Item Response Patterns: Applications for Educational Practice

Delwyn L. Harnisch


Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28198322%2920%3A2%3C191%3AIRPAFE%3E2.0.CO%3B2-R

Journal of Educational Measurement is currently published by National Council on Measurement in Education.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ncme.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
ITEM RESPONSE PATTERNS: APPLICATIONS FOR EDUCATIONAL PRACTICE

DELOWYN L. HARNISCH*
University of Illinois, Urbana-Champaign

In most uses of achievement tests, the primary focus is on a few summary scores that are based on the number of items answered correctly. The focus on the number of correct answers or some transformation of it (e.g., a grade-equivalent score) is natural. Certainly, the overall level of performance in a content area such as arithmetic is a major consideration. Furthermore, the interrelatedness of subsets of achievement test items (e.g., addition and subtraction items) makes it difficult to obtain reliable information about special strengths and weaknesses of individuals or groups of students in addition to a global summary score.

In situations when the ranking of students on a global dimension is emphasized, the role of specific test content receives relatively little attention. While it may be useful to know that the average score of students in an instructional program is well below the national mean on a fourth-grade mathematics test, this information does not indicate the types of arithmetic skills the students have. As shown by Porter, Schmidt, Floden, and Freeman (1978; See also Schmidt's article in this issue), there is wide variation in content coverage of the fourth-grade mathematics tests from the four most widely used achievement test batteries. Furthermore, the level of performance can be expected to be influenced by the degree of overlap between the instructional content and the content of the test (e.g., Armbruster, Steven, & Rosenshine, 1977; Leinhardt & Seewald, 1981; Madaus, Kellaghan, Rakow, & King, 1979). Indeed, the specific match between the format of instructional exercises and the format of test questions can have a substantial impact on test scores (Alderman, Swinton, & Braswell, 1979; House, Glass, McLean, & Walker, 1978).

The limitations of global summary scores can also be seen by considering their mathematical foundation. The same number-right score can be obtained in many different ways. Even on a short test of just five items, a score of three can be obtained by ten different combinations of correct and incorrect responses. The number of possible combinations of right and wrong answers expands rapidly with increases in test length. On a 10-item test, for example, there are 252 different response patterns that yield the same number-right score of five.

Although the 252 distinct patterns of item responses can be identified, it obviously is not feasible to provide different interpretations for each unique pattern. Responses to individual items are too unreliable. Many of the possible patterns will not be encountered in practice and the frequency of occurrence of any particular pattern will usually be too small to support any clear generalizations beyond those that can be made from the simple number-right score.

The difficulties in using response patterns to obtain diagnostic information not

*Appreciation is due to Sam Kuo and Rosalie Torres for their assistance in developing the computer program package entitled the Student Problem Package which generated Figures 1–4 and to Bob Linn for his comments on earlier drafts.
contained in the number-right score explain their relatively limited use to date. Recently, however, new and potentially more powerful techniques have been developed for identifying atypical response patterns. The introduction of these techniques has led to a renewed interest in using information contained in response patterns to identify individuals with unusual patterns (See, for example, Harnisch & Linn, 1981a, 1981b, 1982; Levine & Rubin, 1979; Miller, 1981; Sato, 1975, 1981; Tatsuoka & Linn, in press; Tatsuoka & Tatsuoka, 1982, in press; van der Flier, 1977, 1982).

The focus of this manuscript is on the recent psychometric developments in the analysis of item response patterns. Suggestions for analysis, interpretation and reporting of achievement data for students based on their item response patterns are provided.

Several researchers have explored the use of an index which indicates the extent to which an individual's response pattern is unusual. The identification of persons for whom special caution is required in interpreting total scores is important. Potentially even more important, however, is the possible diagnostic value of the response pattern itself. In the next section of the paper, various possible uses of the information based on the pattern of item responses is provided.

USES OF ITEM RESPONSE PATTERNS

The most obvious use of the information based on the pattern of item responses is simply to identify students or groups of students for whom special caution is needed in interpreting their total correct scores. Other uses include the identification of large school-to-school variations on different subsets of items; the identification of differences between classrooms, especially in instructional coverage, as reflected in the patterns of item responses; and the identification of items that are stable across samples. These latter uses would provide designers of school effect studies and large scale evaluations the opportunity to construct scales which are more sensitive to group differences on instructional and program related variables.

Group Differences

Donlon and Rindler (1979) and Harnisch and Linn (1981b) found interesting sex and ethnic/race defined group differences in average values of indices which describe the randomness of an item response pattern. Similarly, Harnisch and Linn (1981b) found large ethnic group differences on an index of inconsistent response patterns for 13-year-old students participating in the National Assessment of Educational Progress (NAEP) for mathematics. Specifically, blacks across each of the quintiles of the test score distribution exhibited atypical response patterns more frequently than non-blacks.

