ANALYSIS OF ITEM RESPONSE PATTERNS:
QUESTIONABLE TEST DATA AND DISSIMILAR
CURRICULUM PRACTICES

DELWYN L. HARNISCH AND ROBERT L. LINN
University of Illinois, Urbana-Champaign

The same number-right score on a test can disguise the fact that students' incorrect answers may yield important information about student preparation. Indeed, response patterns associated with each number-right score can vary quite widely even on relatively short tests. There are, for example, 184,756 possible patterns that yield a score of 10 on a 20-item test. While many possible patterns are never observed in practice, those which are found can help researchers discover relationships between specific tests and curricula.

It has been common practice to simply use a single total score although some attempts have been made to use wrong response patterns diagnostically. This is partially attributable to the fact that there are limited amounts of data available for particular response patterns, and because there is both low reliability of individual items and a tendency to attribute a large fraction of the common variance of the items to a single dimension. Recently, however, several approaches have been suggested which are intended to identify unusual response patterns.

There are a wide variety of factors that could lead to an unusual response pattern. While a question may be very easy for most students, unique background experiences may make that same item very difficult for others. For example, the child who has never gone camping may find a reading passage about a camping experience more difficult than children who have had the experience. Other individual differences in motivational disposition, for instance, test anxiety, may make a normally simple item very difficult for some people. Students' exposure to different subject matter, and the way in which that subject matter has been stressed, will influence how they perform on tests. This may result in some measurable variation in scores from class to class. Items which are generally difficult for most students may be relatively easy for students who have been in classes where that particular content was emphasized. Such variation from the norm may lead to the systematic over- or under-estimation of an individual's or group's level of achievement, distorting the measurement results.

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

There are two major types of indices of the degree to which an individual's pattern of responses is unusual. First, there are the “appropriateness” indices which are based upon

We thank the Illinois State Board of Education for making available the data used in this study, and Mervin M. Brennan, Leslie J. Fyans, Jr. and Thomas Kerins of the staff of the State Board for their cooperation in making this study possible.
item response theory (IRT) as described by Levine and Rubin (1979) and modifications of these indices as suggested by Drasgow (1978). The $\chi^2$ test of person fit that is sometimes used with applications of the Rasch model (e.g., Wright, 1979) is another example of an IRT based index. Second, there are the indices which are based directly upon the pattern of right and wrong answers, such as the "caution" index proposed by Sato (1975), the $U'$ index by van der Flier (1977), the personal biserial by Donlon and Fischer (1968), the norm-conformity index by Tatsuoka and Tatsuoka (1980), and the agreement and disagreement indices discussed by Kane and Brennan (1980).

The focus of this manuscript is on the latter type of indices, which do not require the use of item response theory. Our primary purposes are (1) to provide a comparative analysis of these indices, and (2) to illustrate one potentially important use of such indices. We will review briefly the indices and compare some of their algebraic properties. We will then discuss the empirical interrelationships among the indices and their relationships with total number-right scores for two tests used in a statewide assessment program. Finally, we will examine school and regional differences on one of the indices, and content differences in items which contribute to school-to-school differences in the index.

**DEFINITION AND COMPARISON OF INDICES**

For purpose of defining and comparing indices of the degree to which an individual's response pattern is unusual, it is convenient to start with a consideration of a matrix of zeros and ones. A row of the matrix is associated with each examinee and a column with each item. Ones are recorded for correct responses and zeros for incorrect responses. Rows and columns of the data matrix are permuted so that the items (columns) are arranged from left to right in ascending order of difficulty, and examinees (rows) are arranged from top to bottom in descending order of total number of correct answers. The resulting matrix has been called Student-Problem (S-P) Table by Sato (1975). (See Tatsuoka 1978 for a description in English.)

If the items formed a perfect Guttman Scale (Guttman, 1941) the S-P Table would consist of a section with all ones in the upper left-hand corner and all zeros in the lower right-hand corner. A single step-like boundary line would separate the ones and the zeros. In other words, anyone who responded correctly to a difficult item would also answer all easier items correctly. There would be no unusual response patterns in the sense that is used for the indices described in this article because everyone who had a given total score would have the same pattern of responses. Perfect Guttman scores cannot, however, be expected on achievement test items. Consequently, the S-P Table in this case will be characterized by a predominance of ones in the upper left-hand corner and zeros in the lower right-hand corner, although there will be many exceptions to the pattern, e.g., ones where mostly zeros are expected and vice versa.

A small hypothetical example of an S-P Table with 18 examinees and 5 items is shown in Table 1. The solid and dashed lines in Table 1 are known as the S-curve and P-curve, respectively. The S-curve (solid line) is obtained by drawing a vertical line for each row that has $n_i$ items (columns) to the left of it, where $n_i$ is the total number of correct responses for the $i^{th}$ examinee. The P-curve (dashed line) is obtained by drawing a horizontal line in each column such that there are $n_{ij}$ examinees who answer item $j$ correctly.

