Order Number 9129703

A study of selected response aberrance indices as alternatives to the standard error of measurement in interpreting individual test scores

Shishido, Judy A., Ph.D.

University of Hawaii, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A STUDY OF SELECTED RESPONSE ABERRANCE INDICES AS ALTERNATIVES TO THE STANDARD ERROR OF MEASUREMENT IN INTERPRETING INDIVIDUAL TEST SCORES

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION
OF THE UNIVERSITY OF HAWAII IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN EDUCATIONAL PSYCHOLOGY

MAY 1991

By

Judy A. Shishido

Dissertation Committee:

Harold I. Ayabe, Chairman
Mary E. Brandt
Michael Heim
Morris K. Lai
Machiko Netsu
Dorothy T. Shibano

ACKNOWLEDGMENTS

I wish to acknowledge, with thanks, the assistance and aloha that I received from the faculty, staff, and students of the University of Hawaii at Manoa's Department of East Asian Languages and Literatures. In particular, I wish to thank Evelyn Nakanishi and Ray Kaneyama, as well as Dennis Ogawa, James Unger, Agnes Niyekawa, Kyoko Hijirida, Elizabeth Kishimoto, Myra Taketa, and Shirley Kim.

I would also like to thank the members of my committee, not only for guiding my dissertation research to a happy conclusion but also for their many kindnesses along the way: to Dorothy Shibano and Machiko Netsu for their insights into foreign language teaching and assessment of students' foreign language knowledge, and their advice on data gathering in the East Asian Languages and Literatures Department setting; to Betsy Brandt for guidance on research design and on gathering qualitative data; to Michael Heim for assistance with the complex mathematical concepts in this study and for facilitating the graduate seminars from which most of the ideas in this study evolved; to Morris Lai for his careful critique of this document and for his advice on the myriad of practical considerations from instrument design to choice of computer software; and to my chairperson, Harold Ayabe, for investing several years of his life into overseeing every aspect of my doctoral program.

ABSTRACT

The purpose of this study was to discover any relationships between three aberrant test response indices, the Modified Caution Index (MCI), the Person Average R (PAR), and the item-response theory-based standardized logistic maximum likelihood function (Z3); and four hypothesized reasons for aberrant response patterns, curricular differences, differences in test-taking skills, differences in motivation, and differences in consistency of academic performance. Two secondary purposes of this study were to evaluate the effectiveness of the aberrance indices as aids in interpreting test scores, and to investigate the occurrence of sandbagging on a placement test. Data from the University of Hawaii at Manoa's Japanese Language Placement Test was used, along with survey data obtained from students taking the Placement Test, students enrolled in UH Manoa Japanese language courses, Hawaii high school Japanese language teachers, and UH Manoa Japanese language instructors. Relationships were found between MCI and PAR and curricular differences and differences in consistency of academic performance, and between all three indices and differences in motivation. It was estimated that between four and ten percent of test takers engaged in sandbagging on the Placement Test. The use of either MCI or PAR was recommended as aids in interpreting Placement Test scores.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

One objective in administering a test is to obtain a measure of an individual's ability level in order that some decision regarding that individual can be made. Among the many kinds of test-based decisions are those made for educational classification and placement into special programs (Reschly, 1981), for admission into college (Hargadon, 1981), for employment (Tenopyr, 1981), and for licensure and certification (Shimberg, 1981).

In classical test theory, use of the standard error of measurement (SEM) is recommended for interpreting the reliability of individual test scores (Anastasi, 1982, pp. 125-127; Nunnally, 1978, pp. 218-219; Payne, 1974, pp. 268-269). SEM is the expected standard deviation of scores, due to errors of measurement, for an individual if that individual were to take a large number of randomly parallel tests (Nunnally, 1978, p. 218). It is used to set a confidence zone about the individual's estimated true score. Thus, for example, if an individual with an estimated true score of 20 were to take many, many parallel tests with a calculated SEM of 3, one would expect that individual to earn a raw score between 14 and 26 on about 95 percent of those parallel tests.

The common practice is to apply a single SEM in interpreting all individual test scores. However, it has been demonstrated that SEM varies at different score levels (Mollenkopf, 1949; Lord, 1955), and indeed a recent issue of *Standards for Educational and Psychological Testing* (Committee of AERA, APA, and NCME to Develop Standards for Educational and Psychological Testing, 1985) recommended that "standard errors of measurement should be reported at critical score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported for score levels at or near the cut score" (Standard 2.10, p. 22). One researcher (Lord, 1985) even described a method of finding the social cost associated with errors of measurement. He

concluded that "social losses arising from errors of measurement will be high for examinees near the cutting score. Social losses will be near zero for examinees far from the cutting score, since decisions about these examinees will not be changed by small errors in their scores." He recommended that resources be mobilized to accurately measure ability levels where most people were to be found. Thus researchers agree that an average SEM is sometimes inadequate and a better interpretation of individual test scores can be achieved by using SEMs estimated at different score levels.

Methods for estimating SEM at different score levels have been suggested (Thorndike, 1951; Lord, 1955; Livingston, 1982; Jarjoura, 1986). Lord's (1955) binomial formula, for example, follows:

$$SE = \left[\frac{X\ (K-X)}{K - 1}\right]^{1/2}$$

where X= number right score

K= number of test items

This formula yields a SEM at each score level for a test of K items. Calculation of this particular formula requires only the knowledge of the number of test items. Plotting the score-level SEMs from this formula results in a graph with the shape of a parabola, concave down, similar to a graph of p x q (where p = the proportion correct and $q = 1 - p$). A 54-item test, for example, would have the shape shown in Figure 1. At the extreme score levels the SEMs are smaller than for middle score levels because there is less room for variation in error at very high or very low scores.

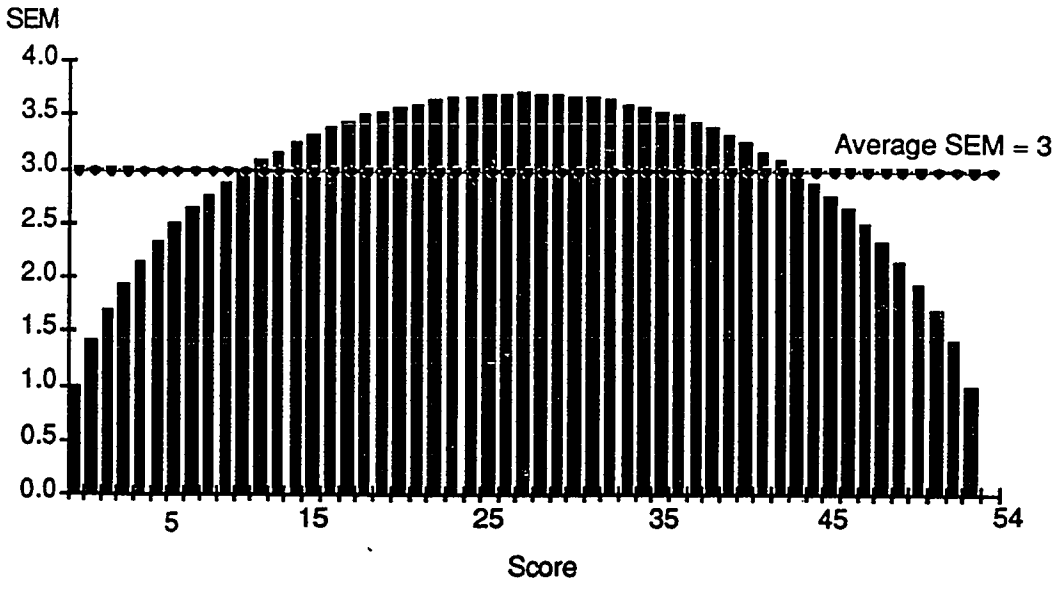


Figure 1 . SEM calculated for a 54-item test using Lord's Binomial

If the test in the above example had a reliability coefficient of about .92, the average SEM would be approximately 3. A horizontal line at SEM=3 has been drawn in on Figure 1. As the reader can see, application of the score level SEM is likely to result in better accuracy than applying the average SEM of 3 to the raw scores of all test takers.

Some researchers have done comparative studies of various methods for calculating SEM at different score levels (Feldt, Steffen, & Gupta, 1985; Blixt & Shama, 1986). What these SEM calculation methods have in common is that they all yield one SEM for each score level, but none can estimate an SEM for each individual. If there were 50 test takers who scored 20, the above methods would apply a single SEM to all of those 50 people.

Since SEM has been shown to vary by score level, the next question is whether or not a single SEM within the score level is adequate. The author believes that it is not. Test theorists acknowledge that some people are less reliable than others and various theorists have discussed factors which affect the reliability of persons. In his discussion of sources of error, for example, Thorndike (1949) included "momentary 'set' for a particular test," "fluctuation and idiosyncrasies of human memory," and "unpredictable fluctuations in attention and accuracy..." as temporary and specific characteristics of the individual which contribute to the variance of his or her test scores. In his discussion of the meaning of test reliability, Cronbach (1947) used a hypothetical example of scores obtained from numerous retesting of an individual. From this example he delineated within-persons sources of error variance which may be attributable to "momentary inattention, guessing, and other random variables" (Cronbach, 1947, p. 4). Lumsden (1977, 1978) theorized about three kinds of change which affect a person's ability over time: long-term developmental trend; short term "swells" which are temporary elevations or depressions of ability and which may occupy hours or days; and, "tremors" which are rapid momentary fluctuations in the level of ability and are essentially random. According to Lumsden (1977, 1978), plots of the proportion of items passed at different item difficulty levels by each subject enable comparisons of person reliabilities.

Some attempts by researchers interested in personality assessment have been made to define individual reliability as a trait and to measure it as such. As with researchers concerned with achievement tests, these researchers questioned whether test results for an extremely inconsistent individual can be interpreted in the same way as test results for an individual whose responses are consistent (Glaser, 1949). Fiske and Rice (1955), for example, defined intra-individual variability on personality tests as the difference between two responses of an individual at two points in time where the individual is exposed to the same stimuli and the testing conditions are the same on both occasions, i.e., test-retest

stability of an individual's responses. They also assumed that intra-individual response variability is not random, but rather a lawful, cross-situational trait. Consistency was used to mean stability over time and across situations. Fiske's research, however, led him to conclude that there is no evidence of a general factor of variability (Mitra & Fiske, 1956).

Other researchers in personality testing also found no evidence of consistency as a cross-situational trait (Rundquist, 1950; Rorer, 1965; Goldberg 1978; Mischel & Peake, 1983), while others continue to argue persuasively for the existence of consistency as a personality trait (Bem & Allen, 1974; Bem, 1983). One of the reasons for the conflicting research results is that a variety of methods have been used in this area as well as a variety of instruments. Block (1968) discussed four reasons for the apparent inconsistencies in the field: first, behaviors selected for study by the researcher may not have been significant or salient for the subjects; second, formulations of personality which are relatively context blind might have resulted in behaviors which appeared inconsistent; third, behaviors which are being related may not be mediated by the same underlying variables; and fourth, when an individual reaches certain personal limits, previous behavioral consistencies may break down, causing the individual to begin to appear inconsistent.

Though researchers in this area do not agree on the existence of consistency as a cross-situational personality trait, they do agree on the stability of individual differences in consistency in retest situations (Glaser, 1952; Mitra & Fiske, 1956; Raine & Hills, 1959; Weksel & Ware, 1967; Hendel & Weiss, 1970; Parker, 1971; Green, 1979; Whitely, 1978; Michel & Peak, 1983; Bond, 1986). For example, in his research, Glaser (1952) administered a test three times to his subjects and counted the number of answer choices changed from the first to the second testing, and from the second to the third testing. He used these counts as a measures of consistency. The two counts were correlated at about r=.65. Other researchers (Weksel & Ware, 1967; Hendel & Weiss, 1970) have shown a relationship between test-retest reliability and consistency as measured by circular triad

scores. Glaser's measure as well as circular triad scores are too cumbersome for applied use. What is needed is an index that measures individual consistency which can be calculated from a single testing. Such indices have been proposed over the past twenty years and will be discussed in the next section.

Individual Reliability

What the practitioner might find useful is a measure of an individual's personal error variance. There are, at present, many such indices. Among those loosely described as group-dependent aberrant response pattern indices are the personal biserial (Donlon & Fischer, 1968), the agreement coefficient (Kane & Brennan, 1980), the agreement index (McQuitty, 1956), Sato's caution index (Harnisch & Linn, 1981), the norm-conformity and the individual consistency indices (Tatsuoka & Tatsuoka, 1982, 1983), and the modified caution index, Harnisch and Linn's (1981) modification of Sato's caution index. There are also many indices which are based on Item Response Theory (IRT). These include the unweighted and the weighted total fit mean square, U1 and W1 (Wright & Panchapakesan, 1969), the unweighted and the weighted Birnbaum model fit mean square, U3 and W3 (Rudner, 1983), the response pattern likelihood using the Birnbaum model, L3 (Rudner, 1983), the dichotomous and the polychotomous logistic likelihood functions, l_o and l_{oh} (Drasgow, Levine, & Williams, 1985), the standardized dichotomous and the standardized polychotomous logistic likelihood functions, Z3 and Z_h (Drasgow, Levine, & Williams, 1985), the optimal index (Drasgow & Levine, 1986; Levine & Drasgow, 1988), the normalized Jackknife variance estimate, JK (Drasgow, Levine, & McLaughlin, 1987), the item-option variance, IOV (Drasgow, Levine, & McLaughlin, 1987), the extended caution indices (Tatsuoka & Linn, 1983), and the standardized extended caution indices (Tatsuoka, 1984). There is also the Person Average R (PAR), based on classical test theory, which was recently introduced by Ayabe and Heim (1987).

Though many indices have been proposed, much more research is needed before any of the indices can be used by the practitioner. One study, done by Harnisch and Linn (1981), was a comparative study of a number of group-dependent indices using data from the 1978 Illinois statewide assessment program. They found that except for the agreement index, all indices studied were highly intercorrelated. However, with the exception of the modified caution index (MCI), all of the indices were also correlated with total score. As Harnisch and Linn (1981) found MCI to be least confounded with total score, they selected MCI for further analyses.

According to Harnisch and Linn (1981), MCI is an aberrant response pattern index because it can be used to identify individuals whose response patterns deviate from an expected Guttman-scale pattern (Guttman, 1941), that is, one in which an individual who responds correctly to a particular item will answer all easier items correctly. The measure of an item's difficulty in this case is based on the number of test takers who answer that item correctly. An item's difficulty may vary from one group of test takers to another. As such, the MCI is group dependent.

IRT-based appropriateness indices on the other hand, "measure the goodness of fit between the individual examinee's item-by-item pattern of responses and a specific psychometric model. In general, these indices indicate the extent to which examinees of equal ability differ in their pattern of responses" (Harnisch, 1983, p. 194). Specific psychometric models referred to in this definition include the one-parameter (Rasch) model and the three-parameter logistic model, among others.

The similarities between the group-based aberrant response pattern indices and the IRT-based approach were noted by Tatsuoka and Linn (1983) who discussed the correspondences between the IRT and Sato's approaches to identifying individuals with unusual response patterns.

Of the many IRT-based indices available, Z3 has been shown to be one of the most effective among appropriateness indices for detecting spuriously high test scores (Drasgow & Levine, 1986).

Another index of interest is the one proposed by Ayabe and Heim (1987). Results of a preliminary study conducted by the author indicate that PAR may be a better index than MCI. In the preliminary study, PAR was calculated on the sample data set in Harnisch and Linn (1981) and correlated with the measures of MCI for the sample data set which was presented in that article. A correlation coefficient of -.98 was found, but it was apparent in a plot of the two indices that PAR provided finer discrimination among the 18 subjects in the sample data set, than did MCI. Furthermore, the correlation between PAR and total score was .25, and the correlation between MCI and total score was -.25, the fairly low correlations indicating that these indices measure something other than what is measured by total score.

MCI

The Modified Caution Index (MCI) is an example of a group-dependent aberrant response pattern index. The following is the equation for calculating MCI:

$$C_i^* = \frac{\sum_{j=1}^{n_i.} (1-u_{ij})n_{.j} - \sum_{j=n_i.+1}^{J} u_{ij}n_{.j}}{\sum_{j=1}^{n_i.} n_{.j} - \sum_{j=J+1-n_i.}^{J} n_{.j}}$$

where

i = 1, 2, ... I, indexes the examinee

j = 1, 2, ... J, indexes the item

u_{ij} = { 1 if examinee i answers item j correctly,
0 if examinee i answers item j incorrectly}

$n_{i.}$ = total correct for the ith examinee

$n_{.j}$ = total number of correct responses to the jth item

The following example will perhaps help to illustrate that MCI contains information about the test taker which is not contained in his or her total score.

Table 1
MCI Example
Five-Item Test Administered to Seven People

Person No.	Response Pattern	Total Score	MCI
1	1 1 1 1 0	4	0.00
2	1 1 1 0 0	3	0.00
3	1 1 1 0 0	3	0.00
4	1 1 0 0 0	2	0.00
5	1 0 1 0 0	2	0.17
6	0 0 0 1 1	2	1.00
7	0 1 0 0 1	2	0.50

In the example above, the items on the five-item test are ordered from left to right in order of difficulty so that the response patterns of the first four people in the example conform to a Guttman scale where the subjects have correctly answered the items, until the number of their total scores. These four people all have MCIs of 0. Person 5, however, answered the easiest item correctly, missed the next item and then correctly answered the third item. Person 5's MCI is .17. Person 6 missed the first three easiest items but correctly answered the last two, most difficult items. This suspicious response pattern is marked with a MCI of 1.00 which indicates the aberrant pattern. Person 7 also has a suspicious response pattern which is so indicated by a MCI of .5. From this example it can be seen that although persons 4 through 7 all have the same total score, calculation of MCIs alerts the practitioner to the unusualness of the scores of persons 6 and 7.

MCI differs from Sato's Caution Index in that it yields an index which has a lower bound of 0 and an upper bound of 1, with larger values of MCI indicating individuals with aberrant response patterns. In their 1981 study, Harnisch and Linn used a MCI cutoff point of .3. They found MCI to be useful in identifying differences in curriculum among schools.

Other uses of aberrant response indices described by Harnisch (1983) include identification of students for whom special caution is needed in interpreting their total correct score, and identification of differences in instructional coverage between classrooms. He also described a system of classifying students using total score and MCI in the following combinations:

Type A = high test performance (greater than 50% of items correct)
and low MCI (less than or equal to .3)

Type B = high test performance (greater than 50% of items correct)
and high MCI (greater than .3)

Type C = low test performance (less than or equal to 50% of items correct) and low MCI (less than or equal to .3)

Type D = low test performance (less than or equal to 50% of items correct) and high MCI (greater than .3)

Classified thusly, a student who can be described as type A is performing satisfactorily; student type B is making careless mistakes; student type C needs to study more or has poor study habits; and student type D is guessing or has some knowledge of topics which are generally harder but does not know some easier topics. This type of information would be useful to the classroom teacher.

Z3

Z3 is an example of an Item Response Theory (IRT)-based appropriateness index. IRT centers on the relationship between the observed score and the underlying ability or trait as measured by the test. An example of an IRT test model is the one described by Lord (1952) which centers on the assumption that "the probability that an examinee will answer an item correctly is a normal ogive function of his ability" (p. 4).

Other test theorists besides Lord have written about IRT (Lazarsfeld, 1950; Baker, 1965; Wright & Panchapakesan, 1969; Samejima, 1969; Birnbaum in Lord & Novick, 1968), but general interest had been low until about fifteen years ago. Hambleton and Cook (1977) suggested five reasons for this, the first of which is the complex math required in IRT models. Other reasons are that most work in IRT had been addressed to theoreticians and not to practitioners; that there was a lack of fast and ready-to-use computer programs; that many researchers questioned the gains possible with latent trait theory; and that strong assumptions were required in IRT models. In the past fifteen years, however, there has been such considerable interest in IRT that Thissen and Steinberg

(1986) recently proposed a taxonomy of IRT models. IRT computer programs such as LOGIST, ANCELLES, ASCAL and BILOG have become available, and some researchers have been doing comparative studies of those computer programs (Harrison, 1986; Vale & Gialluca, 1988; Mislevy & Stocking, 1989). Other researchers have been investigating the robustness of item and ability estimation in IRT models to violations of the strong assumptions (McKinley & Mills, 1985; Harrison, 1986; Wainer & Thissen, 1987).

The reason that IRT is attractive is that it is purported to be sample free and item free; that is, item parameter estimates are independent of the sample used for item calibration, and examinee ability estimates are independent of the particular subset of items chosen from the entire set of calibrated items; and because the precision of ability estimates are known (Hambleton & Swaminathan, 1985, p. 11). IRT can be used to detect item bias (Lord, 1980; Ironson, 1982; Burrill, 1982; McCauley & Mendoza, 1985; Thissen, Steinberg, & Gerrard, 1986; Tatsuoka, Linn, Tatsuoka, & Yamamoto, 1988), to develop parallel tests (Lord, 1980; Skaggs & Lissitz, 1986), to create adaptive tests for different ability levels (McBride, 1977; Samejima, 1977; Lord, 1980), to minimize decision risks (Van der Linden & Mellenberg, 1977, 1978; Wilcox, 1978; Drasgow & Guertler, 1987) and to identify inappropriate test scores (Levine & Drasgow, 1982; Smith, 1986).

It is this last use of IRT which is of interest in the current paper. A number of IRT-based appropriateness indices have been proposed and comparative studies done (Levine & Rubin, 1979; Levine & Drasgow, 1982; Drasgow, 1982; Rudner, 1983; Tatsuoka & Linn, 1983; Birenbaum, 1985; Drasgow, Levine & Williams, 1985; Drasgow & Levine, 1986; Nelson & Chatman, 1986; Tomsic & Mittman, 1986; and Drasgow, Levine, & McLaughlin, 1987). Many of these researchers used computer generated data, but some used experimental subjects. Birenbaum (1985), for example, used 1,864 10th graders from 77 schools in Tel Aviv. He used 18 20-item parallel tests to see whether appropriateness indices could be used to distinguish "cooperative" from "noncooperative"

test takers, and found that this use of appropriateness indices is feasible. He studied nine indices: three nonstandardized and three standardized extended caution indices, the unstandardized and the standardized likelihood functions, and the unweighted total fit mean square.

Other researchers who used experimental subjects were Nelson and Chapman (1986) and Tomsic and Mittman (1986). Nelson and Chapman (1986) used 207 students in two sections of an undergraduate educational psychology course to study whether appropriateness indices could be used to detect test takers who guess frequently. They concluded that guessing is not what is measured by appropriateness indices. The two indices they studied were the mean square fit statistic associated with the Rasch model (Wright & Stone, 1979) and the fourth standardized extended caution index, ECIZ4 (Tatsuoka, 1984). Tomsic and Mittman (1986) used 1,300 third and seventh graders to study the stability of appropriateness indices. They calculated the four standardized extended caution indices (Tatsuoka, 1984) on students in the third and seventh grades, and then again in the following year on the these same students when they were in the fourth and eighth grades, and correlated the two sets of appropriateness index measures. They reported correlations ranging from .05 to .30 for the lower grade level and somewhat higher correlations for the higher grade level (but did not provide a correlation range for the higher grade level).

Many researchers of IRT-based appropriateness indices have used either actual test results from the Graduate Record Examination, Verbal Test (GRE-V) or the Scholastic Aptitude Test, Verbal Section (SAT-V) (Drasgow, 1982; Levine & Drasgow, 1982; Drasgow, Levine, & Williams, 1985), or computer simulated data, based on item parameters previously estimated from the GRE-V or the SAT-V (Levine & Drasgow, 1982; Rudner, 1983; Drasgow & Levine, 1986; Drasgow, Levine, & McLaughlin, 1987). In these studies, records with aberrant response patterns were created by rescoring a

percentage of the test items. To create a spuriously high test score, a randomly selected percentage of test items were rescored as correct; and to create a spuriously low test score, a randomly selected percentage of test items were rescored as incorrect. Thus, in these studies, because the records with aberrant response patterns were constructed, the researchers were able to report the hit rate, or the percentage of records with aberrant response patterns that were so identified with the use of appropriateness indices.

In earlier studies appropriateness indices were found to be related to the IRT statistic which measures the subject's ability (theta, symbolized θ). This problem, however, was resolved by Drasgow, Levine, and Williams (1985) who proposed a method for standardizing appropriateness indices.

Also in the earlier studies, the measures obtained by the various appropriateness indices studied were correlated to discover their relationships. Drasgow and Levine (1986) noted that while intercorrelating indices enabled determination of the best index among those studied, the method did not show whether any of the indices are good enough for operational use. They introduced a method for evaluating the statistical power of any given appropriateness index. Their method consists of comparing the detection rate of the appropriateness index being studied with the detection rate of the optimal index.

Optimal indices are not feasible for practical use, but they are useful in research for evaluating practical indices because they yield the highest detection rates that can be obtained from item responses (Drasgow, Levine, & McLaughlin, 1987). An optimal index is a likelihood ratio:

$$LR = P_{Aberrant}(\mathbf{u})/P_{Normal}(\mathbf{u})$$

where $P_{Normal}(\mathbf{u})$ is the likelihood of a response vector $\mathbf{u}$ by an examinee of ability θ and is calculated under the null hypothesis that the response pattern is nonaberrant, and $P_{Aberrant}$

(**u**) is the likelihood of a response vector **u** by an examinee of ability θ and is calculated under the hypothesis that the response pattern is aberrant. This ratio can be used to test the simple null hypothesis that the response pattern is normal against the simple alternative hypothesis that the response pattern is aberrant. As the Neyman-Pearson Lemma states that maximum power is achieved by a likelihood ratio test, when used as an appropriateness index, the LR statistic becomes the most powerful index that can be computed (Levine & Drasgow, 1988).

