DOCUMENT RESUME

ED 111 692 SE 019 730

AUTHOR Anderson, Edwin R.
TITLE The Certainty of Information in Instructional Decision Making. Report No. 76-4.
INSTITUTION Washington Univ., Seattle. Educational Assessment Center.
PUB DATE Aug 75
NOTE 25p.; Educational Assessment Center Project 503.

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS College Mathematics; *Computer Science Education; Evaluation; Higher Education; Item Analysis; *Mathematics Education; *Research; *Testing; Test Reliability; *Test Results

ABSTRACT

This study measures the stability of performance exhibited where different computer programing classes study the same material. By focusing standard measurement techniques on the item difficulty (the proportion of students answering an item correctly), it was determined that up to two-thirds of the reliable variance of a classroom test is held in common with identical tests given in similar classes. The particular wording of the test item measuring a concept was shown to be a critical factor in knowledge assessment. Classes were given identical terms measuring common concepts and changed items measuring a different set of common concepts. The correlations between classes of item difficulties for identical items is approximately .70, whereas the correlation for changed items is approximately .35. Suggestions are made for utilizing the high correlation between identical items in instructional decision making. (Author/SD)

Educational Assessment Center

University of Washington

August 1975

76-4

The Certainty of Information in Instructionsl Decision Making[1]

Edwin R. Anderson

Abstract

This study measures the stability of performance exhibited where different classes learn the same material. By focusing standard measurement techniques on the item difficulties, i.e. the proportion of students answering an item correctly, of items common to several classrooms, it was determined that up to two-thirds of the reliable variance of a classroom test is held in common with identical tests given in similar classes. The particular wording of the test item measuring a concept was shown to be a critical factor in knowledge assessment. Classes were given identical terms measuring common concepts and changed items measuring a different set of common concepts. The correlations between classes of item difficulties for identical items is approximately .70 whereas the correlation for changed items is approximately .35. Suggestions are made for utilizing the high correlation between identical items in instructional decision making.

Educational Assessment Center Project: 503

The Certainty of Information in Instructional Decision Making

Effective decisions are based on the ability to predict the outcomes of future events with some degree of success. The decision maker is happiest when he can predict future events with total certainty, but, in the absence of such good fortune, he will look for the best statistical advantage allowed by his available information. For example, the registrar's office of most universities makes use of the positive relationship ($r \approx .5$) between high school grades and college grades to accept a sample of applicants who will have the best prognosis for college success. The purpose of this paper is to assess the certainty of information available to the teacher within his own classroom for his instructional decision making.

Rosenshine (1970) reviewed studies examining the consistency of teacher effects in classroom or classroom-like situations and found only nine studies which attempted to make such a consistency check. The results of these studies were disappointing in that when student achievement was the dependent variable very little consistency of effect was demonstrated. These studies tested many classes taught by many teachers (24 to 106) with a standardized test and correlated mean student achievement for a given teacher's class with the same mean in the same class taught at a later time. Thirteen correlations obtained in five long term studies of this type ranged from -.08 to .53 with a mean of .28. This approach to assessing classroom data stability has two major disadvantages: (a) it requires large numbers of teachers and students and (b) it does not provide the individual teacher with the detailed information needed for instruction improvement decisions. The remaining studies reviewed assessed consistency in teacher effects for short (30 min.) lectures. Positive results were shown but again the magnitude of consistency was not great.

What may prove to be a better approach to assessing information certainty in the classroom is suggested by research involving paired-associate (PA) learning. Coleman (1970) reviewed performance data collected from children given reading exercises in PA format. The words the children were learning to read were rank ordered on the dimension of item difficulty. The rank orders from two or more experiments using the same words were then correlated; 31 of the correlations reported fell between .69 and .98 while the remaining two were .33 and .31. More

recently researchers (Atkinson, 1972; Atkinson & Paulson, 1972; Laubsch, 1969) have successfully used item difficulties gathered from one group of subjects to provide the basis for decisions about which PA item to present next in sequence of instruction experiments. The PA experiments suggest that consistency of effect in the classroom might be better demonstrated through the use of item difficulties computed for tests common to several classes.