Ideally, the S- and P-curves would coincide. The degree of divergence provides an
indication of the degree of homogeneity of the response patterns. Sato (1975) has developed an index based on the area between the S- and P-curves which is potentially useful in evaluating the homogeneity of the test. (See Tatsuoka, 1978.) Of greater interest for our present purposes, however, is Sato's caution index.

Sato's Caution Index \((C_i)\)

Sato's Caution index, \(C_i\) for the \(i^{th}\) examinee, may be defined as follows:

\[
C_i = \frac{\sum_{j=1}^{n_i} (1 - u_{ij})n_{.j} - \sum_{j=n_i+1}^{J} u_{ij}n_{.j}}{\sum_{j=1}^{n_i} n_{.j} - n_i \left( \frac{\sum_{j=1}^{J} n_{.j}}{J} \right)},
\]

(1)

where

- \(i = 1, 2, \ldots, I\), indexes the examinee,
- \(j = 1, 2, \ldots, J\), indexes the item,
- \(u_{ij} = \begin{cases} 1 & \text{if examinee } i \text{ answers item } j \text{ correctly,} \\ 0 & \text{if examinee } i \text{ answers item } j \text{ incorrectly,} \end{cases}\)
- \(n_{i.} = \text{total correct for the } i^{th} \text{ examinee, and}\)
- \(n_{.j} = \text{total number of correct responses to the } j^{th} \text{ item.}\)

A parallel index for the \(j^{th}\) item may be defined by simply reversing the roles of \(i\) and \(j\) in the above equation, but only the person index will be considered in the present paper. Values of \(C_i\), and for all of the indices described below, are listed in Table 1.

The name of the index comes from the notion that a large value is associated with examinees giving unusual response patterns. It notes that some caution may be needed in interpreting a total correct score for an examinee. An unusual response pattern may result from guessing, carelessness, high anxiety, an unusual instructional history or other experiential background, a localized misunderstanding that influences responses to a subset of items, or copying a neighbor's answers to certain questions. The key point is that the caution index provides information about an examinee that is not contained in the total score. A large value of the caution index raises doubts about the validity of the usual interpretation of the total score for an individual.

The Modified Caution Index \((C^*_i)\)

A modified form of Sato's caution index, \(C^*_i\), was introduced to yield a lower bound of 0 and an upper bound of 1. This modified caution index, \(C^*_i\), for the \(i^{th}\) examinee may be defined as follows:

\[
C^*_i = \frac{\sum_{j=1}^{n_i} (1 - u_{ij})n_{.j} - \sum_{j=n_i+1}^{J} u_{ij}n_{.j}}{\sum_{j=1}^{n_i} n_{.j} - \sum_{j=n_i+1}^{J} n_{.j}}.
\]

(2)

By being bounded between 0 and 1, the modified caution index eliminates the extreme scores that are sometimes obtained on the caution index, especially in cases where a very high scoring examinee misses a single very easy item. This is seen as a potential advantage of \(C^*_i\).
Table 1
S-P Table for 18 Examinees and 5 Items
(Hypothetical Example)

