An Educational and Psychological Test Item Data Base System

MASAHIKO KURATA* and TAKAHIRO SATO*

This paper presents a computer-assisted test construction support system of educational and psychological test based on a test item data base. The test item data base consists of a large number of items coded by their characteristics and statistics. A test author can construct a high quality test by selecting items in the data base referring to their characteristics and statistics such as their content area within a curriculum, behavioral objectives and item statistics. This system provides several kinds of check sheets to investigate content validity, and also the predictions of test score distributions and a coefficient of reliability to investigate statistical reliability of a constructed test prior to test administration. If these check sheets or the result of predictions do not satisfy the test author's aim, he can update several items and reform the test.

1. Introduction

A test item data base which can be used to construct a test for educational and psychological measurement consists of a large number of items, each coded by its characteristic and statistic data such as behavioral objectives, grade level and item statistics. These test items in the data base can be drawn on to construct a test as occasion calls for. This paper presents the computer-assisted test construction support system of educational and psychological test based on a test item data base.

The test construction flow diagram and the test construction support system is shown in Fig. 1. In Fig. 1, the area surrounded by dotted lines shows the test construction support system.

The test author constructs the high quality test by selecting some items in the data base. He can inquire items referring to such as their content area within a curriculum, behavioral objectives and item statistics. This system can provide several kinds of check sheets or prescription to investigate content validity and statistical reliability of a constructed test for the test author. The check sheets and prescription contains a specification matrix for item characteristics and a prediction of test score distributions. If the specification matrix for item characteristics or the predictions of test score distributions do not satisfy the test author's aim, he can update several items and reform the test prior to test administration.

2. System Design

This system is based upon an Item file and a Test file which can be referred to by the management program, the reference program, the item analysis program and the validity check programs. The data files are described in Table 1 and the series of programs are described in Table 2.

The Item file contains item characteristics and statistics. The Test file contains test characteristics and statistics. The Item file is made in two parts. One contains fixed data, and another contains accumulated data. Fixed data are characteristic data which are item number, content area within a curriculum, behavioral objectives, item type, related item number, etc. Accumulated data are made up of item performance data which are error rate, discrimination index, number of subjects, etc. The Test file contains test number, teacher's name, applied group name, test score distribution for testee and other satistical indices.

The management program updates an Item file and a Test file. The reference program is used to refer to the desired items and test in the data file. This reference can be interactively operated using the KB-CRT terminal. The item analysis program is used to represent informative reports to the test author and register item/test statistics in these files depending on actual raw test data. The validity check programs provide specifications to prescribe the constructed test.

3. System Operation

3.1 Registration Phase

The management program updates item and test data. There are three different registration methods. One is by using punched cards. This method is used to register a large amount of data, for example accumulated past data classified in a traditional hard-copy card system. The second method is by using a KB-CRT terminal. This is administered under interactive mode