Item difficulty is a notion quite familiar to educational test and measurement specialists. However, the concern of educational measurement has in general been with the reliable assessment of the individual student's knowledge. This translates into estimation of how accurately the student's total score on a test reflects the state of his knowledge. Answers to single items are not particularly reliable estimates of a single examinee's knowledge and so individual item statistics are used in constructing the best possible overall test. Suppose instead that this emphasis were changed to regard the item difficulty, defined as the proportion of students correctly answering a test item, as the statistic of major interest. If instruction is delivered under close to constant conditions and if the same test items are used with successive classes, the product moment correlation between two classes on an item by item pairing should be quite high. This correlation of item difficulties can be used as a means of assessing the stability of instruction efforts in the classroom.

One goal of education, broadly defined, is the development of a state within a person called the learner which is similar to an internal state within a person called the knower. When the learner is in this state, he is said to "understand". The state of understanding is inferred from behavior in relation to a context, i.e. a person who emits situationally appropriate behavior may be said to understand the situation. May be is underlined in the previous sentence because understanding is not inferred from any "particular behavior" (Deese, 1969). Deese writes, "The criteria for understanding are in the potential for an indefinite number of appropriate reactions, some linguistic and some not."

In writing a test item to probe the student's ability to react appropriately, the teacher is constructing one test question from an indefinite set of test questions. Ideally, the student should give the correct response to any member of the indefinite question set if he understands the concept being probed or he should give all incorrect responses if he does not. In practice, we would expect the particular wording of a multiple choice test question to affect the estimate of

the student's competence for at least two reasons: (a) Concepts stored in memory must be retrieved from storage and changed wording of a question could conceivably change the ease of access to the concept needed to answer the question. (b) Changed wording of response alternatives could affect the difficulty of the discriminations needed to identify the correct alternative. One way to assess the impact of specific wordings is by giving two classes learning the same subject matter identical items and items measuring the same concept with changed wording and then correlating the resulting data. If the understanding of the students is key, the correlations of both types of item difficulties (identical items and changed items) will be the same. If item wording is a major factor, the correlation for changed items will be lower than the correlation for identical items.

In addition to correlations between identical test items and between changed items, the data collected from the classrooms described in the methods section allow a number of intentional and natural experimental comparisons. Some of the classes were taught using a workbook specially prepared for the class while others were not (intentional). In one case, the textbook, which was common to all the classes, was changed. Many of the teachers involved in these classes lectured during class periods while others used the classroom primarily for testing and assisting students with problems. In some of the classes, students were given multiple choice test items written by the same professor who wrote the multiple-choice items of a common final examination while in other classes the students were given essay and problem quizzes designed by a different instructor before receiving the common multiple-choice final. Data on these comparisons are included with the correlation data in the results section.

Methods

Description of classes and subjects. At the University of Washington the introductory FORTRAN IV computer programming classes are handled by the general engineering department. Engineering 141, as the course is labeled, has 10 to 12 sections each quarter with between 15 and 30 students in each section. The sections usually have roughly equal percentages of upper and lower classmen. The students are drawn from the general university population, but there does tend to be a larger number of engineering students in each section than would be expected from a random sample of the student body. The course is a four-credit

course which normally meets for four, one-hour periods per week, but on occasion it meets for two, two-hour sessions. Students in all of the sections are given access to the University CDC 6400 computer in order to test and run their practice programs.

Course reading materials. All classes involved in the data collection from Autumn Quarter, 1974, and Winter Quarter, 1975, used a common textbook, Fortran IV Programming by Rule, Finkinaur, and Patrick (1973). Data was collected from a single course in the Spring of 1975 and that class used a different text, Fundamental of Fortran Programming by Nickerson (1975). In addition to the textbook, three Autumn classes and the one Spring class used a workbook prepared locally by Professor W. Dunn of the Civil Engineering department. The workbook has 13 sections corresponding to topics in Fortran programming, e.g. DO loops and subscripted variables. Each section has two types of problems, short answer essay questions and multiple-choice questions, and in addition, many of the sections have matching exercises. Answers are included for all of the questions.

Test items. Three classes from Autumn Quarter and three classes from Winter Quarter were given weekly quizzes (13 to 30 items) from the second through the ninth week of the quarter. The quizzes given to the three classes during the same week tested the same concepts, sometimes with identical multiple-choice items and sometimes with changed, multiple-choice items. An item was considered identical if the wording of the question stem remained unchanged between two classes and if the wording of the four response alternatives was unchanged; reordering of the response alternatives was allowed under the identical condition. Changed items had at least one word changed in the question stem, the response alternatives, or in both stem and alternatives. Problems having the same words but new numbers were considered changed items.