The optimal detection rate of aberrant response patterns is obtained by "classifying **u** as aberrant when

$$P_{\text{Aberrant}}(\mathbf{u}) \geq \text{Constant}\ P_{\text{Normal}}(\mathbf{u}),$$

where the Constant is selected to achieve a specified Type I error rate (i.e., a specific rate of misclassifying normal response patterns as aberrant)" (Drasgow & Levine, 1986, p. 61).

In two studies (Drasgow & Levine, 1986; Drasgow, Levine, & McLaughlin, 1987) where various indices were compared against the optimal index, Z3 was recommended as performing closest to the optimal index. In both studies, parameters estimated from the 1975 SAT-V were used to generate response vectors. In the first study (Drasgow & Levine, 1986), a subset of computer generated response vectors were rescored to simulate 10 percent spuriously high scores, i.e., 10 percent of the items were randomly selected and if they were correct, the items were unaltered but if they were incorrect, the items were rescored as correct. A second subset was rescored to simulate 10 percent spuriously low scores. On the average, 10 percent rescoring resulted in four altered items on the 85-item SAT-V. Using Z3, the researchers were able to detect about 20 percent of the spuriously high scores and about 11 percent of the spuriously low scores at the .05 error rate. The maximum detection rate, as calculated by the optimal index was 24 percent for the spuriously high scores and about 17 percent for spuriously low scores, resulting in a

detection rate for Z3 which was 85 percent of the maximum for spuriously high scores and 67 percent of the maximum for spuriously low scores (Drasgow & Levine, 1986). In the second study, detection rates of appropriateness indices were studied in relation to ability levels. The researchers found Z3 to be effective for detecting spuriously high scores for low ability response vectors and for detecting spuriously low scores for high ability response vectors but not as effective for detecting aberrance at average ability response vectors. It was, however, among the most effective appropriateness indices currently available (Drasgow, Levine, & McLaughlin, 1987).

The following is the equation for Z3:

$$Z3 = \frac{l_o - M(\widehat{\theta})}{[S(\widehat{\theta})]^{1/2}}$$

where

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

where

u_i is the dichotomously scored (1=correct, 0=incorrect) response for item i (i=1, 2, 3 ...n), and l_o is the natural logarithm of the three-parameter logistic likelihood function evaluated at the maximum likelihood estimate $\widehat{\theta}$ of θ. Other elements of the equation for Z3 are as follows:

$$P_i(\theta) = \widehat{c}_i + \frac{1 - \widehat{c}_i}{1 + \exp[\, - D\widehat{a}_i(\theta - \widehat{b}_i)]}$$

$P_i(\theta)$ is an estimate of the probability that a person at ability level θ will answer item i correctly

$$Q_i(\theta) = 1 - P_i(\theta)$$

$\hat{a}_i$ = item discrimination parameter

$\hat{b}_i$ = item difficulty parameter

$\hat{c}_i$ = guessing parameter

θ = ability estimate (theta) of the test taker

D = 1.702, a scaling constant associated with the ogive curve when working with natural logarithms

exp = exponential

$$\exp\left[- D\hat{a}_i (\theta - \hat{b}_i)\right] = e^{- D\hat{a}_i (\theta - \hat{b}_i)}$$

and

$$M(\hat{\theta}) = \Sigma\left[P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + Q_i(\hat{\theta}) \ln Q_i(\hat{\theta})\right]$$
$$S(\hat{\theta}) = \Sigma P_i(\hat{\theta}) Q_i(\hat{\theta}) \left\{\ln\left[P_i(\hat{\theta}) / Q_i(\hat{\theta})\right]\right\}^2$$

The maximum likelihood method was developed by R.A. Fisher and is widely used in statistics for estimating parameters. It is based on the principle that the best estimates of parameters are those which maximize the likelihood function (Arley & Buch, 1950, p. 150). This principle is applied to IRT, where the item parameters for discrimination (a), difficulty (b), and guessing (c), as well as the ability estimate of the test taker (θ, theta) are estimated. Based on the maximum likelihood method, the best estimates of these parameters are those which maximize the logistic likelihood function (l_o).

This same logistic likelihood function can be used as a measure of appropriateness if an appropriate test response pattern is defined as one which is representative of the group whose abilities are being measured by the test, and an inappropriate response pattern is

defined as one which does not contribute much to maximize the likelihood function. The appropriateness index defined in the equation for l_o above is the unstandardized measure of the examinee's contribution to the likelihood function (Birenbaum, 1985, p. 525). In the equation for Z3, $M(\theta)$ is the conditional expected value of l_o given $\theta = \hat{\theta}$ and $S(\theta)$ is the conditional expected variance of l_o given $\theta = \hat{\theta}$. Thus, Z3 is simply the standardized l_o (Drasgow, Levine, & McLaughlin, 1987). This transformation of l_o to Z3 reduces its dependence on ability (Drasgow, Levine, & Williams, 1985).

PAR

The third index of interest which will be studied in this paper was proposed by Ayabe and Heim (1987). They called their index the Person Average R (PAR) for which Heim (1989) presented various computational forms. Basically, PAR is obtained by calculating the average correlation of each individual's response to each test item with those of the n-1 other people taking the test. It is based on a method proposed by Kuder and Richardson (1937) for estimating the reliability of an individual item on a test. In their derivation of KR20, Kuder and Richardson suggested using the average correlation of item i with the n-1 other items in the test as one way of estimating the reliability of a particular item. By transposing the item by person matrix of test item responses and correlating each test taker with every other test taker, it is possible to calculate the average correlation for each individual. It follows that individuals with very low PARs are individuals whose test response patterns are markedly different from those of their fellow test takers.

In a preliminary study, PAR was found to have some stability with a test-retest correlation of .6 calculated on 21 fourth graders who took a 60-item Metropolitan Achievement Test once in fall and once in spring of the 1987-88 academic year (Heim, 1988).

The equation for PAR is as follows:

$$PAR_j = \frac{\sum_{i=1}^{N} r_{ij}}{N - 1}, \; i \neq j$$

In the equation above, PAR for the jth person is equal to the mean of the correlations of item responses of person j with the item responses of the N-1 other people taking the test. Thus, a person with an item response pattern similar to that of the majority of the people in the sample will have a high PAR. A person with a unique item response pattern will have a low PAR.

Reasons for a Unique Response Pattern

Why might a person have a unique response pattern? One possible reason is curricular differences (Harnisch & Linn, 1981). For example, in the area of foreign language learning, a student who transfers in midyear from another school may have a unique test response pattern because the textbook and coverage at his or her former school were different. Similarly, a student who has lived in the country of the target foreign language or who has had contact with native speakers of the language may have a unique test response pattern compared to his or her classmates who have not had those experiences.

According to Harnisch (1989), though MCI can be used to detect aberrant response patterns, it cannot be used to distinguish the reasons for the aberrance, although he has reported a relationship between MCI and curricular differences (Harnisch & Linn, 1981; Harnisch, 1983).

Curricular differences probably have the greatest impact on response patterns. There may, however, be other factors that have an impact on whether or not a test taker answers an item correctly.

Though studies on such factors are scarce, some researchers have studied aberrant response pattern indices and appropriateness indices in connection with guessing (Nelson & Chapman, 1986), and motivation (Birenbaum, 1985). Researchers have also hypothesized connections between the indices and cheating and poor test-taking strategies (Levine & Rubin, 1979) and carelessness (Smith, 1986).

Birenbaum's (1985) study seems particularly important in illuminating one of the shortcomings of some research designs, a shortcoming which limits the applicability of the results of research studies. In Birenbaum's study, subjects were high school students who were asked to take a test which would have no impact on their course grade. Subjects were asked to write their names on their test papers and those who did so were categorized as cooperative and those who failed to do so were categorized as noncooperative. Birenbaum defined cooperative subjects as those having appropriate response patterns and uncooperative subjects as those having inappropriate patterns. A third group consisted of randomly generated test response patterns. Nine indices were calculated on test response patterns and used to classify the test responses into the three groups, cooperative, non-cooperative and random pattern groups. Birenbaum showed that eight of the indices studied could be used to distinguish among the three groups, and that Z3 was one of the three found to be superior for detecting inappropriate response patterns.

Researchers in tests and measurement sometimes fail to adequately consider the significance of subjects' motivation. Methods sections of research studies may include information on whether or not research subjects were volunteers or received points toward their course grades but further implications on the generalizability of the findings are rarely discussed. Atkinson (1980) criticized traditional test theorists for not accommodating the

role of motivation in what he calls the overly simplistic "obtained score equals true score plus error score" equation of classical test theory. Raynor (1974) demonstrated the role of motivation on test taking behavior in his work on future orientation and academic performance, and Fiske (1957) described how in his research uncooperative subjects gave the impression of greater intraindividual consistency than cooperative subjects, thereby confounding the results of his study.

Test takers' motivation is a particularly important consideration in the practical test-taking situation. Precautions against opportunities for cheating on achievement tests immediately come to mind, but in some testing situations "faking" is also a concern. For example, in personality testing, detection of clever patients who sometimes "fake" answers in order to appear normal is of great interest to researchers (Burkhart, Christian, & Gynther, 1978; Snyter & Allik, 1981; Rigby, 1987; Worthington & Schlottmann, 1986). Another example of "faking" is "sandbagging" where the student deliberately answers items on a placement test incorrectly, in order to enroll in an easy course and thereby gain an advantage over his/her classmates. In the case of foreign students, "sandbagging" may be occurring for reasons concerning visa obtainment. For example, the length of time an East West Center Grantee has to finish a graduate degree depends in part on how well he or she does on an English language proficiency test. A promising student with a poor grasp of English is often granted a longer period of time to complete his or her degree than a student who has an excellent command of English. Thus, it is conceivable that applicants from politically and/or economically unstable countries might sandbag on their English proficiency test in order that they might be granted longer stays in America. Since Birenbaum (1985) showed that cooperation is related to aberrant response pattern indices, it might be possible to use those indices to identify test takers who deliberately sandbag on tests.

Levine and Rubin (1979) have suggested differences in test-taking skills as another possible reason for aberrant test response patterns. Researchers have shown the effects of testwiseness on test performance (Callenbach, 1973; Crehan, Koehler, & Slakter, 1974; Rowley, 1974; Diamond, Ayrer, Fishman, & Green, 1976; Bajtelsmit, 1977; Fagley, 1987). It is possible that poor test-taking skills may be among the reasons for an aberrant response pattern. A poor test taker is someone who, for example, may select an incorrect answer to a multiple-choice question because he or she has not bothered to read all of the answer options, or someone who wastes time on a difficult item and is therefore unable to finish the test (Graham & Robinson, 1984; Good & Brophy, 1986). One can see how someone who knows the answer might fail test items simply because he or she ran out of time and never got to the items. Conversely, a very good test taker may sometimes answer test items correctly without knowing the answer if there are grammatical or other clues in the way the test items are worded. Carter (1986) discussed five common test item faults in multiple-choice tests, which testwise test takers may use as clues. Among these common faults is the more than chance probability of choice "c" being the correct option in a multiple-choice test.

Another possible reason for aberrant response patterns, carelessness, was hypothesized by Smith (1986). In terms of academic performance, a careless person is someone who exhibits a greater degree of randomness than a careful person. One would expect a careful person to perform more consistently at his or her academic level than a careless person. One would, therefore, expect a careless person who performs inconsistently at his or her level to produce the kind of test response patterns which are detectable by aberrant response pattern indices.

The purpose of this study was to discover any relationships between the three aberrance indices, MCI, Z3, and PAR, and four hypothesized reasons for unusual test response patterns in a real testing situation. A Japanese Language Placement Test which is

currently used to place students was used to investigate four possible reasons for unusual test response patterns: curricular differences, differences in test-taking skills, differences in motivation, and differences in consistency of academic performance. Secondary purposes of the study were to evaluate the feasibility of applying the indices for placement purposes, and to investigate the extent of sandbagging, or deliberate failure, on an actual placement test.

II. METHOD

The purpose of this study was to discover any relationships between three aberrance indices (the Modified Caution Index (MCI), the Person Average R (PAR) and the standardized logistic maximum likelihood function (Z3)) and four hypothesized reasons for aberrant response patterns (Curricular Differences, Differences in Test-Taking Skills, Differences in Motivation and Differences in Consistency of Academic Performance.) Two secondary purposes were to evaluate the feasibility of using aberrance indices for placement decisions and to investigate the occurrence of sandbagging on a placement test.

Subjects

Subjects were 221 high school seniors who were interested in enrolling in a Japanese language course at the University of Hawaii at Manoa (UH Manoa) in Fall 1988, and who had previously studied the language. Such students are regularly screened by the East Asian Languages and Literatures Department (EALL). Of the 367 students who were screened by the EALL Department in March 1988, 221 agreed to participate in the study.

Instruments

Japanese Language Placement Test (JP Test). The JP Test was developed by UH Manoa's EALL Department for placement of students in lower division courses and has been in use since 1977 to place 700-800 students per year in Japanese language courses. This estimate is about three-fourths of all students wishing to enroll in a Japanese language course for the first time at UH Manoa, and yields a cumulative estimate of 8,000 to 10,000 students who have taken the test since it was implemented 12 years ago.

The last review of the effectiveness of the JP Test was done in 1982-83 by an ad hoc committee which was composed of Peter Tanaka, Yaeko Habein and Machiko Netsu,

and appointed by Hiroshi Miyaji, then chairman of the Department. The committee found the test appropriate for placing students in lower division courses.[1]

More recently, in Fall 1987, the Department conducted a survey of its students who had taken the JP Test.[2] Of the 373 surveyed, 88.7 percent thought that their placement level was appropriate, 4.6 percent thought it was too low and 6.7 percent thought it was too high (Table 2). In other words, 11.3 percent of the students surveyed felt that they were not appropriately placed.

The JP Test seems to be an acceptable test from a measurement point of view. The Grammar Section, for example, has an internal consistency coefficient of .92 and a standard error of measurement of three points, based on the March 1986 test data. A test-retest correlation is not available, but in its place coefficient alpha can be used as a meaningful substitute (Cronbach, 1947). ("In practice, the coefficient of equivalence or the coefficient of stability may be used meaningfully where the reliability is called for" p. 15.)

Because the primary use of the test is placement of students, in addition to the question of how effective the JP Test is as a test instrument, the question regarding the prevalence of "sandbagging" must be addressed. Sandbagging is defined as deliberately performing poorly on the placement test in order to be placed in a lower level so that one will have an easy time getting through the course. It is possible that more than 4.6 percent of students surveyed in Fall 1987 felt that their placement level was too low, but students who sandbagged are not likely to admit, much less complain, about getting something for which they had hoped. Sandbagging can be considered a form of cheating because its aim is to gain an unfair advantage for the sandbagger over his or her classmates. The topic,

---

[1] Memo to Dr. Hiroshi Miyaji, Chairman, from Ad Hoc Committee on Japanese Placement Test (EALA3-050) dated May 26, 1983.

[2] Ray Kaneyama, EALL Placement Officer, personal communication, March 1988.

Table 2
Japanese Placement Test Survey Results
Conducted by UHM EALL Department
Fall 1987

Course		Okay*		Too Low*		Too High*	
	N	N	%	N	%	N	%
100	(137)	125	91.2%	4	2.9%	8	5.8%
101	(60)	58	96.7%	2	3.3%	—	—
102	(57)	45	78.9%	3	5.3%	9	15.8%
201	(67)	61	91.0%	2	3.0%	4	6.0%
202	(27)	21	77.8%	3	11.1%	3	1.1%
301/302	(16)	14	87.5%	2	12.5%	---	----
401/403	(9)	7	77.8%	1	11.1%	1	11.1%
Total	(373)	331	88.7%	17	4.6%	25	6.7%

*"The course I was placed in was (please check one of the following):

__ okay __ too low __ too high."

however, has not received attention in the field of tests and measurement, and there is very little research reported in the literature.

The JP Test consists of four parts, but only Part I, the Grammar Section of the test will be used in this study. It is the section of the test which is most heavily depended upon in making placement judgments.

Background Information Sheet. The Department's Background Information Sheet (BIS) requests demographic data including number of years of Japanese language study in high school, in language school, years of residence in Japan, and access to native speakers of Japanese. Answers to these four items on the BIS provided measures on the differences in exposure to the Japanese language. The BIS is required of all students taking the JP Test. (See Appendix A.)

Questionnaire 1. Questionnaire 1 (for students) requested responses to 10 specific and one general question regarding test-taking skills (see Appendix B). Eight of the 10 specific questions were taken directly from the Learning and Study Strategies Inventory (LASSI) scale on test-taking skills (Weinstein, 1987). In addition, Questionnaire 1 requested the respondent to guess his number-correct score on the four sections of the JP Test and the point range of confidence. The 10 specific and one general question on test-taking skills provided measures of test-taking skills. The difference between the individual's obtained number-correct score and his guessed score provided a measure of the accuracy of the student's (asserted) knowledge of Japanese.

Questionnaire 2. Questionnaire 2 for high school Japanese language teachers requested teachers to rate the appropriateness of each student's placement level, the motivation of each student to do well in Japanese language study, each student's test-taking skills and each student's margin of error (see Appendix C). The margin of error was defined as the consistency of a student's performance at his or her ability level. Teachers were asked to rate a student whose performance was excellent one day and poor the next as

having a wide margin of error and to rate a student who performed consistently from day to day as having a narrow margin of error. Thus, Questionnaire 2 provided measures of each student's appropriateness of placement level, motivation, test-taking skills, and margin of error.

Questionnaire 3. Questionnaire 3 for students requested students to provide responses to eight questions dealing with test-taking strategies as applied to the JP Test. One question asked directly whether the statement, "I didn't want to do too well on the Placement Test," was true or not true for the respondent. Thus the eight test-taking strategy questions were also measures of motivation and sandbagging.

Questionnaire 3 also asked two additional questions to measure sandbagging. It asked respondents 1) to estimate the percentages of his or her classmates who seemed to be too advanced, those who seemed to be too slow, and those seemed to be appropriately placed for the course level, and 2) whether it is true that students who take the JP Test deliberately do poorly in order to enroll in an easy Japanese language course.

In addition, Questionnaire 3 asked appropriateness of placement questions such as the percentage of course coverage which was review for the respondent, difficulty level of the course for the respondent, and whether or not the respondent was keeping up with the course work. It also asked the respondent to rate his or her study habits in Japanese language study (see Appendix D).

Questionnaire 4. Questionnaire 4 requested experimental subjects' college Japanese language instructors to rate their students on five scales: the appropriateness of each student's placement level, each student's motivation to do well in Japanese language study, and each student's test-taking skills, study habits and margin of error. As in Questionnaire 2 which was completed by high school Japanese language teachers, the margin of error was defined as the consistency of a student's performance at his or her ability level.

Instructors were asked to rate a student whose performance was excellent one day and poor the next as having a wide margin of error and to rate a student who performed consistently from day to day as having a narrow margin of error. Each student's grade as of the survey day was also requested as a measure of appropriateness of placement as well as progress in the course (see Appendix E).

Quizzes 2 and 3. Quizzes 2 and 3 for Japanese course level 102 (Jpn 102), produced and administered by the Department, provided objective measures of achievement for a subset of the experimental subjects. The quizzes were 10 points each in value, with 8 points of common items across all sections of Jpn 102. The remaining two points were based on items written by individual instructors and varied by section and/or instructor.

Procedure

JP Test data were collected in March and April 1988 during the EALL Department's spring testing of incoming freshmen. All students who took the JP Test completed the Department's Background Information Sheet (BIS). Students who agreed to participate in the study completed Questionnaire 1 immediately after completing the JP Test. Two hundred, twenty-one of the 367 students who took the test completed Questionnaire 1. In May 1988 the high school Japanese language teachers of the 221 subjects were asked to complete Questionnaire 2.

The Department provided an electronic data file of the Spring 1988 JP Test results, which was edited to delete student names and social security numbers after identification numbers were assigned to identify students who agreed to participate in the study. MCI, PAR and Z3 were calculated on the entire pool of examinees.

A database was created which included the calculated indices, the demographic data from the BIS, and for students who agreed to participate in the study, test-taking skills

measures from Questionnaire 1 and high school Japanese language teacher's ratings from Questionnaire 2.

Study participants were tracked into their Fall 1988 UH Manoa Japanese language courses. Class time was granted by the Department to administer Questionnaire 3 in early October. Jpn 101 through 202 classes in which at least one experimental subject enrolled were surveyed. In all, thirty classes were surveyed.

The Japanese language instructors of the study participants were asked to complete Questionnaire 4 in early October, immediately after Questionnaire 3 was administered to their classes.

Photocopies of Quizzes 2 and 3 from Jpn 102 were provided by the Department in October, with names of nonexperimental subjects deleted.

Data from Questionnaires 3 and 4, as well as from Quizzes 2 and 3 were added to the database mentioned earlier.

Statistical analyses were performed using SAS for Personal Computers (PC-SAS) (SAS Institute, 1988). In order to determine the existence of relationships between the indices of interest and the four possible reasons for unusual response patterns, correlations between the indices and the variables measuring the four reasons were calculated. Table 3 is a summary of the four reasons for unusual response patterns and the variables which were used to measure those reasons. The 31 variables listed in Table 3 which were significantly correlated to the indices were also regressed on the indices to find the variance in the indices accounted for by those variables of interest.

In order to evaluate the effectiveness of the three indices as aids in interpreting test scores, the indices were correlated to three ratings of appropriateness of placement level: the high school Japanese language teacher's rating, the college Japanese language

Table 3
Possible Reasons for Unusual Test Response Patterns
And Variables Selected to Measure Them

	Variable Code	Source*
Curricular differences		
Years of Japanese studied in high school	HY	BIS
Years of Japanese studied in language school	JY	BIS
Years of residence in Japan	RY	BIS
Access to native speakers of Japanese	NS	BIS
Differences in test-taking skills		
10 test-taking questions taken from the LASSI scale	X1-X10	Q1
Test-taking score, sum of 10 LASSI items	TTT	Q1
Test-taking skills self rating by student	X11	Q1
Absolute value of score minus guessed score	OMG	Q1
Test-taking skills rating by by high school Japanese language teacher	SD	Q2
Test-taking skills rating by college Japanese language instructor	C4	Q3
Differences in motivation		
Motivation to do well in Japanese language study, rating by high school Japanese language teacher	SC	Q2
Motivation to do well in Japanese language study, rating by college Japanese language instructor	B4	Q4
Motivation to do well on the JP Test, student's self-asserted ratings	V2-V9	Q3
Differences in consistency of academic performance		
Margin of error rating by high school Japanese language teacher	SE	Q2
Margin of error rating by college Japanese language instructor	E4	Q4

*BIS Background Information Sheet, completed by students during the Placement Test session in March 1988.
Q1 Questionnaire 1, administered to students during the Placement Test session.
Q2 Questionnaire 2, completed by high school Japanese language teachers, May 1988.
Q3 Questionnaire 3, administered to students in Japanese language course in October 1988.
Q4 Questionnaire 4, completed by college Japanese language teachers in October 1988.

instructor's rating and the student's self rating. In addition, a system for classifying students using total test score and aberrance indices as described by Harnisch (1983) was applied to the data in order to evaluate the feasibility of using the indices as aids in interpreting JP Test scores.

To ascertain whether students' motivation to do poorly on the JP Test in order to enroll in an easy course level may have affected the distribution of appropriateness indices calculated on the Placement Test, the MCI and PAR were calculated on Quizzes 2 and 3 at the Jpn 102 course level. The distributions of MCI and PAR, calculated on the quizzes, were compared against those calculated on the JP Test. Patterns of correlations with selected variables were also examined. (Z3 was not calculated on Quizzes 2 and 3 because quizzes were not multiple-choice tests and it was not possible to calculate Z3 on them.) Students' perception of the degree of sandbagging on the JP Test was also gauged with items on Questionnaire 3 and will be discussed in the next two chapters.

III. RESULTS

The purpose of this study was to discover any relationships between three aberrance indices (the Modified Caution Index (MCI), the Person Average R (PAR), and the standardized logistic maximum likelihood function (Z3)) and four hypothesized reasons for aberrant response patterns (Curricular Differences, Differences in Test-Taking Skills, Differences in Motivation, and Differences in Consistency of Academic Performance.)

Two secondary purposes of this study were to evaluate the effectiveness of the aberrance indices as aids in interpreting test scores, and to investigate the occurrence of sandbagging on the Japanese Placement Test (JP Test).

Descriptive statistics on the indices will be presented first in this chapter. Included in this first section will be descriptive statistics on MCI, PAR, and Z3 calculated on the JP Test as well as those indices calculated on Quizzes 2 and 3.

Next, the relationships between the indices calculated on the JP Test and the four hypothesized reasons for aberrant response patterns will be presented in the form of correlations between the indices and the variables measuring the four hypothesized reasons. Since some of the indices and many of the variables which measure the four hypothesized reasons were related to JP Test Score, this will be followed by a section on correlations between the indices and the variables measuring the four hypothesized reasons with JP Test score partialed.

The fourth and last section in this chapter will be a presentation of the data on the effectiveness of the indices as aids in interpreting test scores, and evidence of the occurrence of sandbagging on the JP Test.

