ANALYSIS OF TEST SCORE PATTERNS: THE STUDENT-PROBLEM (S-P) TECHNIQUE

David L. McArthur

CSE Report No. 218
1983

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

Definition of the model. A system of analyzing patterns of student responses called Student-Problem (S-P) score table analysis has been developed over the last decade by a group of educational researchers in Japan (Sato, 1974, 1975, 1980, 1981a, 1981b; Sato & Kurata, 1977; Kurata & Sato, 1981; Sato, Takeya, Kurata, Morimoto & Chimura, 1981). While the mathematics associated with derivative indices in this system are relatively complex, the S-P system itself is predicated on a simple reconfiguring of test scores. Rather similar analyses of student performance on educational tests can be found in the professional literature of a half-century ago, but recent developments by Sato and colleagues represent significant improvements both in concept and execution. The method appears to hold a number of possibilities for effective and unambiguous analysis of test score patterns across subjects within a classroom, items within a test, and, by extension, to separate groups of respondents. It is a versatile contribution to the field of testing, containing minimal requirements for sample size, prior scoring, item scaling, and the like. The S-P model lends itself to extensions into polychotomous scoring analysis of multiple patterns, and analysis of patterns of item bias.

Test scores are placed in a matrix in which rows represent individual respondents' responses to a set of items, and columns represent the responses given by a group of respondents to a set of items. The usual (and most convenient) entries in this matrix are zeros for wrong answers and ones for correct answers. Total correct

Figure 1
S-P Chart for a Six Item Test Administered to 20 Students

Items in ascending order of difficulty

rank	1	2	3	4	5	6
item #	1	5	4	2	3	6

Average passing rate p = .425
Discrepancy D* = .525

Students, in descending order of total score									Total Correct	Caution Index for Students
Rank	I.D.#									C_i^*
1	02	1	1	1	1	0	0:		4	0.000
2	04	1	1	1	1	0	0:		4	0.000
3	05	1	1	1	0	1:	0		4	0.000
4	11	1	1	0	1	1:	0		4	0.034
5	12	1	0	1	0	1:	1		4	0.552 *
6	14	1	1	1	1	0:	0		4	0.000
7	20	1	1	1	0	1:	0		4	0.000
8	22	1	1	1	0	1:	0		4	0.000
9	23	1	1	1	0	0:	1		4	0.276
10	07	1	1	0	0	1:	0		3	0.033
11	17	1	1	0:	0	1	0		3	0.033
12	19	1	1:	0	1	0	0		3	0.033
13	27	1	1:	0	1	0	0		3	0.033
14	29	0	1:	1	0	1	0		3	0.433 *
15	03	1	0:	0	0	1	0		2	0.276
16	06	0	1:	0	0	1	0		2	0.448 *
17	08	1	1:	0	0	0	0		2	0.000
18	10	1	1:	0	0	0	0		2	0.000
19	15	1:	0	1	0	0	0		2	0.241
20	16	1:	0	0	1	0	0		2	0.276
21	21	1:	0	0	1	0	0		2	0.276
22	28	1:	0	1	0	0	0		2	0.241
23	01	:1	0	0	0	0	0		1	0.000
24	09	:0	1	0	0	0	0		1	0.238
25	13	:1	0	0	0	0	0		1	0.000
26	18	:0	0	0	1	0	0		1	0.619 *
27	24	:0	0	0	1	0	0		1	0.619 *
28	25	:0	1	0	0	0	0		1	0.238
29	26	:1	0	0	0	0	0		1	0.000
ITEM TOTALS:		**2**	**1**	**1**	**1**	**1**	**0**			
		3	**8**	**1**	**0**	**0**	**2**			
C_j^* caution index for items		0	0	0	0	0	0			
		.	.	.	.	.	.			
		1	1	1	4	2	0			
		6	4	1	2	3	0			
		7	8	1	3	1	0			
					*					

* High caution index for unusual response pattern.

scores are calculated for each respondent, and total number of correct responses are tallied for each item. Rows are reordered by descending total number of correct responses; columns are reordered by ascending order of difficulty of items. The resulting matrix has several aspects which are particularly convenient for a detailed appraisal of respondents or items, singly or collectively. A short example, annotated and indexed with several computations to be explained below, is shown on Figure 1.