The items from all of the weekly quizzes were written by Professor Dunn, as were the test items used for the final examinations. Five sections of Engineering 141 were given a common, 44-item final examination at the end of the Autumn Quarter and eight sections were given a common, 54-item final at the end of the Winter Quarter. All tests were machine scored at the University of Washington Educational Assessment Center. The computer printout of the scoring includes an item by item analysis which gives the proportion of students making the correct response to an item.

The multiple-choice questions of the workbook were a parallel form of the weekly quizzes. The same concepts were tested on weekly quizzes as were covered by the workbook quiz with items which were in the majority of cases (55%) identical to those of the workbook. Except for a small number of items included in the Spring Quarter final examination, none of the items from the final examinations were identical to items given during the quarter.

Teaching methods, Autumn. Three classes during the Autumn Quarter used the same textbook, the same workbook, and parallel forms of the weekly quizzes. All three of these classes were taught using a semi-mastery instruction method which allowed each student scoring below 90% on the weekly quiz the first time it was given to retake a parallel form of the quiz. The student was allowed to study his first test results to determine his errors before taking the second quiz; all students were scheduled for the first and second testing sessions during a week at the same time. Mastery instruction typically allows self-pacing, hence, the use of the term "semi-mastery" in describing the method. Class time was used to handle details of course administration and to answer student questions on an individual basis. Very little lecturing was done in these classes. The two additional classes given the common final in the Autumn Quarter were taught more traditionally with lectures during class and single try test sessions.

Teaching methods, Winter. During the Winter Quarter three instructors were again compared on the weekly quizzes, two instructors used the semi-mastery method and their instructor adopted a lecture approach. This third instructor placed special emphasis on structured programming (Dijkstra, 1973) in the hope of improving the programming skills of his students. Five additional instructors used the final test; their instructional methods are best described as traditional lecture. Some of these classes were given weekly quizzes composed of programming problems designed by the instructor of the section.

Results

The primary data reported are the correlations of item difficulties among classes for identical items and the correlations among classes for changed items. The reader should bear in mind that the items contained in the identical set and the changed set are not the same for each correlation reported, e.g. the identical item set between class one and class two does not match the identical item set between class one and class three. In some cases items were discarded from the

tests by instructors in one or more of the sections because of dissatisfaction with the items; all item discards were made before the tests were scored. Each correlation reported is followed by the number of items included in the correlation, e.g. .76 (33). Note that the number in parenthesis is the number of items included in the comparison and is not the number of subjects used in computing the item difficulties. The number of subjects used to determine item difficulty is always between 15 and 30.

The measurement theorist usually begins from a two-dimensional data matrix in which one dimension is a listing of the individual subjects and the second dimension is a listing of the test items. Each subject-item cell in the matrix is filled with a one if that subject responded to that item correctly and with a zero if an incorrect response occurred. The formulas derived from test theory for the manipulation of this data matrix are designed to estimate the reliability of the test in measuring the student's knowledge. Throughout the results section there is a shift from this perspective. In the standard approach the test items are seen as measuring the student; in the analysis performed here the students as a class are seen as measuring the difficulty of the test items. The same data matrix is used in the shifted perspective, but the formulas used in computation with the data are analogs of the standard formulas. For example, coefficient alpha, in the case of dichotomous items, takes the following form (Nunnally, 1967).

$$r_{kk} = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} p_j q_j}{\sigma_S^2}\right)$$

Where p is the proportion of students getting an item correct, q is the proportion getting the same item incorrect, k is the number of items, and σ_S^2 is the variance of the subjects' total scores. Coefficient alpha is computed as follows under the changed perspective.

$$r_{nn} = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} c_j e_j}{\sigma_i^2}\right) \quad (1)$$

Where n is the number of subjects, c is the proportion of the items a subject correctly answers, e is the proportion of the items the same subject answers incorrectly (e = 1 - c), and 6_i^2 is the variance of the total scores of the items. The second form of coefficient alpha is a measure of the reliability of the item difficulty estimates within a single class.

