing to complete test/Random Guessing**. This behavior will result in both unexpected correct and incorrect responses. Usually it will result in an over-estimate of the persons ability.
7. **Disinterest/Boredom**. General disinterest in the entire test or sections of the test will often result in unexpected incorrect responses, resulting in an under-estimate of the persons ability.
8. **Fatigue**. Long tests can often introduce a fatigue factor which causes unexpected incorrect responses late in the test often resulting in an under-estimate of the persons ability.

sheet to check if the items are miscoded. It is easy to note that the total score of 5 in these four cases does not represent adequately the performance of these examinees.

3. Types of Person(Item) Aberrant Response Patterns

The detection of aberrant response patterns is considered an important issue in the measurement process because such aberrant response patterns may produce an inaccurate estimate of the persons true ability. Smith(1982, p.127 128) identified two types of measurement disturbances. The first type of measurement disturbances, presented in <Table 2>, is associated with the person. The second type of measurement disturbances, involving the interaction between items and persons, is listed in <Table 3>.

As shown in <Table 1>, the process of ordering items from easiest to hardest and investigating how each examinee responded to the items, provides insights into the person fit problem. The simple examples shown in <Table 1> suggest that it is necessary to study not only what score an examinee receives, but also how he/she achieves the score, and why he/she produces a deviant score pattern. In practice, however, it is difficult to find many of the aberrant response patterns which are described in <Table 2> and <Table 3>. Meijer(1996) explained the difficulty and provided the following reasons:

". . . (a) due to the probabilistic nature of the procedure underlying the item response because item score patterns may not convincingly reflect the underlying aberrant behavior, and (b) because aberrant behavior may only play a role with a small number of items from the test. In addition, even if a pattern is statistically identified as aberrant, the researcher cannot always be sure of the kind of aberrance underlying test performance because different forms of aberrant behavior may result in the same kind of item score pattern" (p.7).

With this in mind, it can easily be inferred that the best person fit statistic may not exist to explain each type of aberrant

<Table 3>

Measurement Disturbances involving the Interaction between Items and Persons(Smith,1982, p.128)

1. **Guessing when the correct answer is not known**. This usually occurs on items that are very difficult for the person and results in an over-estimate of the persons ability.
2. **Sloppiness/Excessive Carelessness**. This usually occurs on items that are very easy for the person and results in an under-estimate of the persons ability.
3. **Item Content/Person Interaction**. This usually occurs when one of the skills or topics included on the test is over-learned or under-learned. This may result in an over- or under-estimate of the persons ability.
4. **Item Type/Person Interaction**. This usually occurs when one of the item formats differentially favors a person. It may result in an over or under estimate of the persons ability.
5. **Item Bias/Person Interaction**. This usually occurs when a subset of items differentially favors an ethnic group, age group, sex, curriculum, or cognitive style. This may result in an over or under estimate of the persons ability.

response pattern. Nonetheless, it should be noted that the basic idea of indicating person fit comes from identifying both invalid test scores and specific types of aberrant response patterns.

II. Review of Fit Indices

Various indices for analyzing aberrant response patterns have emerged and been developed(van der Flier, 1977; Sato, 1975; Harnisch & Linn, 1981a; Donlon & Fisher, 1968; Tatsuoka & Tatsuokaa, 1982a,1982b; Kane & Brennan, 1980; Wright & Panchapakesan, 1969; Levine & Rubin, 1979; Tatsuoka & Linn, 1983).

There are two major types of indices. The *first group* of indices is based on observed patterns of responses(for a brief review, see Harnisch & Linn, 1981a, p.134-139; Meijer & Sijtsma, 1995, p.266-267). These indices include Sato's Caution Index(1975), van der Flier's U Index(1977), Tatsuoka and Tatsuoka's Norm Conformity Index(1982), Donlon and Fisher's Personal Biserial(1968), Kane and Brennan's Agreement and Disagreement/Dependability Index(1980), Harnisch and Linn's Modified Caution Index (1981a), and the D'Costa's B and W indices (1993). Typically, the frame of reference for detecting/judging aberrance is the examinees group. Hence, these indices can be referred to as group-based indices.

An advantage of Sato's(1975) or D'Costa's (1993) approach is that it is easy to interpret the results and use them in situations such as small classroom settings. Therefore, these indices are meaningful to classroom teachers. Teachers understand that a difficult item answered correctly is surprising and worthy of caution. Also, an easy item that is missed provides reason for concern and is worthy of caution. One disadvantage concerning these indices is that the distributional properties of these indices are not known. Thus, it may not be appropriate to establish absolute critical values to classify aberrant response patterns. For example, although a value higher than 0.5(Sato, 1975), or 0.3(Harnisch & Linn, 1981a), was suggested as a cut-off value for aberrance for the Sato Caution Index, no rationale has been provided for these critical values(D'Costa, 1993). Also, since the distribution of these indices is not known, it is impossible to apply a statistical test to determine if an index value deviates significantly from its expected value.