Schooling Differences

In general, patterns in item responses can be used in various schooling contexts. For example, a state department may use the techniques to identify schools in which the curriculum or the instruction does not match the test content. The techniques described by Harnisch and Linn (1981a) offer state departments a method to examine how test items function across districts, schools within districts, and classrooms within schools, or the uniformity of instruction across districts, schools, or classrooms relative to the items.

The above example illustrates how item response patterns can be used to identify
group-to-group differences in instructional coverage and emphasis. When schools (classrooms, districts) vary in their content coverage, they would be expected to vary not only in their overall performance as measured by the total score, but also in the relative difficulty of particular subsets of items. The demonstration of systematic between-school differences in response patterns and the association of these differences with instructional practices has potentially important implications for test use in evaluation studies (Harnisch & Linn, 1981a; Miller, 1981). The indices described in this paper can potentially indicate when comparisons between groups (classes, schools, districts) solely on the basis of total scores may be misleading. The clusters of items that contribute to high values on these indices provide alternative comparisons and possibly more meaningful interpretations of observed performance outcomes.

Patterns in the errors students make can also reflect schooling differences. When the pattern of incorrect answers to a set of homogeneous items is the same for groups of students in a classroom, school or district, then several questions might be asked. Were the students taught the topic? Were the students given adequate instruction on the topic? Did students have opportunity to practice with similar problems on this topic? Were students taught enough information to answer items on this topic? Will students from a given class/school or district answer this item correctly?

Finally, the information from the pattern of item responses can be used to supplement the instruction. For example, by measuring areas of content coverage and the errors on specific topics, suggestions for changes in the instructional sequence are possible. An example of the kind of question that could be addressed is does teaching with concrete examples lead to the desired type of problem solving skills? Or does teaching which emphasizes abstract reasoning skills lead to the desired type of problem solving skills?

**INDICES OF RESPONSE PATTERNS**

Various approaches to analysis of item response patterns have emerged recently. These approaches can be broadly classified according to whether a constructed or a selected response is analyzed. In the constructed response approach, unusual response patterns are found by intuitive error analysis or by extensive clinical interviews. This approach relies heavily on the level of experience and on undifferentiated human judgment in classifying the item responses. It can also be time consuming and costly.

Unusual response patterns also can be identified by analytical methods when students are asked to select from a group of response alternatives. When student responses are scored as correct or incorrect, indices based on the pattern of right and wrong answers can be used to describe the degree to which an individual’s pattern of responses is unusual. Alternatively, patterns in the choice of distractors can provide diagnostic information about the type of misunderstanding the student has on various item/topics. This latter use is an integral feature of recent investigations of test responses from a cognitive psychology perspective (Brown & Burton, 1978; Glaser, 1981; Nitko, 1980; Tatsuoka & Tatsuoka, 1982).

In the remaining part of this section, the main points about the indices based on selected response options will be discussed. Two sets of indices based on the pattern of correct/incorrect responses are reviewed first followed by a discussion of approaches used for analysis of distractors.
Appropriateness Indices

There are two major sets of indices of the degree to which an individual's pattern of item responses is unusual. The first set are the "appropriateness" indices as described by Levine and Rubin (1979) and modifications of these indices as suggested by Drasgow (1978, 1982). These indices are based upon item response theory (IRT). The chi square test of person fit that is sometimes used in applications of the Rasch model (e.g., Wright, 1977) is another example of an IRT-based index.

Appropriateness indices 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.

Group-Dependent Indices

The second set of indices, group-dependent indices, are based directly upon the pattern of right and wrong answers. Specific indices include the "caution" index proposed by Sato (1975), the modified caution index by Harnisch and Linn (1981a), the U' index by van der Flier (1977), the personal biserial by Donlon and Fischer (1968), the norm-conformity index by Tatsuoka and Tatsuoka (1982), and the agreement and disagreement indices discussed by Kane and Brennan (1980). These indices indicate whether an individual examinee's pattern of responses is atypical relative to the response patterns in the group of examinees.

The two sets of indices recently have been linked by Tatsuoka and Linn (in press). Specifically, they extended the concepts of the Student-Problem (S-P) curve theory (Sato, 1975) and the caution index to the continuous domain of the IRT models through the use of discrete summary statistics.

Rudner (1982) compared the detection rates of the two sets of indices based on the pattern of item responses by conducting a Monte Carlo study. The indices based on IRT consistently out-performed the indices based on sample statistics. The IRT indices, although more powerful with large sample sizes, are not applicable in practice. Rudner found that most of these techniques identified higher proportions of students with altered response patterns than students with unaltered response patterns. Rudner further notes that information provided by any of these statistics can be extremely helpful if a practitioner is interested in screening students. He recommends that testing programs consider incorporating an index which describes the accuracy of the individual assessment.