The item difficulty correlations obtained during the Autumn Quarter among the three semi-mastery instruction sessions are shown in Table 1. The mean correlations from Table 1 are .65 for identical items and .33 for changed items. The mean difficulty of the items, the standard deviations of the items, and the dependent t test values between the classes compared are shown in Table 2. Three t tests are reported instead of one analysis of variance because the item sets vary from comparison to comparison. Note that the mean item difficulty from the test items tends to be high (approximately 85%). The range of item difficulties is restricted and the correlations reported in Table 1 may underestimate the magnitude of relationship that actually exists between classes (Minium, 1970, p. 190). Data from the final examination given to five sections of the programming class in the Autumn Quarter is shown in Table 3. The mean correlation from Table 3 is .73. See Table 4 for the mean item difficulties and standard deviations of the five classes. An items X classes repeated measures analysis of variance done for the 39 items of the final examination that all classes answered shows significant variability among the classes ($F_{4,152} = 11.4$, $p < .001$). Orthogonal contrasts show the mean item difficulty of class three to be significantly greater than the mean item difficulty of class two ($F_{1,152} = 5.27$; $p < .05$). Class five and class four also show a significant difference ($F_{1,152} = 9.95$, $p < .01$). Any interpretation of the significant orthogonal contrasts in terms of instruction received is confounded by the facts that class three contained 80 percent upperclassmen whereas the normal class contains approximately 50 percent upperclassmen and that class five was told in advance that scores on the final examination would not be included in calculations of their course grade.

During the winter quarter, comparisons were made among two semi-mastery courses and a third course which emphasized the structured approach to program writing (See tables 5 and 6 for the data from these comparisons). The average correlation among the three classes is .70 for identical items and .39 for changed items. The correlations from the winter quarter replicate the autumn quarter results. Table 7 shows the intercorrelations of eight classes taking

Table 1

Item difficulty correlations of three semi-mastery courses for identical and changed items

Class Number	Class Number 1	2	3
Identical Items			
1	-	.68(78)[a]	.70(47)
2		-	.57(47)
3			-
Changed items			
1	-	.44(45)	.34(28)
2		-	.21(28)
3			-

[a]The number in parenthesis is the number of test items used in calculating the correlation.

Table 2

Mean item difficulties, standard deviations, and t test values for comparisons among three semi-mastery classes

Classes Compared	Statistics				
	First Mean	Second Mean	Sd1	Sd2	t
Identical Items					
1 & 2	86.09	85.55	12.16	11.65	.50
1 & 3	84.19	86.83	12.99	14.71	-1.66
2 & 3	85.70	86.83	12.32	14.71	-.61
Changed Items					
1 & 2	77.11	80.07	16.53	14.43	-1.20
1 & 3	79.46	90.14	15.73	7.78	-3.77*
2 & 3	83.04	90.14	13.62	7.78	-2.64*

*$p < .01$

Table 3

Item difficulty correlations from the common, autumn quarter final examination

Class Number	Class Number				
	1	2	3	4	5
1	-	.76(44)[a]	.86(41)	.79(42)	.68(44)
2		-	.73(41)	.71(42)	.64(44)
3			-	.71(39)	.65(41)
4				-	.77(42)
5					-

Note. Classes one, two, and three in this table are the same as classes one, two, and three of table 1.

[a]The number in parenthesis is the number of test items used in calculating the correlation.

Table 4

Final examination mean item difficulties and standard deviations for five autumn classes

Class Number	Statistics	
	Mean	Standard Deviation
1	67.86	23.08
2	70.09	19.79
3	75.61	19.39
4	68.29	24.02
5	59.89	24.68

Note. Classes one, two, and three in this table are the same as classes one, two, and three of table 1.

an identical, 54-item final examination in the winter quarter. The mean correlation from table 7 is .71; this value is very close to the value (.73) obtained for the fall classes. An items X classes repeated measures analysis of variance done with seven of the classes showed significant variability among the class means ($F_{6,307} = 3.35$, $p < .01$). Item difficulties were not present for 11 of the cells in the data matrix; their values were determined using a missing data estimation procedure recommended by Myers (1966, p. 171). Class eight, the structured programming section, had 13 missing item difficulties. Since the mean of this section was near the grand mean of all sections, the decision to exclude this section from the analysis because of the missing data probably produces a slight inflation of the F statistic (the between means variance estimate is high). An orthogonal contrast of the semi-mastery instruction sections one (mean item difficulty = 70.13) and three (mean item difficulty = 63.58) shows significant difference ($F_{1,307} = 7.75$, $p < .01$) as does a contrast of the high and low traditionally taught sections ($F_{1,307} = 6.76$, $p < .01$). An orthogonal comparison of semi-mastery classes and traditional classes shows no significant difference associated with the type of class ($F_{1,307} = 1.93$, $p < .10$). Table 8 shows the means and standard deviations of the eight classes taking the winter final examinations.