The *second group* of indices includes those based on Item Response Theory(IRT). They have been called 'fit statistics' by Wright and Panchapakesan(1969), 'appropriateness index' by Levine and Rubin(1979), and 'Extended Caution Indices(ECIs)' by Tatsuoka and Linn(1983). These indices measure the goodness of fit between an individual response pattern and the IRT stochastic model. Real data can never fit the IRT model exactly because of the strict assumptions imposed by IRT, or because of real life measurement disturbances(e.g., unidimensionality, local independence, guessing, test anxiety etc.). Therefore, IRT-based indices are based on probability. Response patterns with a low probability in the context of the IRT model are classified as aberrant.

Rudner(1983) compared the aberrance

detection rates of nine indices(point biserial correlation, biserial correlation, norm conformity index, modified caution index, weighted Rasch-model fit mean-square, unweighted

Rasch model-fit mean-square, unweighted 3-parameter model-fit mean-square, weighted 3-parameter-model fit mean-square, and likelihood index using 3-parameter model). He reported that indices based on IRT models showed better detection of aberrant response patterns than other types of indices.

This study will focus on IRT-based indices because of the advantages of IRT over classical test analysis procedures(see Wright & Stone, 1979; Andrich, 1988; Baker, 1992, p.114-170; Hambleton, Swaminathan & Rogers, 1991; Wilson, 1994; Suen, 1990, p. 83-129). The next section presents research literature in which the three IRT-based fit indices approaches are examined or reviewed. Emphasis was placed on reviews of research and empirical studies utilizing the three IRT-based fit indices approaches in terms of their sensitivity for detecting person/item aberrances.

1. Residual-Based Approaches

The basic idea of residual-based approaches to person or item fit indices comes from the concept of departure from expectation. In IRT models, the probability of a correct response can be described as the interaction of a persons ability and the difficulty of an item, which indicates an expected response. Given an expected response, it is possible to compare an observed response with an expected response.

The difference between the observed and the expected response will produce a residual, or departure from expectation. These residuals can be squared and summed over items or people and used as the basis for evaluating person-fit or item-fit in the model.

Wright(1979) reports the unweighted and weighted total fit index. The formula for the unweighted total mean square, MS(UT), is given as

$$MS(UT) = \frac{1}{n} \frac{\sum_{i=1}^{n}(U_{ij} - P_{ij})^2}{\sum_{i=1}^{n}[P_{ij}(1 - P_{ij})]}$$

where U_{ij} is the observed response of person j to item i, P_{ij} is the probability of correct response, and n is the number of items.

Mathematical derivations concerning mean and standard deviation of MS(UT) can be found in Smith(1982, p.55-58). This fit index can be standardized to an approximate unit normal(0,1) by a cube root transformation, which is given by the following formula:

$$UW_Z = (MS^{\frac{1}{3}} - 1)(3/S) + (S/3)$$

where S is the standard deviation of MS given above.

This index is referred to as the standardized OUTFIT statistic in the BIGSTEPS (Linacre and Wright, 1993) manual. This fit index is sensitive to outliers. For example, unexpected incorrect responses by high ability persons to easy items, or unexpected correct responses by low ability persons to hard items, will produce larger index values than expected. Therefore, a weighted total fit index was developed to lessen the effect of outliers. The weighted total mean square, MS(WT), is expressed as

$$MS(WT) = \frac{\sum_{i=1}^{n} w_{ij} \sum_{i=1}^{n}(U_{ij} - P_{ij})^2}{\sum_{i=1}^{n} w_{ij}}$$

$$= \frac{\sum_{i=1}^{n}(U_{ij} - P_{ij})^2}{\sum_{i=1}^{n} w_{ij}}$$

where U_{ij} is the observed response of person j to item i, P_{ij} is the probability of correct response, n is the number of items, and $w_{ij} = P_{ij}(1 - P_{ij})$.

This index is also standardized with a cube-root transformation to approximate the unit normal (0,1) distribution. The *weighted* standardized residual-based index is referred to as INFIT statistic in the BIGSTEPS (Linacre and Wright, 1993) manual. The weighted total-fit index is less affected by unusual response patterns by persons with ability far from the difficulty level of the item.

According to Smith *et al.*(1994c), OUTFIT expected values have been shown to be closer to the expected value of 0.0 than INFIT expected values. However, there have been some criticisms of these statistics.

Smith(1991b) summarizes these criticisms as follows:

"First, the observed response and the expected response are not independent since the observed response is used in estimating both the item and person estimates to get the expected response. Second, the observed response for dichotomous items is a discrete variable while the expected response is continuous one"(p.153-155).

In order to apply asymptotic theory to the residual-based approach, first, the expected score and the observed score must be independent; and second, they must both be continuous variables(Smith, 1988, p.659; Smith, 1991a, p.547-548). In residual-based approaches, however, the asymptotic theory might not be appropriate for dichotomous items. According to Divgi(1986, p.293) who claim that the use of the normal approximation to a binomial distribution in this situation will produce misleading results. Therefore, some authors(Anderson, 1973; Gustafsson, 1980; van den Wollenberg, 1982) maintain that the likelihood ratio chi-square, rather than the Pearson chi-square, should be used, since the true distributional properties of residual-based IRT approaches are not known.