In the remainder of the paper the emphasis will be on group dependent indices. Research efforts describing the use of these indices to link testing and instructions are provided.

Use of Group-Dependent Indices

Harnisch and Linn (1981a) investigated ten group-dependent indices and found them to be highly intercorrelated. Not only were large individual differences identified by using one of the indices but classroom and school differences were also found. Harnisch and Linn (1981a) found large school-to-school variations on certain categories of items after adjusting for overall school performance on the mathematics test of the Illinois Statewide Assessment program. For example, students at one school performed consistently better than expected on items using geometric figures to represent fractions while performing
worse than expected on figural representation of fraction items. A third school did well on story problems dealing with money (e.g., “Mary earned $1.00 raking leaves. Candy bars cost $.15. How many candy bars can she buy with her money?”) but relatively poorly on calculation items. Even within the category of story problems, there were sizeable school-to-school contrasts in the relative performance on those involving money and those involving other applications (e.g., time or amounts of physical quantities).

Miller (1981) argues that the caution indices could also be used as indicators of a group phenomenon at the classroom level. He suggests that instructional group (e.g., a classroom) with a high mean caution index level most likely indicates that the group does not approach the material covered on the test in the same manner as the rest of the sample. For example, in his study of class level performance on a fifth grade fractions test, Miller (1981) notes that “the pattern of correct item responses on the posttest clearly showed a relationship to instructional coverage and emphasis that were not visible prior to instruction. In addition, the differences in coverage and emphasis are best described at the item level” (p. 157).

Harnisch and Linn’s (1981a) and Miller’s (1981) linkage of observed subgroup differences in patterns of responses to variation in instruction is far removed from most present day practice. The present day practice of using the total or subtest scores on standardized achievement tests or on locally made competency tests do not satisfy a common request of students which is, according to Davis (1979) “Tell me what I’m doing wrong” (p. 6). The error, a “bug” in the student’s procedures (Brown & Burton, 1978) systematically leads to the wrong answer on similar problems (Davis, Jockusch, & McKnight, 1978). In their study of signed number arithmetic problems, Tatsuoka and Tatsuoka (in press), demonstrated procedures which successfully diagnosed hundreds of bugs. Such procedures could be very useful in the near future in the improvement of teaching and the design of new instructional materials. Additional research on the potential applications of these techniques for evaluating student performance on computer adaptive testing activities would be beneficial.

Distractor Analysis

The multiple-choice test items found on most standardized achievement tests require the addition of distractors very similar to the correct answer. Such distractors are popular and give support to the intuition of the item writer. Items of this nature contribute to discrimination, but do not tell the student what he or she is doing wrong. Nearly 15 years ago, Guttman (1969) advocated having the distractors constructed systematically for sets of items and, thus, have the items provide information about what a student was doing wrong. To assess student performance and to diagnose student errors, teachers or administrators need more than a single total test score which is merely the number of items correct on a test. Total test scores by themselves can be very misleading. Birenbaum and Tatsuoka (1980), for example, identified several students who had errors of varying degrees of seriousness even though they had the same total test score on a test of addition and subtraction of signed numbers. Their investigation provides ample evidence of why total test scores are not useful in diagnosing either the nature of the errors or the degree of seriousness of the errors (Tatsuoka, 1981). Additional research is needed here to develop a general strategy for classifying the degree of correctness for item responses.

Tatsuoka and Tatsuoka (in press) demonstrate the benefits of systematically constructing error-diagnostic tests. In addition to the total test score, scores are generated based on
their pattern of responses to unique item sets to assess the seriousness of specific error types. However item construction for an error-diagnosing test is quite different from that of other traditional achievement tests. Careful selection of items and distractors is required on error-diagnostic tests so that each item contributes uniquely to determining the seriousness of the erroneous rules committed by the student.

Crucial to a diagnostic system is the ability to determine the extent to which the pattern of item responses is "typical" or "consistent." Consistency of a response pattern may refer to either the average response pattern of a group or an individual's response pattern over repeated testing periods. Tatsuoka and Tatsuoka (1982) have developed distinct indices for the two different situations. Their norm conformity index was one of the ten group-dependent indices compared by Harnisch and Linn (1981a). Tatsuoka and Tatsuoka's individual consistency index (ICI) measures the degree to which an individual's response pattern remains invariant over time. The ICI indicates the extent to which a response pattern remains relatively stable from one subtest to another. A limitation of the ICI is that it requires multiple items on each concept to identify the rules of misunderstanding. Since tests of this nature seldom exist in practice, a generalized caution index developed by Tatsuoka and Linn (in press) which removes the requirement of multiple items on each concept may be more useful in practice.