<table>
<thead>
<tr>
<th>Examinee</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
<th>Total n_i</th>
<th>Sato's Index C_i</th>
<th>Modified Index C_i</th>
<th>Personal Point Biserial r_i</th>
<th>Personal Agreement Index A_i</th>
<th>Disagreement Index D_i</th>
<th>Index of Dependability not corrected for chance θ_i</th>
<th>Index of Dependability corrected for chance θ ci</th>
<th>van der Flier's Index U_i</th>
<th>Tatsuoka and Tatsuoka's NCI_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>.00</td>
<td>.00</td>
<td>.73</td>
<td>1.05</td>
<td>1.94</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>.65</td>
<td>.33</td>
<td>.26</td>
<td>.36</td>
<td>1.78</td>
<td>.17</td>
<td>.91</td>
<td>.35</td>
<td>.25</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>.00</td>
<td>.00</td>
<td>.81</td>
<td>1.02</td>
<td>1.61</td>
<td>.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>.16</td>
<td>.08</td>
<td>.68</td>
<td>.86</td>
<td>1.56</td>
<td>.06</td>
<td>.97</td>
<td>.84</td>
<td>.17</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>.65</td>
<td>.31</td>
<td>.29</td>
<td>.36</td>
<td>1.39</td>
<td>.22</td>
<td>.86</td>
<td>.35</td>
<td>.33</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1.13</td>
<td>.54</td>
<td>-.10</td>
<td>-.13</td>
<td>1.22</td>
<td>.39</td>
<td>.76</td>
<td>-.13</td>
<td>.50</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>.00</td>
<td>.00</td>
<td>.88</td>
<td>1.12</td>
<td>1.22</td>
<td>.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>.00</td>
<td>.88</td>
<td>1.12</td>
<td>1.22</td>
<td>.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.00</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>.44</td>
<td>.23</td>
<td>.49</td>
<td>.63</td>
<td>1.06</td>
<td>.17</td>
<td>.86</td>
<td>.56</td>
<td>.17</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>.59</td>
<td>.31</td>
<td>.36</td>
<td>.46</td>
<td>1.00</td>
<td>.22</td>
<td>.82</td>
<td>.41</td>
<td>.33</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>.74</td>
<td>.38</td>
<td>.23</td>
<td>.30</td>
<td>.94</td>
<td>.28</td>
<td>.77</td>
<td>.26</td>
<td>.33</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>.88</td>
<td>.46</td>
<td>.10</td>
<td>.10</td>
<td>.89</td>
<td>.33</td>
<td>.73</td>
<td>.12</td>
<td>.50</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.00</td>
<td>.00</td>
<td>.70</td>
<td>1.00</td>
<td>.67</td>
<td>.00</td>
<td>1.00</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.00</td>
<td>.00</td>
<td>.70</td>
<td>1.00</td>
<td>.67</td>
<td>.00</td>
<td>1.00</td>
<td>.00</td>
<td>1.00</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.45</td>
<td>.22</td>
<td>.38</td>
<td>.55</td>
<td>.56</td>
<td>.11</td>
<td>.83</td>
<td>.55</td>
<td>.25</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1.14</td>
<td>.56</td>
<td>-.10</td>
<td>-.14</td>
<td>.39</td>
<td>.28</td>
<td>.58</td>
<td>-.14</td>
<td>.50</td>
</tr>
<tr>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1.36</td>
<td>.67</td>
<td>-.26</td>
<td>-.36</td>
<td>.33</td>
<td>.33</td>
<td>.50</td>
<td>-.36</td>
<td>.75</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1.36</td>
<td>.67</td>
<td>-.26</td>
<td>-.36</td>
<td>.33</td>
<td>.33</td>
<td>.50</td>
<td>-.36</td>
<td>.75</td>
</tr>
</tbody>
</table>

n i 12 10 7 6 3
The Personal Point-Biserial Index ($r_i$)

The modified caution index, $C_i^*$, can be shown to be closely related to a third index, the personal point-biserial correlation (Brennan, 1980). The personal point-biserial, $r_i$, is simply the product moment correlation between the item scores for individual $i$ (i.e., $u_{ij} = 0$ or 1 for each $j$) and the number of people who responded correctly to each item (i.e., $n_{ij}$). The $n_{ij}$ could, of course, be replaced by the item difficulties, $p_{ij} = n_{ij}/I$ to compute $r_i$. A high value of $r_i$ shows that person $i$ tends to answer correctly items that are easy for the group as a whole and miss items that are difficult for the group. Thus, it is an index of agreement between the individual responses and the group determined difficulties.

A disadvantage of $r_i$ is that its boundaries depend on an individual's number right score. With the $n_{ij}$ values for the five items shown in Table 1, for example, the maximum possible $r_i$ is $.70$, $.88$, $.81$, and $.73$ for individuals with $n_i = 1$, $2$, $3$, and $4$ respectively. The minimum possible values are $-.73$, $-.81$, $-.88$, and $-.70$ for $n_i = 1$, $2$, $3$, and $4$ respectively. Thus, the fact that person 7 with $n_i = 2$ has $r_i = .88$, while person 13 with $n_i = 1$ has $r_i = .70$, may be considered misleading (see Table 1). In both cases, the individual responses agree with the item difficulty order to the maximum extent possible. But $r_i$ gives the appearance that the agreement with the norm is better for person 7 than that for person 13 and neither appears to be as good as it might be since $r_i = 1.0$. Both $C_i$ and $C_i^*$, on the other hand are $.00$, for both person 7 and person 13, indicating no disagreement.

The exact relationship between $C_i^*$ and $r_i$ is given by

$$C_i^* = \frac{\max(r_i) - r_i}{\max(r_i) - \min(r_i)}, \quad (3)$$

where $\max(r_i)$ and $\min(r_i)$ are the maximum and minimum possible values of $r_i$ respectively, for a given $n_i$ (Brennan, 1980).

Personal Biserial Correlation ($r^*$)

The fourth index that will be investigated is the personal biserial correlation (Donlon & Fischer, 1968). This is the biserial correlation for each individual corresponding to $r_i$, based on the assumption of a normally distributed continuous variable underlying $u_{ij}$.