At this point, however, Smith(1990, p.79) argues that all applications of chi-square are approximations since real data never fit any ideal model. Smith and Hedges(1982) have compared the likelihood ratio chi-square (Anderson, 1973; Gustafsson, 1980; van den Wollenberg, 1982) and Pearson chi-square tests of fit in the case of the Rasch model. They found that the Pearson and likelihood chi-square fit statistics are highly correlated, almost 0.99. MacKinly and Mills(1985), on the other hand, compared four chi-square approaches and concluded that the likelihood ratio chi-square fit statistics(Bishop *et al.*, 1975) produced the least erroneous rejection of the null hypothesis than the other chi-square procedures(Bock, 1972; Yen, 1981; Wright & Mead, 1977), when the data fit the Rasch model. Reise(1990) reports that although the likelihood ratio chi-square test(Levin & Rubin, 1979; Drasgow *et al.*, 1985) is closer to the normal distribution than the chi-square test(Bock, 1972), the chi-square test is more sensitive to misfit under the three-parameter logistic model.

Waller(1981) suggests that in small sample

sizes, the likelihood chi square tests would perform better than the Pearson chi square tests.

Despite the problems concerning the residual based approach(Smith, 1991b, p.153 155), Wright and Stone(1979) state that the standardized residual is distributed more or less normally with a mean of about 0 and a variance of about 1.

Smith(1986, 1988, 1991)conducted a simu lation study to investigate the distributional properties of the standardized residuals and the sensitivity of the standardized residual to detect aberrant response patterns. In his study, he found that, when the data fit the Rasch model the standardized residuals are approximately normally distributed, and that Type I error rates can be used to identify aberrant response patterns. Based on his simulation study, Smith (1990) argued that:

". . .even though these statistics are not true chi-squares. . . Studies of the distributional properties of WP(Wright-Panchapakesan)statistics show that the tails of their distributions are regular enough to identify outliers reliably. There is no practical reason to use anything more complicated" (p.79).

Kogut(1988) investigated the asymptotic distribution of the between subtests person fit index(denoted BF)in which BF can be defined as a multiplication of Smith's(1985) UB(Unweighted Between fit) index by a constant:BF=(J 1)UB. Kogut proves algebraically that BF is asymptotically chi square in both cases in which the persons ability is known or is estimated by the ML(Maximum Likelihood) method from his or her response pattern.

Rogers and Hattie(1987) argued that the person fit indices derived from the Rasch model (the mean square residual, total and between t statistics) are inadequate measures of fit, and showed that these person fit indices lack sensitivity to detect aberrant response patterns from the Rasch model. Molenaar and Hoijtink(1990), however, indi cated that:

". . . it is not recommended to use person fit indices for tests of less than 15 or 20 items, because in very short tests the probabilities fluctuations will dominate the systematic information about the fit"(P.78).

Reise and Due(1991) also state that with short tests, fewer than 20 items, it is hard to find better detection for unusual response patterns. Hence, Rogers and Hattie's results (1987) should be considered tentative since they only used 15 items to investigate the sensitivity of person fit indices.

Smith *et al.*(1995) conducted a simulation study to investigate the effect of sample size on both these item fit mean squares, MS(UT) and MS(WT), and the t transformations of those mean squares. They found, for the item fit mean squares, that both unweighted and weighted fit mean squares did not perform well. In their study, for the unweighted mean square to get a Type I error rate of .04, a critical value of 1.3 would be needed with 150 persons, 1.2 with 500 persons, and 1.1 with 1000 persons. For the weighted mean squares, the Type I error rate(i.e., the probability of misclassification of a person as aberrant) in the null situation(i.e., no aberrant persons) would approximate 0.005 if a critical value of 1.2 were to be used.

This is similar to the findings reported in Smith *et al.*(1994c), in which they observed that a ±2 critical value, for the INFIT statistic, would yield a Type I error rate of approximately 0.015. This implies that too many aberrant persons will remain undetected. In conclusion, these results show that the critical value for the fit-mean square is sensitive to both the type of mean square and the sample size. Furthermore, the mean squares are more sensitive to sample size than the t-transformation, although a critical value for the t-transformation(also see Smith *et al.*, 1994c) also produces different Type I error rate for the two types of indices(e.g., unweighted vs. weighted), sample size(e.g., 150, 500, and 1000), and test lengths(e.g., 20 vs. 50).

As a consequence, the use of the mean square as the critical value to indicate aberrant responses implies that no aberrant persons will be classified as aberrant or vice versa.

Smith(1991a) observed that with different test-lengths both the unweighted and weighted fit indices appear to be distributed approximately normal as the test items increase. These fit indices increase negatively as the sample size increases. The differences, however, appear to be due to sampling variation. Furthermore, he conducted studies of the sensitivity of the unweighted and weighted fit indices in which he created three kinds of aberrant response patterns: random guessing, start-up problems, and differential item familiarity(item bias). For the guessing and start-up problems, the unweighted fit index seems to be more sensitive than the weighted fit index. For the item bias, both indices appear to be deficient in detecting those items.

Therefore, Smith(1991) recommends a combination of the two fit indices to detect various types of measurement disturbances.

