DOCUMENT RESUME

ED 221 556 TM 820 575

AUTHOR Miller, M. David; Burstein, Leigh
TITLE Multilevel Properties of Test Items: An Exploratory Study.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO CSE-R-181
PUB DATE 81
NOTE 58p.; Appendix A contains small print that may not reproduce clearly.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Achievement Tests; Elementary Secondary Education; *Item Analysis; *Norm Referenced Tests; *Regression (Statistics); *Research Design; Standardized Tests; Test Format; *Test Items
IDENTIFIERS Beginning Teacher Evaluation Study; International Evaluation Education Achievement; *Unit Of Analysis Problems

ABSTRACT

The multilevel characteristics of test item data are considered as a method for examining the characteristics of standardized norm-referenced tests. A theoretical rationale for examining multilevel characteristics is presented. It can be used as an aid to understand why program and instructional effects on measures constructed from individual-level psychometric data are weak, to improve instructional sensitivity and program relevance of tests, and to indicate what features of a test will increase the sensitivity of the test to instructional program variables. An empirical example form the International Evaluation of Education Achievement study and an analysis of one test from the Beginning Teacher Evaluation Study are examined to better understand how multilevel analysis can lead to more informed use of item data. It was found that between-student analysis fails to take into account the instructional context and its effect on student item response. This failure has two effects. First, the relationship between item response and other variables cannot be explained since it is a conglomerate of two different processes. Second, the between-student analysis may give a distorted view of whether an effect does or does not exist. (Author/PN)

MULTILEVEL PROPERTIES OF TEST ITEMS: AN EXPLORATORY STUDY

M. David Miller
Leigh Burstein

CSE Report No. 181
1981

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

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.

Multilevel Properties of Test Items: An Exploratory Study

M. David Miller and Leigh Burstein

Because of the belief that schooling affects student outcomes, the largely negative results from school effects studies and large scale evaluations of the relationship of school inputs to student outcomes have caused educational researchers to reexamine the statistical techniques and models which have traditionally been used to arrive at these conclusions. One methodological issue that has received much criticism has been the use of standardized norm-referenced achievement tests as the sole measure of educational outcomes.

Rather than abandon norm-referenced tests, an analysis of the tests may reveal ways to improve them. A possible method for examining the characteristics of standardized norm-referenced tests might be a multilevel examination of test items. Cronbach (1976) was the first to discuss the possible utility of multilevel item analysis:

> Once the question of units is raised, all empirical test construction and item-analysis procedures need to be reconsidered. Is it better to retain items that correlate across classes? Or items that correlate within classes? A correlation based on deviation scores within classes indicates whether students who comprehended one point better than most students also comprehended the second point better than most -- instruction being held constant. A correlation between classes indicates whether a class that learned one thing learned another, but this depends first and foremost on what teachers assigned and emphasized. It is the items that teachers give different weight to that have the greatest variance across classes. This (differential emphasis) leads us to regard

> the between-group and within-group correlations of items as conveying different information, and makes the overall correlation for classes pooled an uninterpretable blend. (Cronbach, 1976, pp. 9.19-9.20)

The effects that Cronbach cites need to be better understood. By considering the multilevel characteristics of test item data, test developers and users could potentially become better informed about test development, analysis, usage, interpretation, and reporting. For example, some test items may be more sensitive to background effects (e.g., prior knowledge or socioeconomic status), while other items may be more sensitive to instructional and program variables (e.g., time allocated per content area, time spent on high or low success tasks). By learning what variables an item is sensitive to, test developers will be better equipped to guard against unknowingly selecting items which are influenced by irrelevant characteristics of the environment in which the test is administered (irrelevant to the purposes for which the test is developed). Perhaps test constructors will also be able to better select items for a test which are more sensitive to the variable of interest (e.g., amount learned). At the least, multilevel analyses of test items will help test developers to better describe the statistical properties of the test and its items.

This report will be divided into four sections. In the first section, a theoretical rationale for examining the multilevel characteristics will be sketched. In the second section, an empirical example from the International Evaluation of Educational Achievement study (IEA) will be examined. Next, a preliminary analysis of one test from the Beginning Teacher Evaluation Study (BTES) will be presented. Finally, the potential utility of of multilevel item analysis and some possible directions for further research will be discussed.