The Agreement ($A_i$) and Disagreement ($D_i$) Indices

Brennan (1980) has pointed out that the modified index is closely related to the agreement and disagreement indices discussed by Kane and Brennan (1980). In particular,

$$C_i^* = \frac{\max(A_i) - A_i}{\max(A_i) - \min(A_i)}, \quad (4)$$

where $A_i = \sum_{j=1}^I u_{ij} p_{ij}$, is an index of agreement between an individual's responses and the group-determined difficulties, and $\max(A_i)$ and $\min(A_i)$ are the maximum and minimum values of $A_i$ for a given $n_i$. As further noted by Brennan (1980), other measures of similarity might also be considered. Hence, definitions of $A_i$ and of other indices that are derived from $A_i$ are possible. Only the above definition of $A_i$ will be considered here, however. An equivalent form of Equation 4 is:
where \( D \) refers to disagreement. More specifically, \( D_i = \max(A_i) - A_i \) and \( \max(D_i) = \max(A_i) - \min(A_i) \). Although \( A_i \) and \( D_i \) were not proposed by Kane and Brennan for use in identifying individuals with unusual response patterns, they are obviously related at least to the modified caution index and could be candidates for this purpose. Thus, they were included as the fifth and sixth indices in the analyses. It is recognized, however, that \( A_i \) will typically be highly related to the total number-right score, and for that reason is not a very viable candidate as an index for detecting unusual response patterns.

**The Dependability Indices \( \theta_i, \theta_{ci} \)**

Kane and Brennan also discussed two dependability indices labeled \( \theta \) and \( \theta_{ci} \), which could be defined for individual examinees and used to detect people with less “dependable,” i.e., more unusual, response patterns. The definition of \( \theta \) for individual \( i \) is

\[
\theta_i = \frac{A_i}{\max(A_i)}
\]

The division of \( A_i \) by the maximum agreement possible, given the person’s total score, would be expected to yield an index that is less highly related to total score. The \( \theta \) was used as the seventh index in the analysis.

Kane and Brennan’s \( \theta_{ci} \) is a “corrected for chance” index. For an individual examinee, it will be denoted \( \theta_{ci} \) and is defined by

\[
\theta_{ci} = \frac{A_i - A_{ci}}{\max(A_i) - A_{ci}}
\]

where \( A_{ci} \) is the value of \( A_i \) that would be expected if the \( i \)th individual had a probability of \( n_i / J \) of answering each item correctly. That is, \( A_{ci} \) is the “chance” level of agreement given that the total number of correct answers is \( n_i \). Although originally considered as another possible index, \( \theta_{ci} \) provides exactly the same information as Sato’s caution index provided that similarity is measured as in the expression given above for \( A_i \). With this definition of similarity, \( \theta_{ci} \) and \( C_i \) are functionally related by

\[
\theta_{ci} = 1 - C_i
\]

Since \( \theta_{ci} \) and \( C_i \) obviously would have identical correlations, except by a reverse of sign with any other variable, \( \theta_{ci} \) was not included as a separate index in the correlational analyses. Of course, with other definitions of similarity, however, \( \theta_{ci} \) would not have a perfect inverse relationship with \( C_i \).

For purposes of calculation, the form of \( \theta_{ci} \) in Equation 7 is more convenient than that of \( C_i \) in Equation 1. This observation along with Equation 8 suggests the following calculational form for \( C_i \),

\[
C_i = \frac{J \sum_{j=1}^{n_i} n_{i,j} - \sum_{j=1}^{J} u_i n_{i,j}}{J \sum_{j=1}^{n_i} n_{i,j} - n_i \sum_{j=1}^{J} n_{i,j}}
\]

This equation assumes that the items are ordered in descending order of \( n_{i,j} \).
The van der Flier Index \( (U_i) \)

The ninth index of the degree to which an examinee's response pattern is unusual was proposed by van der Flier (1977).\(^1\) Using the order of the items and the subjects as in the S-P Table, the deviation from the S-curve can be quantified by adding the number of 1's to the right of every 0 which is called \( U \), resembling the Mann Whitney \( U \). The minimum value of \( U \) is 0 and the maximum value equals the number of correct answers multiplied by the number of incorrect answers. The range of \( U \) values are made equal for different total scores by dividing them by this maximum value. The resulting measure \( (U'_i) \) has a lower bound of 0 and an upper bound of 1.