Two cumulative ogives are drawn over the matrix to form the framework for further analysis. Because the data are discrete, the ogives take on a stair-step appearance, but both can be thought of as approximations to curves which describe in summary form the two distinct patterns embedded in the data. The first is a curve reflecting respondents' performance as shown by their total scores; the second is a similarly overlaid ogive curve reflecting item difficulties. In one special circumstance, the two curves describe only one pattern: if the matrix of items and respondents is perfectly matched in the sense of a Guttman scale, both of the curves overlap exactly. All of the correct responses would be to the upper left while all of the incorrect responses would be to the lower right. However, as the occurence of either unanticipated errors by respondents with high scores or unanticipated successes by respondents with low scores increases, or as the pattern of responses becomes increasingly random, the respondent or student curve (S-curve) and the item or problem curve (P-curve) become increasingly discrepant. Sato has developed an index which evaluates the degree of discrepancy or lack of conformation between the S- and P-curves. This index will be

zero in the special case of perfectly ordered sets, and will approach 1.0 for the case of totally random data.

For any respondent, or for any item, taken individually, the pattern of scores reflects that row or column in relation to the pattern established by the configuration of sorted rows and columns. For any given individual respondent or single item, the response pattern may be "perfectly ordered" in the sense used above. The row or column shares a symmetry with the associated row or column marginal; in the case of dichotomous data this symmetry is seen in a high positive point-biserial correlation. As the match between patterns declines-- that is, as the row or column under consideration shares less and less in common with the associated marginal formed from all rows or all columns--the point-biserial also declines. Unfortunately, r_{pbis} is not independent of the proportions within the data and never reaches 1.0 in practice. Cases of complete "symmetry" between row or column and the corresponding marginal which happen to differ in proportions do not yield the same correlation coefficients.

An index which is stable across differing proportions is Sato's Caution Index C, which gives a value of 0 in the condition of "perfect symmetry" between row or column and row marginal or column marginal. As unanticipated successes or failures increase and "symmetry" declines, the index increases (a modification of the Caution Index, called C*, has an upper bound of 1.0). Thus a very high index value is associated with a respondent or item for which the pattern of obtained responses is very discrepant from the overall pattern established by all members of the set.

Harnisch and Linn (1982) present the modified Caution Index as follows:

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

where $i = 1,2,\ldots,I$ indexes the examinee,

$j = 1,2,\ldots,J$ indexes the item,

$u_{ij} = 1$ if the respondent i answers item j incorrectly,

0 if the respondent i answers item j incorrectly,

n_i = total correct for the i^{th} respondent, and

n_{ij} = total number of correct responses to the j^{th} item.

Harnisch and Linn explain that the name of the index comes from the notion that a large value is associated with respondents that have unusual response patterns. It suggests that some caution may be needed in interpreting a total correct score for these individuals. An unusual response pattern may result from guessing, carelessness, high anxiety, an unusual instructional history or other experiential set, a localized misunderstanding that influences responses to a subset of items, or copying a neighbor's answers to certain questions.

A large value may also suggest that some individuals have acquired skills in an order which is not characteristic of the whole group. The index says nothing about the most able respondents with perfect total scores, because the "symmetry" condition is met. More importantly, if a respondent gets no item correct whatsoever, both the

total score and the caution index will be zero since, again, the "symmetry" condition is met; in this situation the available information about the respondent is insufficient to make any useful diagnosis. Most persons, though, will achieve total scores between the extremes and for them the caution index provides information 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.

A related development is a modification of the Caution Index to examine patterns of responses to clusters or subtest scores and an "ideal" pattern of scores of individual subtests, the perfect Guttman pattern (Fujita & Nagaoka, 1974, in Sato, 1981).

Sato has developed an index of discrepancy to evaluate the degree to which the S and P curves do not conform either to one another or to the Guttman scale. Except in the case of perfectly ordered sets there is always some degree of discrepancy between curves. The index is explained as follows:

$$D^* = \frac{A(I,J,p)}{A_B(I,J,p)}$$

where the numerator is the area between the S curve and the P curve in the given S-P chart for a group of I students who took J-problem test and got an average problem-passing rate p, and $A_B(I,J,p)$ is the area between the two curves as modeled by cumulative binomial distributions with parameters I,J, and p, respectively (Sato, 1980, p. 15; indices rewritten for consistency with notation of Harnisch & Linn).