The correlations between different sections of engineering 141 should be compared with the values of coefficient alpha for the sections (See equation 1). These coefficients indicate the reliability of item difficulty within each section and represent the maximum correlation that could be expected between sections. Table 9 presents coefficient alpha for the eight winter quarter classes. The average correlation (.71) from the intercorrelation matrix should be compared with the average value of coefficient alpha ($\bar{r}_{nn} = .86$) instead of with the maximum possible product moment correlation, i.e. 1. The square of $\bar{r}_{nn}$ when the square is multiplied by 100 is an estimate of the percent of the total variance within the classes that is reliably measured by the test instruments. The reliably measured variance accounts for 74% of the total variance whereas the variance common to the classes is approximately 50% of the total variance. A combination of these two figures suggests that up to two-thirds of the reliable variance from the measuring instruments is common to the eight classes.

One instructor was followed through the autumn, winter, and spring quarters.

Table 5

Correlations of item difficulties among two semi-mastery(SM) and one structured programming(SP) classes

Class	Class		
	SM1	SM2	SP
Identical items			
SM1	-	.77(74)[a]	.74(65)
SM2		-	.59(64)
SP			-
Changed items			
SM1	-	.47(29)	.34(32)
SM2		-	.37(26)
SP			-

[a]The number in parenthesis is the number of test items used in calculating the correlation.

Table 6

Mean item difficulties, standard deviations, and dependent t test values for comparisons among two semi-mastery(SM) and one structured programming(SP) classes

Classes Compared	Statistics				
	First Mean	Second Mean	Sd1	Sd2	t
Identical items					
SM1 & SM2	81.97	72.78	15.72	19.30	6.34**
SM1 & SP	82.18	77.60	15.94	21.09	2.61*
SM2 & SP	74.69	79.63	18.83	18.64	-2.34*
Changed items					
SM1 & SM2	81.57	73.97	16.66	21.91	2.02
SM1 & SP	80.13	77.25	17.59	19.57	.76
SM2 & SP	74.88	73.69	19.69	20.78	.27

*p < .05

**p < .01

Table 7

Correlations of item difficulties from a Winter Quarter final examination given to eight classes

Class Number	Class Number							
	1	2	3	4	5	6	7	8
1	-	.72(54)[a]	.68(48)	.70(52)	.81(54)	.78(51)	.76(54)	.73(41)
2		-	.69(48)	.71(52)	.78(54)	.77(51)	.79(54)	.68(41)
3			-	.66(47)	.73(48)	.77(47)	.73(48)	.60(39)
4				-	.66(52)	.64(49)	.57(52)	.70(40)
5					-	.86(51)	.83(54)	.61(41)
6						-	.81(51)	.61(41)
7							-	.55(41)
8								-

Note: Classes one, three, and eight are identical to classes SM1, SM2, and SP in Table 5.

[a]The number in parenthesis is the number of test items used in calculating the correlation.

Table 8

Final examination mean item difficulties and standard deviations for eight winter classes

Class Number	Statistics	
	Mean	Standard Deviation
1	70.13	20.30
2	68.33	21.99
3	63.58	22.05
4	66.25	24.69
5	66.48	23.67
6	61.96	22.35
7	62.28	26.93
8	64.80	24.63

Note: Classes one, three, and eight are identical to classes SM1, SM2, and SP in Table 5.

Table 9

Coefficient alpha for eight Winter Quarter classes

Class Number	Coefficient Alpha	Number of Students
1	.78	19
2	.83	21
3	.86	25
4	.86	20
5.	.88	24
6	.86	25
7	.92	22
8	.89	23