Smith(1988) reported similar findings concerning the effect of test lengths and sample size on the Rasch standardized residuals. The values for the standardized residuals appear to be independent of the number of items, persons, and the dispersions of the item difficulties.

2. Likelihood-Based Approach

The next index is based on a likelihood function. To describe likelihood function in terms of the item response function, consider the probability

$$P(U_i|\theta) = P_i^{U_i} Q_i^{1-U_i}$$

where $Q_i = 1 - P_i$.

Then, the joint probability of responses can be expressed as

$$P(U_1,U_2,U_3,.....,U_n \mid \theta) = \prod_{i=1}^{n} P_i^{U_i} Q_i^{1-U_i}$$

The logarithm of the likelihood function is given by

$$\ln L(U_1,U_2,U_3,.....,U_n \mid \theta) = \prod_{i=1}^{n} [\, U_i \ln P_i + (1 - U_i) \ln Q_i \,]$$

Based on the above expression it is possible to compute the logarithm of the likelihood function at the maximizing value of θ, which is given by

$$L(\theta) = \sum_{i=1}^{n} \{ U_i [\ln P_i(\theta)] + (1-U_i) [\ln Q_i(\theta)] \}$$

where U_i is the dichotomous item response, $P_i(\theta)$ is the probability of a correct response

given θ, $Q_i(\theta) = 1 - P_i(\theta)$, and n is the number of items.

Suppose that an examinee with high ability responds to relatively easy items incorrectly, or an examinee with low ability responds to several difficult items correctly. Then, the likelihood of that response pattern will be small, which will produce a small value for $L(\theta)$, indicating an aberrant response pattern. Thus, a high value for $L(\theta)$ indicates good fit, whereas a low value for $L(\theta)$ indicates poor fit.

Levin and Rubin(1979) defined an 'appropriateness' index to measure how well a response pattern fits the model. Drasgow *et al.*(1987) have also advocated the use of the appropriateness index, noting that:

". . . The Neyman-Pearson Lemma asserts that maximum power is achieved by a likelihood ratio test. More specifically, let LN(x) and LA(x) denote the likelihoods of the data x under the null and alternative hypotheses, respectively. Then the Neyman-Pearson Lemma states that of all tests with a Type I error rate of α , none is more powerful than a test obtained from the likelihood ratio LA(x) / LN(x)"(p.61).

Drasgow *et al.*(1985), however, observed that mean $L(\theta)$ rises as ability increases, which means that the distribution of $L(\theta)$ is dependent on ability θ. Therefore, they proposed a standardized likelihood index, denoted as Lz

$$L_z = \frac{L(\theta) - E[L(\theta)]}{SD[L(\theta)]}$$

where

$E[L(\theta)]$ is the expected value of $L(\theta)$, and $SD[L(\theta)]$ is the standard deviation of $L(\theta)$.

The expected value of $L(\theta)$ is given by

$$E(L|\theta) = \sum_{i=1}^{n} \{ P_i(\theta)[\ln P_i(\theta)] + Q_i(\theta)[\ln Q_i(\theta)] \}$$

and the variance is

$$V(L|\theta) = \sum_{i=1}^{n} P_i(\theta)Q_i(\theta)\{ \ln [P_i(\theta) / Q_i(\theta)] \}^2$$

Mathematical proofs of these formulas can be found in Drasgow *et al.*(1985). From the above definition, the standardized likelihood index Lz is expected to be distributed approximately unit normal, and is independent of ability level.

Besides, from L_z, it is possible to identify two types of aberrant response patterns, inconsistent and hyper consistent response patterns(Reise &Due, 1991, p.219). That is, it is interesting to note that negative values of Lz will indicate response patterns that are unlikely, given the IRT model and the ability estimate; and positive values of Lz will indicate response patterns that are more consistent than the IRT model expected around the mean 0(Reise, 1990, p.129).

Drasgow *et al.*(1987)stressed two criteria, standardization and relative power, to evaluate the appropriateness index. In their words:

". . . *standardization* . . . It refers to the extent to which the conditional distribution (given particular values of the latent trait)of an index are invariant across levels of the latent trait . . . well-standardized indices . . . high rates of detection of aberrant response patterns . . . the

second consequence of the independence of ability and measured appropriateness for a well-standardized index is that it is easy to use in practice because index scores for individuals with different standings on the latent trait can be compared directly . . . *relative power* . . . Given a particular rate of misclassification of normal response patterns as aberrant(Type I error rate) . . . If a well-standardized index has acceptable power, then it can be used in operational settings" (p.60-61).

The above paragraphs, if we have a well-standardized index, suggest that Type I error rate can be used to identify examinee response patterns as aberrant or nonaberrant.

Several studies have been conducted to investigate the distributional properties of appropriateness indices and their sensitivity to detect aberrant response patterns(see Drasgow, 1982; Levine & Drasgow, 1982; Drasgow *et al.*, 1984; Drasgow & Levine, 1986; Rudner, 1983; McKinley & Mills, 1985; Reise, 1990; Reise & Due, 1991). Drasgow *et al.*(1984) indicate that the distribution of the standardized appropriateness index, denoted Lz, is close to the standard normal distribution at all ability levels. Furthermore, Drasgow *et al.*(1987) show that the standardized appropriateness index has a high rate of detection for aberrant response patterns with low or high ability levels.