The Norm-Conformity Index \( (NCI_i) \)

The final index to be considered here is one that was developed by Tatsuoka and Tatsuoka (1980). Their index is called the norm-conformity index, \( NCI \). The \( NCI \) provides a measure of the degree of consistency between the response pattern of an individual and the difficulty ordering of the items for a norm group. Tatsuoka and Tatsuoka have shown that the \( NCI \) is closely related to Cliff's (1977) group consistency indices. In their work, however, Tatsuoka and Tatsuoka have focused on consistency for individuals rather than overall indices of group performance. \( NCI \)'s may range from \(-1.0\) to \(1.0\). A value of 1.0 indicates perfect consistency with the norm-group ordering of item difficulties, that is, correct responses to the \( n_i \) easiest items and incorrect responses to the remaining \( J-n_i \) items. Thus, an \( NCI = 1.0 \) corresponds to values of 0.0 on \( C_i, C_i^* \) and \( U'_i \). There is a perfect negative correlation between \( NCI_i \) and \( U'_i \). Specifically,

\[
NCI_i = 1 - 2U'_i. \tag{10}
\]

A total of ten indices have been briefly described. Each index is intended to provide an indication of the degree to which an individual's response pattern either diverges from or conforms to a norm. There are some close algebraic links among some of the indices (e.g., \( C_i^* \) with \( r_i \), \( A_i \), or \( D_i \)), and two pairs of the indices are functionally related by a linear transformation \( (U'_i \) with \( NCI_i \) and \( C_i \) with \( \theta_i \)). It is not therefore surprising that several of the ten indices tend to order the hypothetical examinees illustrated in Table 1 in a similar way. Because of the perfect inverse relationships between \( U'_i \) and \( NCI_i \) and between \( C_i \) and \( \theta_i \), only eight indices were included in the correlational analyses. These are \( C_i, C_i^*, r_i, r_i^*, A_i, D_i, \theta_i \) and \( U'_i \).

DATA SOURCE AND PROCEDURE

Test results from the 1978 Illinois statewide assessment program were used to compare the indices and to investigate the potential utility of the modified caution index for distinguishing among schools. This annual statewide survey, known as the Illinois Inventory of Educational Progress (IIIEP), provides data for 6,300 students from a random sample of approximately 110 schools at the fourth, eighth, and eleventh grade levels. At each of the attendance centers, 18–22 students are randomly selected for the grade level being tested at the school. This gives an approximate statewide sample of 2,100 students per grade level. The total IIIEP state survey sampling, then, involves some

\(^1\)Van der Flier's index has been shown to be identical to an index denoted \( \gamma \) by Sato (1971) prior to his introduction of the caution index, \( C_i \), defined above in Equation 1 (Sato, 1981).
6,300 students spread evenly across the elementary (4th), junior high (8th), and high school (11th) grades being tested.

The data set includes item response data for tests of reading and mathematics. Also included are questionnaire responses which provide data on student background, self-report of test anxiety, their attributions for success and failure, and other specific and general motivational variables. The results presented below are based on the tests of reading and mathematics that were administered to students in the fourth grade.

The fourth grade mathematics test contains 40 items, while the fourth grade reading test contains 28 items. All eight of the indices described above were computed on each

Table 2

Intercorrelations Among Eight Response Pattern Indices\textsuperscript{a} and Total Test Score

(Math Test Results Above the Diagonal and Reading Test Below the Diagonal)

<table>
<thead>
<tr>
<th>Index</th>
<th>C or $(-\theta_C)$</th>
<th>C*</th>
<th>RPB</th>
<th>RB</th>
<th>A</th>
<th>D</th>
<th>$\theta$</th>
<th>U' or $(-\text{NCI})$</th>
<th>Total Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C or $(-\theta_C)$</td>
<td>.99</td>
<td>.97</td>
<td>.99</td>
<td>.28</td>
<td>.84</td>
<td>.76</td>
<td>.96</td>
<td>.17</td>
<td></td>
</tr>
<tr>
<td>C*</td>
<td>.96</td>
<td>.95</td>
<td>.96</td>
<td>.13</td>
<td>.82</td>
<td>.65</td>
<td>.93</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>RPB</td>
<td>-.96</td>
<td>-.90</td>
<td>.97</td>
<td>.30</td>
<td>.77</td>
<td>.74</td>
<td>-.93</td>
<td>.18</td>
<td></td>
</tr>
<tr>
<td>RB</td>
<td>-.95</td>
<td>-.89</td>
<td>.96</td>
<td>.39</td>
<td>.84</td>
<td>.80</td>
<td>-.97</td>
<td>.28</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-.54</td>
<td>-.34</td>
<td>.63</td>
<td>.73</td>
<td>-.33</td>
<td>.77</td>
<td>-.40</td>
<td>.99</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>.92</td>
<td>.94</td>
<td>.88</td>
<td>.47</td>
<td>-.72</td>
<td>.85</td>
<td>-.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta$</td>
<td>-.93</td>
<td>-.79</td>
<td>.90</td>
<td>.93</td>
<td>.74</td>
<td>.80</td>
<td>-.80</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>U' or $(-\text{NCI})$</td>
<td>.94</td>
<td>.89</td>
<td>.93</td>
<td>.64</td>
<td>.89</td>
<td>.90</td>
<td>--</td>
<td>-.30</td>
<td></td>
</tr>
<tr>
<td>Total Score</td>
<td>-.42</td>
<td>-.21</td>
<td>.50</td>
<td>.63</td>
<td>.99</td>
<td>-.36</td>
<td>.65</td>
<td>-.54</td>
<td>--</td>
</tr>
</tbody>
</table>