PRACTICAL APPLICATIONS

Much of the recent psychometric work on item response patterns requires further development and clarification before practical applications beyond a few specific instances is warranted. Among existing indices and their applications, only the analysis of item response patterns based on the Student-Problem (S-P) curve theory (Sato, 1975) appears to be sufficiently well-developed for broad application in current day testing practice. Sato's methods are already widely used throughout Japan.

In the remainder of this section, we will illustrate how S-P curve theory might be used in practice. An example of an S-P chart1 with description to aid in understanding the performance of 24 fourth-grade students on a test of 44 items is used in our illustration.

S-P Chart

The S-P Chart is a matrix in which students (rows) are arranged from top to bottom in descending order of total test scores, and the problems or test items (columns) are arranged from left to right in ascending order of difficulty. An S-P Chart is given in Figure 1. This chart associates students with (1) their test scores; (2) their Modified Caution Index (MCI) (see Harnisch & Linn, 1981a, for the computational formula used) and Modified Caution Signal which is based on their performance level and response pattern; and (3) their responses to each item. The first column of the chart contains the student identification number. The second and third columns, RAW and %, contain each

---

1Figures 1–4 in this paper were generated by a computer package of subroutines entitled the Student-Problem Package (SPP). Main frame and microcomputer versions of the SPP have been developed along with complete documentation. Information on the programs is available from the author at the Office of Educational Testing, Research and Service, Institute for Child Behavior and Development, University of Illinois, Urbana-Champaign, 51 Gerty Drive, Champaign, Illinois 61820.
student’s test scores. RAW refers to the total number of items correct and % refers to the percentage of items correct. For example, the second student in the S-P chart, student 220, has a raw test score of 31 and has answered 70 percent of the items correctly.

The fourth column of the chart, IND, contains the Modified Caution Index. It is based upon information about an individual’s response pattern. An MCI value of 0 corresponds to an “ideal” student, one whose response pattern is known from the test score.

MCI’s which approach 1 indicate that a student appears to be responding in such a manner as missing some easy items while answering correctly some difficult items. (See Harnisch and Linn, 1981a for further discussion of the MCI.) For example, student numbers 224 and 212 both score 19 on the test but have different modified caution indices of .35 and .03, respectively. The greater modified caution index for student 224 indicates that the response pattern was more random and that more caution should be used in interpreting the score of 19 for student 224 than for student 212.

From examination of the item response patterns found in the S-P chart, the following

---

Figure 1. Example of a Student-Problem (S-P) Chart for a classroom of 24 students on 44 problems.
observations can be made regarding these two students. Of the 14 easiest items on the test, student 224 is answering six of them incorrectly while student 212 responds correctly to all 14 of the easiest items. On the other hand, the item response pattern of these two students on the 21 most difficult items reveals that student 224 answered seven items correctly while student 212 responded incorrectly to all of the 21 most difficult items.

Classification of Students

The fifth column, SGN, contains the Modified Caution Signal. It represents the results of classifying students according to both their total test score and their response pattern. Specifically, the Modified Caution Signal refers to each student’s classification with respect to test performance (high or low) and MCI (high or low). A total of four classifications3 are used here:

- **Signal A** = high test performance (greater than 50% of items correct) and low MCI (less than or equal to .3)
- **Signal B** = high test performance (greater than 50% of items correct) and high MCI (greater than .3)
- **Signal C** = low test performance (less than or equal to 50% of items correct) and low MCI (less than or equal to .3)
- **Signal D** = low test performance (less than or equal to 50% of items correct) and high MCI (greater than .3)

Sato (1975) suggests that classification in each of these four respective cells identifies students who are doing everything fine (A), are making careless mistakes (B), are in need of more study or possess sporadic study habits (C) or have insufficient readiness (D). For example, student 207 with a performance of 21 on the test or 48 percent of the items correct along with a modified caution index of .32 would be given a modified caution signal “D.” Since she or he answered a number of difficult items correctly but missed easy ones, she or he apparently is guessing well or has knowledge of some topics that are harder in general but doesn’t know some easier topics. Similarly, schools or classrooms can be identified as having in general more of one of the above categories of students.