*C & C Systems Research Laboratories, NEC Corporation

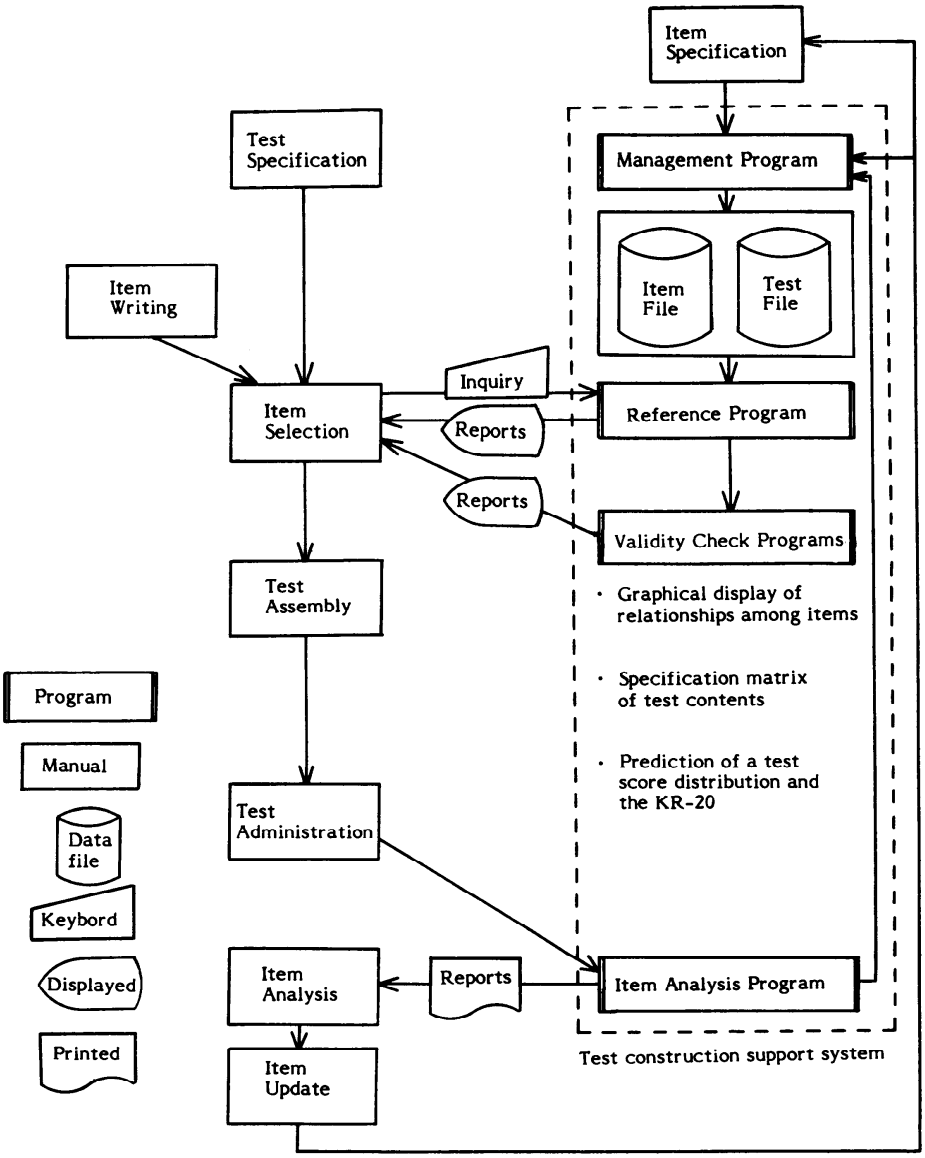

Fig. 1 Test construction flow diagram.

with the operator and is mainly used when a few items are updated. The third method is the direct registration method which is used to register statistical data with the analysis program.

3.2 Reference Phase

The reference program is used to refer to and display items/tests in the data file, and to call validity check programs. This is operated using several reference commands. These commands are simple conversational language and are easily used by a test author. In Table 3, some main commands are listed. The author can effectively build up the test using these commands and some operational data files. The relational flow between commands and operational data files is shown in Fig. 2.

An example of item reference procedure appears in Fig. 3. The underlined part in the following CRT display field shows keyboard input by a test author.

4. System Characteristic

4.1 Graphical Display of Relationships among Items

The relationships among items are hierarchically

Table 1 Data files.

Data file	Description		Created by
Item file	Fixed data	contains item characteristic data: item number, item content area number, content area within a curriculum, behavioral objectives, item type, grade level, key words, etc. contains relational data among items: relational item number and item relational index.	·Management program
	Accumulated data	contains item statistic data: error rate, discrimination index, number of subjects, used frequency, etc.	·Management program ·Item analysis Program
Test file	Fixed data	contains test characteristic data: test number, teacher/author name, test aim, etc.	·Management program
	Accumulated data	contains test performance data: applied date and location, test score distribution, other statistical indices, etc.	·Management program ·Item analysis Program

Table 2 Programs.

Program	Description
Management program	updates the item and test file.
Reference program	performs item and test reference.
Item analysis program	performs item and test analysis. subordinate to Management program.
Validity check programs	provide several kinds of check reports which suggest the test constructing process. subordinate to Reference program.

Table 3 Command list.

Command	Function
RETrieve	refers to a data file and retrieves items/tests which meet the indicated conditions.
DISplay	displays the content of a data file designated by the RET command.
RELationship	displays the relationships among items using registered relations between items.
SAVe	saves data from the current file to save file.
CALI	calls validity check programs.