In contrast, Harnisch and Tatsuoka's data(1983) suggest that, although the Lz index follows the normal distribution, it appears to be related to the total score. Reise(1990) reports that the Lz index appears to have a weak relationship with both ability and item difficulty levels. This finding implies that, because of the independence of the Lz index and ability distribution, the Lz index value for individuals with different total score(or different ability levels) can be compared directly. Rudner(1983), in a Monte Carlo study, also reports that the likelihood statistic, L_z, appears to be performing well in identifying individuals with aberrant response patterns.

Although the standardized likelihood index has been utilized widely with the 3-parameter model, this index can be used in any of the logistic, normal ogive, or other parametric models(Drasgow & Levine, 1986, p.60). Further, Lz can also be used to study item-fit indices(Reise, 1990, p.129).

Reise and Due(1991) examined the effects of test length on the detection of aberrant response patterns with the standardized likelihood index Lz based on the 3-parameter IRT model. They found that hit rates (identifying aberrant response patterns) increase substantially as test-length increased. Especially, detecting aberrant response patterns was problematic with less than 20 items. They also observed that the detection is greater for positive values of θ. Further, they demonstrated that the specific IRT model, the distribution of difficulty, and ability level all influence Lz sensitivity to detect aberrant response patterns.

It is interesting to note that in their simulation research, Lz appears to be a more powerful detector for data that fit the 2-parameter model rather than the 3-parameter model. However, the Reise and Due(1991) findings are quite different compared with Reise's previous study(1990). Reise compared a χ^2 fit index with a loglikelihood-based index under the three-parameter model. Reise observed that

both the Lz and χ^2 index are not related with person ability and item difficulty. Reise found that the loglikelihood index is closer to the normal distribution than the χ^2 fit index; although, in terms of power, the χ^2 fit index is more sensitive to aberrant response patterns. In other words, the χ^2 fit index identified more items and examinees as misfitting than did the Lz fit index.

However, Reise also suggested that long tests will tend to inflate the χ^2 statistic. Noonan *et al.*(1992) examined the effect of test length and IRT model on the three indices; Lz , ECI4, and W in which they used two kinds of IRT models(2-parameter and 3-parameter model)and two test lengths(40 items and 80 items). They, using Pearson correlations, found that the three indices are not dependent on the ability θ. They performed multivariate analyses to investigate the effect of test length and IRT model on the three indices at the three false-positive rates, 0.01, 0.05, 0.10 in which they found that Lz index was highly affected by test length and IRT model, significant at the 0.01 level. Their findings partially support the results of Reise and Due(1991), in terms of test length and choice of IRT model. Further, they found a high relationship between Lz and W ranging from -0.94 to -0.95. This relationship was consistent with the findings of Smith and Hedges(1982).

3. Extended-Caution Index Approaches

The third category of indices, Extended Caution Indices, is referred to as mathematical extensions of the Sato Caution Index(1975) utilizing item response theory. The Sato Caution index, for dichotomously scored items, can be illustrated using a student-by-item response matrix(raw x columns) and drawing what is called the Student-Problem or S-P curve, in which students are located in descending order according to their total scores, and items are arranged in ascending order from left to right according to difficulty level(For details, see Tatsuoka & Linn, 1983). The Sato Caution Index can be expressed as the ratio of two covariances(Tatsuoka, 1984).

$$C_i = 1 - \frac{cov(y_i, y_.)}{cov(x_i, y_.)}$$

where

$y_i = (y_{i1}, \ldots, y_{in})$ is the binary-scored observed response vector i,

$x_i = (x_{i1}, \ldots, x_{in})$ is the reversed Guttman vector with the total score of examinee i,

$y_. = y_{.1}, \ldots, y_{.n}$

Equation described above compares the similarity between observed vector y_i and its Guttman scale vector x_i , respectively with the column sum vector of correct answers for the items. If the response patterns yield substantial deviations from Guttman scaling patterns, the index will produce a high index value. No theoretical justifications for the critical value for aberrant patterns, however, have been provided, although the cut value of 0.5 was originally suggested by Sato(1990). Further, the Sato Caution Index is group dependent, which implies that the index value will not be invariant across different samples of examinees and items. Therefore, Tatsuoka and Linn(1983) proposed another set of indices called Extended Caution Indices(ECIs). ECIs are an extension of the Sato's Caution

Index using the IRT model. Tatsuoka and Linn(1983) showed certain correspondence between the student-problem curve(S-P curve)and test response curve(TRC) and group response curve(GRC). In that study, they found that person response curves(PRC), for the Rasch model, have shown monotonically decreasing functions along the difficulty level. However, for the 2-parameter and 3-parameter model, the PRC did not appear to be represented by a smooth, monotonically decreasing curve(Tatsuoka & Linn,1983, p. 88). This is a major advantage of the Rasch model because the extended caution indices may be viewed as a linear transformation of the covariance or correlation between a persons response pattern and a theoretical curve(Tatsuoka & Linn,1983, p.95)(viz, the PRC, as in the case of ECI4, or the GRC, as in the case of ECI2).