Descriptive Statistics on Indices

The MCI calculated on the JP Test ranged from .056 to .703, with a mean of .251 and a standard deviation of .098 (see Table 4). This distribution is within the range for MCI which is zero to one (Harnisch & Linn, 1981).

Since PAR is the average of correlation coefficients, the maximum range for PAR is -1 to 1. PAR calculated on the JP Test ranged from -.044 to .264, with a mean of .148 and a standard deviation of .057, and is well within the maximum range.

Z3 ranged from -2.329 to 2.823, with a mean of .236 and a standard deviation of 1.027. It is within the expected range for Z3, which as a standardized logistic function, has a range of -3 to 3 for practical purposes.

These statistics for the indices calculated on the JP Test were all within the expected maximum ranges and therefore seem reasonable.

Table 4
Mean, Standard Deviation and Range
For the Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Logistic Maximum Likelihood Function (Z3)
Calculated on the Japanese Placement Test (JP TEST)

	Mean	Std Dev	Range Min	Range Max	*n*
MCI	.251	.098	.056	.703	365
PAR	.148	.057	-.044	.264	365
Z3	.236	1.027	-2.329	2.823	365

Besides these statistics on the indices calculated on the JP Test, similar descriptive statistics were obtained on indices calculated on Quizzes 2 and 3. These are presented in Table 5.

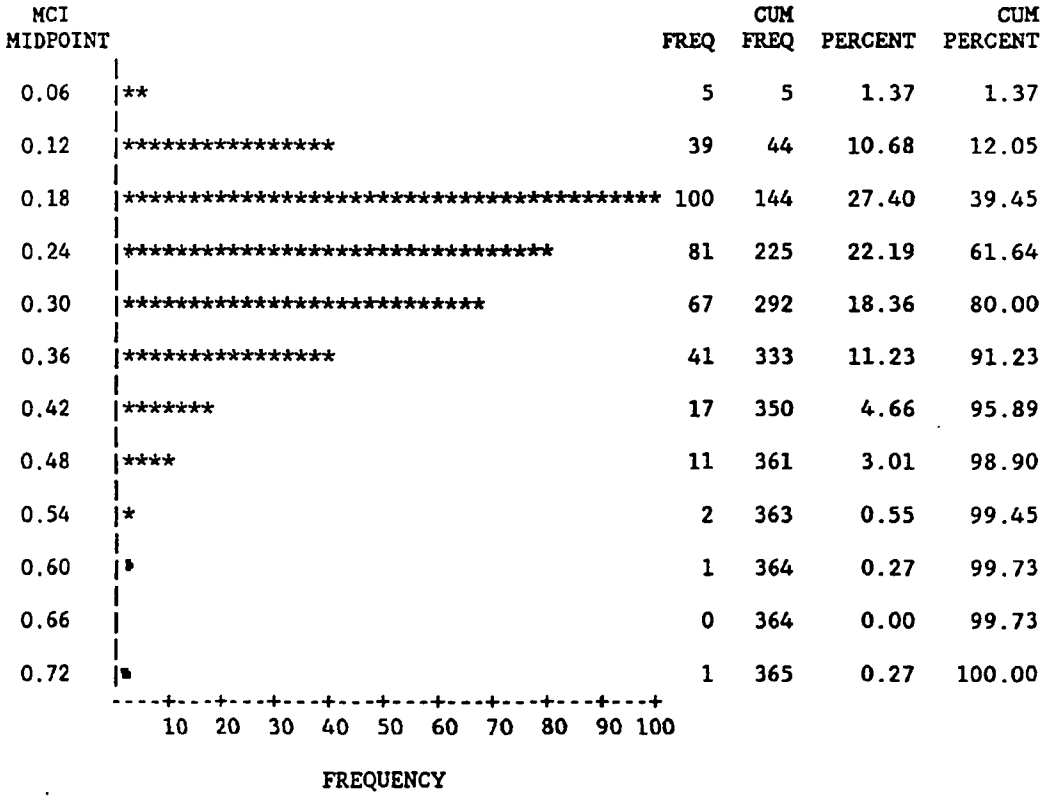
Quizzes 2 and 3 for students enrolled in Japanese language course level 102 (Jpn 102) were collected as objective measures of achievement. Quizzes in lower division Japanese language courses at the University of Hawaii at Manoa are coordinated so that about 80 percent of the items are uniform across all sections of the courses. The remaining 20 percent of quiz items may differ by instructor. Photocopies of the two quizzes for all sections of Jpn 102 were provided by the Department of East Asian Languages and Literatures. This enabled the author to calculate MCI and PAR on the common items on the quizzes. Z3, however, was not calculated because the quiz items were not in multiple-choice format. To maintain confidentiality, names on the photocopies of the quizzes, except those of participants in this study, were obliterated by the Department of East Asian Languages and Literatures. It was therefore not possible to correlate indices across quizzes.

There were 19 common items on Quiz 2 and 26 common items on Quiz 3. The mean MCIs calculated on Quiz 2 (.190) and Quiz 3 (.189) were lower than the mean MCI calculated on the JP Test (.251), and the mean PARs on Quiz 2 (.243) and Quiz 3 (.221) were higher than the mean PAR calculated on the JP Test (.148) (see Tables 4 and 5). Since low MCIs and high PARs indicate nonaberrant response patterns, it seems that there were more aberrant test response patterns on the JP Test than on the quizzes.

Table 5
Means, Standard Deviations and Ranges
For the Modified Caution Index (MCI) And
The Person Average R (PAR)
Calculated on Quiz 2 and Quiz 3

	Mean	Std Dev	Range Min	Range Max	*n*
QZ2 MCI	.190	.113	0	.645	278
QZ3 MCI	.189	.100	.004	.536	236
QZ2 PAR	.243	.090	-.046	.432	278
QZ3 PAR	.221	.080	-.040	.372	236

Graphic representations of the frequency distributions of MCI, PAR, and Z3 calculated on the JP Test are presented in Figures 2A, 2B, and 2C. These are followed by frequency distributions of MCI calculated on Quizzes 2 and 3 which are presented in Figures 3A and 3B, and frequency distributions of PAR calculated on Quizzes 2 and 3 which are presented in Figures 4A and 4B.



MCI MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
0.06	**	5	5	1.37	1.37
0.12	**********	39	44	10.68	12.05
0.18	**********	100	144	27.40	39.45
0.24	********	81	225	22.19	61.64
0.30	********	67	292	18.36	80.00
0.36	**********	41	333	11.23	91.23
0.42	*******	17	350	4.66	95.89
0.48	****	11	361	3.01	98.90
0.54	*	2	363	0.55	99.45
0.60	*	1	364	0.27	99.73
0.66		0	364	0.00	99.73
0.72	*	1	365	0.27	100.00

10 20 30 40 50 60 70 80 90 100

FREQUENCY

Figure 2A. Frequency distribution of the Modified Caution Index (MCI) calculated on the Japanese Placement Test (JP Test).

PAR MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
-0.045	*	1	1	0.27	0.27
-0.015	*	1	2	0.27	0.55
0.015	***	8	10	2.19	2.74
0.045	*******	18	28	4.93	7.67
0.075	**********	31	59	8.49	16.16
0.105	**********	55	114	15.07	31.23
0.135	**********	65	179	17.81	49.04
0.165	**********	64	243	17.53	66.58
0.195	**********	77	320	21.10	87.67
0.225	**********	34	354	9.32	96.99
0.255	****	11	365	3.01	100.00

10 20 30 40 50 60 70 80 90 100

FREQUENCY

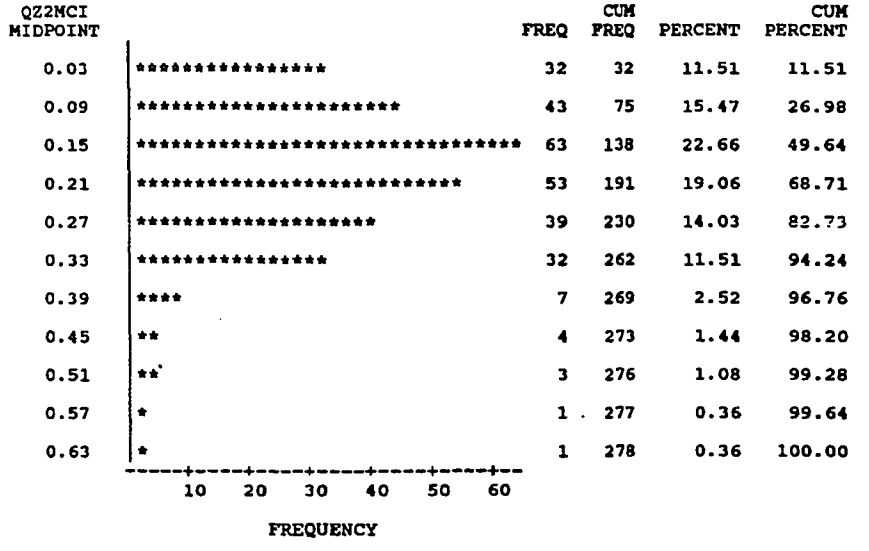
Figure 2B. Frequency distribution of the Person Average R (PAR) calculated on the Japanese Placement Test (JP Test).

Z3 MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
-2.25	***	8	8	2.19	2.19
-1.75	****	10	18	2.74	4.93
-1.25	**********	24	42	6.58	11.51
-0.75	********************	50	92	13.70	25.21
-0.25	**********************	54	146	14.79	40.00
0.25	************************	61	207	16.71	56.71
0.75	********************************	78	285	21.37	78.08
1.25	*****************	43	328	11.78	89.86
1.75	********	19	347	5.21	95.07
2.25	*****	13	360	3.56	98.63
2.75	**	5	365	1.37	100.00

10 20 30 40 50 60 70 80 90 100

FREQUENCY

Figure 2C. Frequency distribution of the Standardized Logistic Maximum Likelihood Function (Z3) calculated on the Japanese Placement Test (JP Test).

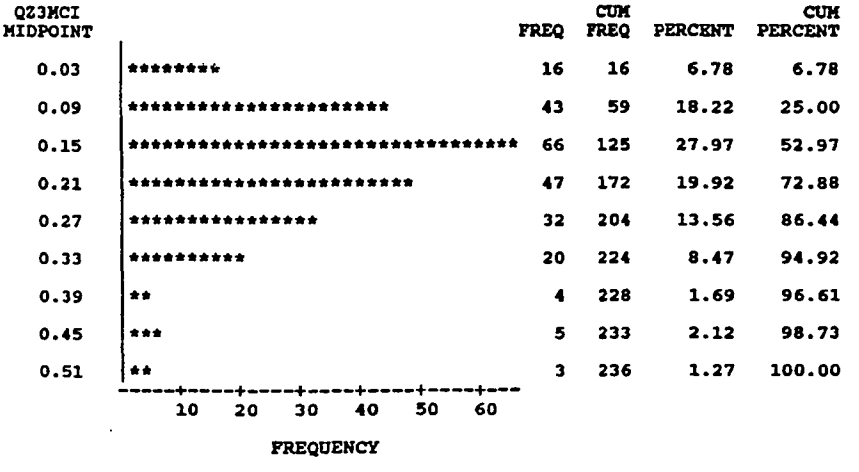


QZ2MCI MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
0.03	**********	32	32	11.51	11.51
0.09	**********	43	75	15.47	26.98
0.15	**********	63	138	22.66	49.64
0.21	**********	53	191	19.06	68.71
0.27	**********	39	230	14.03	82.73
0.33	**********	32	262	11.51	94.24
0.39	****	7	269	2.52	96.76
0.45	**	4	273	1.44	98.20
0.51	**	3	276	1.08	99.28
0.57	*	1	277	0.36	99.64
0.63	*	1	278	0.36	100.00

10 20 30 40 50 60

FREQUENCY

Figure 3A. Frequency distribution of the Modified Caution Index (MCI) calculated on Quiz 2.



QZ3MCI MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
0.03	********	16	16	6.78	6.78
0.09	**********	43	59	18.22	25.00
0.15	**********	66	125	27.97	52.97
0.21	**********	47	172	19.92	72.88
0.27	**********	32	204	13.56	86.44
0.33	**********	20	224	8.47	94.92
0.39	**	4	228	1.69	96.61
0.45	***	5	233	2.12	98.73
0.51	**	3	236	1.27	100.00

10 20 30 40 50 60

FREQUENCY

Figure 3B. Frequency distribution of the Modified Caution Index (MCI) calculated on Quiz 3.

QZ2PAR MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
-0.025	*	2	2	0.72	0.72
0.025	***	6	8	2.16	2.88
0.075	*****	9	17	3.24	6.12
0.125	**********	26	43	9.35	15.47
0.175	**********	45	88	16.19	31.65
0.225	**********	51	139	18.35	50.00
0.275	**********	62	201	22.30	72.30
0.325	**********	47	248	16.91	89.21
0.375	**********	22	270	7.91	97.12
0.425	****	8	278	2.88	100.00

10 20 30 40 50 60

FREQUENCY

Figure 4A. Frequency distribution of the Person Average R (PAR) calculated on Quiz 2.

QZ3PAR MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
-0.025	*	2	2	0.85	0.85
0.025	****	7	9	2.97	3.81
0.075	******	11	20	4.66	8.47
0.125	**********	25	45	10.59	19.07
0.175	**********	42	87	17.80	36.86
0.225	**********	49	136	20.76	57.63
0.275	**********	65	201	27.54	85.17
0.325	**********	28	229	11.86	97.03
0.375	****	7	236	2.97	100.00

10 20 30 40 50 60

FREQUENCY

Figure 4B. Frequency distribution of the Person Average R (PAR) calculated on Quiz 3.

The correlations among the JP Test score and indices calculated on the JP Test are shown in Table 6. MCI and PAR were highly correlated (r=-.92, $d.f.$=363, p<.01). PAR and Z3 (r=.68, $d.f.$=363, p<.01), and MCI and Z3 (r=-.59, $d.f.$=363, p<.01), were also significantly correlated. Of the three indices, only PAR was uncorrelated with JP Test score. However, as can be seen in Figure 5B, the relationship between PAR and JP Test score appears to be a curvilinear one, as does the one between MCI and JP Test score (Figure 5A). Curvilinear regression analysis was done according to Pedhazur (1982, pp. 420-424), and a quadratic relationship was indeed found between MCI and JP Test score and between PAR and JP Test score. The relationship between Z3 and JP Test score, however, was linear. The following equations describe the relationships between JP Test score and the indices.

MCI	= .492 - .017 JPT + .0003 JPT^2	R^2 = .242	R = .492
PAR	= .0014 + .012 JPT - .0002 JPT ^{2}	R^2 = .311	R = .558
Z3	= .988 - .03 JPT	R^2 = .102	R = .319

Table 6
Correlations Among
Japanese Placement Test Score (JP Test),
Modified Caution Index (MCI), Person Average R (PAR) and The
Standardized Logistic Maximum Likelihood Function (Z3)
Calculated on the JP Test
n=365

	JP Test Score	MCI	PAR	Z3
JP Test Score	1.0	-.32**	0.01	-.32**
MCI		1.0	-.92**	-.59**
PAR			1.0	.68**
Z3				1.0

** $p < .01$.

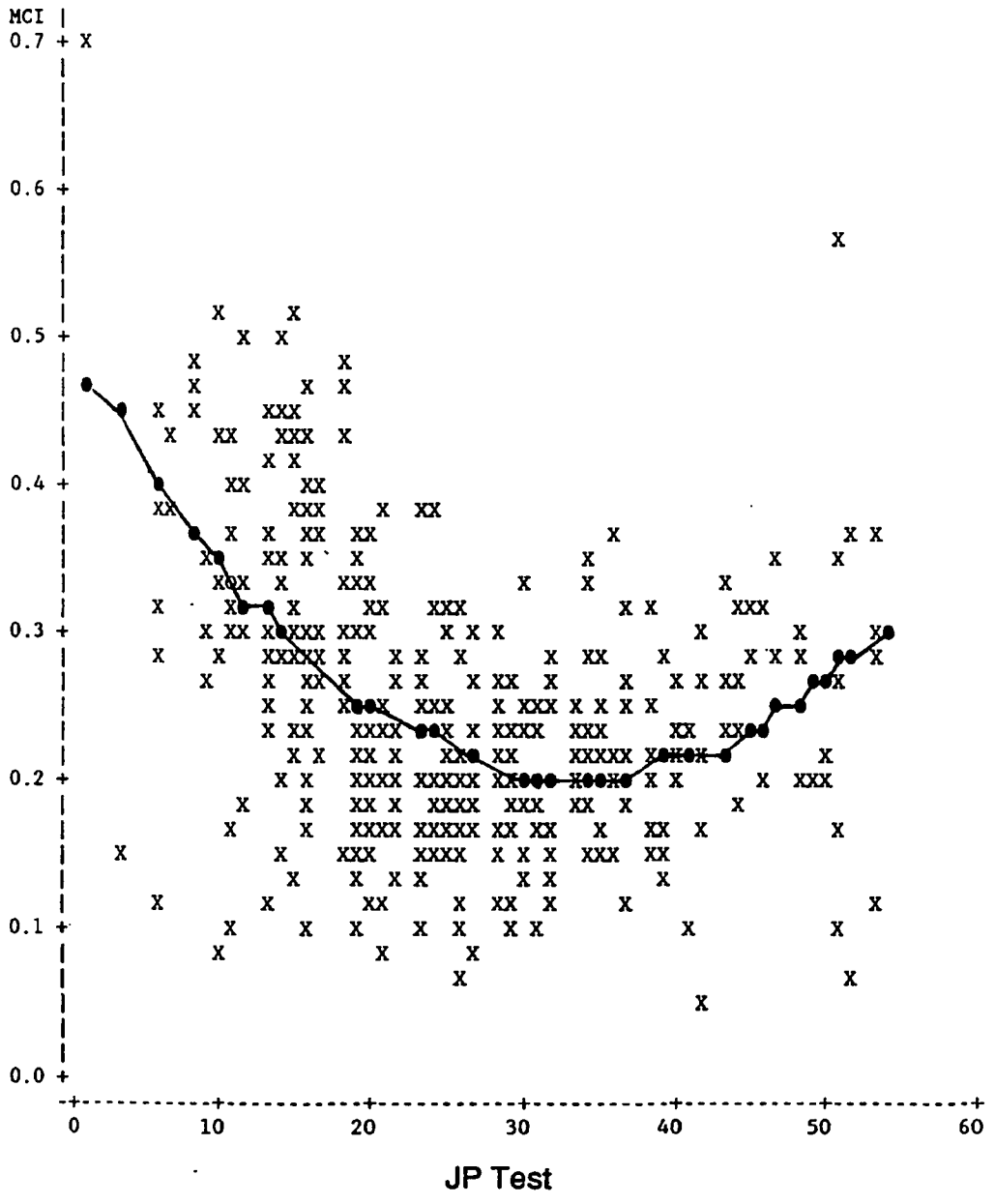


Figure 5A. Bivariate plot of the Modified Caution Index (MCI) with the Japanese Placement Test Score (JP Test).

(MCI = .492 - .017 JPT + .0003 JPT^2 $R^2 = .242$, $R = .492$, $n = 365$, $p < .01$)

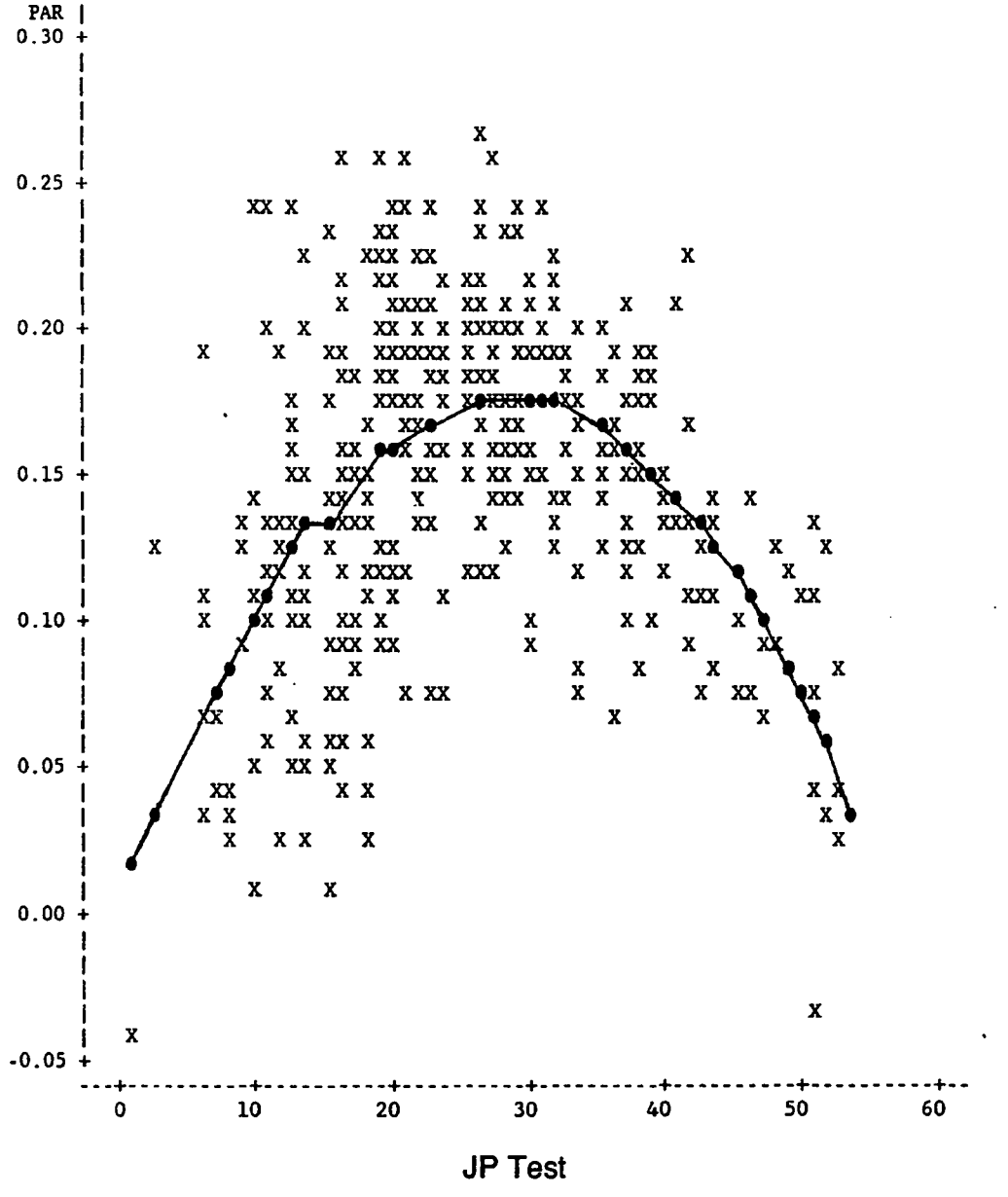


Figure 5B. Bivariate plot of the Person Average R (PAR) with the Japanese Placement Test Score (JPTest).

(PAR = .0014 + .012 JPT - .0002 JPT ^{2} $R^2 = .311$, $R = .558$, $n = 365$, $p < .01$)

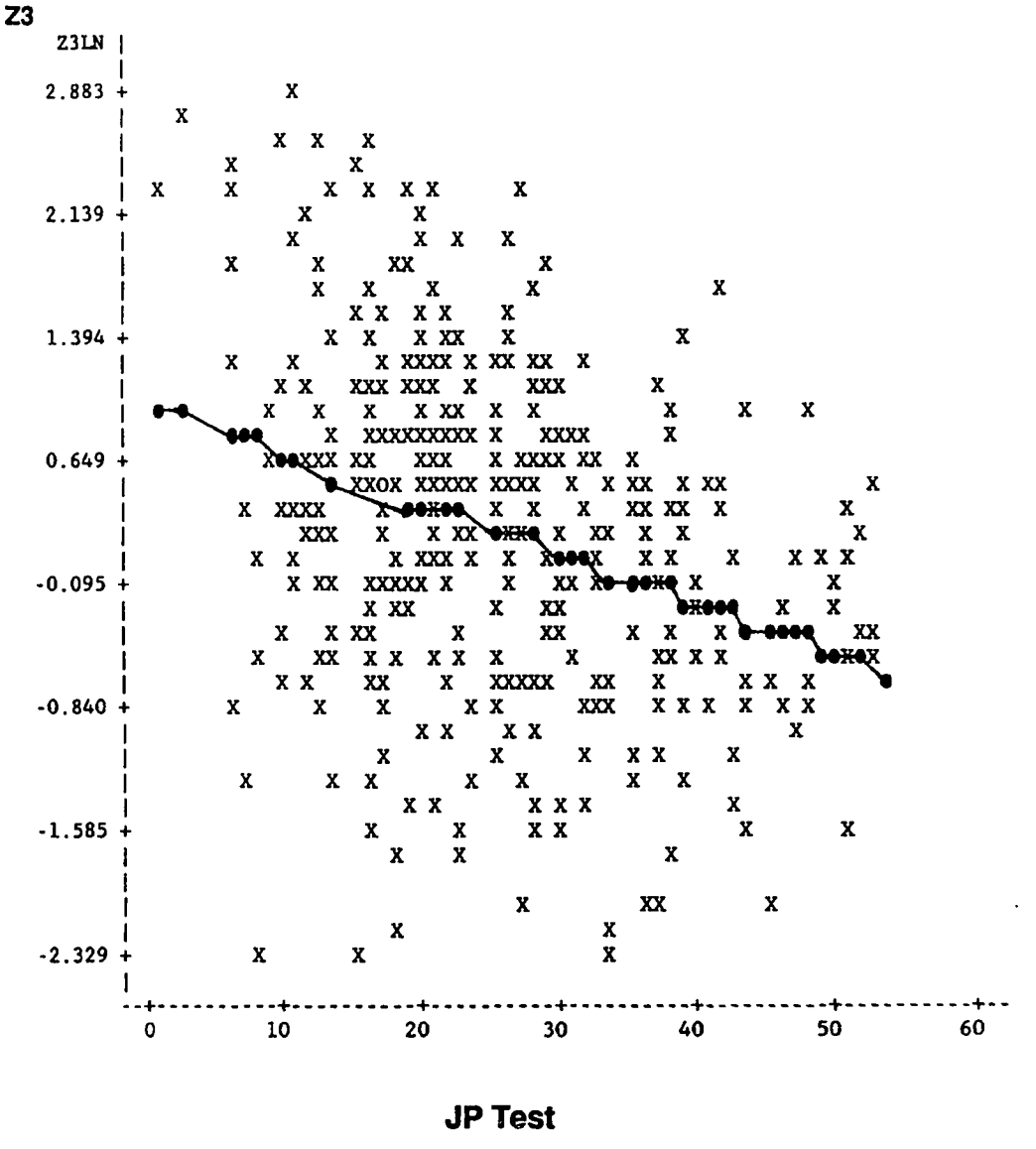


Figure 5C. Bivariate plot of the Standardized Logistic Maximum Likelihood Function (Z3) with the Japanese Placement Test Score (JPTest).