\textsuperscript{a}The indices are

$C = \text{Sato's (1975) caution index,}$

$\theta_C = \text{corrected for chance dependability index based on Kane and Brennan (1980),}$

$C* = \text{modified caution index,}$

$\text{RPB = personal point-biserial,}$

$\text{RB = Donlon & Fischer's (1968) personal biserial,}$

$A = \text{agreement index based on Kane & Brennan (1980),}$

$D = \text{disagreement index based on Kane & Brennan (1980),}$

$\theta = \text{dependability index based on Kane & Brennan (1980),}$

$U' = \text{van der Flier's (1977) index, and}$

$\text{NCI = Tatsuoka & Tatsuoka's (Note 2) norm conformity index.}$
test for the 2,094 students at the fourth grade. Correlations among the eight indices and between the indices and total score were computed. Scatterplots for pairs of indices and for each index with total score were obtained and inspected.

We compared the results based on the correlations both among the indices and their relations to the total score. We then investigated the potential use of an unusual response pattern index to identify school and regional differences. The modified caution index \( (C^*_i) \) was chosen for the latter analysis, primarily because it had the least relationship to total score. As a first step, hierarchical analysis of variance (ANOVA) was used to disentangle the variance attributable to regions, schools within regions, and students nested within schools within regions.

RESULTS

Interrelationships

The intercorrelations among the eight response pattern indices and total score are reported in Table 2. The correlations above the diagonal are based on the math test while those below the diagonal are based on the reading test. Except for the agreement index, \( A_i \), the indices are highly intercorrelated. For the math test, the \( \theta_i \) index has somewhat lower correlations with the other indices than was found for the reading test.

The extremely high correlation between \( A_i \) and total test score (.99 for both tests) is not surprising. The agreement index could also be viewed as a weighted test score where items are weighted by difficulty. It is well known that positively weighted composites of substantial numbers of positively intercorrelated items are generally very highly related. (See Wang & Stanley, 1970.)

A previously mentioned disadvantage of the personal point-biserial is the tendency for it to be related to total score. At least minimum values change as a function of total number right. Although the correlation between the personal point-biserial and total score is substantial for the reading test (.50), both this value and the much lower one (.18) for the math test underestimate the degree of relationship. The nonlinearity of the relationship is quite apparent in the case of the math test. This can be seen in the scatterplot in Figure 1. Using total score and total score squared as predictors, the multiple correlation with the personal point-biserial is .41 for the math test and .55 for the reading test.

The personal biserial has a more nearly linear relationship with total score than was true of the personal point-biserial, which accounts for the higher correlations in Table 2. Both the personal point-biserial and the personal biserial are more confounded with total score than would be desirable for an index designed to identify individuals or groups with unusual response patterns. The correlations of \( U_i \) with total score are also higher than would be desirable. This is also true of \( \theta_i \) which correlated .70 and .65 with total score on the math and reading tests, respectively. The disagreement index is somewhat better in this regard, as are \( C_i \) and \( C^*_i \). For these two tests, the modified caution index is the least related to total test score. Because \( C^*_i \) is least confounded with total score, it was selected for use in the analysis reported below.

School and Regional Differences on Modified Caution Index \( (C^*_i) \)

The reading and math modified caution indices were used as dependent variables in hierarchical ANOVA's, with school as the unit of analysis. The first factor is the five
different regions of the state. School is the second factor. This factor is nested in regions, while students, the third factor, are nested within schools within regions. There are 50, 14, 18, 15, and 13 schools in regions one through five, respectively. The results of these analyses are summarized in Table 3. The schools within Regions 1, 2, and 5 have significantly different modified caution indices on the math test and there are significant school differences within Regions 1 and 5 on the reading test.

The mean caution indices on both tests for the 13 schools within Region 5 are reported in Table 4. The range of the modified caution index is .18 to .32 on the math test and .13 to .26 on the reading test. The relatively wide range of modified caution indices for schools within Region 5 suggest a high degree of variability of item response patterns. In addition, the large differences found for schools within Region 5 may well have been a function of the curriculum. The school effects are important for all indices, which reveals that certain schools may not have covered segments of the content sampled on the test, or that they have given less than typical emphasis to some of the content.