The denominator is a function which expresses a truly random pattern of responses for a test with a given number of subjects, given number of items, and given average passing rate, while the numerator reflects the obtained pattern for that test. As the value of this ratio approaches 1.0, it portrays an increasingly random pattern of responses. For the perfect Guttman scale, the numerator will be 0 and thus D* will be 0. The computation of D* is functionally derived from a model of random responses, but its exact mathematical properties have not been investigated thoroughly.

Also available, but not yet studied in detail, is an index of "entropy" associated with distributions of total scores for students choosing different answers to the same question. This index explores the particular pattern of responses (right answer and all distractors included), in the context of overall correct score totals for these responses.

While most of the published work using the S-P method has concentrated on binary data (0 for wrong answer, 1 for right answer), and calculations are most tractable in that form, the indices developed from the configuration of S- and P-curves are not limited to such data. The technique can be extended to multi-level scoring (see Possible Extensions to the model, below).

Measurement philosophy. A precursor to the S-P method is the concept of "higgledy-piggledy" (or "hig" for short) suggested by Thomson about 1930 and elaborated by Walker in a trio of contributions (1931, 1936, 1940), but evidently carried no further by educational researchers at that time. Walker examined right/wrong answers to a

set of independent items with particular reference to score-scatter, which had been a focus of attention since the early twenties. Where scatter reflects random behaviors on the part of examinees, "hig" is said to be present. However,

> By a test being unig (the converse of hig) we mean that each score x is composed of correct answers to x easiest questions, and therefore to no other questions. Hig implies a departure from this composition. Note that it is not sufficient for our purposes to define unig by stipulating that every score x is identical in composition--there must be added the condition that it is composed of the x easiest items; in other words the score $x + 1$ always compromises the x items of the score x, and one more. Now if hig is absent, that is each score is unig, it is easy to show that an exact relationship exists between the n's of the answer-pattern and the N's of the score scatter (1931, p.75).

The parallel to Guttman scaling, while the latter is far more mathematically rigorous, is obvious; Sato's indices appear to address the same underlying concepts.

Guttman's (1944) statistical model for the analysis of attitudinal data was formulated to solve scaling problems in the context of morale assessment for the U.S. Army. While the initial approaches were not at all technically sophisticated and involved much sorting of paper by hand, Guttman's conceptualization was powerful; the scalogram approach, and especially its mathematical underpinnings, received extensive development during the 1950's. But by 1959, Maxwell had expressed rather strong disappointment with the narrow range of application these procedures had enjoyed, and suggested two general statistics which might serve to dissolve the arbitrary distinction between qualitative and quantitative scales, and, at the same time, reduce some of the cumbersome calculations. (One of these statistics is a regression coefficient developed from the residual between observations and perfect patterns of responses to a given set

of items, which bears some conceptual resemblance to Sato's D*.) However, the primary audience for these technical contributions appears to have been educational statisticians and researchers. Only infrequently was attention given to simplifying the techniques for a broader potential audience; Green's (1956) contribution is one exception, although published in a highly sophisticated journal.

Many of the publications by Sato and colleagues in Japan seem geared directly to end-users, teachers in the classroom who, with the S-P method and handscoring or microcomputer processing, can analyze their own instructional data for purposes of understnading their students' comprehension and modifying their own instruction. The overarching concern of the Educational Measurement and Evaluation Group at the Nippon Electric Company's Computer and Communication Systems Research Laboratories has been development and dissemination of readily understandable and adaptable procedures. Evidently it has proved popular in a variety of classroom settings in Japan, and has been applied to the following areas:

- test scoring and feedback to each examinee about his/her own performance on a test
- feedback to the instructor about both individual and group performance
- analysis of types of errors made by students
- analysis of instructional process and hierarchies of instructional units
- item analysis, rating scale analysis, questionnaire analysis
- test score simulations
- development of individual performance profiles across repeated testings

Two characteristics are shared by all of these approaches: first, the central focus of the study is the degree to which items and/or respondents are heterogeneous, and second, the actual element of raw data (say, 0 or 1) is assumed to be best understood in terms of its position in a matrix with orderly properties. Interestingly, the article by Green (1956) noted above forms the only overt link between the S-P method and earlier work in English on analysis of response patterns.

Where the S-P method diverges from its predecessors can be seen in the very reduced role played by probability theory, and the absence of anything resembling tests of statistical significance (a shortcoming addressed below). Much of the work on the S-P method is either in Japanese or in English-language journals not generally available in the West. In the U.S. the small number of research presentations using the S-P method to date is small (Harnisch, 1980; Harnisch & Linn, 1981, 1982; McArthur, 1982; Tatsuoka, 1978; Tatsuoka & Tatsuoka, 1980).