Response Pattern

The remaining columns of the table contain individual responses (correct = +, incorrect = distractor chosen such as A, B, C, D, or E) to each problem/item on the test. Information in the wrong responses can be used by classroom teachers to examine the questions and responses given by their students. For example, the easiest item on the test, number 29 as noted by its position in the S-P chart, had only distractor A chosen by the students answering the item incorrectly. This test item read, “What is the next larger odd number after 5?” with 6 being the response corresponding to distractor A. With this example it is clear that the careless mistake made by students has to do with misreading the question and hence responding with the next larger number rather than the next larger odd number. Similarly, one could evaluate the most difficult item, number 52, and examine the characteristics of the test item with reference to the two distractors, C and E,

3The selection of the cut scores for classification is dependent on the context. For example, the proportion of items correct might just as well be .75 for a classroom test. Similarly, we also don’t know enough about the distribution of modified caution indices to say that .3 is always a reasonable cut point. Although the notion of a signal system is helpful, it is important to choose boundaries pertinent to a specific context.
chosen by the students from this classroom to understand what may be influencing the students to select these distractors.

**S-Curve**

There are two lines drawn on the S-P chart. The S-curve (solid line on Figure 1) is drawn by placing a vertical line over each S whose position corresponds to the total test score earned by the student represented by that row. The S-curve is completed by starting at the bottom of the chart and connecting the top end point of each vertical line segment to the bottom end point of the line segment on its right. The S-curve on the S-P chart tells us at a glance the consistency of the response pattern for an individual student by examining a row at a time. For a student who has not selected any incorrect distractors to the left of the S-curve, we have an "ideal" one in the sense that the correct responses can be determined from the test score. A greater number of distractors to the left of the S-curve indicates that the student is answering correctly some of the more difficult items.

**P-Curve**

The P-curve (dotted line on Figure 1) is drawn in a similar manner except that the roles of students and items are reversed. A horizontal line segment is drawn over each P whose position corresponds to the number of students correctly answering the item represented by that column. The P-curve is completed by starting at the bottom of the chart and connecting the right end point of each line segment to the left end point of the line segment above it.

Examination of the pattern of item responses with respect to the P-curve provides us with information about each respective item. An item with no distractors above the P-curve would indicate an "ideal" item which is as Tatsuoka (1978) notes "an item failed by no one who got a total score higher than anyone who passed it" (p. 14). An unusual item is one that has a large number of better than average students answering it incorrectly while an equal number of less than average students answer it correctly. Such an item can be classified as an item in need of revision or it may be an item that is measuring a different ability factor for the students.

**Identification of Unusual Items**

By use of the S-P chart, items may also be identified for whom the students in the school/classroom have not had opportunity to cover the material. Studies which examine whether the test content was covered in the course of instruction received by students are very important in documenting the relationship between what is tested and what is taught. Lewy (1972) and Anderson (1975) both illustrate that opportunity to learn defined in terms of teacher's reports of material taught has a significant positive association with student achievement. Leinhardt and Seewald (1981) recently reported on efforts to use teacher's ratings of material covered in class as a predictor of student test performance. They found that teacher's ratings of whether students have been taught information required to answer test items and have been exposed to the format of items explain significant amounts of variance in posttest scores above and beyond that explained by pretest scores.

An MCI for items can be calculated in the same manner as the MCI for students, except that the roles of student and item are reversed. High MCI's for items indicate an unusual set of responses by students of varying ability, and thus these items should be examined closely by the test constructors and the classroom teacher.
Figure 2. Item characteristics for each of the 44 problems. (Read vertically for each of the problems as they are ordered from the easiest to the most difficult.)

The Modified Caution Signal for items refers to each item’s classification with respect to item difficulty (difficult or easy) and MCI (high or low). A total of four classifications⁴ are possible.

Signal W = difficult item (50% or fewer students answered correctly) and low MCI (less than or equal to .3)

Signal X = difficult item (50% or fewer students answered correctly) and high MCI (greater than .3)

Signal Y = easy item (greater than 50% of students answered correctly) and low MCI (less than or equal to .3)

Signal Z = easy item (greater than 50% of students answered correctly) and high MCI (greater than .3)

Summary statistics for the 44 items are given in Figure 2. The items are ordered from easy to difficult based on the performance for the group of 24 students. For example, the easiest item, number 29, was answered correctly by 83 percent of the students and had a modified caution index of .20. The item was classified as a “Y” item based on the above classification procedures. Items relatively difficult for this group of students are found near the right side of Figure 2.

Classroom Statistics

Summary statistics for the classroom of students, the 44 mathematics items, and the mathematics test are given in Figure 3. Sato (1975) defined a measure which indicates the amount of disparity between the S- and P-curves. The S- and P-curves are widely separated from each other when certain situations occur. For example, a large disparity coefficient would be generated when we have a classroom of students that are

⁴As with the signals for students, the setting of the boundary score for determining easy items from difficult items is dependent on the context. One may set the boundary score at .5 for testing conducted at the beginning of a school year while later setting it at .8 for testing at the end of a school year. In addition, not enough is known about the distribution of modified caution indices for items to say that .3 is always a reasonable value.
heterogeneous in ability. Another example of having a large disparity coefficient results when the students are at various levels of understanding on a specific topic. The disparity coefficient ranges from 0 in the “ideal” situation when the S- and P-curves coincide to 1 when the item responses are random.