(Z3 = .988 - .03 JPT $R^2 = .102,\ R = .319,\ n = 365, p < .01$)

By way of comparison, correlations among quiz score and MCI and PAR calculated on quizzes for Quizzes 2 and 3 are presented in Tables 7 and 8. The correlations between quiz score and PAR calculated on Quiz 2 (r=-.40, $d.f.$=276, p<.01) and on Quiz 3 (r=-.51, $d.f.$=236, p<.01) were much higher than the correlation between JP Test score and PAR calculated on JP Test (r=.01, $d.f.$=363, n.s.). Other relationships between quiz score and indices calculated on quizzes were similar to those between JP Test score and indices calculated on JP Test score.

Bivariate plots of MCI calculated on quizzes with quiz score for Quizzes 2 and 3 are presented in Figures 6A and 6B. The relationship between Quiz 2 score and MCI calculated on Quiz 2 was linear but the relationship between Quiz 3 score and MCI calculated on Quiz 3 was quadratic. The following equations describe the relation between Quiz scores and MCIs calculated on quizzes.

$$\text{QZ2MCI} = .087 + .008\ \text{QZ2SCORE} \qquad R^2 = .048 \quad R = .219$$

$$\text{QZ3MCI} = .231 - .015\ \text{QZ3SCORE} + .001\ \text{QZ3SCORE}^2 \qquad R^2 = .112 \quad R = .335$$

Bivariate plots of PAR calculated on quizzes with quiz scores for Quizzes 2 and 3 are presented in Figures 7A and 7B. The relationships between Quiz 2 score and PAR calculated on Quiz 2 and between Quiz 3 score and PAR calculated on Quiz 3 were quadratic relationships. The following equations describe these relationships.

$$\text{QZ2PAR} = .088 + .042\ \text{QZ2SCORE} - .002\ \text{QZ2SCORE}^2 \qquad R^2 = .238 \quad R = .488$$

$$\text{QZ3PAR} = .207 + .014\ \text{QZ3SCORE} - .0007\ \text{QZ3SCORE}^2 \qquad R^2 = .297 \quad R = .545$$

Not only was the relationship between Quiz 2 MCI and Quiz 2 score a linear one, but the curvilinear relationship between Quiz 3 MCI and Quiz 3 score was less pronounced

(Figure 6B) compared to the relationship between MCI and JP Test score (Figure 5A). A similar observation can be made for PAR: the curvilinear relationship between Quiz PAR and quiz scores was less pronounced (Figures 7A, 7B) than the relationship between PAR and JP Test score (Figure 5B).

It can be seen in Figures 5A and 5B that a large number of test takers who earned low JP Test scores had high MCIs and low PARs. However, almost all quiz takers who earned low quiz scores had low MCIs calculated on quizzes and high PARs calculated on quizzes. In other words, a large number of low scorers had aberrant test-response patterns on the JP Test, but on the quizzes, the low scorer was more likely to have a nonaberrant test-response pattern.

Table 7
Correlations Among Quiz 2 Score,
The Modified Caution Index (MCI) Calculated on Quiz 2
And the Person Average R (PAR) Calculated on Quiz 2
n=278

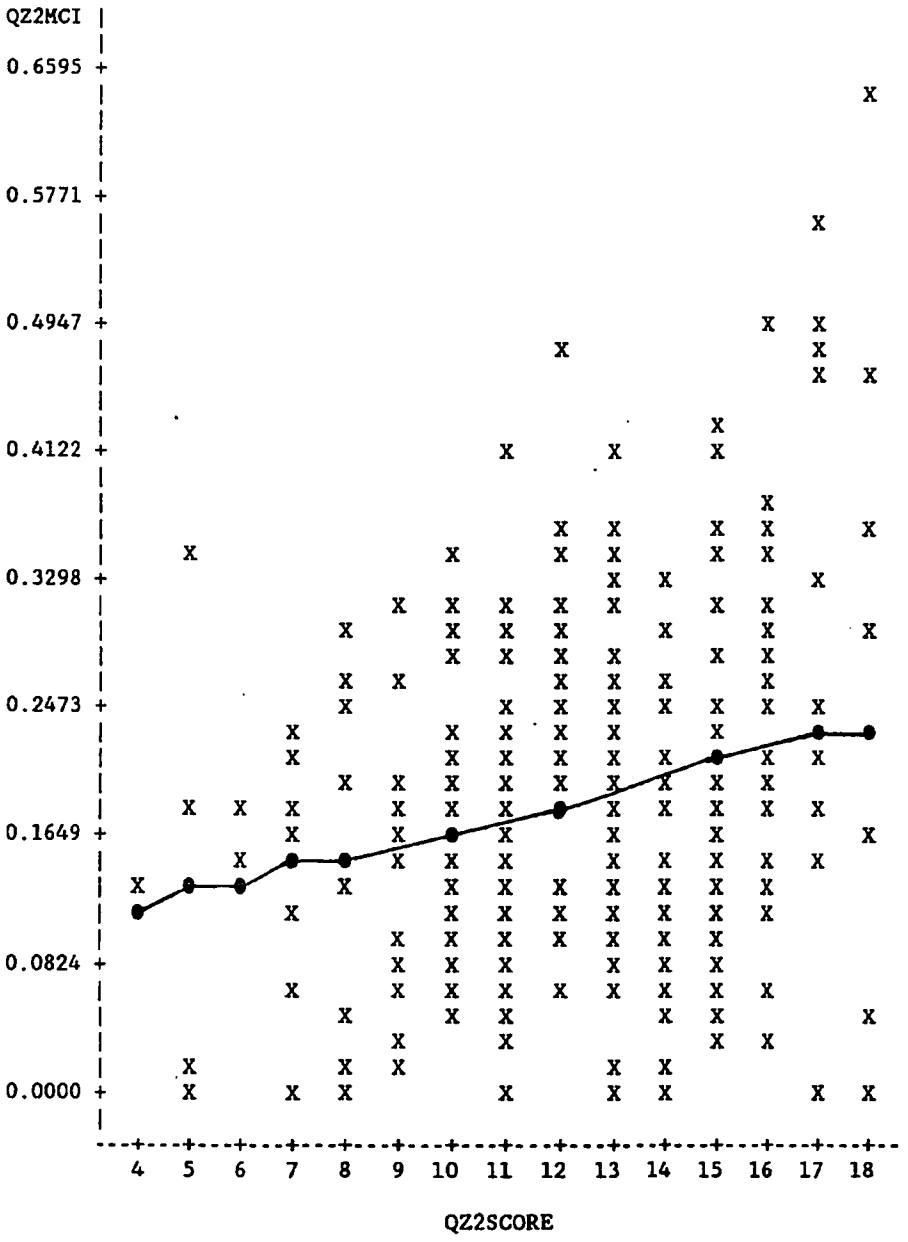
	Quiz 2 Score	Quiz 2 MCI	Quiz 2 PAR
Quiz 2 Score	1.0	.22**	-.40**
Quiz 2 MCI		1.0	-.93**
Quiz 2 PAR			1.0

* $p < .05$.
** $p < .01$.

Table 8
Correlations Among Quiz 3 Score,
The Modified Caution Index (MCI) Calculated on Quiz 3
And the Person Average R (PAR) Calculated on Quiz 3
n=236

	Quiz 3 Score	Quiz 3 MCI	Quiz 3 PAR
Quiz 3 Score	1.0	.30**	-.51**
Quiz 3 MCI		1.0	-.96**
Quiz 3 PAR			1.0

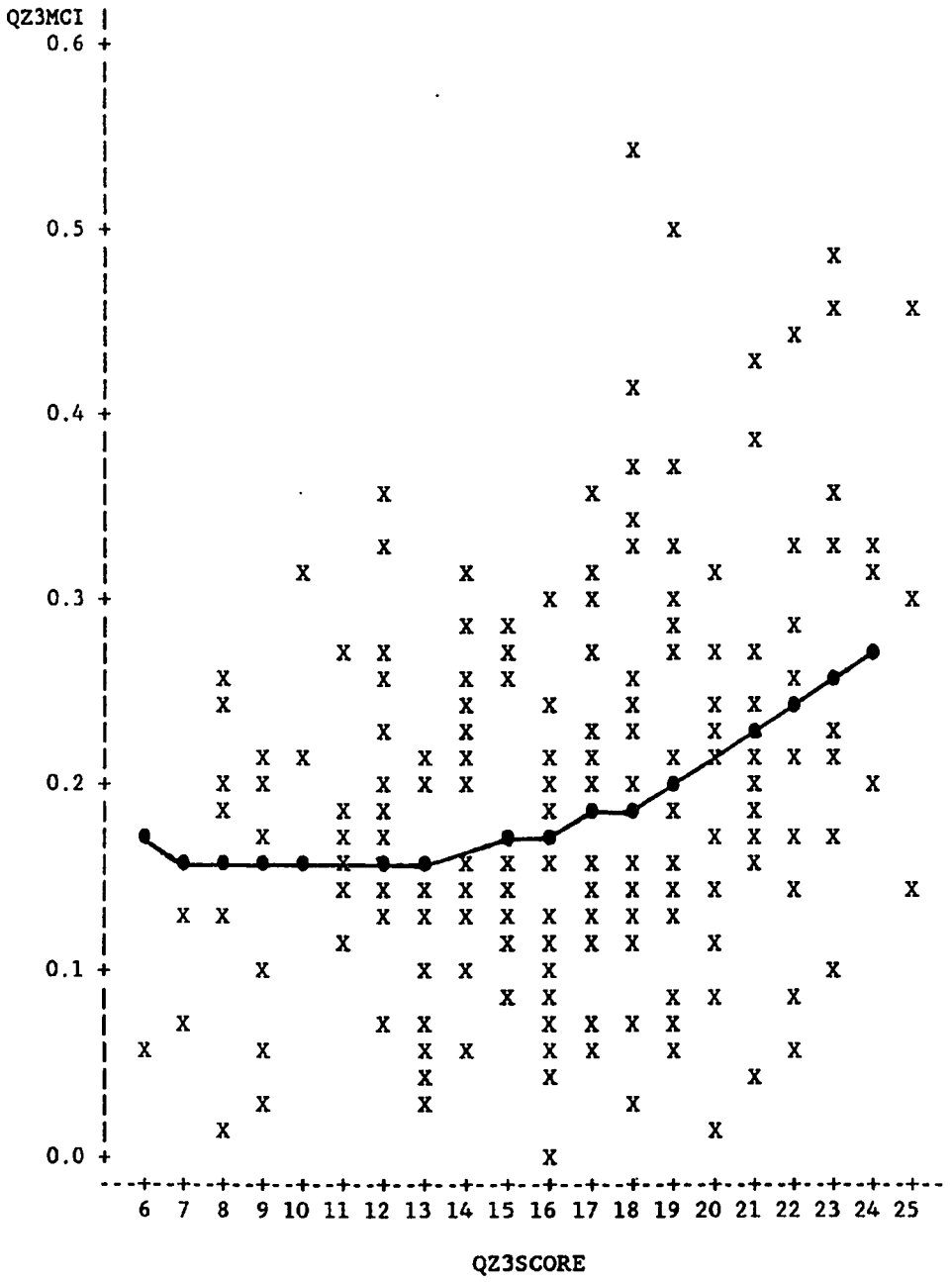
* $p < .05$.
** $p < .01$.



NOTE: 374 obs hidden.

Figure 6A. Bivariate plot of the Modified Caution Index (MCI) calculated on Quiz 2 with Quiz 2 Score.

(QZ2MCI = .087 + .008 QZ2Score $R^2 = .048$, $R = .219$, $n = 278$, $p < .01$)



OTE: 294 obs hidden.

Figure 6B. Bivariate plot of the Modified Caution Index (MCI) calculated on Quiz 3 with Quiz 3 Score.

$$(\text{QZ3 MCI} = .231 - .015\ \text{QZ3Score} + .001\ \text{QZ3Score}^2$$

$$R^2 = .112,\ R = .335,\ n = 236,\ p < .01)$$

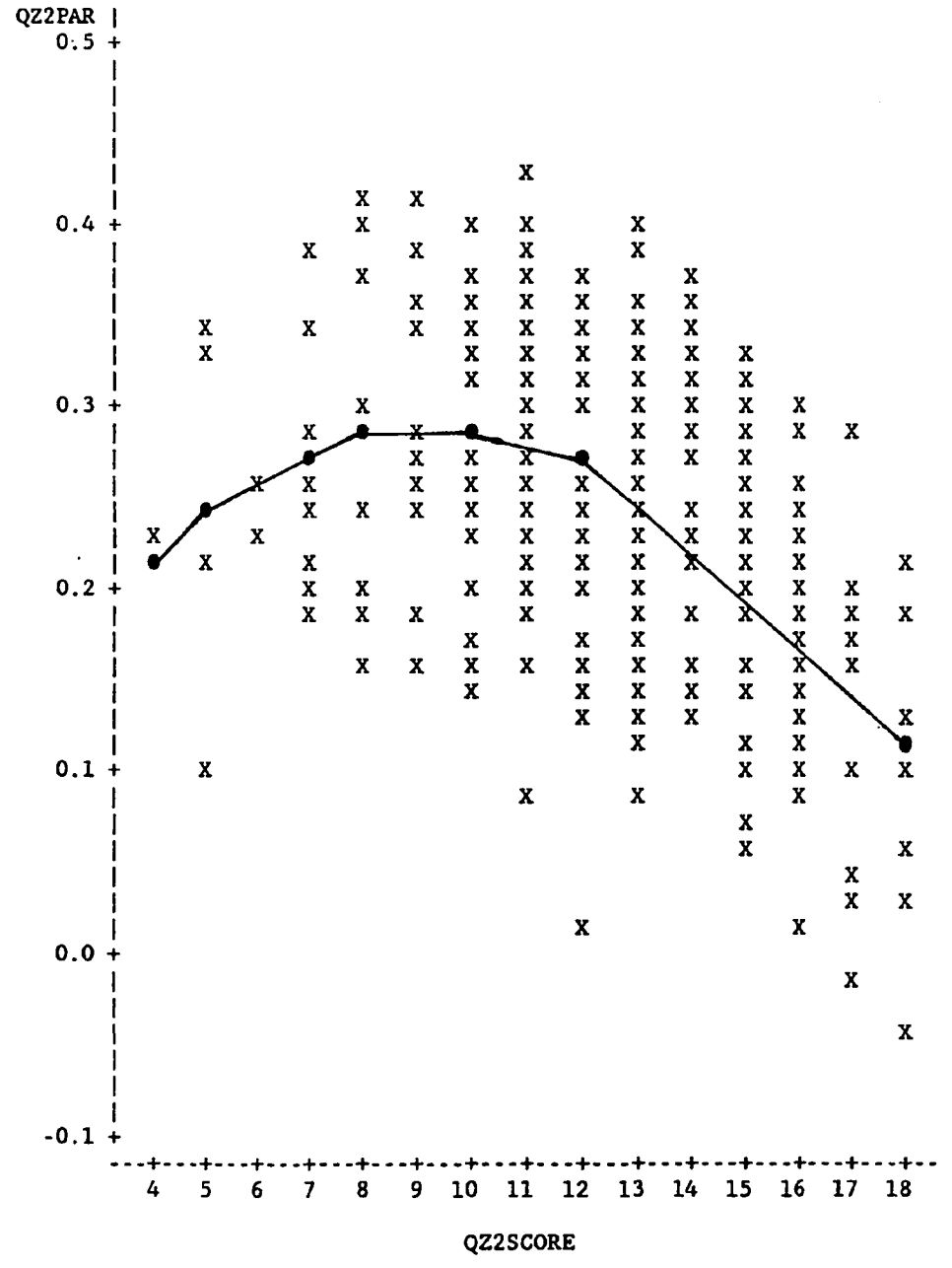


Figure 7A. Bivariate plot of the Person Average R (PAR) calculated on Quiz 2 with Quiz 2 Score.

$$(Z2PAR = .088 + .042\ QZ2Score - .0002\ QZ2Score^2$$

$$R^2 = .238,\ R = 488,\ n = 278,\ p < .01)$$

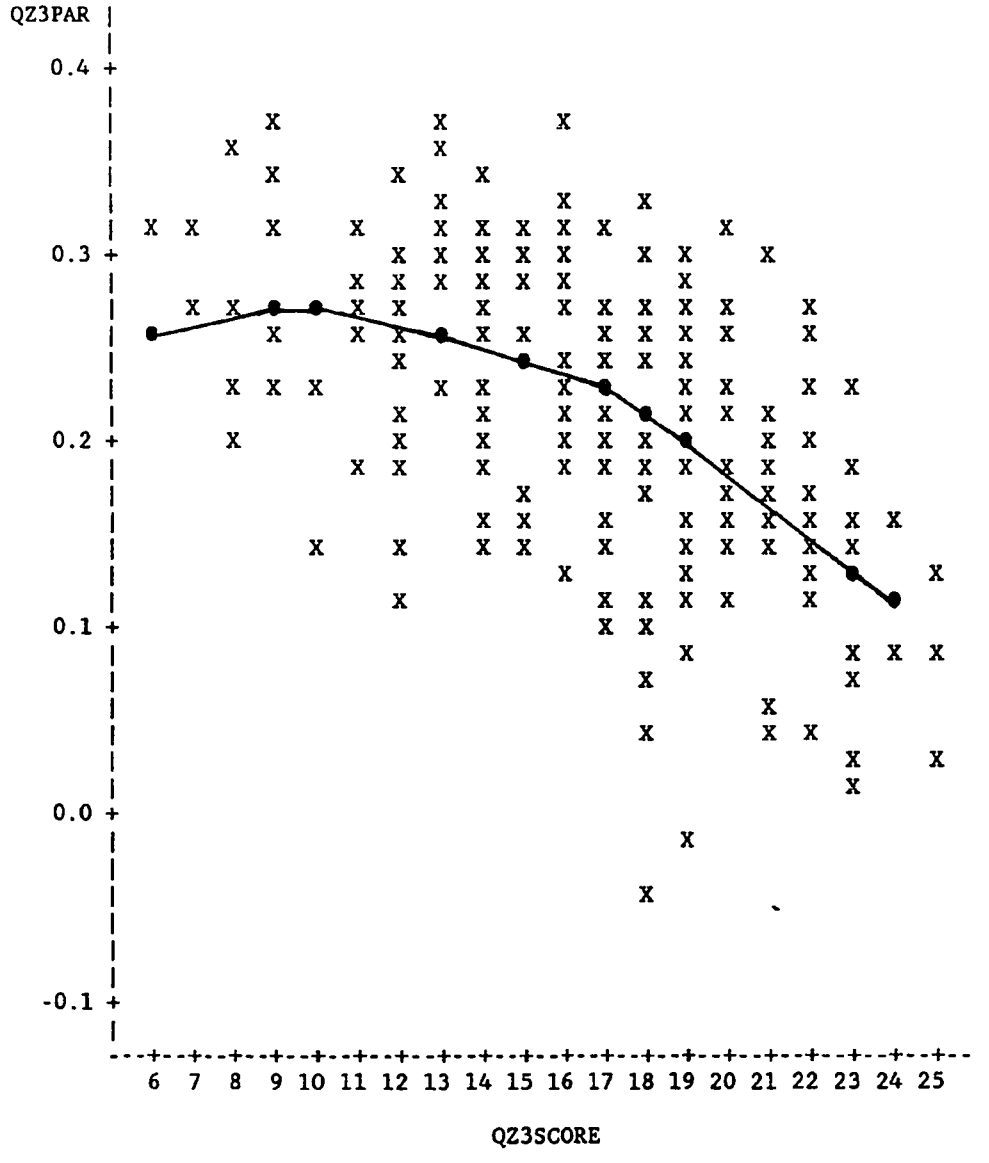


Figure 7B. Bivarite plot of the Person Average R (PAR) calculated on Quiz 3 with Quiz 3 Score.

(QZ3PAR = .207 + .014 QZ3Score - .0007 QZ3Score^2

$R^2 = .297,\ R = .545,\ n = 236,\ p < .01$)

Correlations Between MCI, PAR, and Z3 with the Four Reasons for Aberrant Response Patterns

In this section the correlations between the aberrance indices, MCI, PAR, and Z3 and the four hypothesized reasons for aberrant response patterns, Curricular Differences, Differences in Test-Taking Skills, Differences in Motivation, and Differences in Consistency of Academic Performance, will be presented. The correlations between the indices and the variables measuring the four hypothesized reasons are listed in Tables 9A through 9D.

Curricular Differences. Of the four variables which measure Curricular Differences, two were related to two of the indices and two were related to one index each (Table 9A).

One of the variables that was related to two indices was years of Japanese language studied in high school (HY). HY was related to MCI (r=-.26, $d.f.$=363, p<.01) and to PAR (r=.20, $d.f.$=363, p<.01), such that the more years of Japanese language studied, the less aberrant a student's response pattern tended to be. HY was not, however, related to Z3.

The other variable that was related to two indices was access to native speakers of Japanese (NS). NS was related to PAR and to Z3, but not to MCI. NS was related to PAR (r=-.14, $d.f.$=363, p<.01) and to Z3 (r=-.15, $d.f.$=363, p<.01). Test takers who had access to native speakers tended to have response patterns which were different from their fellow test takers as measured by PAR. Similarly, test takers with access to native speakers tended to have low Z3 indices.

Table 9A
Correlations Between The
Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Three Parameter Logistic Function (Z3)
And Variables Which Measure Curricular Differences

Variables		MCI	PAR	Z3	*n*
Years of Japanese language study in high school	HY	-.26**	.20**	-.09	365
Years of Japanese language study in language school	JY	-.15**	-.01	-.10	365
Years of residence in Japan	RY	.01	-.15**	-.09	365
Access to native Japanese language speakers	NS	-.02	-.14**	-.15**	365

* $p < .05$.
** $p < .01$.

Two variables which were related to only one index each were years of Japanese language studied in language school (JY) and years of residence in Japan (RY). Years of Japanese language studied in language school was related to MCI only (r=-.15, $d.f.$=363, p<.01). The more years Japanese was studied in language school, the less aberrant the response pattern tended to be, as measured by MCI. JY was not related to PAR or to Z3. Years of residence in Japan (RY) was also related to only one index. RY was related to PAR (r=-.15, $d.f.$=363, p<.01) but not to MCI or to Z3. As with access to native speakers, years of residence in Japan was associated with an aberrant response pattern.

Stated in another way, of the four variables measuring Curricular Differences, two were statistically significantly related to MCI, three were related to PAR and only one was related to Z3. MCI was related to years of Japanese language studied in high school (HY)

and to years of Japanese language studied in language school (JY). PAR was related to HY, to years of residence in Japan (RY) and to access to native Japanese language speakers (NS). Z3 was related only to access to native Japanese language speakers.

Differences in Test-Taking Skills. Though there were 15 variables which measured test-taking skills, only one variable was related to both MCI and PAR (Table 9B). The college Japanese language instructor's rating of the student's test-taking skills (C4) was related to MCI (r=-.25, $d.f.$=98, p<.05) and to PAR (r=.25, $d.f.$=98, p<.05) but it was not related to Z3. A rating of poor test-taking skills was associated with aberrant response patterns as measured with MCI and PAR. None of the other test-taking skills variables were related to either MCI or PAR.

Three test-taking skills variables, however, were related to Z3. Two were specific test-taking skills items. The student's response to, "I enjoy taking tests which are challenging," (X2) was related to Z3 (r=-.21, $d.f.$=219, p<.01) and the student's response to, "My mind wanders a lot when I take a test," (X6) was related to Z3 (r=-.14, $d.f.$=219, p<.05). In other words, test takers who enjoy challenging tests tended to have low Z3 indices and test takers with wandering minds tended to have high Z3 indices. Since Z3 is a standardized index, very high and very low Z3s indicate aberrant response patterns. Middling Z3s indicate nonaberrant response patterns.