Although Tatsuoka and Linn(1983)described five ECIs, the present study will consider only two of the ECIs(ECI2z and ECI4z) since Tatsuoka(1984) suggested that:

". . . ECI4 and ECI6 have identical standardized forms . . . the relationship between ECI1 vs ECI2 . . . . correlate very highly . . . we drop ECI1 . . . and recognize ECI2z and ECI4z . . . as representative indices among the family of ECI indices"(p.104).

Therefore, the two types of extended caution indices selected for this study are Extended Caution Index Two(ECI2) and Extended Caution Index Four(ECI4). ECI2 and ECI4 are given by (Tatsuoka, 1984, p.98)

$$ECI2 = 1 - \frac{cov(y_i, G)}{cov(P_i, G)}$$

$$ECI4 = 1 - \frac{cov(y_i, P_i)}{cov(G, P_i)}$$

where

y_i is the observed response vector for i

P_i is the probability vector for the ith row, and

G is the GRC vector, which is the average of the column-sum vector of P_{ij}.

ECI2 compares the similarity of the group response curve(GRC) with observed response and probability vector. On the other hand, ECI4 compares the similarity of the person response curve (PRC)with observed response and group response vectors.

Like other unstandardized indices, the above two extended caution indices are dependent on ability level, which makes it impossible to compare two values obtained from two students at two different ability levels.

According to the study conducted by Tatsuoka and Tatsuoka(1982b), unstandardized extended caution indices appear to be functions of ability and have U-shaped trend curves. Therefore, the standardized caution indices, ECI2z and ECI4z, mathematically derived by Tatsuoka (1984) are given by the following formulae (Drasgow *et al.*, 1987, p. 65):

$$ECI2_z = \frac{\sum_{i=1}^{n} [P_{ij}(\theta) - U_{ij}][G_i - \bar{G}]}{[\sum_{i=1}^{n} P_{ij}(\theta)Q_{ij}(\theta)(G_i - \bar{G})^2]^{\frac{1}{2}}}$$

$$ECI4_z = \frac{\sum_{i=1}^{n} [P_{ij}(\theta) - U_{ij}][P_{ij}(\theta) - \bar{P}_j]}{[\sum_{i=1}^{n} P_{ij}(\theta)Q_{ij}(\theta)(P_{ij} - \bar{P}_j)^2]^{\frac{1}{2}}}$$

where

i = Item (1. . . n), j = Person (1. . . N),

U_{ij} is the observed response,
P_{ij} is the probability of a correct response, $Q_i(\theta) = 1 - P_i(\theta)$.

$$G_i = \frac{1}{N}\sum_{j=1}^{N} P_{ij}(\theta)$$

$$\overline{G} = \frac{1}{n}\sum_{i=1}^{n} G_i$$

$$\overline{P}_j = \frac{1}{n}\sum_{i=1}^{n} P_{ij}(\theta)$$

Some studies(Tatsuoka & Tatsuoka, 1982b; Tatsuoka & Linn, 1983; Harnisch & Tatsuoka, 1983; Tatsuoka, 1984) have shown that ECI2z and ECI4z follow a distribution that is close to the normal distribution and show little relationship with ability level.

Birenbaum(1985)compared nine indices based on the 2-parameter logistic model in which Lz, ECI2z, and ECI4z performed well, compared to the other indices in terms of ability-level variations and measurement disturbances. Drasgow *et al.*(1987), however, reported that although ECI2z shows better detection of aberrant response patterns than other indices, the ECI2z and ECI4z indices are not well-standardized across ability-level groups. Harnisch and Tatsuoka(1983) have examined correlations among fourteen indices in which the correlation between ECI2z and ECI4z was revealed to be 0.94. However, the standardized ECI2z yielded the least relationship with the total score.

Tomsic *et al.*(1987) studied the distributions of ECI2z and ECI4z over 1437 students. In that study, their specific research question was whether nonnormal distributions will tend toward normality when the worst fitting items are removed from the test. Unfortunately, the removal of the worst fitting items did not appear to move the distributions toward the normal. One possible explanation of this finding is that the ECI distributions may not be normal. Noonan *et al.*(1992), however, reported that ECI4z has shown to be the closest to the normal distribution, showing less skewness and kurtosis than Lz and W. Further, using multivariate analysis of variance, both test length and IRT model did not significantly affect ECI4z at the .05 false positive rate.

According to Harnisch and Tatsuoka's study(1983), both ECI2z and ECI4z yielded high intercorrelations with Lz, -0.92 and -0.91 respectively; and least intercorrelations with the unweighted Rasch fit mean-square index, namely 0.08 and 0.05 respectively. Also, ECI2z revealed the least relationship with the total score among 14 indices analyzed. On the other hand, Birenbaum(1985) reported that the unstandardized likelihood index denoted Lz showed the least relationship(r=.007) with the total score; whereas ECI4z revealed the highest relationship(r=.223) with the total score among 9 indices.