Sato (1975) and his associates have proposed some rules of thumb for interpreting the disparity coefficient. Based on examination of large amounts of data of various types, Sato notes that disparity coefficient values of around .4 are acceptable while values greater than .6 are “danger signals.” When the disparity coefficient exceeds .6, the test items may be heterogeneous or the classroom of students may consist of subgroups at various stages of learning. The value of the disparity coefficient in Figure 3 is approximately .67 indicating that both the students and the items are relatively heterogeneous. This large disparity coefficient is not surprising since this data set is from a test designed for a statewide survey and not a test for assessing individual progress on a classroom test.

Disparity coefficients generally increase in value when one proceeds from S-P charts prepared based on responses to teacher-made drill sheets to chapter tests to commercially

```
***** STUDENT SUMMARY *****

AVERAGE RAW SCORE = 19.83
STANDARD DEVIATION OF RAW SCORE = 6.39
AVERAGE PERCENT OF ITEMS CORRECT = 45.08 %
AVERAGE MODIFIED CAUTION INDEX = .27
STANDARD DEVIATION OF MODIFIED CAUTION INDEX = .11

***** PROBLEM SUMMARY *****

AVERAGE ITEM DIFFICULTY = 45.08 %
STANDARD DEVIATION OF ITEM DIFFICULTY = .18
AVERAGE MODIFIED CAUTION INDEX = .30
STANDARD DEVIATION OF MODIFIED CAUTION INDEX = .12

***** TEST SUMMARY *****

AVERAGE OVERALL STUDENT PERFORMANCE ON TEST = 45.08 %
RELIABILITY COEFFICIENT measures the internal consistency of the test (CRONBACH'S ALPHA) = .79
DISPARITY COEFFICIENT indicates discrepancy between s- and p-curves = .67

Figure 3. Summary statistics for the classroom of 24 students, the 44 items and the test.
prepared achievement tests. Sato (1975) and his associates note that the disparity coefficient has provided new and valuable information which assist in the analysis of tests and even sometimes in the identification of student groups. The information contained in the disparity coefficient is valuable to the test designer as an index describing quantitatively the correspondence between curriculum coverage (as reflected in the test items) and the performance of the classroom of students being examined.

**Classification and Frequency Tables**

Classification and frequency tables for students and items are given in Figure 4. The first table classifies students with respect to (1) test performance: high (Signal A or B) or low (Signal C or D); and (2) MCI: high (Signal B or D) or low (Signal A or C). The cells of the table give the number (frequency) of students in each of the four possible classifications.

The table for students is a good example of how the modified caution index is put to diagnostic use. For the sake of simplicity, both the modified caution index and the student's total score are each split into two groups forming the four cells in the contingency table. The following comments might be made of the students in each of the cells: A, everything is fine for 7 students; B, no students of high ability made many careless mistakes; C, 11 students need to apply greater effort so that they may master
more of the problems; and D, 6 students showed insufficient readiness for the test or some inconsistent work-study patterns based on their response pattern.

The second table in Figure 4 classifies items with respect to (1) item difficulty: difficult (Signal W or X) or easy (Signal Y or Z); and (2) MCI: high (Signal X or Z) or low (Signal W or Y). The cells of the table give the number (frequency) of items in each of the four possible classifications.

The classification of items on the basis of their difficulty and modified caution index results in the four cells of the lower table. The following comments might be made of the items in each of the respective cells: W, 13 difficult items were discriminating high test performers from low test performers; X, 14 items possibly contain ambiguous phrases or attractive distractors or are items that measure content not covered uniformly; Y, 12 items help to discriminate among low test performers and possibly contain hints about the correct answer; and Z, 5 items in possible need of revision since numerous low test performers answered them correctly.

In general, teachers and counselors have found S-P charts helpful as a means of providing diagnostic information, as in the identification of particular skills and knowledge which individuals have yet to master. Administrators, researchers, counselors and classroom teachers have indicated that the S-P charts are very easy to understand. Furthermore, teachers have indicated numerous advantages of the S-P chart reporting form, where correct responses on the S-P chart are represented with a + and incorrect responses by the selected distractor, over the more traditional S-P reporting forms.