Multilevel Analysis

The educational system is inherently multilevel. That is, schools are nested within districts; classes are nested within schools; and students are nested within classrooms. Data analysis can be conducted both between and within each of the various levels of the educational system. Furthermore, analysis between and within different levels can have different substantive meanings (Burstein, 1978; Burstein, Fischer, and Miller, 1979; Cronbach, 1976). Recognizing the importance of the choice of a unit of analysis, major evaluations, such as Follow Through (Haney, 1974) and the National Day Care Study (Singer and Goodrich, 1979), have considered this issue in some detail. Since education can affect students between and within all levels of the educational system, it has been argued that evaluations of educational data should look at more than one level of analysis for a more complete understanding of the determinants of student achievement.

Cronbach (1976) has argued that the "majority of studies of educational effects -- whether classroom experiments, or evaluations of programs or surveys -- have collected and analyzed data in ways that conceal more than they reveal. The established methods have generated false conclusions in many studies" (Cronbach, 1976, p. 1). Schooling effects studies have traditionally selected one unit of analysis, such as the individual or the school, and have used a between unit analysis. However, given the intact nature of educational data, single-level analyses are often inappropriate; the individual-level analysis can be decomposed into a between-group analysis and within-group analysis. It has been shown that the correlation of two variables at the individual-level is a weighted combination of the between-group correlation and the pooled within-group correlation (Knapp, 1977; Robinson, 1950):

$$\rho_{XY} = \eta_X \eta_Y \rho_{\bar{X}\bar{Y}} + \sqrt{1 - \eta_X^2}\,\sqrt{1 - \eta_Y^2}\,\rho_{(X-{}^{-}Y)(Y-\bar{Y})}$$

where ρ_{XY} is the correlation of X and Y across individuals; $\rho_{\bar{X}\bar{Y}}$ is the correlation of X and Y for the weighted group means; $\rho_{(X-\bar{X})(Y-\bar{Y})}$ is the correlation of the individuals deviations from their group means on X and Y; and η_X^2 and η_Y^2 are the proportion of variance in X and Y, respectively, that is attributable to group differences.

It is also true that the individual-level regression coefficient can be decomposed into a weighted combination of a between-group coefficient and a pooled within-group regression coefficient (Duncan, Cuzzort, and Duncan, 1961):

$$\beta_t = \beta_b \eta_X^2 + \beta_w (1 - \eta_X^2)$$

where β_t is calculated by regressing the individual level dependent measure (Y) on the individual level independent measure (X); β_b is calculated by regressing the weighted group means of the dependent measure ($\bar{Y}$) on the weighted group means of the independent measure ($\bar{X}$); β_w is calculated by regressing the dependent measure deviations from the group means ($Y - \bar{Y}$) on the independent measure deviations from the group means ($X - \bar{X}$); and η_X^2 is as defined above. As would be expected, when the influence of the group is weak, η_X^2 approaches zero and β_T approaches β_w. Conversely, when the differences on the independent measure are largely attributable to group differences, η_X^2 approaches 1.0 and β_T approaches β_b.

Often the decomposition of the student-level analysis in educational research has been ignored. This falure to take into account the multilevel properties of the data has often caused educational researchers to arrive at misleading conclusions about the effects of various determinants of educational achievement (Burstein, 1978; Burstein, Linn, and Capell, 1978; Burstein and Miller, 1978; Cronbach, 1976; Cronbach and Webb, 1975). It is possible that the examination of data from a multilevel perspective, which has too often been absent in other aspects of school effects studies and program evaluations, might also help us to better understand why program and instructional effects on measures constructed from individual-level psychometric data are weak. Perhaps a multilevel perspective applied to test development and interpretation will help to improve the instructional sensitivity and program relevance of tests. It is possible that the multilevel characteristics of item data will show what features of a test will increase the sensitivity of the test to instructional and program variables.

Item Analysis

In order to better understand the effects mentioned by Cronbach (1976) and what might be gained from a multilevel analysis of item data, it is important to be aware of classroom and background processes and how they effect differences in between-class and within-class achievement. Cronbach (1976) suggested that items that correlate highly across classes should be indicative of instructional and program effects. If some teachers emphasize a given content area, such as fractions, and others do not, one would expect high correlations across classes of items from a test measuring that given content area. On the other hand, if items correlate positively within classes, it indicates that students who do well on an item relative to the other members of the class will also do well on other items. This effect might be due to differences in students along such dimensions as ability or motivation.

The variance of an item can also be partitioned into two independent components -- between-class variation and within-class variation. The between-class variation of an item can be indicative of instructional and program variables. If teachers spend different amounts of time in a specific content area or they differ in their enthusiasm for that content area, there could be a net effect on the class which could increase the between-class variance of an item from that content area. Similarly, the within-class variance could be affected by instructional and program variables, but with a different substantive interpretation. While the between-class variability can be influenced by the net effect of classroom and instructional variables averaged across students, the within-class variability could represent differential sensitivity of students within a classroom to instructional and program variables. For example, students who are active participants in their learning might learn more from a given program than students who are passive learners. Additionally, within-class variability might represent differences in an instructional or program variable within the class. For example, teachers may spend more time with some students than others, or time on task may vary within the classroom.