III. Recommendations

In the light of the finding of this study, the following recommendations may be useful in carrying out similar studies in the future.

First, it is recommended that this study be replicated to investigate the properties of fit indices based on a polytomous scoring model. As pointed out in the literature review, various fit measures of aberrant response patterns have emerged and have been developed. However, the distributional properties, sensitivity and interpretation of fit

indices for the polytomous model is a recent development(For example, see Smith, 1994b, 1996). Therefore, this issue needs to be addressed by future studies.

Second, The natures of IRT based fit indices should be taken into consideration when a choice of one of them is made in a specific situation. For example, Tatsuoka and Linn(1983) suggested that the $ECI4_z$ index can be an useful index in selected diagnostic situations. Additional work is needed for selecting the appropriate index that fits the specific needs of person(item) response aberrance under various types of measurement disturbances.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123 140.

Andrich, D.(1988). *Rasch models for measurement*. Sage University Paper series on Quantitative Applications in the Social Science, series no. 07 068. Beverly Hills:Sage Publications.

Baker, F. B.(1992). *Item response theory: parameter estimation techniques*. New York: Marcel Dekker, Inc.

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

Bishop, Y., Fienberg, S., & Holland, P.(1975). *Discrete multivariate analysis: Theory and practice*. Cambridge MA: The MIT Press.

Blixt, S. L., & Dinero, T. E.(1985). An initial look at the validity of diagnoses based on Satos caution index. *Educational and Psychological Measurement*, 45, 293 299.

Bock, R. D.(1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29 51.

Croker, L., & Algina J.(1986). *Introduction to classical and modern test theory*. Fort Worth: Holt, Rinehart, and Winston, Inc.

D'Costa, A.(1993). *Validity of the W, B, and Sato Caution Indexes*. Paper presented to the Seventh International Objective Measurement Workshop, Atlanta, Ga.

Divgi, D. R.(1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, 23, 283 298.

Donlon, T. F., & Fischer, F. E.(1968). An index of an individuals agreement with group determined item difficulties. *Educational and Psychological Measurement*, 28, 105 113.

Donlon, T.F., & Rindler, S. E.(1979). *Consistency of item difficulty for individuals and groups in the Graduate Record Examination*. Paper presented at the annual meeting of the American Educational Research Association. San Francisco.

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

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 Psycholo*

gical Measurement, 11, 59 79.

Drasgow, F., Levine, M. V., & Williams, E. A.(1984). *Appropriateness measurement with polychotomous item response models and standardized indices*. Urbana, Illinois: Model Based Measurement.

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 Psycho logy*, 38, 67 86.

Glaser, R.(1949). A methodological analysis of the inconsistency of responses to test items. *Educational and Psychological Measurement*, 9, 721 739.

Glaser, R.(1952). The reliability of incon sistency. *Educational and Psychological Measurement*, 60 64.

Gustafsson, J E.(1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, 33, 205 233.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J.(1991). *Fundamentals of item response theory*. Newbury park, California: SAGE Publications, Inc.

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

Harnisch, D.L., & Linn, R.L.(1981b). *Identification of aberrant response patterns*(Final Report on Grant No. G 80 0003). Washington. D. C. : National Institute of Education.

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*. Vancouver, B.C.: Educational Research Institute of British Columbia.

Kane, M. T., & Brennan, R. L.(1980). Agreement coefficients as indices of dependability for domain referenced tests. *Applied Psychological Measurement*, 4, 105 126.

Kogut, J.(1988). Asymptotic distribution of an IRT person fit index(Project psychometric aspects of item banking No. 38, Research Report 88 13). Netherlands: Twente Univ. Dept.of Education(ERIC Document No. ED 309 186).

Lee, P., & Suen, H. K.(1994). Consequences of removing subjects in item calibration. In Wilson,M. (Ed.), *Objective measure ment: theory into practice (2)*. Norwood, NJ: Ablex Publishing Corporation.

Levine, M. V., & Drasgow, F.(1982). Appro priateness 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.

Linacre, J. M., & Wright, B .D.(1993). *BIGSTEPS: Rasch Model Computer Program*. Chicago: MESA Press.

McKinley, R. L., & Mills, C. N.(1985). A comparison of several goodness of fit statistics. *Applied Psychological Measure ment*. 9, 49 57.

Meijer, R.R.(1996). Person fit research: an introduction. *Applied Measurement in Education*, 9, 3 8.

Meijer, R.R., & Sijtsma, K.(1995). Detection

of aberrant item score patterns: a review of recent developments. *Applied measure ment in education*, 8, 261 272.

Mislevy, R. & Bock, R. D.(1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, 42, 725 737.

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

Mosier, C. I.(1940). Psychophysics and mental test theory: fundamental postulates and elementary theorems. *Psychological Review*, 47, 355 366.

Noonan, B. M., Boss, M. W., & Gessaroli, M. E.(1992). The effect of test length and IRT model on the distribution and stability of three appropriateness indexes. *Applied Psychological Measurement*, 16, 345 352.

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.

Reise, S. P., & Due, A. M.(1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, 217 226.