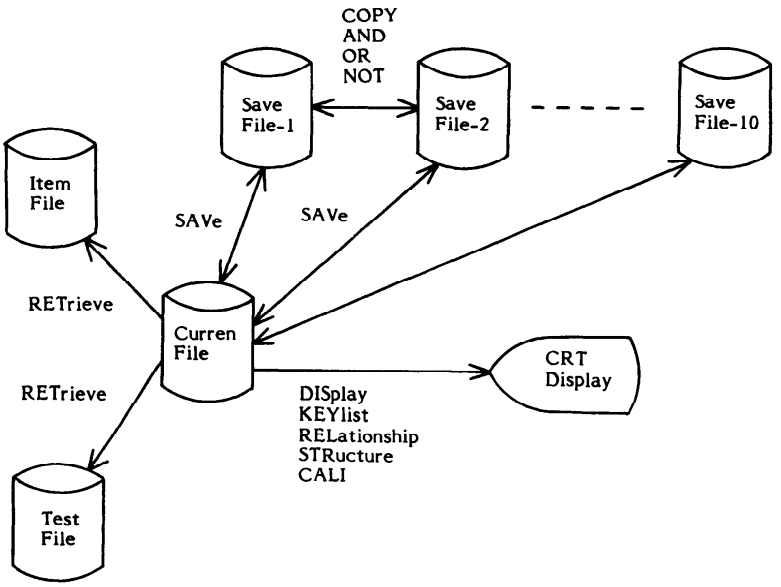

Fig. 2 Commands and operational data files.

displayed to examine content validity of the constructed test. Direct relationships between items, such as implication, prerequisite and subordination are expressed as a digraph. An example of a graphical display of relationships among items is shown in Fig. 4, where numeric numbers correspond to item identification numbers.

4.2 Content Validity Examination

In order to examine content validity of the constructed test, some kinds of check sheets for content analysis of the test are provided for the test author. These check sheets show to what kind of behavioral objectives and content area the items selected in the constructed test belong. The test author exchanges some items for adequate items by examining these prescriptions so that the test may have higher content validity. An example of check sheets for content validity examination is presented in Table 4. In Table 4, a specification matrix of distribution of items between content area and behavioral objectives for the constructed test is represented. This shows that the number of items is five whose content area is addition and whose behavioral objectives is computation. Balance of content area and behavioral objectives is one important aspect of quality in a test.

CRT Display field	Remarks
COMMAND: RET	A test author chooses the RETrieve command, refers
FILE: IF	to the Item File, and retrieves the items whose content
CONDITION: CONT='DIVISION' AND DISC > 0.75	area is 'DIVISION' and whose DISCrimination index is greater than 0.75.
"FOUND 45 ITEMS"	The system finds 45 pertinent items which meet the condition. (These 45 items are registered in the current file.)
COMMAND: DIS	He chooses the DISplay command and the Current
FILE: CF	File.
CONDITION: I.NO,CONT,BEHA,DISC.	He orders the list of item numbers, contents, behavioral objectives and discrimination indices.

** ITEM LIST **

ITEM NO.	CONTENT	BEHAVIOR OBJE.	DISC.
102513	DIVISION	COMPUTATION	0.82
102514	DIVISION	COMPUTATION	0.77
103600	DIVISION	KNOWLEDGE	0.89

The system represents the item list.

Fig. 3 An example of item reference procedure.

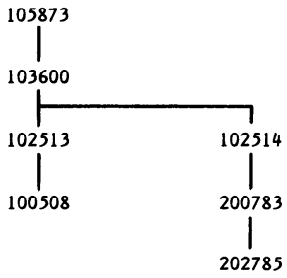

Fig. 4 Graphical display of relationships among items.

Table 4 Specification matrix.

CONTENT	BEHAVIOR: COMPUTATION	KNOWLEDGE	UNDERSTANDING	APPLICATION OF UNDERSTANDING	COMPREHENSION	TOTAL
ADDITION	5		2			7
SUBTRACTION	3		2			5
MULTIPLICATION	2			5		7
DIVISION	2			5		7
FRACTION	2	3	4		5	14
TOTAL	14	3	8	10	5	40

4.3 Statistical Reliability Examination

In order to examine statistical reliability of the test, the mean and the variance of the tests scores and the KR-20, as a coefficient of reliability, are predicted prior to test administration. Moreover, the test score distribution is predicted by applying the beta binominal model (Keats and Lord, 1962). The beta binominal model has only two parameters. These parameters can be easily estimated by using the difficulty level of each item plus one parameter. Moreover this model is applicable to

various categories of test score distribution curves. The estimation method is as follows. The value

$$\bar{P}_i - \bar{P}_{ij} - \frac{n}{n-1}\sigma^2$$

is constant according to the author's experience, under the following conditions (Sato, Kurata and Ikeda, 1978). Where $\bar{P}_i$ is the mean of observed right answer ratio, $\bar{P}_{ij}$ is the mean of simultaneous right answer ratio, σ^2 is the variance in observed right answer ratio and n is the number of items in the test. The condition is that the students are under control and the items are homogeneous. For example, the same teacher administers the same kind of test in a classroom every year. If the preceding value is constant, the student test score distribution is estimated only by the given item difficulty using the model. A predicted score distribution is represented as an S-P chart in Fig. 5 (See the appendix). Two curves on an S-P chart correspond to the cumulative item difficulty distribution (for "P-curve") and the cumulative testee's score distribution (for "S-curve"). An example is shown in Fig. 6. In this figure the solid curve shows the predicted score distribution and the dotted curve shows the observed score distribution respectively. This figure suggests that the predicted distribution agrees closely with the observed one. The above predictions are effectively used to examine statistical reliability of a constructed test.