Finally, the between-class and within-class components may also be affected by background variables. The between-class component may be due to differing community characteristics (e.g., socioeconomic status), such as the effects of differences across classes in the abilities and backgrounds of students. The within-class component may reflect the differing abilities of students within the class, differences in learning rates, or differences in the students' reactions to different instructional methods.

Empirical Examples

In order to better understand how multilevel analysis can lead to more informed use of item data, data from two sources will be examined. The first example involves a biology subtest from the International Association for the Evaluation of Educational Achievement (IEA) Six Subject Survey. These data were analyzed previously by McLarty (1979). The second example is drawn from the Beginning Teacher Evaluation Study (BTES).

IEA Biology Items

IEA collected data from 21 countries across six subject areas. Science was considered because it was a subject that was potentially less influenced by sources outside of the school environment. Information on the development of the science test items is available in Comber and Keeves (1973). Data on the results of the science test in the United States is also available in Wolf (1977).

In order to narrow the focus of the analysis, the data from Form B of the Biology subtest for Population II (14 year olds) in the United States were examined. The nine test items (numbered 2 through 10, as on the test) are contained in Appendix A. For data management and economic reasons, McLarty (1979) selected a random sample of schools (schools with less than 20 cases were eliminated first). The sample actually used for multilevel item analysis included 1210 students in 50 schools.

The descriptive statistics for the items are contained in Table 1. Since this test was developed using traditional psychometric techniques, the individual differences are maximized rather than the school differences (i.e., SD (within) > SD(between)). This has the likely effect of yielding small n^2. Note that the proportion of variation accounted for by the schools ranges from 6 percent on item 10 to only 10 percent on item 8.

The item intercorrelations are contained in Table 2. As McLarty (1979) points out, the low within-school correlations are probably due to the nature of the construct being measured. Biology covers a wide range of subjects including botany, zoology, and chemical processes involved in the life cycle (e.g., photosynthesis). Thus, it is conceivable that a student might learn the material necessary for one item and not another, depending on what area of biology the student is interested in. So that knowing one biology item is not necessarily related to knowing another. What relationship there is between items seems to be due to between-school differences. The school average on one item seems to be highly related to the school average on another item. These results are also confirmed in the point biserial correlations of Table 3.

In Tables 4 through 9, item responses were regressed on six variables. The following independent measures were used:

1. Student Sex - 1=male and 2=female;
2. Raw Word Knowledge - score on 40 item vocabulary test;
3. Liking of Biology - five point ascending scale of a student's rating of each school subject;
4. Books in the Home - 1=none, 2=1-10; 3=11-25, 4=26-50, and 5= 51 or more;
5. Hours of Biology Instruction per Week - 1=do not take, 2=less than 1 hour, 3=less than 3 hours, 4=less than 5 hours, and 5= more than 5 hours; and

6. Hours of Biology Homework per Week - 1=none, 2=less than 1 hour, 3=less than 3 hours, 4=less than 5 hours, and 5=more than 5 hours.

Each table contains three rows of regression coefficients corresponding to two regression equations:

$$Y = a + b_T X$$

$$Y = a^* + b_1 \bar{X} + b_2 X$$

The first row of regression coefficients (total) is derived from the first equation - the regression of student item scores (Y) on the student variables mentioned above (X). The remaining two rows come from the second equation - the regression of student item scores (Y) on the school means for the variable ($\bar{X}$) and the student level measure (X). The two coefficients b_1 and b_2 are interpreted as the between-school effect after controlling for the individual level measure and the within-school effect, respectively (see Alwin, 1976; Burstein, 1978; Firebaugh, 1978 for evidence on the interpretation).

The implications from Table 4 are fairly straightforward. For items 3, 4, 8, and 9, males will score higher than females in the same school. For item 5, the opposite effect was found (within the same school, females will score higher than males). Furthermore, for items 2, 4, 7, and 8, the between-school coefficients suggest that schools with a higher ratio of males to females will perform higher than schools with a lower ratio. These coefficients may represent sex bias at different levels. Scientists have traditionally been viewed as a male role. Possibly, this expectancy of different roles for males and females can be seen through differences in instruction. Classes with more males may receive more science instruction and encouragement. In addition, within the classroom, males may receive more help and encouragement from the teacher than their female classmates.