**Categorized S-P Chart**

The "categorized S-P chart" is another form of applying the S-P curve theory to report item response patterns to classroom teachers. This form of reporting can be applied with tests that have multiple items classified to a particular content category. First, the categories of items are arranged from easy to hard. Caution indices for students are given based on the ordering of items from easy to hard across categories. Figure 5 provides a categorized S-P chart for 20 students on 10 problems representing three item categories.

New information is gained from the categorized S-P chart which helps to further diagnose students' understanding of specific skills. This simple rearrangement of the response pattern helps to better understand student strengths and weaknesses and also the areas of the curriculum in need of greater coverage and emphasis. For example, the response pattern for student 10012 represents a student with mastery of the skills for items in category 2 and 3 but nonmastery of the items in category 1, the easiest of the three test item categories. On the other hand, student 10017 also with a total score of 7 has a response pattern which indicates complete mastery of the items in the two easiest item categories while showing nonmastery of the items from category 2. The interesting difference between these two students with similar total test scores of 7 is the difference in their modified caution index values. Student 10012, with a modified caution index of .62 varies dramatically from student 10017 whose modified caution index is .08.

**LIMITATIONS AND ADVANTAGES**

One limitation of S-P curve theory for examining item response patterns is that the results are dependent solely on the test or subset of items used in the analysis. If the validity of the test is questioned, then the S-P results based on such a test would have to be used with caution. Another concern expressed about this approach is that the indices
### Categorized S-P Chart

<table>
<thead>
<tr>
<th>Student Number</th>
<th>Test Score (Raw) (%)</th>
<th>Modified Caution Ind/Sgn</th>
<th>Problem Number</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>10018</td>
<td>10 100</td>
<td>.00 A</td>
<td>136 9570 842</td>
<td>111 3333 222</td>
</tr>
<tr>
<td>10009</td>
<td>9 90</td>
<td>.83 B</td>
<td></td>
<td>111 0111 111</td>
</tr>
<tr>
<td>10019</td>
<td>9 90</td>
<td>.17 A</td>
<td></td>
<td>110 1111 111</td>
</tr>
<tr>
<td>10020</td>
<td>9 90</td>
<td>.00 A</td>
<td></td>
<td>111 1111 111</td>
</tr>
<tr>
<td>10010</td>
<td>8 80</td>
<td>.50 B</td>
<td></td>
<td>110 0111 111</td>
</tr>
<tr>
<td>10011</td>
<td>8 80</td>
<td>.40 B</td>
<td></td>
<td>101 1111 011</td>
</tr>
<tr>
<td>10003</td>
<td>7 70</td>
<td>.00 A</td>
<td></td>
<td>111 1110 100</td>
</tr>
<tr>
<td>10012</td>
<td>7 70</td>
<td>.62 B</td>
<td></td>
<td>000 1111 111</td>
</tr>
<tr>
<td>10017</td>
<td>7 70</td>
<td>.08 A</td>
<td></td>
<td>111 1111 000</td>
</tr>
<tr>
<td>10001</td>
<td>6 60</td>
<td>.00 A</td>
<td></td>
<td>110 1100 110</td>
</tr>
<tr>
<td>10004</td>
<td>6 60</td>
<td>.53 B</td>
<td></td>
<td>111 0000 111</td>
</tr>
<tr>
<td>10002</td>
<td>5 50</td>
<td>.06 C</td>
<td></td>
<td>111 1100 000</td>
</tr>
<tr>
<td>10008</td>
<td>5 50</td>
<td>.63 D</td>
<td></td>
<td>101 0010 101</td>
</tr>
<tr>
<td>10005</td>
<td>4 40</td>
<td>.27 C</td>
<td></td>
<td>100 1001 100</td>
</tr>
<tr>
<td>10013</td>
<td>3 30</td>
<td>.62 D</td>
<td></td>
<td>001 1001 000</td>
</tr>
<tr>
<td>10014</td>
<td>3 30</td>
<td>.08 C</td>
<td></td>
<td>110 1000 000</td>
</tr>
<tr>
<td>10015</td>
<td>3 30</td>
<td>.69 D</td>
<td></td>
<td>000 1000 011</td>
</tr>
<tr>
<td>10006</td>
<td>2 20</td>
<td>.10 C</td>
<td></td>
<td>100 0100 000</td>
</tr>
<tr>
<td>10016</td>
<td>2 20</td>
<td>.30 D</td>
<td></td>
<td>010 1000 000</td>
</tr>
<tr>
<td>10007</td>
<td>1 10</td>
<td>.33 D</td>
<td></td>
<td>000 0100 000</td>
</tr>
</tbody>
</table>

Problem Total

<table>
<thead>
<tr>
<th></th>
<th>111</th>
<th>1111</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>520</td>
<td>4300</td>
<td>109</td>
</tr>
</tbody>
</table>

Percent Correct by Category

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 5. Categorized Student-Problem (S-P) Chart for 20 students on 10 problems in 3 content categories.
computed for individuals are based on groups of students, for example, students in a
classroom/school or district, and not on an individual student. At a more technical level,
the statistical properties and standard errors of these indices are not well understood.