5. System Utilization

A high school teacher constructed a test item bank. He prepared about 500 items for high school level physics, such as statics, optics, electricity. All items were classified into content area and their behavioral objectives, which are classified into five phases, such as comprehension, knowledge, calculation, graph drawing

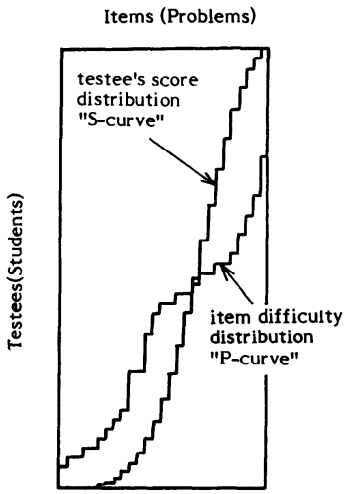

Fig. 5 Predicted distribution result representation.

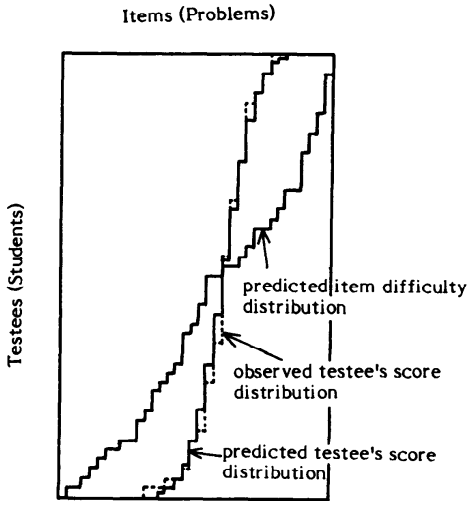

Fig. 6 Predicted distribution and observed distribution example.

and applied question. He administered five different test to total of 1000 students a year. Each test was made up of about 30 test items. Item and test statistic data were stored after every test administration.

The usage of this system is described in more detail. Before a test author assembles the test, his test purpose has been determined. For example, the test author should consider that the test score distribution curve has two peaks if he wants to divide a group of students into two groups according to the results of a test. To the contrary, uniform distribution is desirable if he wants to determine the student ability gradation sequence. A test author selects a candidate of items referring to item contents. It is necessary to determine whether or not the constructed test based on selected items is appropriate to his test purpose. He can check content validity of the constructed test examining the check sheets described in section 4.2. If the constructed test is suitable judging from balance of content area and behavioral objectives, it is examined whether statistical reliability of the constructed test is high or not. In order to examine statistical reliability of the constructed test, the mean, the variance and the test score distribution are predicted prior to test administration. If the test author is satisfied with the prediction, he administers the test. On the contrary, if he is not satisfied with the prediction of statistical reliability, he can update several items and reform the test until he is satisfied with the prediction results. After test administration the item analysis program represents statistical reports including the item analysis informative reports and the list of test items which should be improved. The test author updates the contents of the data files through test results according to his own judgement.

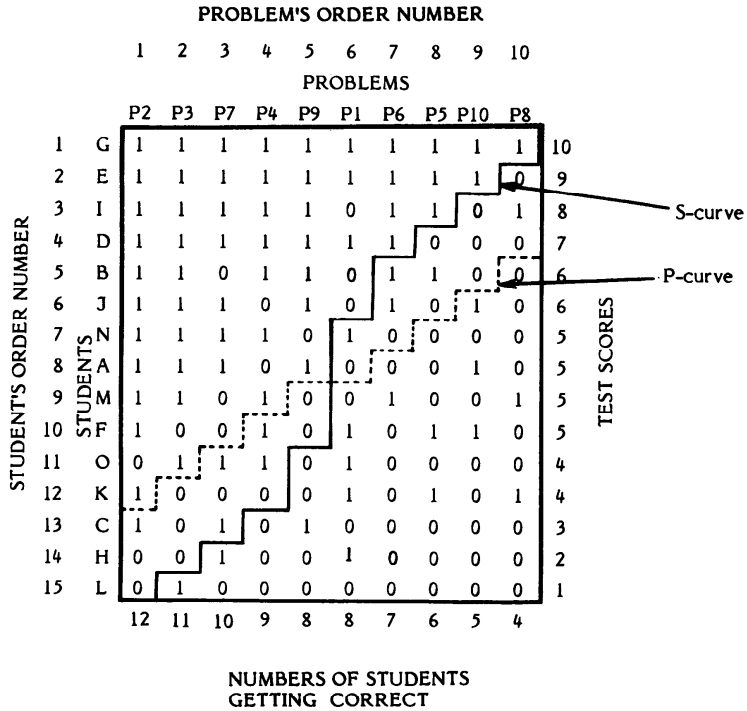

Fig. 7 An example of S-P chart.
The step graph in dotted lines is the P-curve; that in solid lines is the S-curve.