The third variable which was related to Z3 was a general measure of test-taking skills. The student was asked to guess his/her score on the JP Test. The absolute value of the student's actual JP Test score minus his/her guessed score (OMG) measured the accuracy of the student's assessment of his/her own test performance. OMG was related to Z3 (r=.17, $d.f.$=211, p<.05), such that the more inaccurate a student's guessed score, the higher his or her Z3 index.

Table 9B
Correlations Between The
Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Three Parameter Logistic Function (Z3)
And Variables Which Measure Test-Taking Skills

Variables		MCI	PAR	Z3	*n*
Test-taking skills scale	TTT	-.06	.00	-.06	221
"When studying, I try to think up test questions."	X1	-.02	.03	.03	221
"I enjoy taking tests which are challenging."	X2	-.02	-.08	-.21**	221
"In taking tests, I misunderstand directions, and lose points."	X3	-.02	.01	-.02	221
"I'm so good at taking tests, I get a higher score than I deserve."	X4	-.04	.04	.08	221
"I use a process of elimination on multiple choice tests."	X5	-.09	.05	-.06	221
"My mind wanders a lot when I take a test."	X6	.05	-.10	-.14*	221
"I get so nervous and confused that I don't do well on tests."	X7	-.12	.08	.07	221
"I skip over difficult test questions and come back if I have time."	X8	-.08	.05	.06	221
"I look for clues in the way that a test questions is stated."	X9	-.06	.04	-.03	221
"I read all answer options before selecting an answer."	X10	.06	-.03	.06	221
Student's self rating of test-taking skills	X11	-.03	-.04	-.07	210
Observed JP Test score minus student's guessed score	OMG	.00	.11	.17*	213
High school Japanese language teacher's rating of student's test-taking skills	SD	-.09	.13	.08	143
College Japanese language instructor's rating of student's test-taking skills	C4	-.25*	.25*	-.02	100

* $p < .05$.
** $p < .01$.

In other words, of the 15 variables listed in Table 9B, MCI was related to one, PAR was related to one and Z3 was related to three variables.

Differences in Motivation. There were ten variables which measured motivation. Two were ratings of the student's motivation to do well in Japanese language study; one by the high school Japanese language teacher (SC) and one by the college Japanese language instructor (B4). The eight other motivation variables were specific statements regarding the JP Test. The student was asked to respond whether each of these eight statements were true or false for them.

Of the ten motivation variables, one variable was related to all three indices. V6, the student's response to, "I marked any old answer on most of the questions on the JP Test," was related to MCI (r=-.30, $d.f.$=95, p<.01), to PAR (r=.27, $d.f.$=95, p<.01), and to Z3 (r=.20, $d.f.$=95, p<.05). The student for whom V6 was true tended to have an aberrant response pattern.

Two other motivation variables were related to MCI and PAR but not to Z3. B4, the college Japanese language instructor's rating of the student's motivation to do well in Japanese language study, was related to MCI (r=-.24, $d.f.$=98, p<.05) and to PAR (r=.22, $d.f.$=98, p<.05). As one would expect, the poorly motivated student tended to have an aberrant response pattern. The other motivation variable which was related to MCI and to PAR but not to Z3 was V9, the student's response to, "I wanted to do the best I could on the JP Test." V9 was related to MCI at r=.29 ($d.f.$=95, p<.01) and to PAR at r=-.28 ($d.f.$=95, p<.01), such that students for whom V9 was true tended to have a nonaberrant response pattern.

In addition, one motivation variable each was related to MCI only, to PAR only, and to Z3 only. The variable that was related to MCI only was V7, the student's response to, "I answered only the questions for which I knew the correct answer." It was related to MCI at r=-.22 ($d.f.$=95, p<.05). Students who answered only questions for which they

knew the correct answer tended to have aberrant response patterns as measured by MCI. The variable that was related to PAR only was SC, the high school Japanese language teacher's rating of student's motivation to do well in Japanese language study. SC was related to PAR at r=.23 ($d.f.$=142, p<.01), with poorly motivated students tending to have aberrant response patterns. The variable that was related to Z3 only was V2, the student's response to, "I just skipped many of the questions on the JP Test." V2 was related to Z3 at r=-.30 ($d.f.$=95, p<.01). Students for whom V2 was true tended to have high Z3 indices.

In all, MCI was related to four of the ten motivation variables. It was related to the college instructor's rating of the student's motivation to do well in Japanese language study, and to three JP Test specific motivation variables, V6, V7, and V9.

PAR was also related to four motivation variables. It was related to both the high school Japanese language teacher's and the college Japanese language instructor's ratings of the student's motivation to do well in Japanese language study. In addition, PAR was also related to V6 and to V9.

Table 9C
Correlations Between The
Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Three Parameter Logistic Function (Z3)
And Variables Which Measure Motivation

Variables		MCI	PAR	Z3	*n*
High school Japanese language teacher's rating of student's motivation to do well in Japanese language study	SC	-.16	.23**	.15	144
College Japanese language instructor's rating of student's motivation to do well in Japanese language study	B4	-.24*	.22*	.05	100
"I just skipped many of the questions on the JP Test."	V2	-.05	.02	-.30**	97
"I guessed on some of the questions if I sort of knew the answer."	V3	.13	-.07	-.10	97
"I didn't want to do too well on the JP Test."	V4	-.09	.10	-.05	97
"I tried to answer every question on the JP Test."	V5	.11	-.08	.09	97
"I marked any old answer on most of the questions on the JP Test."	V6	-.30**	.27**	.20*	97
"I answered only the questions for which I knew the correct answer."	V7	-.22*	.19	.05	97
"I tried to get the highest score I could on the JP Test."	V8	.09	-.09	.01	97
"I wanted to do the best I could on the JP Test."	V9	.29**	-.28**	-.19	97

* $p < .05$.
** $p < .01$.

Z3 was related to only two of the ten motivation variables. Z3 was related to V2, which was related neither to MCI nor to PAR. Z3 was also related to V6, which was related to both MCI and to PAR.

Differences in Consistency of Academic Performance. There were two variables which measured Consistency of Academic Performance, a rating by the high school Japanese language teacher and a rating by the college Japanese language instructor. One of them was related to both MCI and PAR (Table 9D). E4, the college instructor's rating of the student's consistency of academic performance, was related to MCI (r=.38, $d.f.$=98, p<.01) and to PAR (r=-.37, $d.f.$=98, p<.01), such that a rating of inconsistency was associated with aberrant response patterns. Neither SE nor E4 was related to Z3.

Table 9D
Correlations Between The
Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Three Parameter Logistic Function (Z3)
And Variables Which Consistency in Academic Performance

Variables		MCI	PAR	Z3	*n*
High school Japanese language teacher's rating of student's consistency	(SE)	.13	-.11	-.07	136
College Japanese language instructor's rating of consistency	(E4)	.38**	-.37**	-.06	100

* $p < .05$.
**$p < .01$.

To summarize this section on the relationships of the aberrance indices to the four hypothesized reasons for aberrant response patterns, Curricular Differences and Motivation seem to be related to MCI and to PAR and related somewhat to Z3. Test-Taking Skills seems to be related to Z3 but not to MCI and PAR. Consistency of Academic Performance seems to be related to MCI and PAR but not to Z3.

Three of the four Curricular Differences variables were related to PAR, two were related to MCI and one of the four variables was related to Z3.

Only one of the 15 Test-Taking Skills variables was related to MCI and to PAR. Three specific variables, however, were related to Z3.

Four of the ten Motivation variables were related to PAR and MCI and two of the ten variables were related to Z3. In addition, of the 31 variables studied, the only variable which was related to all three indices was V6, a motivation variable (the student's response to, "I marked any old answer on most of the questions on the JP Test," see Table 9C).

Consistency of Academic Performance was related to MCI and PAR, but not to Z3. Though there were only two variables measuring Consistency of Academic Performance, one of the them was related to both MCI and PAR, and those correlations were the highest among the 31 variables studied.

Regression analysis was done to find the variance in the indices accounted for by the variables studied. The following are the indices and the variables which were related to each of the indices and the corresponding statistics for those regression models:

MCI:	HY JY C4 B4 V6 V7 V9 E4	$R^2 = .32$	$R = .57$	$n = 94$	$p < .01$
PAR:	HY RY NS C4 SC B4 V6 V9 E4	$R^2 = .37$	$R = .61$	$n = 40$	n. s.
Z3:	NS X2 X6 OMG V2 V6	$R^2 = .12$	$R = .37$	$n = 61$	n. s.

Correlations Between MCI, PAR, and Z3 with the Four Reasons for Aberrant Response Patterns, with JP Test Score Partialed

Since JP Test score was related to many of the 31 variables studied as well as to the indices, the correlations between the indices and the variables of interest were recalculated with JP Test score partialed. The results are presented in Tables 10A through 10D, which are similar in format to Tables 9A through 9D, except for the column labelled "JP Test." The numbers under the JP Test column are the correlations between JP Test score and the variables listed to the left.

Curricular Differences. In the previous section, statistically significant correlations were found between MCI and two variables (HY and JY); between PAR and three variables (HY, RY and NS); and between Z3 and one variable (NS). With JP Test score partialed, MCI and PAR were found to be related to the same three variables (HY, RY, and NS) and Z3 was related to none of the variables.

All three correlations were slightly higher between PAR and the three variables than between MCI and the three variables. As can be seen in Table 10A, all four Curricular Differences variables were related to JP Test score, with the highest correlation being between JP Test score and years of language study in Japanese language school (JY). With JP Test score partialed, however, JY was the only variable which was not related to any of the indices.

Table 10A
Correlations Between The
Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Three Parameter Logistic Function (Z3)
And Variables Which Measure Curricular Differences
WITH JAPANESE PLACEMENT TEST (JP TEST) SCORE PARTIALED

Variables		JP Test	MCI	PAR	Z3	*n*
Years of Japanese language study in high school	HY	.40**	-.16**	.21**	.04	365
Years of Japanese language study in language school	JY	.45**	-.01	-.01	.05	365
Years of residence in Japan	RY	.29**	.11*	-.16**	.00	365
Access to native Japanese language speakers	NS	.39**	.11*	-.16**	-.03	365

* $p < .05$.
** $p < .01$.

Differences in Test-Taking Skills. With JP Test score partialed, of the 15 variables which measure test-taking skills, the only variable that was related to any of the indices was the student's observed minus guessed score (OMG). OMG was related to PAR at r=-.15 ($d.f.$=211, p<.05), with the direction of the correlation indicating that students who guessed their test scores inaccurately tended to have aberrant response patterns.

In the previous section, MCI and PAR were related to college Japanese language instructor's rating of the student's test-taking skills (C4) and Z3 was related to OMG and to two specific test-taking skills questions, X2 and X6 (see Table 9B). These four variables, C4, OMG, X2 and X6, were statistically related to JP Test score (Table 10B). With JP Test score partialed, the relationships between these variables and the aberrance indices were no longer statistically significant.

Table 10B
Correlations Between The
Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Three Parameter Logistic Function (Z3)
And Variables Which Measure Differences in Test-Taking Skills
WITH JAPANESE PLACEMENT TEST (JP TEST) SCORE PARTIALED

Variables		JP Test	MCI	PAR	Z3	*n*
Test-taking skills scale	TTT	.17*	-.01	.00	.00	221
"When studying, I try to think up test questions."	X1	.01	-.02	.03	.03	221
"I enjoy taking tests which are challenging."	X2	.27**	.06	-.09	-.13	221
"In taking tests, I misunderstand directions, and lose points."	X3	.11	.02	.00	.02	221
"I'm so good at taking tests, I get a higher score than I deserve."	X4	-.03	-.05	.04	.07	221
"I use a process of elimination on multiple choice tests."	X5	.14*	-.05	.04	-.01	221
"My mind wanders a lot when I take a test. "	X6	.16*	.10	-.10	-.09	221
"I get so nervous and confused that I don't do well on tests."	X7	.09	-.10	.08	.10	221
"I skip over difficult test questions and come back if I have time."	X8	.01	-.08	.05	.07	221
"I look for clues in the way that a test questions is stated."	X9	.08	-.04	.03	.00	221
"I read all answer options before selecting an answer."	X10	-.12	.03	-.03	.02	221
Student's self rating of test-taking skills	X11	.20**	.03	-.04	.01	210
Observed minus student's guessed score	OMG	-.26**	.10	-.15*	-.12	213
High school Japanese language teacher's rating of student's test-taking skills	SD	.06	-.08	.13	.11	143
College Japanese language instructor's rating of student's test-taking skills of student's test-taking skills	C4	.35**	-.11	.17	.05	100

* $p < .05$.
** $p < .01$.

Differences in Motivation. In the previous section, PAR was found to be related to four motivation variables (high school teacher's and college instructor's ratings of motivation, and the student's response to two motivation statements, V6 and V9); MCI was also found to be related to four variables (college teacher's rating of motivation, and the student's response to three motivation variables, V6, V7, and V9); and Z3 was found to be related to two variables (the student's response to two motivation statements, V2 and V6). With JP Test score partialed, three of the relationships which were statistically significant in the previous section were no longer important. However, two correlations which were unimportant in the previous section were statistically significant when JP Test score was partialed.

The college Japanese language instructor's rating of the student's motivation to do well in Japanese language study (B4) was related to MCI and PAR in the previous section (Table 9C) but when JP Test score was partialed, the relationships were no longer significant. Similarly, V7 which was related to MCI in the previous section, was no longer related when when JP Test score was partialed.

One of the two relationships which were not statistically significant in the previous section but which was significant when JP Test score was partialed, was the high school Japanese language teacher's rating of the student's motivation to do well in Japanese language study (SC). SC was related only to PAR in the previous section, but when JP Test score was partialed, SC was also related to Z3. The correlation between SC and Z3 increased from a statistically nonsignificant r=.15 to r=.17 ($d.f.$=142, p<.05) when JP Test was partialed.

In addition, the correlation between V9 and Z3 increased from a nonsignificant r=-.19 to a statistically significant correlation of r=-.23 ($d.f.$=95, p<.05) when JP Test score was partialed (Table 10C). Thus, in the previous section, only one motivation variable (V6) was related to all three indices, but with JP Test score was partialed, two of

the ten motivation variables were related to all three aberrance indices. In addition to V6, the student's response to, "I marked any old answer on most of the questions on the JP Test," V9, the student's response to, "I wanted to do the best I could on the JP Test" was also related to all three indices (Table 10C).

The relationships between the indices and the variables measuring Differences in Motivation can be summarized in the following way: two variables (V6 and V9) were related to all three indices, one variable (SC) was related to PAR and Z3 but not to MCI, and one variable (V2) was related to Z3 only.

Students who marked any old answer and students who did not want to do their best on the JP Test tended to have aberrant response patterns. Students who were rated as poorly motivated by their high school Japanese language teachers tended to have aberrant response patterns. And students who said that they skipped many of the JP Test questions tended to have a higher Z3 index. As mentioned in the previous section, very high or very low Z3 indices indicate aberrant response patterns.

Table 10C
Correlations Between The
Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Three Parameter Logistic Function (Z3)
And Variables Which Measure Differences in Motivation
WITH JAPANESE PLACEMENT TEST (JP TEST) SCORE PARTIALED

Variables		JP Test	MCI	PAR	Z3	*n*
High school Japanese language teacher's rating of student's motivation to do well in Japanese study	SC	.02	-.16	.23**	.17*	144
College Japanese language instructor's rating of student's motivation to do well in Japanese study	B4	.30**	-.11	.14	.11	100
"I just skipped many of the questions on the JP Test."	V2	.45**	.17	-.11	-.23*	97
"I guessed on some of the questions if I sort of knew the answer."	V3	-.20	.06	-.02	-.15	97
"I didn't want to do too well on the JP Test."	V4	.25*	.02	.04	.01	97
"I tried to answer every question on the JP Test."	V5	-.26*	.00	-.01	.03	97
"I marked any old answer on most of the questions on the JP Test."	V6	.17	-.25*	.24*	.25*	97
"I answered only the questions for which I knew the correct answer."	V7	.28**	-.11	.12	.13	97
"I tried to get the highest score I could on the JP Test."	V8	-.15	.03	-.05	-.03	97
"I wanted to do the best I could on the JP Test."	V9	-.11	.27**	-.26*	-.23*	97

* $p < .05$.
** $p < .01$.

Table 10D
Correlations Between The
Modified Caution Index (MCI), the Person Average R (PAR)
And the Standardized Three Parameter Logistic Function (Z3)
And Variables Which Measure Consistency of Academic Performance
WITH JAPANESE PLACEMENT TEST (JP TEST) SCORE PARTIALED

Variables		JP Test	MCI	PAR	Z3	*n*
High school Japanese language teacher's rating of student's margin of error	SE	-.14	.10	-.10	-.14	136
College Japanese language instructor's rating of student's margin of error	E4	.37**	.26**	-.30**	-.14	100

* $p < .05$.
** $p < .01$.

Differences in Consistency in Academic Performance. There was no change in the pattern of correlations between the three indices and the two variables which measure Consistency of Academic Performance when JP Test score was partialed. However, the degree of the relationships decreased slightly between college instructor's rating of the student's consistency of academic performance (E4) and MCI (from r=.38 in Table 9D to r=.26, $d.f.$=98, p<.01 in Table 10D) and between E4 and PAR (from r=-.37 in Table 9D to r=-.30, $d.f.$=98, p<.01 in Table 10D).

To summarize this section on the correlations between the indices and the four reasons for aberrant response patterns with JP Test score partialed, one can say that Curricular Differences and Differences in Consistency of Academic Performance were related to MCI and to PAR but not to Z3, Differences in Motivation was related to all three indices, and Differences in Test-Taking Skills was minimally related to PAR only.

Effectiveness of MCI, PAR, and Z3

In this section, results regarding the effectiveness of using MCI, PAR, and Z3 as aids in interpreting individual test scores will be presented.

The effectiveness of using MCI, PAR, and Z3 as aids in interpreting individual test scores was evaluated by first correlating these indices with three ratings of appropriateness of placement level; the high school Japanese language teacher's rating of the appropriateness of the student's placement level (HST), the college Japanese language instructor's rating of the appropriateness of the student's placement level (CT), and the student's self rating of the appropriateness of his or her placement level (ZZ10). Only the correlations between the high school Japanese language teacher's rating with MCI and PAR were statistically significant (Table 11). HST was related to MCI at r=.24 ($d.f.$=115, p<.05) and to PAR at r=-.25 ($d.f.$=115, p<.01), such that a rating of inappropriate placement was associated with a tendency toward aberrant response patterns.

Further analysis using the high school teacher's rating of appropriateness was done.

Table 11
Correlations between the Modified Caution Index (MCI), The Person Average R (PAR) and the Standardized Logistic Maximum Likelihood Function (Z3), And Appropriateness of Placement Ratings By High School Teachers (HST), College Instructors (CT) And Students

	MCI	PAR	Z3
Rating by high school teacher (HST, n=117)	.24*	-.25**	-.05
Rating by college instructor (CT, n=100)	-.04	.04	.05
Student's self rating (ZZ10, n=97)	-.08	.05	-.01

* $p < .05$.
** $p < .01$.

One hundred seventeen students whose placements were rated by their high school Japanese language teachers enrolled in Japanese language courses at UH Manoa. Of these, 87 students had placements which their high school teachers considered appropriate and 30 had placements which their teachers considered inappropriate. Ten of the inappropriate placements were considered too high and 20 were considered too low. Thus, without the use of aberrance indices, according to the high school teacher, 74 percent of placements were appropriate and 26 percent were inappropriate.

A method using MCI for identifying students for whom special caution is needed in interpreting their test scores was applied (see page 10 of the Introduction). The JP Test cut scores used by the Department of East Asian Languages and Literatures for UHM Japanese

language course levels are as follows: 20 for differentiating between Jpn 101 and Jpn 102 levels, 30 for differentiating between Jpn 102 and Jpn 201 levels, 40 for differentiating between Jpn 201 and Jpn 202 levels, and 50 for differentiating between Jpn 202 and Jpn 301 levels. Following recommendations by Harnisch (1983), quadrants were delineated within each of these cutting score intervals using the score interval midpoint and MCI = .3. These quadrants are labelled A, B, C, and D in Figure 8. The five cutting score intervals resulted in five sets of quadrants. Within each cutting score interval, students in quadrant A had high JP Test scores and low MCIs; students in quadrant B had high JP Test scores and high MCIs; students in quadrant C had low JP Test scores and low MCIs; and students in quadrant D had low JP Test scores and high MCIs. The number of cases within the quadrants, by appropriateness ratings of high school Japanese language teachers, were summed across the five sets of quadrants. The results are presented in Table 12A.

The row totals of quadrants A and C together comprised 81 percent of the data in Table 12A. As measured by MCI, these 81 percent represent students with nonaberrant response patterns. The combined 19 percent of students in quadrants B and D had MCIs greater than .3, indicating that they had aberrant response patterns.

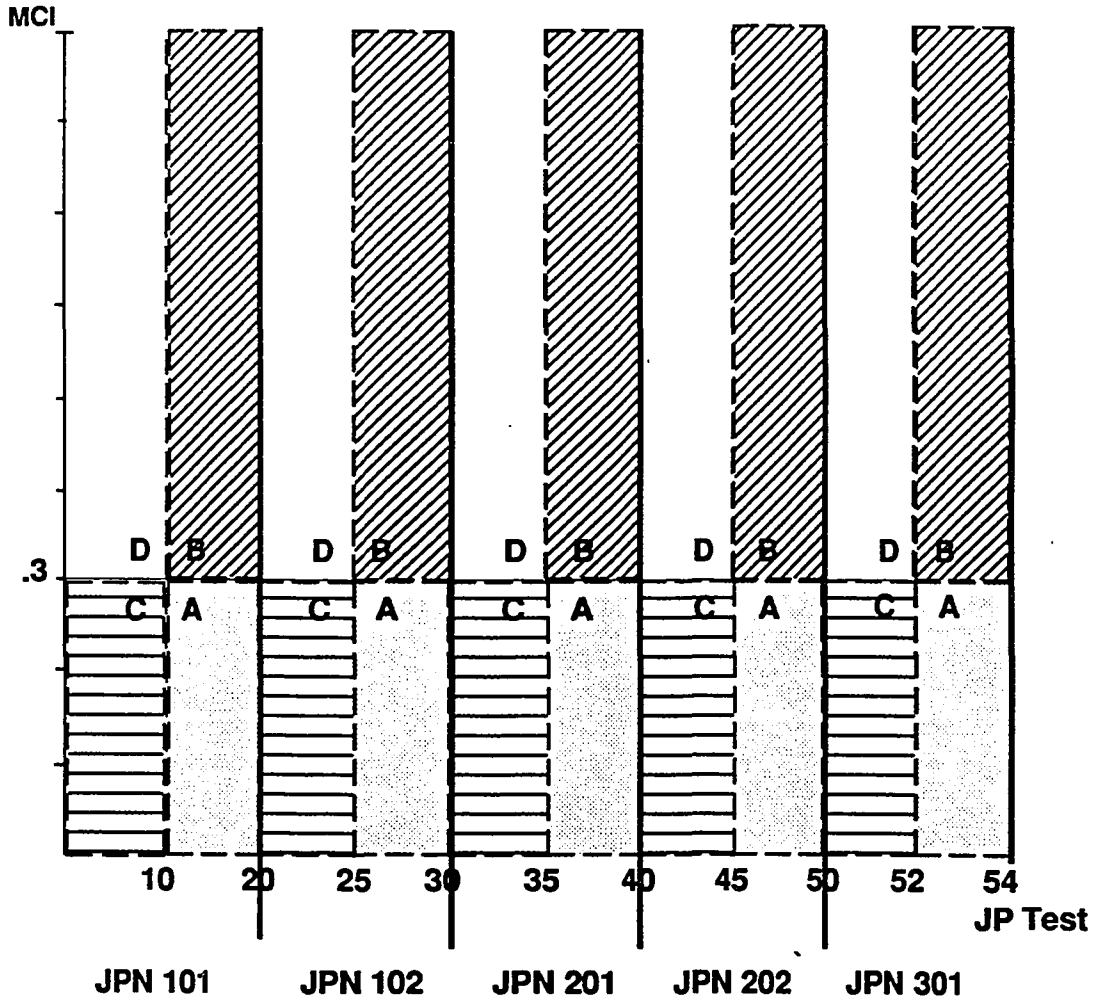


Figure 8. Quadrants within each cutting score interval were drawn using the interval midpoint and MCI = .3. For each cutting score interval,

A = High JP Test score, low MCI

B = High JP Test score, high MCI

C = Low JP Test score, low MCI

D = Low JP Test score, high MCI

The percentages under the Appropriate and Inappropriate columns in Table 12A are percentages of the row totals. Thus, for example, the seven students in quadrant B whose placements were rated as inappropriate by their high school Japanese language teachers represent 54 percent of the 13 students in quadrant B. Because the overall percentage of inappropriate placements was 26 percent, special attention should be paid to the placement levels of students who fall in quadrant B. Attention should also be paid to students who fall in quadrant C where 28 percent of the placement levels were considered inappropriate by high school Japanese language teachers.