The significant schools-within-region effects denote the high degree of variability of
Table 3

Summary of F Ratios for Hierarchical Analyses of Variance for the Modified Caution Index on Two Tests

<table>
<thead>
<tr>
<th>Effect</th>
<th>Math Test</th>
<th>Reading Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>.57</td>
<td>2.12</td>
</tr>
<tr>
<td>Schools within Region</td>
<td>1.64**</td>
<td>1.74**</td>
</tr>
<tr>
<td>Schools within Region 1</td>
<td>1.69*</td>
<td>1.97**</td>
</tr>
<tr>
<td>Schools within Region 2</td>
<td>2.78**</td>
<td>1.73</td>
</tr>
<tr>
<td>Schools within Region 3</td>
<td>.46</td>
<td>.66</td>
</tr>
<tr>
<td>Schools within region 4</td>
<td>1.37</td>
<td>1.35</td>
</tr>
<tr>
<td>Schools within Region 5</td>
<td>2.17*</td>
<td>2.83**</td>
</tr>
</tbody>
</table>

*p < .01

**p < .001

student performance in schools within certain regions of the state. Curriculum offerings may very well contribute to these large differences. To explore this possibility, we conducted a more detailed analysis of the response patterns of students at schools with high means on the caution indices. This study was designed to identify the subset of items which contribute most to the caution indices for the identified schools. The analyses of these subsets of items were used to describe unique patterns of performance by item content. Such patterns suggest differences in content coverage that make the test less appropriate for some schools than others.

Discussion of Schools with Large Modified Caution Indices

Schools with high mean modified caution indices on the math test, .30 or greater, were evaluated for unusual response patterns. The p-values, i.e., the proportion of students who answered an item correctly, were computed for each school on each of the 40 items. Since school means performance on the test is directly related to the p-values on the items, a linear regression was performed on the p-values for each school with the p-values from the state sample. The regression equation was used to compute the expected proportion correct on each item for each school. Residual scores were computed simply by subtraction of expected from observed proportion correct on each item for each school.

Items were categorized in terms of their content and format in order to find clues about the possible reasons for the large differences in the residuals. The mean of the residuals
Table 4
Means and Standard Deviations on Modified Caution
Indices on Two Tests for Schools Within Region 5

<table>
<thead>
<tr>
<th>School Number</th>
<th>Sample Size</th>
<th>Math Test Mean</th>
<th>S.D.</th>
<th>Reading Test Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>.23</td>
<td>.10</td>
<td>.15</td>
<td>.08</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>.27</td>
<td>.07</td>
<td>.23</td>
<td>.10</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>.26</td>
<td>.06</td>
<td>.14</td>
<td>.08</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>.24</td>
<td>.09</td>
<td>.20</td>
<td>.12</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>.32</td>
<td>.07</td>
<td>.24</td>
<td>.13</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>.24</td>
<td>.07</td>
<td>.13</td>
<td>.10</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>.29</td>
<td>.08</td>
<td>.24</td>
<td>.12</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>.31</td>
<td>.10</td>
<td>.22</td>
<td>.08</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>.23</td>
<td>.07</td>
<td>.22</td>
<td>.09</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>.23</td>
<td>.10</td>
<td>.26</td>
<td>.13</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>.23</td>
<td>.09</td>
<td>.18</td>
<td>.08</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>.18</td>
<td>.08</td>
<td>.17</td>
<td>.12</td>
</tr>
</tbody>
</table>

for each category was then standardized by dividing by the standard error of estimate. Finally, the standardized mean residuals were multiplied by the square root of the number of items in the content category as a means of weighting the standardized mean residuals according to the number of items in the category. The resulting weighted standardized mean residuals, which are analogous to critical ratios, were used to compare the items in different categories. These results are reported in Table 5.

An entry in Table 5 greater than 2.0 in absolute value indicates that items in that particular category are much easier or much harder for students in that school than would be expected from their overall performance and the relative difficulty of these items for the statewide sample as a whole. Four of the content areas have entries in Table 5 greater than the 2.0 absolute value for one or more schools. These categories are items using figures to represent fractions, story problems dealing with money,\(^2\) numeration questions, and items involving the metric system. The large negative entry (−2.3) for

\(^2\)An example of a story problem dealing with money is: "Mary earned $1.00 raking leaves. Candy bars cost 15¢. How many candy bars can she buy with her money?"
### Table 5