Rogers, H. J., & Hattie, J. A.(1987). A Monte Carlo investigation of several person and item fit statistics for item response model. *Applied Psychological Measurement*, 11, 47 57.

Rudner, L. M.(1983). Individual assessment accuracy. *Journal of Educational Measure ment*, 18, 171 182.

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

Sato, T.(1990). *An introduction to educational information technology*. NEC: Technical College, Kanagawa, Japan.

Smith, R. M.(1982). *Detecting measurement disturbances with the Rasch model*. Ph. D. dissertation, University of Chicago.

Smith, R. M.(1983). *Test fairness is a personal issue !*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Smith, R. M.(1985). Validation of individual test response patterns. *International Encyclopedia of Education*(5410 5413). Oxford: Pergamon Press.

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

Smith, R. M.(1988). The distributional properties of Rasch standardized residuals. *Educational and Psychological Measure ment*, 48, 657 667.

Smith, R. M.(1990). Theory and practice of fit. *Rasch Measurement Transactions*, 3(4), 78 79.

Smith, R. M.(1991a). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measure ment*, 51, 541 565.

Smith, R. M.(1991b). *IPARM: item and person analysis with the Rasch model*. computer program, Chicago: MESA Press.

Smith, R. M.(1994a). A comparison of the power of Rasch total and between item fit statistics to detect measurement disturbances. *Educational and Psycholo gical Measurement*, 54(1), 42 55.

Smith, R. M.(1994b). Detecting item bias in the Rasch rating scale model. *Educational and Psychological Measurement*, 54(4), 886 898.

Smith, R.M.(1996). Polytomous Mean Square fit statistics. *Rasch Measurement Transac*

tions, 10(3), p.516 517.

Smith, R. M., & Hedges, L. V.(1982). A comparison of likelihood ratio χ^2 and Pearsonian χ^2 tests of fit in the Rasch model. *Education, Research and Perspectives*, 9, 41 54.

Smth, R. M., Schumaker, R. E., & Busch, M. J. (1994c). *Examining Replication effects in Rasch fit statistics.* Paper presented at the annual meeting of the American Educa tional Research Association, New Orleans.

Smth, R.M., Schumaker, R.E., & Busch, M.J.(1995). *Using item mean squares to evaluate fit to the Rasch model.* Paper presented at the annual meeting of the American Educational Research Associ ation, San Francisco.

Suen, H. K.(1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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 patterns: links between two general approaches and potential applications. *Applied Psycholo gical Measurement*, 7, 81 96.

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

Tatsuoka, K. K. & Tatsuoka, M.M.(1982b). *Standardized extend caution indices and comparisons of their rule detection rates.* Urbana, Illinois: Computer based Education Research Lab.

Tomsic, M., Schellenberg, S., & Mittman, A.(1987). *The effect of poor fitting items on the distributions of extended caution indices.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

van den Wollenberg, A. L.(1982). Two new test statistics for the Rasch model. *Psychometrika*, 47, 123 140.

van der Flier, H.(1977). Environmental factors and deviant response patterns. In Y.H. Poortinga (Ed.), *Basic problems in cross cultural psychology.* Amsterdam: Swets and Seitlinger, B. V.

van der Flier, H.(1982). Deviant response patterns and comparability of test scores. *Journal of Cross Cultural Psychology*, 13, 267 298.

Waller, M. I.(1981). A procedure for comparing logistic latent trait models. *Journal of Educational Measurement*, 18, 119 125.

Wilson, M. (Ed.). (1994). *Objective measure ment: theory into practice* (2). Norwood, NJ: Ablex Publishing Corporation.

Wright, B. D.(1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97 116.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Mead, R. J.(1977). *BICAL: Calibrating items and scales with the Rasch model.* Chicago: MESA Press.

Wright, B. D., & Panchapakesan, N. A.(1969). A procedure for sample free item analysis. *Educational and Psychological Measure ment*, 29, 23 48.

Wright, B. D. & Stone, M. H.(1979). *Best test design.* Chicago: MESA Press.

Yen, W. M.(1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245 262.

초록

문항반응이론에 기초한 적합도 지수의 탐색

설 현 수

(중앙대학교)

지금까지 비정상적인 문항반응을 나열하고 분석하려는 지속적인 연구들이 수행되어왔다. 그 이유중에 하나는 피험자(또는 문항)가 속한 집단의 특성에 대한 이해와 더불어 검사조정과정에서 비정상적인 문항반응을 보인 피험자(또는 문항)을 분리하고 처치함으로 인해서 보다 나은 검사문항 개발을 기할 수 있다는 이점 때문에 많은 연구자들이 관심을 가져 온것이라 볼 수 있다.

본 연구의 목적은 비정상적인 반응형태를 분석하기 위해서 지금까지 제시된 다양한 적합도 지수에 대한 장·단점을 소개함과 아울러서, 비정상적인 문항반응과 관련해서 각각의 적합도 지수들이 얼마나 민감한지를 문헌연구를 통해 밝혀보고자 하는데 목적을 두고 있다.

Key Words : 문항반응이론, Rasch 모형, 적합도 지수, 비정상적인 문항반응.