Table 12A
Number of Students in JP Test Score by Aberrant Index Quadrants By Appropriateness of Placement Ratings By
High School Teachers for Modified Caution Index (MCI)
$n = 117$

	Appropriate Placement	Inappropriate Placement	Row Total (Pct of Total)
Quadrant A High JP Test, Low MCI	48 81%	11 19%	59 (50%)
Quadrant B High JP Test, High MCI	6 46%	7 54%	13 (11%)
Quadrant C Low JP Test, Low MCI	26 72%	10 28%	36 (31%)
Quadrant D Low JP Test, High MCI	7 78%	2 22%	9 (8%)

The overall proportion of placement levels which were rated too high versus too low by high school teachers was 1 to 2. Within each of the quadrants, the proportion of too high to too low was as follows:

Quadrant A	3 : 8
Quadrant B	3 : 4
Quadrant C	4 : 6
Quadrant D	0 : 2

Quadrants A and D had disproportionately high numbers of students with placement levels which were considered too low by their high school teachers.

Harnisch's recommended method was applied to the data using PAR and high school teacher's appropriateness of placement ratings. A cutoff of PAR = .1 was used. The results are presented in Table 12B and are similar to the results obtained with MCI. The proportion of too high to too low using PAR was as follows:

Quadrant A	3 : 10
Quadrant B	3 : 2
Quadrant C	4 : 5
Quadrant D	0 : 3

Compared to the analysis using MCI, the disproportion in the number of students in quadrants A and D with placement levels which were considered too low by their high school teachers was slightly more pronounced in the analysis using PAR.

Harnisch's method was applied to Z3 (Table 12C), using a cutoff point of Z3 within 1.35 standard deviations from the mean in order to approximate the MCI aberrant to nonaberrant proportion.

Table 12B
Number of Students in JP Test Score by Aberrant Index Quadrants By Appropriateness of Placement Ratings By
High School Teachers for Person Average R (PAR)
$n = 117$

	Appropriate Placement	Inappropriate Placement	Row Total (Pct of Total)
Quadrant A High JP Test, High PAR	49 79%	13 19%	62 (53%)
Quadrant B High JP Test, Low PAR	5 50%	5 50%	9 (9%)
Quadrant C Low JP Test, High PAR	26 74%	9 26%	35 (30%)
Quadrant D Low JP Test, Low PAR	7 70%	3 30%	10 (9%)

Table 12C
Number of Students in JP Test Score by Aberrant Index Quadrants By Appropriateness of Placement Ratings by High School Teachers
For Standardized Logistic Maximum Likelihood Function (Z3)
$n = 117$

	Appropriate Placement	Inappropriate Placement	Row Total (Pct of Total)
Quadrant A High JPTest, Z3 within 1.35 s.d.	43 77%	13 23%	56 (48%)
Quadrant B High JPTest, Z3 > 1.35 s.d.	11 69%	5 31%	16 (14%)
Quadrant C Low JPTest, Z3 within 1.35 s.d.	21 72%	8 28%	29 (25%)
Quadrant D Low JPTest, Z3 > 1.35 s.d.	12 75%	4 25%	16 (14%)

Evidence of Sandbagging

Correlations Between Appropriateness Ratings and Indices. Sandbagging occurs on a placement test when a test taker deliberately selects an incorrect answer option in order to earn a low score and thus to gain entry into a course level below his/her ability level. Correlations between appropriateness of placement ratings and JP Test score, quiz scores and indices at the Jpn 102 level were examined for evidence of the occurrence of sandbagging on the JP Test. The numbers of observations are smaller than those in previous sections of this chapter because the analysis was confined to one course level.

Three appropriateness of placement ratings, one by the high school Japanese language teacher, one by the college Japanese language instructor and one by the student, were correlated with the JP Test score, with the two quiz scores, and with MCI and PAR calculated on the JP Test and on the two quizzes. These correlation coefficients are listed in Table 13. Because the numbers of subjects were small, even a correlation coefficient of .41 was not statistically significant. However, the pattern of correlations is worth noting. For example, the high school Japanese language teacher's appropriateness of placement rating was not correlated with JP Test score, but it was correlated with Quiz 2 (r=.41, $d.f.$=21, n.s.) and 3 (r=.54, $d.f.$=20, p<.01) scores. The higher the student's quiz scores, the more likely that his or her placement level was considered inappropriate by his or her high school teacher. This pattern of correlations was repeated, to a lesser degree, between the college Japanese language instructor's appropriateness rating with JP Test and quiz scores.

Similarly, PAR calculated on the JP Test was not correlated with high school teacher's or college instructor's appropriateness of placement rating, but PARs calculated on the two quizzes were correlated to appropriateness ratings. High school Japanese language teacher's appropriateness of placement rating was correlated with Quiz 2 PAR at r=-.51 ($d.f.$=21, p<.05). The college instructor's appropriateness rating was correlated to

Quiz 2 PAR at r=-.40 ($d.f.$=42, p<.01) and to Quiz 3 PAR at r=-.31 ($d.f.$=40, p<.05). In other words, students whose placements were rated as appropriate tended to have non-aberrant response patterns on the quizzes, but not necessarily so on the JP Test.

The appropriateness of placement self rating by students did not have the same pattern of correlations with test scores and indices as the other two ratings of appropriateness.

Table 13
Evidence of Sandbagging
Correlations of Appropriateness of Placement Ratings
By High School Teachers, College Instructors and Students
With Japanese Placement Test (JP Test) Score, Quiz Scores
And Modified Caution Index (MCI) and Person Average R (PAR)
Calculated on the JP Test and Quizzes
Jpn 102

	Appropriateness Ratings					
	By H.S. Teacher		By College Prof		By Student	
	r	*n*	*r*	*n*	*r*	*n*
JP Test Score	.10	25	.02	47	-.14	47
Quiz 2 Score	.41	23	.19	44	.07	44
Quiz 3 Score	.54**	22	.32*	43	-.20	43
MCI on JP Test	.00	25	.02	47	-.04	47
Quiz 2 MCI	.36	23	.27	44	.12	44
Quiz 3 MCI	.19	22	.22	42	.09	42
PAR on JP Test	-.02	25	-.02	47	.04	47
Quiz 2 PAR	-.51*	23	-.40**	44	-.12	44
Quiz 3 PAR	-.27	22	-.31*	42	.03	42

Note: The appropriateness ratings were recoded as 1=appropriate and 2=not appropriate for this analysis.

* $p < .05$.
** $p < .01$.

Student Feedback on Sandbagging. Students' perception of the occurrence of sandbagging was measured with Questionnaire 3 (Appendix D). Students were asked whether it is true that "students who take the Placement Test deliberately do poorly in order to get into an easy Japanese language course." Their responses are summarized in Table 14. Overall, 32 percent responded either "I don't know" or "No, I've never heard of anyone doing that," 43 percent responded "I don't know of anyone who actually did that but I'm sure some students have done it," and 25 percent responded "Yes, I think that it's rather common." The respondents to this question were divided into two groups according to whether or not they had taken the JP Test. Twenty-eight percent of the respondents who had taken the JP Test (placed students) felt that sandbagging on the JP Test was common, compared to 18 percent of respondents who had not taken the JP Test (nonplaced). If the nonplaced students' 18 percent response of "sandbagging is common" is taken as a baseline reflecting speculation as opposed to personal experience, then an estimate of the occurrence of sandbagging on the JP Test might be 28 minus 18, or 10 percent.

Also on Questionnaire 3, respondents were asked to estimate the percentage of their classmates who, in their opinion, were enrolled at the appropriate Japanese language course level, the percentage who were too advanced for the course and the percentage who they felt ought to have enrolled in a lower level. Their responses are summarized in Table 15.

As can be seen in Table 15, overall, students responded that an average of 82 percent of their classmates seemed to belong in the class, 7 percent of their classmates seemed too slow for the class, and 11 percent of their classmates seemed too advanced for the class. If normal distribution of foreign language talent is assumed, then the 11 minus 7, or four percentage points of classmates perceived to be too advanced for the class may be those too advanced due to having engaged in a little sandbagging on the JP Test.

Thus it can be estimated that between four and ten percent of students who take the JP Test deliberately perform poorly in order to gain entry into a relatively easy Japanese language course level.

Table 14
Sandbagging as Perceived by Students
Overall, and by Placed Versus Not Placed

	Overall		Doesn't Occur		Don't Know		Yes, It Occurs		Yes, It's Common	
	N	V%	N	H%	N	H%	N	H%	N	H%
TOTAL	357	100.0	20	5.6	93	26.1	153	42.9	91	25.5
Placed	269	75.4	14	5.2	54	20.1	126	46.8	75	27.9
Not-Placed	88	24.6	6	6.8	39	44.3	27	30.7	16	18.2

Variable V23: **Is it true that students who take the Placement Test deliberately do poorly in order to get into an easy Japanese language course?**

_____ No, I've never heard of anyone doing that.
_____ I don't know.
_____ I don't know of anyone who actually did that but I'm sure some students have done it.
_____ Yes, I think that it's rather common.

Table 15
Average Percentage Students who Are
Perceived by the Classmates As
Too Advanced, Too Slow or Belonging in the Course
By Placed Versus Not Placed

	Too Advanced			Belonging			Too Slow		
	n	%	std dev	n	%	std dev	n	%	std dev
TOTAL	330	11.0	16.0	339	82.3	19.2	325	7.4	12.0
Placed	250	10.6	14.7	257	82.8	19.1	247	7.4	11.6
Not Placed	80	12.2	18.3	82	82.0	19.7	78	7.3	13.2

IV. DISCUSSION

One objective in testing is to obtain a measure of an individual's ability level in order that some decision regarding that individual can be made. The number correct or total score is commonly used, sometimes in conjunction with the standard error of measurement. In this study, aberrance indices were proposed as possible alternative aids in interpreting test scores.

The three indices studied were the Modified Caution Index (MCI), the Person Average R (PAR), and the standardized three parameter logistic maximum likelihood function (Z3) which is based on item response theory (IRT). Though calculated somewhat differently, both MCI and PAR measure the degree to which a given test taker's test response pattern is similar to his or her fellow test takers' test response patterns. Thus, calculation of MCI or PAR provides the examiner with an additional piece of information regarding the test taker's performance on a test. Beyond the information contained in the number correct or total score, MCI and PAR indicate the similarity of the test taker's pattern of right and wrong test items as compared to the test response patterns of the rest of the test takers. It was hoped that this additional piece of information might be useful in interpreting a test taker's number correct score.

Conceptually, Z3 is somewhat similar to MCI and PAR. It is based on the three parameter IRT model in which an appropriate test response pattern is defined as one which is representative of the group whose abilities are being measured by the test. The best estimates of the test parameters are the estimates which maximize something called the logistic likelihood function. As Z3 is the standardized measure of the test taker's contribution to the likelihood function, it can be used to detect test takers with test response patterns which do not contribute much to maximize the likelihood function. It can thereby

be used to detect test takers whose response patterns are inappropriate, or are not representative of the group of test takers from whose tests the parameters were estimated.

The main purpose of the present study was to investigate reasons which underlie an unusual test response pattern. Four reasons were hypothesized: differences in curricular background, differences in test-taking skills, differences in motivation, and differences in consistency of academic performance. Two secondary purposes of the study were to investigate the occurrence of sandbagging in a real placement testing situation, and to assess the feasibility of using aberrance indices as aids in interpreting placement test scores.

Of the 31 variables which together measured the four hypothesized reasons, eight were related to MCI, nine were related to PAR and six were related to Z3. When the indices were regressed on the variables to which they were related, about a third of the variance in MCI was accounted for, 37 percent of the variance in PAR was accounted for, and 12 percent of the variance in Z3 was accounted for. Including JP Test score in the regression analysis for Z3 increased the variance accounted for to 13 percent. Inclusion of JP Test score and the square of JP Test score in the regression analyses for MCI and PAR increased the variance accounted for to 40 percent in MCI and 44 percent in PAR. The squared JP Test score was included in the analyses for MCI and PAR because a quadratic relationship was found between JP Test score and these two indices. Neither of the regression models for Z3 (i.e., excluding and including JP Test) was statistically significant at the .05 level. The regression models for MCI were statistically significant at the .01 level, but the models for PAR were not, probably because there were only 39 degrees of freedom for the PAR models. When the variable SC was omitted from the regression models for PAR, the variance accounted for decreased from 44 to 41 percent, but the degrees of freedom increased to 93, and, the models were statistically significant at

the .01 level. These large percentages of variance accounted for in MCI and PAR indicates that the reasons for aberrant response patterns studied are important ones.

More research, however, is needed to better understand Z3. The range and other descriptive statistics obtained on Z3 in this study seem reasonable. However, when Z3 was regressed on the specific variables to which it was related, the resulting variance accounted for was not statistically significant. Z3 may be related to some other variable which was not considered in this study.

Four Reasons for an Unusual Response Pattern

Curricular Differences. As expected, variables measuring curricular differences were related to MCI, PAR, and Z3, and to MCI and PAR when JP Test score was partialed. With JP Test partialed, MCI and PAR were related to years of Japanese studied in high school ($r = -.16$, $d.f. = 363$, $p < .01$ for MCI and $r = .21$, $d.f. = 363$, $p < .01$ for PAR), such that the more years a student studied Japanese in high school, the less aberrant his or her response pattern tended to be. MCI and PAR were also related to years of residence in Japan ($r = .11$, $d.f. = 363$, $p < .05$ for MCI and $r = -.16$, $d.f. = 363$, $p < .01$ for PAR) and to access to native speakers of Japanese ($r = .11$, $d.f. = 363$, $p < .05$ for MCI and $r = -.16$, $d.f. = 363$, $p < .01$ for PAR). These correlations indicate that students with access to native speakers of Japanese as well as students with more years of residence in Japan tended to have aberrant response patterns compared to their fellow test takers who had neither lived in Japan nor had access to native speakers. Thus, in interpreting aberrance indices, one might suspect a student with an aberrant response pattern, as indicated by a high MCI or a low PAR, to have lived in Japan or to have had access to a native speaker.

It is not surprising that more years of study in high school is related to nonaberrant response patterns as there has always been good articulation between the Hawaii State

Department of Education (DOE) and UH Manoa's Department of East Asian Languages and Literatures (EALL). Not only have many of the DOE Japanese language teachers received their education through UH Manoa, but the Hawaii Association of Teachers of Japanese (HATJ), a professional organization, holds regular meetings and conferences which are well attended by EALL instructors and professors and public and private high school Japanese language teachers. The EALL department is also very open regarding UH Manoa Japanese language course coverage. Over the years a consensus regarding what to teach and in what order may have developed.

Conversely, a test taker who grew up in Japan or who had access to a native speaker may have learned non-standard Japanese, or colloquial Japanese. It is not uncommon for a student with little formal education to have a weak grasp of grammatical rules though he or she may have excellent pronunciation and intonation, and may have learned a few very advanced grammatical forms. It would not be surprising for such a student to have a high MCI or a low PAR, indicating an aberrant response pattern on the JP Test.

Differences in Test-Taking Skills. Of the four hypothesized reasons, differences in test-taking skills was the one which was least related to the aberrance indices. When JP Test was partialed, only one of the 15 test-taking skills variables was related to only one of the three aberrance indices. Observed minus guessed score was related to PAR at $r = -.15$ ($d.f. = 211, p < .05$), such that the greater the discrepancy between the observed score and the guessed score, the more aberrant the student's response pattern tended to be.

One explanation for the absence of relationships between the aberrance indices and the variables measuring test-taking skills may lie in the fact that the test takers in this study were college bound high school seniors. One would expect such students to all have developed relatively good test-taking skills, and therefore differences in test-taking skills may not have been an important reason for unusual test patterns for these students. Based

on a ten-point scale, the standard deviation of the high school teacher's and the college instructor's rating of these students' test-taking skills were both only about two points. If this study were repeated with subjects with a wider range of test-taking skills, then for that group of subjects, such differences might be an important reason for unusual response patterns.

Differences in Motivation. The ten motivation variables studied can be separated into two kinds of motivation: motivation to do well on the JP Test and motivation to do well in Japanese language study. The variables measuring the two kinds of motivation were not statistically related.

The two variables which measured the first type of motivation, that is, the student's motivation to do well in Japanese language study, were the high school Japanese language teacher's rating (SC), and the college Japanese language instructor's rating (B4). With JP Test score partialed, the high school Japanese language teacher's rating of the student's motivation to do well in Japanese language study was related to PAR ($r = .23$, $d.f. = 142$, $p<.01$) and to Z3 ($r = .17$, $d.f. = 142$, $p < .05$) but not to MCI ($r = -.16$, $d.f. = 142$, n.s.). Low motivation as rated by the high school teacher tended to indicate an aberrant response pattern on the JP Test. B4, the college Japanese language instructor's rating, however, was not related to the indices. The reason may be that B4 was measured six weeks into the semester, which may have been insufficient time for the college instructor to have formed an accurate judgement of the student's motivation to do well in Japanese language study. As B4 was related to JP Test score ($r=.30$, $d.f.=95$, $p<.01$), the college Japanese language instructor may have fallen back on an assessment of the student's ability rather than his or her motivation to do well. The high school teacher's rating, SC, in contrast, was not related to JP Test score. SC data was collected in May, at the end of the student's senior year in high school.

Eight of the 10 motivation variables measured the student's motivation to do well on the JP Test. Two of these variables were related to all three aberrance indices: V6, the student's response to the statement, "I marked any old answer on most of the questions on the JP Test," and V9, the student's response to the statement, "I wanted to do the best I could on the JP Test". Students were asked to respond "1" ("very true for me") through "4" ("not true at all for me") for each of the eight items. Test takers who said they marked any old answer tended to have aberrant response patterns (V6 was related to MCI at $r = -.25$, $d.f. = 95$, $p < .05$, to PAR at $r = .24$, $d.f.=95$, $p < .05$, and to Z3 at $r = .25$, $d.f. = 95$, $p < .05$). Similarly, students who said that they didn't want to do their best on the JP Test tended to have aberrant response patterns (V9 was related to MCI at $r = .27$, $d.f. = 95$, $p < .01$, to PAR at $r = -.26$, $d.f. = 95$, $p < .05$, and to Z3 at $r = -.23$, $d.f. = 95$, $p < .05$). The connection between these two variables and the indices may have been the element of carefulness. One might expect a test taker who marked any old answer, or who didn't want to do his or her best, to have been less careful in taking the JP Test than his or her counterpart who marked each item carefully and who wanted to do his or her best. The existence of aberrant response patterns for such students is understandable.

A second interpretation of this relationship is that some students literally marked "any old answer" on the test; they answered the test items randomly. Such students may have seen a disadvantage in "doing their best" on a placement test. This interpretation leads to the conclusion that some students may have intended to sabotage their test results.

In addition to V6 and V9, which were related to all three aberrance indices, V2, the student's response to the statement, "I just skipped many of the questions on the JP Test," was correlated only to Z3 ($r=-.23$, $d.f.=95$, $p<.05$ with JP Test partialed). The test takers who said they skipped many of the JP Test questions tended to have a higher Z3 index. As Z3 is a standardized index, both very low and very high values of Z3 are indications of unusual response patterns. Low negative values of Z3 are associated with response

patterns which are unlikely given the measurement model and high positive Z3 values are associated with response patterns which are more consistent than expected given the probabilistic IRT model (Reise, 1990). Thus, students who skipped many of the questions on the JP Test tended to have response patterns which were over-fitted to the 3-parameter logistic model.

The correlation between V2 and Z3, and the absence of any relationships between V2 and MCI or PAR, may have been due to coding requirements for Z3. In calculating Z3, correct items were coded as 1, incorrect items were coded as 0, skipped items were coded as 2, and never reached items were coded as 3, in accordance with the directions for running LOGIST, the software used in this study. Calculation of MCI and PAR required correct items to be coded as 1 and all other items as 0. Thus, the correlation between V2 and Z3, and the absence of relationships between V2 and MCI or PAR may be a reflection of this difference in coding.

Differences in Consistency of Academic Performance. One of the most interesting findings of this study was the relationship between MCI and PAR and the college Japanese language instructor's rating of the student's consistency of academic performance (E4). Of the 31 variables studied, the correlations between MCI and E4 (r=.38, $d.f.$=98, p<.01) and between PAR and E4 (r=-.37, $d.f.$=98, p<.01) were the highest. E4 was a general measure of the individual student's reliability in terms of academic performance. Instructors were asked to rate a student who performed consistently at his or her level as having a narrow margin of error and a student who performed exceptionally well one day and poorly on another day as having a wide margin of error. A wide margin of error tended to be associated with a test taker with an aberrant response pattern. The strength of the relationship was similar to that found by Weksel and Ware (1967), that is, the average relationship between an individual's test-retest reliability and his or her consistency as measured by a circular triad score was r=-.39.

E4, the college instructor's rating of the student's consistency, was related to JP Test score (r=.37, $d.f.$=98, p<.01) such that a higher test score was associated with a wider margin of error. This seems reasonable. A student who knows very little, as indicated by a low JP Test score, is more likely to perform consistently at a low level than a student who knows a great deal but who may have gaps in his or her knowledge of the Japanese language. The student who has gaps in his or her knowledge is not likely to perform at a high level consistently, and will thus be perceived as having a wide margin of error. In other words, a student who knows very little will seem to perform at a consistently low level because he or she will have no room for gaps in his or her knowledge.

What is puzzling is that the college instructor's rating, E4, was not related to SE, the high school Japanese language teacher's rating of the student's margin of error (r=-.06, $d.f.$=39, n.s.). One possible explanation is that the margin of error concept was not totally clear to the high school Japanese language teachers. SE was more highly correlated with the high school Japanese language teacher's rating of his/her student's motivation (r = -.74, $d.f.$=134, p<.01) and test-taking skills (r = -.71, $d.f.$=134, p<.01) than was E4 with the college instructor's rating of his/her student's motivation (r = -.50, $d.f.$=98, p<.01) and test taking skills (r = -.51, $d.f.$=98, p<.01) . For both high school teachers and college instructors, students with narrow margins of error were perceived to be more highly motivated and as having better test-taking skills than students with wide margins of error. One would expect these ratings to be correlated but if they are too highly correlated one may suspect that the concepts may not have been distinct.

Sandbagging and Applied Use of Indices

Sandbagging. Real testing situations sometimes present challenges which are not covered in textbooks. One of the secondary purposes of this study was to investigate the

occurrence of sandbagging on the JP Test. It was estimated from student feedback that between four and ten percent of students who take the JP Test deliberately perform poorly in order to gain entry into a relatively easy Japanese language course level.

Examination of the indices calculated on the JP Test and on the quizzes also indicate the occurrence of sandbagging on the JP Test. For example, the mean MCI calculated on the JP Test (.251) was higher than the mean MCI calculated on Quiz 2 (.190) and Quiz 3 (.189). amd the mean PAR calculated on the JP Test (.148) was lower than the mean PAR calculated on Quiz 2 (.243) and Quiz 3 (.221). Since low MCIs and high PARs indicate nonaberrant response patterns, it seems that there were more aberrant test response patterns on the JP Test than on the quizzes. One of the possible reasons for this may have been the occurrence of sandbagging on the JP Test. In addition, a large number of test takers who earned low JP Test scores had high MCIs and low PARs calculated on the JP Test. However, almost all quiz takers who earned low quiz scores had low MCIs calculated on quizzes and high PARs calculated on quizzes. In other words a large number of low scorers had aberrant test-response patterns on the JP Test, but on the quizzes, the low scorer was more likely to have a nonaberrant quiz-response pattern.

Unfortunately, the sandbagger is a difficult person to detect. In an effort to identify sandbaggers, eight statements on the student's motivation to do well on the JP Test were included on the student survey which was administered in UHM Japanese language courses in October 1988. Students who had taken the JP Test were asked to respond whether each of the eight statements were very true ("1") for them through not at all true ("4") for them. As discussed earlier, two of them, V6 ("I marked any old answer on most of the questions on the Placement Test") and V9 ("I tried to do the best I could on the Placement Test") were related to all three indices.

Two sandbagging strategies can be delineated. In the first strategy (strategy 1), the sandbagger adopts the nonchalant, "I don't care" attitude reflected in the motivation

variables V6 and V9. Some students may truly not care how they do on the placement test; the sandbagger does care. The person who sandbags on a placement test takes care to answer items a little more randomly and with a little less care than he or she might on, for example, a final exam. The two types of students are indistinguishable, as they both tend to have aberrant response patterns.

A second strategy (strategy 2) that a sandbagger might follow would be to answer correctly the easier items on the placement test and to either skip or to answer incorrectly the more difficult items. A sandbagger following this strategy is likely to have a nonaberrant test response pattern. Since aberrance indices such as MCI and PAR are based on difficulty levels of test items, such indices are unlikely to be useful in detecting sandbaggers who are able to correctly judge the difficulty level of test items. IRT-based indices such as Z3 might be effectively used for detecting sandbaggers to the extent that a sandbagger following such a test-taking strategy has a test response pattern which varies from the pattern predicted by the IRT model. For example, the student who "just skips many of the items on the Placement Test" (V2) has a tendency toward a higher Z3. By using Z3, it may be possible to detect the sandbagger who follows strategy 2 and skips items rather than selects incorrect answers.