Assumptions made by the model. The S-P method starts from a complete matrix of scores, doubly reordered by I rows and J columns. The model applies equally well to the trivial case of a 2 x 2 matrix, and to 2 x J and I x 2 retangular matrices; it also appears to have no functional upper limit on the number of items or respondents. However, missing data cannot be incorporated effectively. That is, each respondent and item must have complete data since all calculations are made with reference to i and j as constant values. For purposes of reordering, if two or more respondents have the same total score their ranks are tied but their positions within the sorted

matrix must be unique, so ties between marginals are resolved arbitrarily (a situation which could cause some small instability in the S and P curves). In respect to both individual scores and sets of scores taken as a whole, no explicit probabilistic formulation is involved, although underlying the analysis of the matrix is a model premised on cumulative binomial or beta binomial distributions, with parameters I (number of cases), J (number of items), and p (average passing rate). No study has been made of how guessing affects the obtained pattern of responses, nor how corrections for guessing might affect the S-P chart. Because of the very small number of assumptions made by the model, its interpretation does not require a strong theoretical background, and in fact can be annotated easily by computer as an aid to the user novice. Indeed, the graphic reordering with overlay of S- and P-curves but no further statistics appears sufficient to allow teachers, with use of a brief nontechnical reference guide, to make well-reasoned instructional decisions.

One implicit assumption deserves special attention. In the derivation of a caution index for item or respondent, the entire existing configuration of I items and J respondents, whether valid or not, enters into consideration. That is, because the frame of reference does not extend beyond the data at hand, the derivative indices are inherently subject to limits on their analytic utility. However, it is important to recognize that for the great bulk of practical testing applications, such limitations in fact may be advantageous. Each index also depends on a linear interpretation of steps between marginal totals, although it is readily demonstrable

that substitution of a highly discriminating item for a weakly discriminating one, or a very able examinee for a poor one, can alter many of the indices for both persons and items. Additionally, the linearity constraint treats all data elements within the matrix equally, despite unknown (and perhaps inestimable) contributions from chance correct responses. On the other hand, without further tests of significance, the resulting statistical uncertainties, which are small under most conditions, have little practical importance in the usual classroom situation.

Strengths and weaknesses. Obvious strengths of the S-P system are its simplicity, wide potential audience, and portability. The code required for computer processing can be exceptionally brief and with the increased availability of microcomputers, can be delivered to the classroom teacher directly. According to Harnisch and Linn (1982), the caution indices compare well with Cliff's (1977) C_{i1} and C_{i2}, Mokken's (1971) H^*_i, Tatsuoka and Tatsuoka's (1980) Norm Conformity Index (NCI), and van der Flier's (1977) U', all of which are harder to calculate as a rule. As an inherently flexible system, it appears to be suitable for a variety of test types, and for a range of analyses even within the same test. The novice user need not master the full range of calculations in order to make excellent use of more elementary portions of the results. A sophisticated user can easily iterate selectively through an existing data set, choosing particular items or persons not meeting some criterion for performance, and recasting the remaining matrix into a revised chart. Under certain conditions, addressed below, the method can be adapted to examination of test bias (McArthur, 1982).

Weaknesses include the following three general criticisms. No substantive body of psychometric or educational theory preceded the development of practical applications of the model because in fact its development was not paradigm-driven. Instead, the S-P techniques arose in response to a perceived need for classroom teachers to have a readily interpretable, minimally complex tool for test analysis. Thus, at present little can be said regarding questions of reliability, validity, true scores, scaling theory, or quality of measurement. No extant work addresses either the problem of signal/noise ratio or of model fit. The absence of a strong theoretical base dampens the development of rationally interconnected research hypotheses, although the method offers ample opportunities for direct investigation of individual performance and item characteristics. The absence of strong theory-derived hypotheses leaves a recognizable gap in the ability to draw strong inferences from the S-P method. That is, in developing a diagnostic interpretation of a student's score pattern, the teacher or researcher must make a conscious effort to balance the evidence in light of some uncertainty about what constitutes critical or significant departure from the expected.