**Weighted Standardized Mean Residuals of**

Within School Item Difficulties by Content Category

<table>
<thead>
<tr>
<th>Content Category</th>
<th>School 1</th>
<th>School 2</th>
<th>School 3</th>
<th>School 4</th>
<th>School 5</th>
<th>School 6</th>
<th>School 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculation</td>
<td>-.07</td>
<td>-1.34</td>
<td>-1.13</td>
<td>.11</td>
<td>.56</td>
<td>-.58</td>
<td>1.32</td>
</tr>
<tr>
<td>Definitions</td>
<td>1.27</td>
<td>.39</td>
<td>-.24</td>
<td>.22</td>
<td>.78</td>
<td>.00</td>
<td>-.32</td>
</tr>
<tr>
<td>Numeration</td>
<td>.00</td>
<td>-.66</td>
<td>-2.13</td>
<td>.78</td>
<td>.29</td>
<td>-.81</td>
<td>-1.64</td>
</tr>
<tr>
<td>Story Problems (general)</td>
<td>-.03</td>
<td>-.79</td>
<td>-.57</td>
<td>.63</td>
<td>-.76</td>
<td>.19</td>
<td>-.60</td>
</tr>
<tr>
<td>Story Problems (money)</td>
<td>-.10</td>
<td>2.60</td>
<td>1.51</td>
<td>-1.91</td>
<td>-1.29</td>
<td>.01</td>
<td>-.30</td>
</tr>
<tr>
<td>Metric System</td>
<td>-.10</td>
<td>-.56</td>
<td>.88</td>
<td>.56</td>
<td>1.08</td>
<td>2.28</td>
<td>2.20</td>
</tr>
<tr>
<td>Figures (fractions)</td>
<td>-2.29</td>
<td>4.16</td>
<td>2.94</td>
<td>-1.08</td>
<td>-1.64</td>
<td>-1.89</td>
<td>4.33</td>
</tr>
<tr>
<td>Unclassified</td>
<td>.93</td>
<td>-.85</td>
<td>.07</td>
<td>-1.44</td>
<td>1.61</td>
<td>.28</td>
<td>-.41</td>
</tr>
</tbody>
</table>

School 1 for the figural representation of fractions stands in marked contrast to the large positive values for Schools 2, 3, and 7 in this category (4.2, 2.9, and 4.3 respectively). This suggests the hypothesis that the use of figures to represent fractions may be quite common in Schools 2, 3, and 7 but rare in School 1. Similar hypotheses are suggested by the other large values in Table 5. Thus, one would expect that special emphasis is placed on the metric system in Schools 6 and 7.

Differences similar to those reported above indicate at the first level the type of items that function differently for various schools. These differences may have greater diagnostic value than total scores. There may be a variety of reasons, however, for schools to have substantially different patterns. Differences in response patterns, for example, could result from attendance patterns and school-to-school variability in content coverage and emphasis. This latter possibility is currently being pursued with data from the National Assessment of Educational Progress where we are attempting to identify school characteristics that are associated with high incidence of unusual response patterns.

**CONCLUSION**

Response pattern indices of a test for an individual could be used in a variety of ways. The indices may identify students for whom the test is inappropriate or schools with curricula that do not match the test content. Sato (1975) has also suggested that the caution index may be used along with the total test score to identify students who need more study, who make careless mistakes, who possess sporadic study habits or
insufficient readiness or who are doing everything fine. Schools can similarly be identified as having in general more of one of the above categories of students.

REFERENCES

BRENNAN, R. L. Personal communication, April 1980.
GUTTMAN, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst, P. Wallin, & L. Guttman (Eds.), The prediction of personal adjustment. New York: Social Science Research Council, Committee on Social Adjustment, 1941.

AUTHORS


LINN, ROBERT L. Address: Department of Educational Psychology, 210 Education, 1310 S. Sixth St., University of Illinois, Champaign, IL 61820. Title: Professor. Degrees: B.A. University of California at Los Angeles, M.A. University of Illinois at Urbana-Champaign, Ph.D. University of Illinois at Urbana-Champaign. Specialization: Psychometric Theory, Educational Influences.
Linked Citations

You have printed the following article:

Analysis of Item Response Patterns: Questionable Test Data and Dissimilar Curriculum Practices
Delwyn L. Harnisch; Robert L. Linn
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28198123%2918%3A3%3C133%3AAOIRPQ%3E2.0.CO%3B2-4

This article references the following linked citations. If you are trying to access articles from an off-campus location, you may be required to first logon via your library web site to access JSTOR. Please visit your library's website or contact a librarian to learn about options for remote access to JSTOR.

References

Measuring the Appropriateness of Multiple-Choice Test Scores
Michael V. Levine; Donald B. Rubin
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28197924%294%3A4%3C269%3AMTAOMT%3E2.0.CO%3B2-L

Differential Weighting: A Review of Methods and Empirical Studies
Marilyn W. Wang; Julian C. Stanley
Stable URL:
http://links.jstor.org/sici?sici=0034-6543%28197012%2940%3A5%3C663%3ADWAROM%3E2.0.CO%3B2-7

Solving Measurement Problems with the Rasch Model
Benjamin D. Wright
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28197722%2914%3A2%3C97%3ASMPWTR%3E2.0.CO%3B2-%23