The quadrant method (Harnisch, 1983) of using aberrance indices in conjunction with the total test scores was applied to the JP Test score. Quadrants were drawn in each interval between cutting scores and the data were summed across each corresponding quadrant. Quadrant A consisted of test takers with high JP Test scores and nonaberrant response patterns. Quadrant B consisted of students with high JP Test scores and aberrant response patterns. Quadrant C consisted of students with low JP Test scores and nonaberrant response patterns. Quadrant D consisted of students with low JP Test scores and aberrant response patterns. A higher than the overall proportion of "too low" ratings of placements were found in quadrants A and D. One might expect sandbaggers following

sandbagging strategy 1, that is, the careless attitude strategy, to fall in quadrant D, particularly if the sandbagging engaged in was blatant. Moderate sandbagging using strategy 1 would land the student in quadrant B. Sandbaggers following sandbagging strategy 2 of correctly answering only easier items might be expected to fall in quadrant A, particularly if the sandbagging engaged in was moderate. Blatant sandbagging using strategy 2 would land the student in quadrant C. The number of cases in this study, unfortunately, was too small, and the degree of sandbagging was not considered to enable the drawing of firm conclusions. Further study of the sandbagging question using the quadrant method is recommended.

Though also based on a small number of cases, the patterns of correlations between appropriateness of placement ratings by high school teachers with students' JP Test and quiz scores, and between appropriateness of placement ratings by college instructors with students' JP Test and quiz scores are promising. What was found was a lack of correlation between the student's JP Test score and the appropriateness of placement rating by the high school teacher and the college instructor; yet there was evidence of a relationship between quiz score and appropriateness of placement ratings. One is led to suspect that the JP Test scores may not accurately reflect the ability level of the test takers. The direction of the correlations between the quiz scores and the appropriateness ratings was such that the higher the quiz score, the more likely was a rating of inappropriate placement. One immediately suspects sandbagging to have occurred.

In one sense, this result is reasonable because the relationships between MCI and JP Test score, and between PAR and JP Test score are quadratic relationships. Aberrant response patterns are to be found at both ends of the parabola which results from graphing JP Test score against MCI or against PAR. One would expect to find ratings of inappropriate placement in the same places that one would find aberrant response patterns. Therefore, since the relationship MCI and JP Test score and between PAR and JP Test

score was not a linear one, one would not expect to find a correlation between appropriateness of placement ratings and JP Test score. Unfortunately, the number of cases was small, and the extent to which the JP Test scores are contaminated by sandbagging is not known.

Further research on the sandbagging question should start with obtaining some baseline statistics on the items in the Japanese Language Placement Test. Individual test items should be embedded in midterm or final examinations of appropriate Japanese language course levels in order to obtain information which is uncontaminated by sandbagging motivations. If placement test items were also embedded in the first quiz of the semester, then another estimate of the occurrence of sandbagging can be derived.

Applied Use of Indices. More study of the sandbagging problem is needed before a method for identifying sandbaggers is found. In the meantime, aberrance indices can be used to efficiently identify groups of students whose placement levels may require extra attention. Using the quadrant method, the placement officer can isolate the ten percent of test takers who fall in quadrant B, that is, students with high JP Test scores and aberrant response patterns. It was found that 50 percent of students who fall in Quadrant B had placement levels which were likely to be considered inappropriate by their high school teachers, compared to an overall inappropriateness rate of 26 percent.

The quadrant cutoff point used for MCI was .3, which was recommended by Harnisch (1983). The cutoff point used for PAR was .1, which was selected because it approximated the 81 percent nonaberrant to 19 percent aberrant proportion which resulted from the MCI cutoff point. Scarcity of previous research necessitated the arbitrary selection of a cutoff point for PAR. A modification of PAR which is worth examining for comparative and practical purposes was proposed by Heim (personal communication, 1989). The index which results from the modification has been dubbed Pargutt. It is the ratio of PAR over the PAR which would have resulted if it were calculated on a response

pattern where the easier items up to the total score were answered correctly and all harder items answered incorrectly. Pargutt was calculated on the data in the present study and preliminary analysis indicate that it may be similar to MCI. A more thorough comparative study of PAR, Pargutt, and MCI may be interesting.

Conclusion

MCI and PAR can be used immediately as aids in interpreting placement test scores. The quadrant method can be applied to efficiently direct the placement officer's attention to the 10 percent of test takers whose placement levels are disproportionately inappropriate compared to the overall group of test takers. For applied use, MCI and PAR are similar. There were some differences in the degree of the relationships between MCI and the variables representing different reasons for aberrant response patterns, and between PAR and the variables representing different reasons for aberrant response patterns, as well as some differences in the relationships found. For example, PAR was related to the high school teacher's rating of the student's motivation to do well in Japanese language study but MCI was not. However, MCI and PAR are similar enough that either can be recommended for practical use. Z3, however, is probably too cumbersome for applied use at this time.

With all three indices further research is recommended, particularly with real data whenever possible. Approximately 40 percent of the variance in MCI and PAR was accounted for with the reasons for aberrant response patterns selected for research in the present study. Further study as to whether these reasons for aberrant response patterns are important in other testing situations is recommended.

APPENDIX A
Background Information Sheet

(Administered regularly by the East Asian Languages and Literatures Department to all students taking a Japanese language course at UH Manoa for the first time.)

BACKGROUND INFORMATION SHEET (For Placement)

Name: _____
Social Security No.: _____ Today's Date: _____
Class Standing: _____ Major: _____

Check if you plan to use Japanese language courses to fulfill:

_____ a foreign language requirement. _____ a major requirement.

List the following information for all Japanese language courses you have taken at any *university, college,* or *communtity college* (including here at UH):

School	Course	Instructor	Semester/year	Grade

List the following information for all Japanese language courses you have taken at any *high school*:

School: _____ Number of years taken: _____ Last year taken: 19___

List the following information for all Japanese language courses you have taken elsewhere (e.g., intermediate/elementary school, Japanese language school, private language institute, private tutor, etc.):

School:_____ Number of years taken:_____ Last year taken: 19___
School:_____ Number of years taken:_____ Last year taken: 19___

Your native language: ____ English ___ other (specify): _____
Other languages you speak fluently: _____

Check if either of your parents or anyone else with whom you are currently living or have lived for a substantial length of time is a native speaker of Japanese:
___ mother ___ father ___ other (specify): _____

If you have lived in Japan or Okinawa for substantial lengths of time, fill in the following:
Total number of years of residence in Japan/Okinawa: _____
From 19__ to 19__ (age ___ to age ___)
From 19__ to 19__ (age ___ to age ___)

APPENDIX B
Questionnaire 1

(Student questionnaire administered in Spring 1988 to students taking the Japanese Language Placement Test.)

Japanese Language Placement Test
Score Prediction Form
Spring 1988

Your name: _____
High School: _____

I. Please predict your score on each of the four parts of the Japanese Placement Test that you just completed and estimate the point-range confidence in your predicted score.

(Example: On a 25-point test, if you think you scored somewhere between 20 and 24, then predict 22 plus or minus 2 points.)

Predicted Score:

Part I	(54 maximum)	_____ within +/- _____	points
Part II	(10 maximum)	_____ within +/- _____	points
Part III	(54 maximum)	_____ within +/- _____	points
Part IV	(17 maximum)	_____ within +/- _____	points

NEXT PAGE PLEASE

II. A. Please tell us about your test-taking skills. Rate how often the following statements are true for you by circling a letter on the scale to the right of each statement, where a= always; b=usually; c=sometimes; d=rarely; and e=never.

1.	When studying for an examination, I try to think up questions that might be on the examination.	a	b	c	d	e
2.	I enjoy tests which are challenging.	a	b	c	d	e
3.	In taking tests, I find I have misunderstood what was wanted and lose points because of it.	a	b	c	d	e
4.	I am so good at taking multiple choice tests that I get a higher score than I deserve.	a	b	c	d	e
5.	On multiple choice test questions I use a process of elimination to improve my chances of getting the right answer.	a	b	c	d	e
6.	My mind wanders a lot when I take a test.	a	b	c	d	e
7.	I get so nervous and confused when taking an examination that I fail to answer questions to the best of my ability.	a	b	c	d	e
8.	I skip over test questions that I can't answer right away and come back to them if I have time.	a	b	c	d	e
9.	I look for clues to the right answer in the way that a test question is stated.	a	b	c	d	e
10.	With multiple choice questions, I read all answer options before selecting an answer.	a	b	c	d	e

II. B. On a scale from 1 to 10, with 1=Awful and 10=Excellent, please rate your test-taking skills: _____

APPENDIX C
Questionnaire 2

(Rating sheet which was completed by high school Japanese language teachers. Distributed in May 1988 with names and placement levels of individual students provided. Cover memo was signed by Ray Kaneyama, Placement Officer and Undergraduate Adviser at the UH Department of East Asian Languages and Literatures.)

Scale Definitions

Scale A: Please rate the appropriateness of the placement level from -5 to +5 in terms of the student's ability with

-5 = placement level is too low, the level i; too easy for the student
0 = placement is appropriate
5 = placement is too high, the level is too difficult for the student

Scale B: Your recommendation level, if any. Please leave blank if the placement level is adequate.

Scale C: Please rate the level of the student's motivation to do well in Japanese language study from 1 to 5, with

1 = very poorly motivated
10 = very highly motivated

Scale D: Please rate the student's test-taking skills from 1 to 10 with

1 = awful
10 = excellent

Scale E: Please rate the student's margin of error from 1 to 5, with

1 = narrow margin
5 = wide margin

Does the student consistently perform at his or her level? If yes, please rate the student as having a narrow margin. Or does the student's performance vary? Does he or she misunderstand directions or make careless mistakes and consequently lose points, or on the other hand, does he or she sometimes perform exceptionally well? If performance varies widely, please rate the student as having a wide margin of error.

Student	Placement Level	A* Appropriateness of Placement (-5 = Placed level too easy 0 = Placement is Appropriate +5 = Placement level is too difficult for the student)	B* Recommended Placement Level Adjustment (if any)	C* Student's Motivation to do well in Japanese (1 = Poorly Motvated 5 = Highly Motivated)	D* Student's Test-Taking Skills (1 = Awful, 10 = Excellent)	E* Margin of Error (1 = Narrow Margin 5 = Wide Margin)

*Please see attached for scale definitions.

APPENDIX D
Questionnaire 3

(Student Questionnaire administered in October 1988. Classtime to administer the instrument was granted by the UH Department of East Asian Languages and Literatures. The instrument was administered to the entire class in which at least one experimental subject was enrolled.)

Student Survey
Measurement Accuracy of the Japanese Placement Test
October 1988

1. Did you take the Japanese Language Placement Test?
   ____ Yes (please continue to Question 2)
   ____ No (please skip to Question 5)

2. Please think back to when you took the Placement Test and answer the following by circling a number on the accompanying scale: (1= very true for me; 2= true for me; 3=not true for me; 4=not true for me at all.)

A. I just skipped many of the questions on the Placement Test.	1	2	3	4
B. I guessed on some of the questions if I sort of knew the answer.	1	2	3	4
C. I didn't want to do too well on the Placement Test.	1	2	3	4
D. I tried to answer every question on the Placement Test.	1	2	3	4
E. I marked any old answer on most of the questions on the Placement Test.	1	2	3	4
F. I answered only the questions for which I knew the correct answer.	1	2	3	4
G. I tried to get the highest score I could on the Placement Test.	1	2	3	4
H. I wanted to do the best I could on the Placement Test.	1	2	3	4

3. Do you feel that you were placed in the appropriate Japanese language level?
   ____ Yes.
   ____ No, I think I should have started at a higher level
   ____ No, I think I should have started at a lower level.

4. How much of the material covered in your first Japanese language course at Manoa was review for you? (If you are currently enrolled in your first Japanese language course, please base your estimate on the material covered so far this semester.) Please make an estimate in terms of a percentage: _____

5. How difficult is the present Japanese language course for you? (Please check one):

   ____ much too difficult for me
   ____ somewhat difficult for me
   ____ just right for me
   ____ somewhat easy for me
   ____ much too easy for me

6. Are you keeping up with the work in your Japanese class? (Please check one):
   ____ I am all already so far behind that I don't think I can catch up.
   ____ I am a little behind but I'm sure that I can catch up.
   ____ I am keeping up with the course work.
   ____ I am keeping ahead of the course schedule by looking al the material before it is covered in class.

7. How many hours per week do you spend in the language lab? ____

8. How many hours per week do you spend on studying Japanese, **excluding class time and time spent in language lab?** ____

9. How many times were you absent from your Japanese class so far this semester, counting both excused and unexcused absences?____

lO. If the semester were to end today, what grade do you think you would get for your Japanese class? (Please check one):

____ A	____ F
____ B	____ CR
____ C	____ NC
____ D	____ No grade, I am an auditor

11. On a scale from 1 to 10 with 1 = Awful and 10 = Excellent, please rate your study habits in the subject area of Japanese language: ____

12 .In your opinion, do the students in your class all belong at that level of Japanese? (Please answer this question by assigning a percentage to the following categories):
   ____ % of the students in my class are too advanced and should be enrolled in a higher level.
   ____ % of the students in my class belong at this level.
   ____ % of the students in my class belong at a lower level.

13. Is it true that students who take the Placement Test deliberately do poorly in order to get into an easy Japanese language course? (Please check one):
   ____ No, I've nevel heard of anyone doing that.
   ____ I don't know.
   ____ I don't know of anyone who actually did that but I'm sure somc students have done it.
   ____ Yes, I think that it's rather common. (If you select this answer, please estimate the percentage of students who deliberately do poorly on the Placement Test in order to get into an easy Japanese languagc course:_____)

THANK YOU FOR PARTICIPATING IN THIS STUDY.
YOUR RESPONSES ON THIS QUESTIONNAIRE ARE CONFIDENTIAL
AND WILL NOT AFFECT YOUR GRADE.

AGREEMENT TO PARTICIPATE IN

Measurement Accuracy of the Japanese Language Placement Test
(Title of Project)

Judy A. Shishido, Wist Annex 2, Rm. 221, 948-7475
(Principal investigator's name, address and phone number)

I certify that I have been told of the possible risks involved in this project, that I have been given satisfactory answers to my inquiries concerning project procedures and other matters and that I have been advised that I am free to withdraw my consent and to discontinue participation in the project or activity at any time without prejudice.

I herewith give my consent to participate in this project with the understanding that such consent does not waive any of my legal rights nor does it release the principal investigator or the institution or any employee or agent thereof from liability for negligence.

If you cannot obtain satisfactory answers to your questions or have comments or complaints about your treatment in this study, contact: Committee on Human Studies, University of Hawaii, 2540 Maile Way, Honolulu, Hawaii 96822. Phone: 948-8658.

_____
Your signature

Date _____

APPENDIX E
Questionnaire 4

(Rating form which was completed by UH Japanese language professors in October 1988. The names of students who were project participants were provided.)

Teacher's name

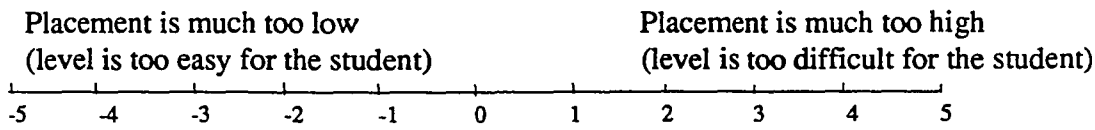
Course

Project Participants*	**Scale A** Appropriately Placed	**Scale B** Motivation	**Scale C** Test-taking Skills	**Scale D** Study Habits	**Scale E** Margin of Error

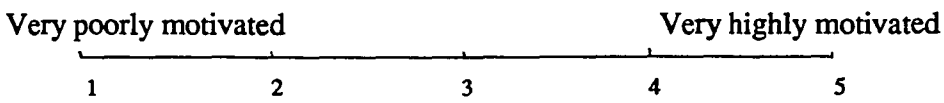
*Also, what grade is this student earning thus far in the semester?

Scale Definitions

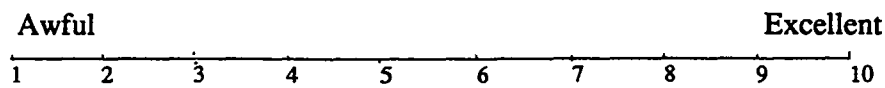
Scale A Please rate the appropriateness of the placement level in terms of the student's ability, from -5 to +5, as follows:



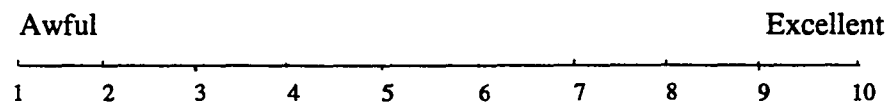
Scale B Please rate the student's motivation to do well in Japanese from 1 to 5:



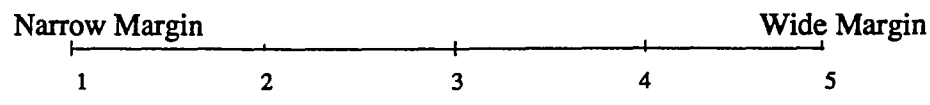
Scale C Please rate the student's test-taking skills from 1 to 10:



Scale D Please rate the student's study habits from 1 to 10:



Scale E Please rate the student's margin of error from 1 to 5:



Does the student consistently perform at his/her level? (If yes, please rate the student as having a NARROW MARGIN.) Or does the student's performance vary? (Does he/she misunderstand directions or make careless mistakes and consequently lose points, and at other times does he/she perform exceptionally well? If performance varies widely please rate the student as having a WIDE MARGIN of error.)

APPENDIX F
Index by Variable Correlations

Variable (r/p/n)	S1	MCI	PAR	Z3LN	HY	JY	RY
S1	1.000 0.000 366	−0.317 0.0001 365	0.011 0.8381 365	−0.319 0.0001 365	0.403 0.0001 366	0.452 0.0001 366	0.290 0.0001 366
MCI		1.000 0.0000 365	−0.920 0.0001 365	−0.585 0.0001 365	−0.264 0.0001 365	−0.152 0.0036 365	0.008 0.8770 365
PAR			1.000 0.0000 365	0.675 0.0001 365	0.195 0.0002 365	−0.008 0.8801 365	−0.152 0.0035 365
Z3LN				1.000 0.0000 365	−0.091 0.0811 365	−0.100 0.0561 365	−0.091 0.0830 365
HY					1.000 0.0000 366	−0.029 0.5805 366	0.031 0.5508 366
JY						1.000 0.0000 366	0.090 0.0845 366
RY							1.000 0.0000 366
NS							
TTT							
X1							
X2							
X3							
X4							
X5							

Variable (r/p/n)	NS	TTT	X1	X2	X3	X4	X5
S1	0.393 0.0001 366	0.173 0.0102 221	0.011 0.8676 221	0.264 0.0001 221	0.112 0.0976 221	−0.033 0.6279 221	0.140 0.0382 221
MCI	−0.023 0.6630 365	−0.058 0.3928 221	−0.020 0.7638 221	−0.024 0.7250 221	−0.018 0.7900 221	−0.036 0.5911 221	−0.088 0.1931 221
PAR	−0.139 0.0077 365	−0.001 0.9872 221	0.030 0.6587 221	−0.079 0.2422 221	0.005 0.9394 221	0.036 0.5963 221	0.045 0.5064 221
Z3LN	−0.151 0.0039 365	−0.061 0.3630 221	0.025 0.7066 221	−0.211 0.0017 221	−0.018 0.7955 221	0.075 0.2700 221	−0.055 0.4198 221
HY	−0.079 0.1329 366	0.071 0.2945 221	0.079 0.2401 221	0.112 0.0972 221	0.080 0.2370 221	−0.084 0.2127 221	0.016 0.8158 221
JY	0.328 0.0001 366	−0.074 0.2713 221	0.010 0.8857 221	−0.018 0.7918 221	−0.053 0.4318 221	−0.086 0.2009 221	0.107 0.1128 221
RY	0.229 0.0001 366	0.063 0.3501 221	0.038 0.5748 221	0.079 0.2440 221	0.010 0.8831 221	0.021 0.7563 221	−0.034 0.6116 221
NS	1.000 0.0000 366	0.078 0.2455 221	−0.107 0.1117 221	0.088 0.1931 221	−0.005 0.9373 221	0.015 0.8193 221	−0.010 0.8819 221
TTT		1.000 0.0000 221	0.453 0.0001 221	0.481 0.0001 221	0.368 0.0001 221	0.311 0.0001 221	0.566 0.0001 221
X1			1.000 0.0000 221	0.088 0.1948 221	−0.035 0.6081 221	−0.014 0.8340 221	0.290 0.0001 221
X2				1.000 0.0000 221	0.028 0.6740 221	0.115 0.0869 221	0.128 0.0572 221
X3					1.000 0.0000 221	−0.120 0.0750 221	0.116 0.0847 221
X4						1.000 0.0000 221	−0.003 0.9590 221
X5							1.000 0.0000 221

Variable *(r/p/n)*	X6	X7	X8	X9	X10	X11	OMG
S1	0.159	0.088	0.008	0.083	−0.122	0.202	0.319
	0.0179	0.1937	0.9098	0.2175	0.0691	0.0033	0.0001
	221	221	221	221	221	210	213
MCI	0.047	−0.116	−0.079	−0.064	0.061	−0.027	−0.012
	0.4877	0.0847	0.2416	0.3436	0.3642	0.7019	0.8603
	221	221	221	221	221	210	213
PAR	−0.097	0.078	0.054	0.036	−0.034	−0.042	−0.128
	0.1488	0.2481	0.4242	0.5941	0.6148	0.5467	0.0614
	221	221	221	221	221	210	213
Z3LN	−0.143	0.065	0.059	−0.029	0.062	−0.070	−0.222
	0.0340	0.3326	0.3860	0.6704	0.3581	0.3149	0.0011
	221	221	221	221	221	210	213
HY	0.027	0.021	0.036	0.076	−0.052	0.162	0.057
	0.6844	0.7596	0.5942	0.2617	0.4418	0.0190	0.4049
	221	221	221	221	221	210	213
JY	−0.056	−0.113	0.076	−0.003	−0.092	0.025	0.162
	0.4085	0.0934	0.2580	0.9611	0.1732	0.0721	0.0177
	221	221	221	221	221	210	213
RY	0.066	−0.031	0.094	0.047	0.051	0.059	0.190
	0.3256	0.6504	0.1620	0.4847	0.4526	0.3928	0.0055
	221	221	221	221	221	210	213
NS	0.137	0.062	−0.005	0.068	0.082	0.024	0.052
	0.0421	0.3629	0.9361	0.3144	0.2228	0.7320	0.4522
	221	221	221	221	221	210	213
TTT	0.568	0.461	0.273	0.556	0.452	0.490	−0.037
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.5927
	221	221	221	221	221	210	212
X1	0.086	0.010	0.183	0.243	0.168	0.188	−0.095
	0.2007	0.8861	0.0065	0.0003	0.0125	0.0063	0.1681
	221	221	221	221	221	210	212
X2	0.182	0.163	0.032	0.202	−0.010	0.314	0.055
	0.0066	0.0150	0.6338	0.0026	0.8770	0.0001	0.4264
	221	221	221	221	221	210	212
X3	0.286	0.309	0.005	0.008	−0.014	0.191	0.072
	0.0001	0.0001	0.9448	0.9048	0.8414	0.0055	0.2964
	221	221	221	221	221	210	212
X4	0.093	0.085	0.152	0.123	0.071	0.232	−0.082
	0.1663	0.2072	0.0236	0.0672	0.2920	0.0007	0.2367
	221	221	221	221	221	210	212
X5	0.190	0.150	0.190	0.342	0.274	0.213	−0.018
	0.0046	0.0267	0.0045	0.0001	0.0001	0.0019	0.7988
	221	221	221	221	221	210	212

Variable *(r/p/n)*	SD	C4	SC	B4	Z2	Z3	Z4
S1	0.061	0.352	0.017	0.304	0.447	−0.200	0.250
	0.4664	0.0003	0.8403	0.0021	0.0001	0.0504	0.0137
	143	100	144	100	97	97	97
MCI	−0.089	−0.247	−0.161	−0.235	−0.049	0.133	−0.090
	0.2902	0.0132	0.0532	0.0188	0.6323	0.1927	0.3807
	143	100	144	100	97	97	97
PAR	0.129	0.254	0.233	0.216	0.019	−0.072	0.100
	0.1244	0.0108	0.0050	0.0311	0.8551	0.4854	0.3817
	143	100	144	100	97	97	97
Z3LN	0.078	−0.019	0.145	0.048	−0.299	−0.101	−0.045
	0.3563	0.8540	0.0831	0.6387	0.0029	0.3235	0.6646
	143	100	144	100	97	97	97
HY	0.120	0.312	0.144	0.096	0.349	−0.090	0.233
	0.1548	0.0016	0.0852	0.3437	0.0005	0.3817	0.0218
	143	100	144	100	97	97	97
JY	0.014	0.001	0.056	0.057	0.146	−0.118	−0.011
	0.8703	0.9939	0.5079	0.5765	0.1530	0.2504	0.9147
	143	100	144	100	97	97	97
RY	−0.093	0.058	−0.181	−0.002	0.031	−0.024	0.115
	0.2712	0.5643	0.0298	0.9825	0.7625	0.8121	0.2634
	143	100	144	100	97	97	97
NS	−0.187	−0.021	−0.153	−0.030	−0.009	−0.084	−0.012
	0.0254	0.8320	0.0670	0.7671	0.9268	0.4107	0.9052
	143	100	144	100	97	97	97
TTT	0.126	0.144	0.015	0.103	0.235	−0.131	0.248
	0.2396	0.2594	0.8896	0.4205	0.0640	0.3050	0.0501
	89	63	89	63	63	63	63
X1	0.228	0.086	0.134	−0.038	0.165	−0.210	0.196
	0.0320	0.5036	0.2091	0.7696	0.1956	0.0979	0.1236
	89	63	89	63	63	63	63
X2	0.066	0.115	−0.078	0.098	0.230	−0.021	0.169
	0.5376	0.3710	0.4688	0.4466	0.0700	0.8714	0.1855
	89	63	89	63	63	63	63
X3	0.126	−0.018	0.149	0.068	0.090	−0.015	0.237
	0.2386	0.8859	0.1630	0.5971	0.4839	0.9089	0.0618
	89	63	89	63	63	63	63
X4	−0.002	0.084	−0.125	0.104	−0.090	0.075	−0.040
	0.9834	0.5126	0.2442	0.4176	0.4827	0.5596	0.7539
	89	63	89	63	63	63	63
X5	0.188	0.148	0.157	0.012	0.120	−0.093	0.049
	0.0775	0.2464	0.1406	0.9235	0.3472	0.4700	0.7009
	89	63	89	63	63	63	63