These weaknesses do not affect the classroom teacher to any major degree. In the classroom, the technique is used for confirming knowledge about individual students gained in the course of interaction with the class, and/or to confirm that items on a particular test are reasonably well suited to the class. From the researcher's viewpoint, the weaknesses constitute rather important blocks to further development. On the other hand, because of some

points of similarity between the S-P technique and less arcane aspects of a number of existing models, hypothesis building tends to proceed anyway. The absence of recognizable criteria for establishing statistical significances for degree of heterogeneity is an important technical problem. Because the various indices appear to share a great deal in common with indices having known statistical properties from other research models, an initial direction for such effort would be to examine these parallels.[1]

Present areas of application. All of the published studies in English to date utilize the S-P method exclusively in the context or right/wrong (1/0) scoring. These studies each use data collected from multiple-choice tests (generally reading or math) administered to primary or secondary level students. In this body of literature the general application is either to the task of individual student analysis, or more frequently, to item analysis. With an appropriate microcomputer--one marketed exclusively in Japan is configured exclusively for the purposes of the S-P method--classroom teachers can use the technique interactively. Science teachers in Japan are evidently the largest cluster of users, although details about acceptance and daily utilization remain sketchy.

A different application arises in the context of large-scale assessment. Harnisch (personal communication) reports that several school districts have contracted for S-P analysis of mid-year and final achievement test scores. Several thousand individuals tested on dozens of items pose no new conceptual or mathematical complexity and in this situation the results can be used to address both item-level and aggregate-level questions.

---

[1] Strong parallels also can be found with aspects of the analysis of planar Wiener processes and spatial patterns, from the domain of mathematical geophysics.

Possible extensions of the model. Three new directions for the S-P method are being explored. The first is the application of iterative procedures, first suggested by Green (1956) in a brief paragraph on p-tuple analysis of Guttman scales. Zimmer (1982) has collected extensive developmental data on children's perception of various tasks and attributions; these data incorporate multiple discrete levels of performance arranged, according to theory, in a logical staircase ascendency. P-tuple iterative analyis by the S-P procedure appears to offer answers to three questions: a) does a broad sample of children respond in an orderly manner to the range of tasks; b) does such order reflect known characteristics of the sample (viz. developmental level as measured on standardized procedures); and c) do deviations from the symmetrical relationship between the developmental complexity of the task and the developmental level of the child reflect consistent support for one or another competing theory of development. For these data, separate S-P analyses were made with the first developmental level scored 0 and all others 1, then the first two levels scored 0 and all others 1, and so on. Stability of person order and item order, uniformity of the staircase intervals, and relationships between item difficulty and item complexity can be studied. Preliminary evidence suggests that the S-P method provides a system of analysis for such multi-level data that exceeds the explanatory power of several extant procedures.

In p-tuple analysis, which makes use of repeated passes through data, some questions of a technical nature are unresolved at this time. For example, it is clear that successive reorderings can perturb the positional stability of any one respondent within the

matrix or any one task within the matrix, to some degree. However, changes in ordering contribute to changes in the S-P indices, and whether such changes, and/or linearity assumptions and violations therein, play an important role is also under study in the context of these developmental data. Another way to think of this problem is to imagine a single matrix of persons x items with the S-P chart from each developmental level overlaid. The most difficult tasks would be accomplished only by the most developmentally advanced individuals, and below a certain competence (i.e. the highest S-curve on this compound chart) virtually no one would be expected to succeed on those tasks. The ordering of those participants who fail at all tasks of that difficulty level is arbitrary, because their total score for these most difficult tasks is zero. But their ordering would not be arbitrary on tasks of moderate or low difficulty, at which more successes might be anticipated and the corresponding S-curves would be located lower on the chart. What constitutes acceptable and interpretable slippage of this kind needs further probing. Perhaps the best analogy is to the term "seiche," drawn from the field of oceanography: it refers to regular, entirely predictable tidal motions occuring within confined bodies of water. Such seiche in a polychotomous S-P chart ought to show itself totally consistent and predictable.

The second area for development of the S-P method is in the realm of scalar data, for which a number of statistical assumptions have been developed. An example is signal detection analysis, in which the "raw element" of data is once again a 0/1 response, this time for absence or presence of perceived stimulus. A variety of complex

statistical techniques have been used to investigate how such stimuli, presented across a range of intensities over a repeated number of trials, are processed by the receiver. The analog in S-P analysis might best be portrayed as a three-dimensional matrix of persons, items, and repeated trials. Items are not necessarily objectively identical from trial to trial, and responses are tempered by not one but several possible orderly progressions. Such three-dimensional and higher-dimensional data challenge the S-P method to provide cohesive summary statistics which can be evaluated probabilistically.