He used the semi-mastery methods of instruction each quarter, but varied the written materials given to the students. Written material in the fall included the Rule, et. al., 1973, text and the workbook, in the Winter Quarter the workbook was removed, and in the Spring Quarter the workbook was reintroduced along with a change in textbook (Nickerson, 1975). The fall-winter correlation for identical items is .61(96) and for changed items is .71(49). The comparison for winter-spring are .24(54) and .09(59). The low correlations for winter-spring are due primarily to the extremely high spring test scores and their consequent lack of variability. A dependent t test comparing autumn and winter results showed no significant difference between the identical item means (autumn mean = 82.58 and winter mean 81.40; $t_{94} = .82$) and a similar test showed no significant difference between the changed item means (autumn mean = 76.96 and winter mean = 75.92; $t_{47} = .47$). These same comparisons were significant between the winter and spring quarters (winter identical = 85.40 and spring identical - 96.95; $t_{52} = -6.56$, $p < .01$; winter changed = 75.71 and spring changed = 85.06; $t_{57} = -2.99$, $p < .01$). Comparison of items common to the fall and winter final examinations shows a correlation of .89(18) and the same comparison for winter-spring shows a correlation of .45(37). For final examinations no significant difference was found between the fall mean (72.05) and the winter mean (68.00) ($t_{16} = 1.45$, $p < .10$) or between the winter mean (68.64) and spring mean (69.89) ($t_{35} = -0.34$). In short, even though the performance of spring quarter students differed from the performance of winter quarter students during the quarter, the changes made within the instrictor's classes did not affect the mean performance of students in different quarters on common final examination items.

Discussion

The high alpha values found in this study can be interpreted as indicating that within a class item difficulty is a very reliable measure. To put it in a more important way, if an item is relatively difficult for one student, it is likely to be difficult for other students. The high correlations resulting from pairing item difficulties from identical item sets clearly indicate a high degree of stability among the classes surveyed. The average correlations of .73 and .71 from the classes taking the autumn and winter final examinations mean that approximately 50% of the performance variance of one classes scores

can be predicted if an item difficulty analysis is available from another classes performance on those same items. Note that this statement holds true when many variables normally thought to influence instruction are ignored. The semi-mastery classes, in addition to a common teaching method, used common learning materials, i.e. same text, same tests during the quarter, same workbook (autumn only) (autumn only), yet the correlations between the semi-mastery classes are not different from the correlations among classes having the textbook as the only common reading source. The correlations of the semi-mastery classes with traditional classes is not different from the correlations among semi-mastery courses themselves. The students in all of these classes were faced with the problem of extracting information about computer programming from written or verbal statements, and they seem to have solved this problem in the same way or at least with the same degree of success in each of the classes. Teacher personality, method of instruction, classroom environment, and any other variables present but unmeasured and unrecognized did not substantially effect the learning of the students. The classes were either constant with respect to such variables and hence equally affected or the variables do not have a major effect on student performance.

Data gathered from item difficulties collected during the fall and winter quarters support the conclusion of high stability between classes when correlations of identical items are used to assess stability ($\bar{r}$ = .65 for autumn quarter and $\bar{r}$ = .70 for winter quarter). However, altering the wording of the test questions used to probe the same students' knowledge of programming concepts substantially lowers the correlation found between classes ($\bar{r}$ = .33 for autumn and $\bar{r}$ = .39 for winter). The assessment of the students' knowledge is related to the particular wording of the test question we use to probe that knowledge. On the other hand the positive correlation that remains after wording changes suggests that item difficulty measures of a common concept will show consistency when compared to the variability of estimates made for different concepts.

The data support the conclusion that treatments aimed at the entire set of concepts the student was to acquire were not effective. The semi-mastery-traditional instruction comparison, the workbook-no workbook comparison, and the structured programming-traditional programming comparison all failed to produce significant differences between classes on the final examinations. This finding is in agreement with a general tendency to find no results in such comparisons

(e.g. Dubin & Taveggia, 1968; Getzels & Jackson, 1963; Stevens, 1967; Wallen & Travers, 1963). The treatments used in classrooms generally do not alter the learning of the students in ways that are detectable in their performance.

The high correlations found between classes present an alternative to the approach of attempting to affect the learning of the entire set of concepts. Since we know a large number of test questions will be readily answered, why not focus the treatment where we know the students will have trouble answering questions? We might, for example, provide the student with a workbook which contains brief explanations and practice problems for concepts we know (from prior data collection) the student is likely to have trouble mastering. Problems related to readily learned concepts would be left out of the workbook entirely. Such a tactic may not change the student's learning strategy, but the selective application may influence the student's allocation of effort. What is being recommended here is the systematic selection of treatment focal points from objective data collection.