On the other hand, the S-P approach is relatively easy to apply with the use of a
computer and the package of Student-Problem programs available for the main frame or
the microcomputer. Furthermore, the information gained from the pattern of item
responses does not require any additional data gathering efforts to examine the coverage
of topics based on the groups of students being analyzed. Information from computer
generated S-P charts is readily comprehensible and is reported in a concise form that
teachers and counselors can use in diagnosing strengths and weaknesses of their students.
This approach is also very useful for the basic researcher concerned about understanding
the relationship of background characteristics with patterns of responses.

REFERENCES

ALDERMAN, D. L., SWINTON, S. S., & BRASWELL, J. S. Assessing basic arithmetic skills
and understanding across curricula: Computer-assisted instruction and compensatory education.

ANDERSON, L. Opportunity to learn, test bias, and school effects. Paper presented at the annual

ARMBRUSTER, B. B., STEVEN, R. O., & ROSENSHINE, B. Analyzing content coverage and
emphasis: A study of three curricula and two sets. Technical Report No. 26, Center for the
Study of Reading, University of Illinois, Urbana-Champaign, 1977.

BIRENBAUM, M., & TATSUMO, K. K. The use of information from wrong responses in
measuring students' achievement (Research Report 80-1 ONR). Urbana, IL: University of

BROWN, J. S., & BURTON, R. B. Diagnostic models for procedural bugs in basic mathematical

DAVIS, R. B. Error analysis in high school mathematics, conceived as information processing
pathology. Paper presented at the annual meeting of the American Educational Research

DAVIS, R. B., JOCKUSCH, E., & MCKNIGHT, C. Cognitive processes in learning algebra. The

DONLON, T. F., & FISCHER, F. E. An index of an individual's agreement with group-

DONLON, T. F., & RINDLER, S. E. Consistency of item difficulty for individuals and groups in
the Graduate Record Examination. Paper presented at the annual meeting of the American

DRASGOW, F. Statistical indices of the appropriateness of aptitude test scores. Unpublished

DRASGOW, F. Choice of test model for appropriateness measurement. Applied Psychological
Measurement, 1982, 6, 297–308.

GLASER, R. The future of testing: A research agenda for cognitive psychology and psychometrics

GUTTMAN, L. Integration of test design and analysis. In Proceedings of the 1969 Invitational

HARNISCH, D. L., & LINN, R. L. Analysis of item response patterns: Questionable test data and


HARNISCH, D. L., & LINN, R. L. Identification of aberrant response patterns in a test of


TATSUOKA, K. K., & LINN, R. L. Indices for detecting unusual patterns: Links between two general approaches and potential applications. Applied Psychological Measurement, in press.


TATSUOKA, K. K., & TATSUOKA, M. M. Spotting erroneous rules of operation by the individual consistency index. Journal of Educational Measurement, in press.


AUTHOR

HARNISCH, DELWYN L. Address: Institute for Child Behavior and Development, University of Illinois, 51 Gerty Drive, Champaign, IL 61820. Title: Assistant Professor. Degrees: B.S., Concordia Teachers College, Seward, NB; M.Ed. & Ph.D., University of Illinois at Urbana-Champaign. Specialization: Applied Statistics, Educational Influences.
You have printed the following article:

Item Response Patterns: Applications for Educational Practice
Delwyn L. Harnisch
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28198322%2920%3A2%3C191%3AIRPAFE%3E2.0.CO%3B2-R

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.

[Footnotes]

2 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%3AAOIRPO%3E2.0.CO%3B2-4

References

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%3AAOIRPO%3E2.0.CO%3B2-4

Overlap: What's Tested, What's Taught?
Gaea Leinhardt; Andrea Mar Seewald
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28198122%2918%3A2%3C85%3AOWTWT%3E2.0.CO%3B2-7

NOTE: The reference numbering from the original has been maintained in this citation list.
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

Detection of Aberrant Response Patterns and Their Effect on Dimensionality
Kikumi K. Tatsuoka; Maurice M. Tatsuoka
Stable URL:
http://links.jstor.org/sici?sici=0362-9791%28198223%297%3A3%3C215%3ADOARPA%3E2.0.CO%3B2-1

Spotting Erroneous Rules of Operation by the Individual Consistency Index
Kikumi K. Tatsuoka; Maurice M. Tatsuoka
Stable URL:
http://links.jstor.org/sici?sici=0022-0655%28198323%2920%3A3%3C221%3ASEROOB%3E2.0.CO%3B2-S

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

NOTE: The reference numbering from the original has been maintained in this citation list.