Variable *(r/p/n)*	Z5	Z6	Z7	Z8	Z9	SE	E4
S1	−0.259	0.171	0.279	−0.151	−0.114	−0.136	−0.365
	0.0105	0.0935	0.0056	0.1386	0.2677	0.1157	0.0002
	97	97	97	97	97	136	100
MCI	0.111	−0.299	−0.215	0.094	0.291	0.130	0.381
	0.2782	0.0030	0.0345	0.3603	0.0038	0.1300	0.0001
	97	97	97	97	97	136	100
PAR	−0.075	0.269	0.185	−0.087	−0.275	−0.105	−0.374
	0.4645	0.0077	0.0704	0.3989	0.0064	0.2255	0.0001
	97	97	97	97	97	136	100
Z3LN	0.085	0.201	0.054	0.008	−0.192	−0.071	−0.059
	0.4061	0.0482	0.6011	0.9343	0.0595	0.4083	0.5588
	97	97	97	97	97	136	100
HY	−0.201	−0.177	0.127	0.054	0.004	−0.035	−0.173
	0.0487	0.0825	0.2151	0.5980	0.9694	0.6879	0.0856
	97	97	97	97	97	136	100
JY	−0.064	0.192	0.171	−0.027	0.028	−0.072	−0.103
	0.5353	0.0589	0.0942	0.7946	0.7889	0.4033	0.3098
	97	97	97	97	97	136	100
RY	−0.141	−0.026	−0.072	−0.020	−0.100	−0.082	−0.010
	0.1695	0.7988	0.4861	0.8448	0.3305	0.3412	0.9215
	97	97	97	97	97	136	100
NS	−0.075	0.015	0.020	−0.036	−0.059	0.093	−0.090
	0.4681	0.8866	0.8465	0.7277	0.5691	0.2835	0.3708
	97	97	97	97	97	136	100
TTT	−0.202	0.180	−0.001	−0.327	−0.361	−0.168	0.036
	0.1132	0.1570	0.9958	0.0090	0.0036	0.1256	0.7823
	63	63	63	63	63	84	63
X1	−0.185	−0.111	0.120	−0.289	−0.349	−0.155	−0.134
	0.1458	0.3867	0.3490	0.0215	0.0051	0.1591	0.2960
	63	63	63	63	63	84	63
X2	−0.217	0.003	−0.075	−0.253	−0.359	−0.016	−0.054
	0.0881	0.9805	0.5612	0.0450	0.0038	0.8881	0.6757
	63	63	63	63	63	84	63
X3	0.033	0.265	−0.137	−0.080	−0.077	−0.188	0.080
	0.7988	0.0359	0.2843	0.5349	0.5502	0.0876	0.5333
	63	63	63	63	63	84	63
X4	0.122	−0.058	−0.003	−0.076	−0.105	0.031	−0.084
	0.3393	0.6498	0.9804	0.5536	0.4138	0.7815	0.5125
	63	63	63	63	63	84	63
X5	−0.041	0.223	0.590	−0.218	−0.235	−0.181	0.054
	0.7493	0.0780	0.6442	0.0854	0.0643	0.0999	0.6769
	63	63	63	63	63	84	63

Variable *(r/p/n)*	HST	CT	ZZ10
S1	0.049	−0.083	0.016
	0.6058	0.4138	0.8741
	117	100	97
MCI	0.235	−0.037	−0.084
	0.0106	0.7175	0.4129
	117	100	97
PAR	−0.249	0.038	0.050
	0.0068	0.7049	0.6234
	117	100	97
Z3LN	−0.050	0.048	−0.012
	0.5899	0.6364	0.9092
	117	100	97
HY	−0.071	0.003	0.031
	0.4455	0.9738	0.7611
	117	100	97
JY	−0.024	0.070	−0.060
	0.7995	0.4885	0.5574
	117	100	97
RY	0.163	−0.074	0.073
	0.0786	0.4632	0.4784
	117	100	97
NS	−0.015	−0.014	−0.084
	0.8690	0.8873	0.4133
	117	100	97
TTT	0.068	−0.011	0.020
	0.5630	0.9333	0.8771
	75	63	63
X1	0.087	−0.030	−0.082
	0.4573	0.8173	0.5252
	75	63	63
X2	0.056	0.127	0.020
	0.6355	0.3194	0.8746
	75	63	63
X3	0.051	−0.580	−0.065
	0.6620	0.6529	0.6140
	75	63	63
X4	−0.067	−0.011	−0.022
	0.5684	0.9312	0.8623
	75	63	63
X5	0.030	0.039	0.090
	0.7990	0.7610	0.4814
	75	63	63

Variable *(r/p/n)*	X6	X7	X8	X9	X10	X11	OMG
X6	1.000 0.0000 221	0.206 0.0020 221	0.133 0.0476 221	0.122 0.0700 221	0.198 0.0031 221	0.312 0.0001 210	−0.028 0.6810 212
X7		1.000 0.0000 221	−0.051 0.4470 221	0.063 0.3526 221	−0.027 0.6891 221	0.345 0.0001 210	−0.045 0.5159 212
X8			1.000 0.0000 221	0.324 0.0001 221	0.210 0.0017 221	0.039 0.5773 210	0.008 0.9073 212
X9				1.000 0.0000 221	0.295 0.0001 221	0.170 0.0136 210	0.067 0.3352 212
X10					1.000 0.0000 221	0.056 0.4157 210	−0.080 0.2440 212
X11						1.000 0.0000 210	−0.083 0.2388 202
OMG							1.000 0.0000 213
SD							
C4							
SC							
B4							
Z2							
Z3							
Z4							

Variable (r/p/n)	SD	C4	SC	B4	Z2	Z3	Z4
X6	0.023 0.8289 89	0.219 0.0852 63	0.013 0.9048 89	0.185 0.1476 63	0.348 0.0052 63	−0.110 0.3929 63	0.195 0.1261 63
X7	0.062 0.5628 89	−0.027 0.8362 63	−0.016 0.8784 89	−0.035 0.7835 63	0.169 0.1850 63	−0.178 0.1617 63	0.284 0.0239 63
X8	0.117 0.2768 89	0.059 0.6449 63	0.033 0.7565 89	0.006 0.9659 63	0.087 0.4963 63	−0.149 0.2430 63	0.041 0.7475 63
X9	0.088 0.4134 89	0.067 0.5993 63	0.077 0.4760 89	0.151 0.2365 63	0.135 0.2906 63	−0.026 0.8426 63	0.024 0.8532 63
X10	−0.249 0.0184 89	0.028 0.8251 63	−0.239 0.0242 89	−0.057 0.6567 63	−0.082 0.5252 63	−0.042 0.7413 63	0.008 0.9503 63
X11	0.266 0.0129 87	0.076 0.5723 58	0.138 0.2038 87	0.033 0.8080 58	0.124 0.3522 58	−0.007 0.9563 58	0.158 0.2348 58
OMG	−0.031 0.7783 85	−0.063 0.6316 61	−0.055 0.6161 85	0.088 0.4980 61	0.002 0.9897 61	−0.139 0.2851 61	0.072 0.5810 61
SD	1.000 0.0000 143	0.124 0.4354 42	0.707 0.0001 143	0.147 0.3431 42	−0.028 0.8623 41	−0.134 0.4029 41	0.261 0.0995 41
C4		1.000 0.0000 100	0.071 0.6567 42	0.693 0.0001 100	0.272 0.0079 94	−0.019 0.8546 94	0.022 0.8339 94
SC			1.000 0.0000 144	0.027 0.8675 42	−0.138 0.3896 41	−0.168 0.2942 41	0.116 0.4709 41
B4				1.000 0.0000 100	0.142 0.1722 94	0.064 0.5402 94	0.094 0.3697 94
Z2					1.000 0.0000 270	0.017 0.7822 269	0.248 0.0001 269
Z3						1.000 0.0000 269	−0.062 0.3112 268
Z4							1.000 0.0000 269

Variable (r/p/n)	Z5	Z6	Z7	Z8	Z9	SE	E4
X6	−0.223	0.067	−0.053	−0.080	−0.121	−0.126	−0.065
	0.0786	0.6008	0.6808	0.5350	0.3438	0.2520	0.6106
	63	63	63	63	63	84	63
X7	−0.172	0.075	−0.003	−0.250	−0.237	−0.083	0.000
	0.1770	0.5576	0.9804	0.0479	0.0617	0.4534	1.0000
	63	63	63	63	63	84	63
X8	0.108	−0.055	0.056	0.024	−0.013	−0.044	0.109
	0.4015	0.6685	0.6646	0.8494	0.9212	0.6883	0.3937
	63	63	63	63	63	84	63
X9	−0.216	0.239	0.070	−0.249	−0.206	−0.165	0.137
	0.0891	0.0589	0.5877	0.0493	0.1055	0.1333	0.2841
	63	63	63	63	63	84	63
X10	−0.032	0.225	0.047	−0.032	0.026	0.166	0.300
	0.8055	0.0765	0.7136	0.8004	0.8413	0.1321	0.0169
	63	63	63	63	63	84	63
X11	−0.157	0.182	−0.032	−0.303	−0.356	−0.160	−0.109
	0.2400	0.1723	0.8091	0.0208	0.0060	0.1520	0.4151
	58	58	58	58	58	82	58
OMG	−0.183	0.047	0.184	−0.032	−0.028	−0.149	0.085
	0.1589	0.7205	0.1560	0.8084	0.8278	0.1859	0.5131
	61	61	61	61	61	80	61
SD	−0.252	0.181	−0.096	−0.311	−0.232	−0.706	−0.128
	0.1125	0.2574	0.5494	0.0479	0.1453	0.0001	0.4200
	41	41	41	41	41	136	42
C4	−0.041	0.025	0.091	0.105	0.108	−0.124	−0.506
	0.6954	0.8124	0.3851	0.3134	0.3009	0.4382	0.0001
	94	94	94	94	94	41	100
SC	−0.065	0.198	−0.033	−0.043	−0.095	−0.741	0.006
	0.6843	0.2157	0.8384	0.7879	0.5554	0.0001	0.9681
	41	41	41	41	41	136	42
B4	−0.022	0.065	0.045	0.084	−0.032	0.034	−0.500
	0.8336	0.5313	0.6682	0.4223	0.7562	0.8350	0.0001
	94	94	94	94	94	41	100
Z2	−0.435	0.052	0.222	−0.215	−0.190	0.044	−0.244
	0.0001	0.3982	0.0002	0.0004	0.0017	0.7898	0.0176
	270	269	268	270	270	40	94
Z3	0.255	0.045	−0.274	0.110	0.146	0.084	0.042
	0.0001	0.4665	0.0001	0.0705	0.0170	0.6071	0.6860
	269	268	267	269	269	40	94
Z4	−0.294	0.153	0.136	−0.589	−0.571	−0.010	−0.222
	0.0001	0.0120	0.0261	0.0001	0.0001	0.9518	0.0314
	269	268	267	269	269	40	94

Variable *(r/p/n)*	HST	CT	ZZ10
X6	0.218	0.045	0.014
	0.0597	0.7239	0.9110
	75	63	63
X7	−0.065	0.034	−0.076
	0.5806	0.7899	0.5522
	75	63	63
X8	−0.081	0.112	0.010
	0.4910	0.3816	0.9355
	75	63	63
X9	0.049	−0.086	0.109
	0.6773	0.5032	0.3959
	75	63	63
X10	−0.109	−0.137	0.152
	0.3519	0.2842	0.2342
	75	63	63
X11	0.115	0.058	0.067
	0.3341	0.6648	0.6183
	73	58	58
OMG	0.152	−0.069	0.152
	0.2066	0.5989	0.2409
	71	61	61
SD	0.038	0.127	−0.125
	0.6821	0.4243	0.4375
	117	42	41
C4	0.085	−0.094	−0.025
	0.6076	0.3502	0.8091
	39	100	94
SC	0.047	0.090	0.231
	0.6112	0.5693	0.1460
	117	42	41
B4	0.173	−0.272	−0.051
	0.2933	0.0061	0.6273
	39	100	94
Z2	0.077	0.063	−0.075
	0.6420	0.5437	0.2187
	39	94	270
Z3	0.173	−0.072	−0.065
	0.2934	0.4904	0.2899
	39	94	269
Z4	−0.039	−0.023	−0.024
	0.8141	0.8230	0.6918
	39	94	269

Variable (r/p/n)	Z5	Z6	Z7	Z8	Z9	SE	E4
Z5	1.000 0.0000 270	−0.018 0.7750 269	−0.275 0.0001 268	0.397 0.0001 270	0.425 0.0001 270	0.146 0.3702 40	0.033 0.7515 94
Z6		1.000 0.0000 269	−0.096 0.1186 267	−0.174 0.0042 269	−0.207 0.0006 269	0.014 0.9334 40	−0.125 0.2312 94
Z7			1.000 0.0000 268	−0.085 0.1673 268	−0.046 0.4496 268	0.088 0.5902 40	−0.073 0.4862 94
Z8				1.000 0.0000 270	0.783 0.0001 270	0.148 0.3605 40	0.048 0.6428 94
Z9					1.000 0.0000 270	−0.012 0.9395 40	0.154 0.1390 94
SE						1.000 0.0000 136	−0.061 0.7029 41
E4							1.000 0.0000 100

Variable *(r/p/n)*	HST	CT	ZZ10
Z5	−0.190	−0.025	−0.091
	0.2466	0.8077	0.1358
	39	94	270
Z6	0.104	−0.014	−0.007
	0.5272	0.8931	0.9033
	39	94	269
Z7	0.280	−0.072	0.043
	0.0842	0.4913	0.4843
	39	94	268
Z8	0.018	−0.043	−0.048
	0.9154	0.6812	0.4360
	39	94	270
Z9	−0.078	−0.031	−0.046
	0.6384	0.7682	0.4479
	39	94	270
SE	−0.090	−0.220	0.146
	0.3429	0.1673	0.3702
	113	41	40
E4	0.014	0.002	0.016
	0.9329	0.9808	0.8822
	39	100	94

BIBLIOGRAPHY

Anastasi, A. (1982). *Psychological testing*. New York: MacMillan.

Arley, N., & Buch, K. R. (1950). *Introduction to the theory of probability and statistics*. New York: Chapman & Hill.

Atkinson, J. W. (1980). Motivational effects in so-called tests of ability and education achievement. In L. J. Fyans (Ed.), *Achievement motivation* (pp. 9-21). New York: Plenum Press.

Ayabe, H. I., & Heim, M. (1987, January). *Is it time for a new internal consistency reliability?* Paper presented at the meeting of the Hawaii Educational Research Association, Honolulu, HI.

Bajtelsmit, J. W. (1977). Test-wiseness and systematic desensitization programs for increasing adult test-taking skills. *Journal of Educational Measurement, 14*, 335-341.

Baker, F. B. (1965). Origins of the item parameters X50 and Beta as a modern item analysis technique. *Journal of Educational Measurement, 2*, 167-180.

Bem, D. J. (1983). Toward a response style theory of persons in situations. In R. A. Dienstbier, & M.M. Page (Eds.), *Nebraska Symposium on Motivation, 1982: Personality - Current Theory and Research* (pp. 201-231). Lincoln: University of Nebraska Press.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81*, 506-520.

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

Blixt, S. L., & Shama, D. D. (1986). An empirical investigation of the standard error of measurement at different ability levels. *Educational and Psychological Measurement, 46,* 545-550.

Block, J. (1968). Some reasons for the apparent inconsistency of personality. *Psychological Bulletin, 70,* 210-212.

Bond, J. A. (1986). Inconsistent responding to repeated MMPI items: Is its major cause really carelessness? *Journal of Personality Assessment, 50,* 50-64.

Burkhart, B. R., Christian, W. L., & Gynther, M. D. (1978). Item subtlety and faking on the MMPI: A paradoxical relationship, *Journal of Personality Assessment, 42,* 76-80.

Burrill, L. E. (1982). Comparative studies of item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press.

Callenbach, C. (1973). The effects of instruction and practice in content-independent test-taking techniques upon the standardized reading test scores of selected second-grade students. *Journal of Educational Measurement, 10,* 25-30.

Carter, K. (1986). Test-wiseness for teachers and students. *Educational Measurement: Issues and Practice, 5,* 20-23.

Committee to Develop Standards for Educational and Psychological Testing of The American Educational Research Association, The American Psychological Association & The National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Crehan, K. D., Koehler, R. A., & Slakter, M. J. (1974). Longitudinal studies of test-wiseness. *Journal of Educational Measurement, 11,* 209-212.

Cronbach, L. J. (1947). "Test reliability": Its meaning and determination. *Psychometrika, 12*(1), 1-16.

Diamond, J. J., Ayrer, J., Fishman, R., & Green, P. (1976). Are inner city children test-wise? *Journal of Educational Measurement, 14,* 39-45.

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

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., & 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,* 59-79.

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.

Fagley, N. S. (1987). Positional response bias in multiple-choice tests of learning: Its relation to testwiseness and guessing strategy. *Journal of Educational Psychology, 79,* 95-97.

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351-361.

Fiske, D. W. (1957). The constraints on intra-individual variability in test reponses. *Educational and Psychological Measurement, 17,* 317-337.

Fiske, D. W., & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin, 52,* 217-250.

Glaser, R. (1949). A methodological analysis of the inconsistency of response to test items. *Educational and Psychological Measurement, 9,* 727-739.

Glaser, R. (1952). The reliability of inconsistency. *Educational and Psychological Measurement, 12,* 93-599.

Goldberg, L. R. (1978). The reliability of reliability: The generality and correlates of intra-individual consistency in responses to structured personality inventories. *Applied Psychological Measurement, 2,* 269-291.

Good, T. L., & Brophy, J. E. (1986). *Educational psychology: A realistic approach* (pp. 779-780), (3rd ed.). New York: Longman.

Graham, K. G., & Robinson, H. A. (1984). *Study skills handbook: A guide for all teachers* (pp. 100-106). Newark, Del.: International Reading Association.

Greene, R. L. (1979). Response consistency on the MMPI: The TR index. *Journal of Personality Assessment, 43,* 69-71.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The predition of personal adjustment* (pp. 319-348). New York: Social Science Research Council.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14,* 75-96.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer Pub.

Hargadon, F. (1981). Tests and college admissions. *American Psychologist, 36,* 1112-1119.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. *Journal of Educational Measurement, 20,* 191-206.

Harnisch, D. L. (1989). Brown Bag Lecture at University of Hawaii at Manoa.

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

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11,* 91-115.

Heim, M. (1988). Working Paper 335. Hawaii Department of Education, Evaluation Section.

Heim, M. (1989, February). *Estimators of person-level measurement error and item response aberrance within the framework of classical test theory.* Paper presented at the meeting of the Hawaii Educational Research Association, Honolulu, HI.

Hendel, D. D., & Weiss, D. J. (1970). Individual inconsistencies and reliability of measurement. *Educational and Psychological Measurement, 30,* 579-593.

Ironson, G. H. (1982). Chi-square and latent trait approaches. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement, 10,* 175-186.

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

Kuder, F. G., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160.

Lazarsfeld, P. F., (1950). The logical and mathematical foundation of latent structure. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (Eds.), *Measurement and prediction* (pp. 362-472). Princeton, NJ: Princeton University Press.

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

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53,* 161-176.

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

Livingston, S. A. (1982). Estimation of the conditional standard error of measurement for stratefied tests. *Journal of Educational Measurement, 19,* 135-138.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph,* No. 7.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15,* 325-336.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1985). Estimating the imputed social cost of errors of measurement. *Psychometrika, 50,* 57-68.

Lord F., & Novick, M. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1*, 477-482.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31*, 19-26.

McBride, J. R. (1977). Some properties of a Bayesian adaptive ability testing strategy. *Applied Psychological Measurement, 1*, 121-140.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9*, 389-400.

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

McQuitty, L. L. (1956). Agreement analysis: Classifying persons by predominant patterns of responses. *British Journal of Statistical Psychology, 9*, 5-16.

Mischel, W., & Peake, P. K. (1983). Analyzing the construction of consistency in personality. In R.A. Dienstbier and M.M. Page (Eds.), *Nebraska Symposium on Motivation, 1982: Personality - Current Theory and Research* (pp. 233-262). Lincoln: University of Nebraska Press.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.

Mitra, S. K., & Fiske, D. W. (1956). Intra-individual variability as related to test score and item. *Educational and Psychological Measurement, 16*, 3-12.

Mollenkopf, W. G. (1949). Variation of the standard error of measurement. *Psychometrika, 14*, 189-229.

Nelson, R. B., & Chatman, S. P. (1986). *The influence of guessing on measures of response aberrance when using the Rasch model.* Paper presented at the meeting of the American Educational Researchers Association, San Francisco, CA.

Nunnally, J. (1978). *Psychometric Theory.* New York: McGraw-Hill.

Parker, G. V. C. (1971). Prediction of individual stability. *Educational and Psychological Measurement, 31*, 875-886.

Payne, D. A. (1974). *The assessment of learning: Cognitive and affective.* Lexington, MA: D.C. Heath.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction.* New York: Holt, Rinehart and Winston.

Raine, W. J., & Hills, J. R. (1959). Critique and notes: Measuring intra-individual variability within one testing. *Journal of Abnormal Psychology, 58*, 264-266.

Raynor, J. O. (1974). Relationships between achievement-related motives, future orientation, and academic performance. In J. W. Atkinson, & J. O. Raynor (Eds.), *Motivation and achievement.* Washington, D.C.: V.H. Winston & Sons.

Reschly, D. J. (1981). Psychological testing in educational classification and placement. *American Psychologist, 36,* 1094-1102.

Rigby, K. (1987). "Faking good" with self-reported pro-authority attitudes and behaviors among schoolchildren. *Personality and Individual Differences, 8,* 445-447.

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin, 63,* 129-156.

Ross, J., & Lumsden, J. (1968). Attribute and reliability. *British Journal of Mathematical and Statistical Psychology, 21,* 251-263.

Rowley, G. L. (1974). Which examinees are most favored by the use of multiple choice tests? *Journal of Educational Measurement, 11*, 15-23.

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

Rundquist, E. A. (1950). Response sets: A note on consistency in taking extreme positions. *Educational and Psychological Measurement, 10*, 97-99.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, No. 17.

Samejima, F. (1977). A use of information function in tailored testing. *Applied Psychological Measurement, 1,* 233-247.

SAS Institute, Inc. (1988). *SAS language guide for personal computers, Release 6.03 edition*. Cary, NC: Author.

Shimberg, B. (1981). Testing for liscensure and certification. *American Psychologist, 36,* 1138-1146.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review Educational Research, 56,* 495-529.

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

Snyter C. M., & Allik, J. P. (1981). The fakability of the personal orientation dimensions: Evidence for a lie profile. *Journal of Personality Assessment, 45,* 533-538.

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 Psychological Measurement, 7,* 81-95.

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

Tasuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221-230.

Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement, 25,* 301-319.

Tenopyr, M. L. (1981). The realities of employment testing. *American Psychologist, 36,* 1120-1127.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 567-577.

Thissen D., Steinberg, L., & Gerrad, M. (1986). The group-mean difference: The concept of item bias. *Psychological Bulletin, 99,* 118-128.

Thorndike, R. L. (1949). *Personnel Selection: Tests and Measurement Techniques.* New York: J. Wiley.

Thorndike, R.L. (1951). Reliability. In E.F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Tomisic, M., & Mittman, A. (1986). *Stability of caution indices.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Vale, C. D., & Gialluca, K. A. (1988). Evaluation of the efficiency of item calibration. *Applied Psychological Measurement, 12,* 53-67.

VanDerLinden, W. J., & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement, 1,* 593-599.

VanDerLinden, W. J., & Mellenbergh, G. J. (1978). Coefficients for tests from a decision theoretical point of view. *Applied Psychological Measurement, 2,* 119-134.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12,* 339-368.

Weinstein, C. E. (1987). *LASSI user's manual for those administering the learning and study strategies inventory.* Clearwater, FL: H & H Pub. Co.

Weksel, W., & Ware, E. E. (1967). The reliability and consistency of complex personality judgements. *Multivariate Behavioral Research, 2,* 537-541.

Whitely, S. E. (1978). Individual inconsistency: Implications for test reliability and behavioral predictability. *Applied Psychological Measurement, 2,* 571-579.

Wilcox, R. R. (1978). A note on decision theoretic coefficients for tests. *Applied Psychological Measurement, 2,* 609-613.

Worthington, D. L., & Schlottmann, R. S. (1986). The predictive validity of subtle and obvious empirically derived psychological test items under faking conditions. *Journal of Personality Assessment, 50,* 171-181.

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

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