An extension of the S-P technique to the study of test bias has been made by McArthur (1982). Where two distinct groups have been tested on the same instrument or on two instruments one of which is an exact translation of the other, S-P analysis offers an interesting alternative to the complex techniques for detection of biased items generally in use. McArthur studied the response patterns for items on the California Test of Basic Skills, administered to both English-speaking and Spanish-speaking children, the latter taking the CTBS-Espanol. Even when proportions of children achieving correct responses to a given item differ between the two language groups, the item may not be biased. However, the D* values for the student-problem matrices calculated separately for the two groups suggest that the Spanish-language group engaged in more random responding than did their English-speaking counterparts. A significantly larger number of items for the former group show that those children above the P-curve (children who in a case of "symmetry" as defined earlier would be expected to do well) who gave the correct response were frequently fewer in number than the corresponding sample

from the English-language group. That is, deleting cases below the P-curve, which are more likely to have engaged in random responding, leaves a finite number of respondents for whom the prediction of success is high. Obviously on easier items this reduced sample is larger than for difficult items because of the shape of the P-curve. Nonetheless, while the p values for a given item may differ significantly between one group and the other, the proportions of right answers above the P-curves can be statistically identical. To establish evidence of bias, the additional requirement is that for students in the disadvantaged group who by their pattern of performance on the test as a whole should have succeeded with a particular item, that item generated erroneous responding for one group more than for another.

REFERENCES

Cliff, N. A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 1977, 42, 375-399.

Fujita, T., & Nagaoka, K. Arbitrary Ho full-marked S-P table. Institute of Electronic Communication Engineers of Japan, 1974 (In Japanese).

Green, B.F. A method of scalogram analysis using summary statistics. Psychometrika, 1956, 21, 79-88.

Guttman, L. A basis for scaling quantitative data. American Sociological Review, 1944, 9, 139-150.

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

Harnisch, D.L., & Linn, R.L. Identification of abberant response patterns. Champaign, Illinois: University of Illinois, 1982. National Institue of Education Grant No. G-80-0003, Final Report.

Kurata, T., & Sato, T. Similarity of some indices of item response patterns based on an S-P chart. Computer and Communication Systems Research Laboratories, Nippon Electric Company, Research Memorandum E181-4, 1981.

Maxwell, A.E. A statistical approach to scalogram analysis. Educational and Psychological Measurement, 1959, 19, 337-349.

McArthur, D.L. Detection of item bias using analyses of response patterns. Paper presented to the Annual meeting of the American Educational Research Association, New York, 1982.

Mokken, R.J. A theory of procedure of scale analysis. The Hague: Mouton, 1971.

Sato, T. A classroom information system for teachers, with focus on the instructional data collection and analysis. Association for Computer Machinery Proceedings, 1974, 199-206.

Sato, T. Analysis of students' pattern of response to individual subtests. Computer and Communications Systems Research Laboratories, Nippon Electric Company, Research Memorandum E181-2, 1981a.

Sato, T. Similarity of some indices of item response patterns. Computer and Communications Research Laboratories, Nippon Electric Company, Research Memorandum E181-1, 1981b.

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

Sato, T. The S-P chart and the caution index. Nippon Electric Company, Educational Informatics Bulletin, 1980.

Sato, T., & Kurata, M. Basic S-P score table characteristics. NEC Research and Development, 1977, 47, 64-71.

Sato, T., Takeya, M., Kurata, M., Morimoto, Y., & Chimura, H. An instructional data analysis machine with a microprocessor -- SPEEDY. NEC Research and Development, 1981, 61, 55-63.

Tatsuoka, M.M. Recent psychometric developments in Japan: Engineers grapple with educational measurement problems. Paper presented at the Office of Naval Research Contractors' Meeting on Individualized Measurement, Columbia, Missouri, 1978.

Tatsuoka, M.M., & Tatsuoka, K. Detection of abberant response patterns and their effects on dimensionality. Computer-based Education Research Laboratory, University of Illinois, Research Report 80-4, 1980.

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

Walker, D.A. Answer-pattern and score-scatter in tests and examinations. British Journal of Psychology, 1931, 22, 73-86; 1936, 26, 301-308; 1940, 30, 248-260.

Zimmer, J.M. Analysis of developmental levels of children. University of California, Santa Barbara, 1982. In preparation.