If we accept Deese's notion that understanding leads to appropriate behavior in response to an indefinite set of related situation, accept the high correlations of item difficulties from identical items given in different classes as an indication of the stability of the item difficulty measures, and accept the premise that low item difficulties indicate a misunderstanding on the part of several class members, we are led to some direct conclusions about instructional improvement. Ideal understanding of a concept would lead to an appropriate response on the part of all students to all items from the indefinite set for the concept. A low item difficulty on any item from the set indicates less than perfect understanding even though the remaining items from the set might be answered correctly. We are thus justified in modifying instruction on the basis of information collected from a single, specific test item. If an improvement is registered in a subsequent quarter on an item receiving focussed attention, we would then reword that item for the next teaching of the class to insure that other members of the indefinite set are also favorably affected. In other words an attempt is made through changes in the course materials to selectively shape student test performance but the possibility of shaping being confined to exact test item wording is avoided by changing item wording and reassessing the understandings drawn by the students in a subsequent version of the course.

The study reported here has several limitations which deserve attention. (a) The data was gathered from courses teaching Fortran programming; there is no

guarantee that the data will be duplicated in courses of a different type, e.g., social science courses. (b) The item difficulties were gathered with a single type of test question, i.e., multiple choice; there is no check made to determine if other testing modes will produce similar results. (c) There is a need to follow more instructors from quarter to quarter, particularly in view of the failure of within course data to replicate between course data (identical item $r = .61$ and changed item $r = .71$). (d) No satisfactory explanation is offered for the significant variability found within instructional methods. (e) And finally, the use of repeated measures item X classes analysis of variance assumes a random sampling from a normally distributed pool of item difficulties which in fact did not occur. This same criticism is, however, also true of many subjects X treatments analysis, particularly when students from a class are treated as randomly assigned to the class.

The stability of item difficulties from quarter to quarter and class to class opens new possibilities for educational research. Since test items can be transferred from class to class, class comparisons can be matched on an item by item basis to provide more sensitive comparisons via dependent t tests and repeated measures designs. Since difficult items can be reliably identified, selective strategies which specifically focus on difficult items can be attempted. Given the difficulty of establishing adequate control in classroom research, the potential of stable item difficulties for the production of more sensitive measurement is welcome.

Footnote

[1]The author wishes to thank Dr. Gerald Gillmore for his critique of an earlier draft of this paper. Special thanks also go to the engineering professors who had their classes participate in this study. Professor W. Dunn was largely responsible for gaining the cooperation of the engineering faculty as well as being responsible for the preparation of the written materials used in this study. Everyone who has occasion to do classroom evaluation research should be blessed with such a willing ally.

References

Atkinson, R. C. Optimizing the learning of a second language vocabulary. Journal of Experimental Psychology, 1972, 96, 124-129.

Atkinson, R. C. & Paulson, J. A. An approach to the psychology instruction. Psychological Bulletin, 1972, 78, 49-61.

Coleman, E. B. Collecting a data base for a reading technology. Journal of Educational Psychology Monograph, 1970, 61, 1-23.

Deese, J. Behavior and fact. American Psychologist, 1969, 24, 515-522.

Dijkstra, E. W. Notes on Structured Programming. In O. J. Dahl, E. W. Dijkstra, & C. A. R. Hoare. Structured Programming, New York: Academic Press, 1973, pp 1-81.

Dubin, R. & Taveggia, T. C. The Teaching-learning Paradox: A Comparative Analysis of College Teaching Methods. Eugene, Ore.: Center for the Advanced Study of Educational Administration, University of Oregon, 1968.

Getzels, J. W. & Jackson, P. W. The teachers personality and characteristics. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963, 506-582.

Laubsch, J. H. An Adaptive Teaching System for Optimal Item Allocation. Doctoral Dissertation: Stanford University, 1970.

Minium, E. W. Statistical Reasoning in Psychology and Education. New York: John Wiley, 1970.

Myers, J. L. Fundamental of Experimental Design. Boston: Allyn and Bacon, 1966.

Nickerson, R. C. Fundamentals of Fortran Programming. Cambridge; Winthrop Publishers, 1975.

Nunnally, J. C. Psychometric Theory. New York: McGraw Hill, 1967.

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40, 647-662.

Rule, W. P., Finkenaur, R. G., & Patrick, F. G. Fortran IV Programming. Boston: Prindle, Weber, & Schmidt, 1973.

Stephens, J. M. The Process of Schooling. New York: Holt, Rinehart, & Winston, 1967.

Wallen, N. E. & Travers, R. M. W. Analysis and investigation of teaching methods. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963, 448-505.