6. Conclusion

In this paper, a computer-assisted test construction support system of educational and psychological test based on a test item data base was presented. This system is written in COBOL and can be administered in a relatively small computer. At present, this system has been employed in Japan Center for Examination Research.

In order to employ the item data base efficiently, it is important to evaluate and update item content continually. That is, a test item data base system can be regarded as a quality control system of testing.

When the contents of test items are stored in a computer data base with characteristic and statistic data, a test paper which contains the selected items is edited and printed out for use.

References

1. KEATS, J. A. and LORD, F. M. A Theoretical Distribution for Mental Test Scores, *Psychometrica* **27**, (Mar. 1962), 59-72.
2. KURATA, M. and SATO, T. Test Construction System Applying Item Statistics, *Proceedings of the International Conference on Cybernetics and Society*, (1978), 368-372.
3. KURATA, M., CHIMURA, H., TAKEYA, M. and SATO, T. TEST ITEM DATA BASE SYSTEM, *Proceeding of IFIP TC-3 3rd world conference on computers in education*, (1981), 823-824.
4. LIPPEY, G., Ed. *Computer-Assisted Test Construction*, Educational Technology Pub., New Jersey, 1974.
5. SATO, T. A Classroom Information System for Teachers; with focus on Instructional Data Collection and Analysis, *Proceedings of ACM'74*, (1974), 199-206.
6. SATO, T. *S-P Table Analysis; Analysis and Interpretation of Test Scores*, Meiji-Tosho Publishing Co., Tokyo, 1975 (Japanese).
7. SATO, T., et al. A Study of Item Banking System (II), Paper of technical group on Educational Technology IECE Japan ET 76-6, (1976), 49-53 (Japanese).
8. SATO, T. and KURATA, M. Basic S-P Score Table Characteristics, *NEC RESEARCH & DEVELOPMENT*, No. 47 (Oct. 1977), 64-71.
9. SATO, T., KURATA, M. and IKEDA, H. Estimation of Statistical Characteristics of Educational Tests, Paper of technical group on Educational Technology IECE Japan ET78-2, (1978), 27-30 (Japanese).
10. TATSUOKA, M. M. Recent Psychometric Developments in Japan: Engineers Grapple with Educational Measurement Problems, paper presented at the ONR Contractor's Meeting on Individualized Measurement, Columbia, Missouri, 1978.

Appendix

An S-P (Student-Problem) chart is essentially a binary data matrix with 1's (for "correct") and 0's (for "incorrect"), in which the students (represented in rows) have been arranged from top to bottom in descending order of their total test scores, and the items (represented in columns) have been arranged from left to right in ascending order of difficulty (Sato, 1975; see Tatsuoka, 1978, for a description in English). An example of the S-P chart is given in Fig. 7.

The two-step-function graphs superimposed on the chart are constructed as follows. A vertical line segment is drawn across each row at the right-hand edge of the column whose number (i.e., the problem's order number shown in the top margin) corresponds to the total test scores -shown in the right margin- earned by the student represented by that row. Thus, starting from the lower left corner of the chart, vertical line segment are drawn at the right-hand edges of every column. The length of each line segment (measured in row-widths) equals the number of students who earned that particular score. After the vertical line segments have all been drawn, the top endpoint of each is connected to the bottom endpoint of the one to its right, thus completing the staircase-like appearance. The step-function graph is called the S-curve (for "student curve"), shown as the solid line graph in Fig. 7.

The graph shows the step-function ogive of the cumulative distribution of test scores for the group of testees.

In the same way, a horizontal line segment is now drawn across each column at lower edge of the row whose number (i.e., the student's order number shown in the left margin) corresponds to the number of students -shown in the bottom margin- who correctly answered the problem represented by that column. The length of each line segment (measured in column-widths) is equal to the number of items that were correctly answered by the number of students indicated by the row number. This step-function graph is called the P-curve (for "problem curve") shown as the dotted line graph in Fig. 7.

(Received December 8, 1983; revised April 13, 1984)