Much of the information from Table 4 is lost through examination of the between-student (Total) coefficients. First, when there was only a between-school difference, the between-student coefficients were not sensitive enough to find any differences (items 2 and 7). Secondly, in one case (item 10), the between-student analysis found an effect from the combination of two nonsignificant effects (between-school and within-school). The interpretation of this and any other significant between-student coefficient is not as straightforward as the interpretation of the multilevel coefficients. As Cronbach (1976) points out, the between-student analysis is often an uninterpretable blend of the between-school and within-school analyses.

The raw word knowledge test used in Table 5 can be interpreted as a measure of verbal ability. The positive coefficients in the table show two things. For items 3 through 10, the within-school coefficients show that students who are higher in verbal ability than their schoolmates are more likely to answer the item correctly. In addition, for items 3, 4, 6, 7, and 9, schools with a higher mean verbal ability did better on the item than schools with a lower mean verbal ability. This suggests that the test may require a high level of verbal ability. An inspection of the items shows that they do require a fairly high level of reading proficiency. The largely verbal format of the test may require as much verbal ability as biology. However, it is also possible that students who excel in one academic area (e.g., verbal ability) also excel in other areas (e.g., biology).

Tables 6 and 7 can be interpreted in a similar manner. In Table 6, liking of biology is an attitude indicator. In Table 7, number of books in the home can be seen as an indicator of socio-economic status.

The following results were drawn from Tables 6 and 7. Schools with higher mean attitude toward biology did better on items 3 and 5. However, most of the items were more sensitive to within-school attitudes (items 3, 4, 6-10). That is, students that liked biology more than their peers were also more likely to respond correctly to the items. Finally, schools with a higher average socio-economic status did better on most items (3-10) and students with a higher socio-economis status than their peers did better on items 3, 4, and 6.

The direction of the regression coefficients is consistent with prior findings about the relationship of socio-economic status, liking of subject matter, and verbal ability to student achievement. That is, schools containing students with a more positive attitude toward the subject matter, a higher mean socio-economic status or a higher mean verbal ability were more likely to exhibit higher achievement. In addition, students who were higher than their peers on the three variables were more likely to achieve higher than their peers. However, items are differentially sensitive to different variables. For example, item 2 is only sensitive to between-school sex differences; whereas, item 4 is sensitive to within-school differences on all four variables and between-school differences on three of the four variables. Also, examination of the between-student coefficients will not reveal the various processes. For example, on item 7, the total coefficient on liking of biology, books in the home, and raw word knowledge represents within-school differences, between-school differences, and the combination of between-school and within-school differences, respectively.

Finally, in Tables 8 and 9, two school variables are used to predict item response: hours of instruction, and hours of homework. As can be seen from items 3, 4, 6, 7, 8, and 10, the more instruction a student receives relative to his/her peers, the higher the student will achieve relative to his/her schoolmates. The amount of homework also had a positive effect both between-school and within-school. Item 3 shows that the more biology homework that is done across the school, the higher the school mean will be for this item. For items 4, 6, 7, 8, and 10, more homework by the student results in higher achievement than his schoolmates with less biology homework. Apparently, the amount of instruction and homework do effect student achievement within the school.

BTES

The Beginning Teacher Evaluation Study was sponsored by the California Commission for Teacher Preparation and Licensing with funds from the National Institute of Education. The study was conducted to examine the relationship between instructional variables and achievement in reading and mathematics in grades 2 and 5. Of particular interest to this paper was the learning of fifth grade mathematics -- a subject area in which a great deal of time and effort are put into teaching fractions. Tests were administered to six student in each of 25 second and 25 fifth grade classes on four occasions -- (A) October, 1976; (B) December, 1976; (C) May, 1977; and (D) September, 1977. In addition to the achievement tests, measures of allocated time, engagement rates, and success rates were obtained. Students were selected for not being extremely low or extremely high in ability (roughly 30 to 70 percentile). This restriction in range of entering student ability, combined with the care taken to measure instructional variables and the development of instructionally sensitive tests, makes this data set an interesting example for examining the relationship between the multilevel characteristics of items and instructional and program variables.

While the IEA data did have some instructional and school process variables, the BTES is especially noteworthy for their efforts to develop instructionally sensitive instruments (BTES, Filby and Dishaw, 1975, 1976). Since the goal of BTES was to understand the relationship between instructional variables and student achievement, special efforts were made to develop tests which would be reactive to instruction. The researchers felt that tests used to evaluate instructional processes must be sensitive indicators of classroom learning. Test items were checked for content validity to be sure that test content and instructional content overlapped. Then, items were checked to see whether gains were related to instruction (Carver, 1974). In their analysis, Filby and Dishaw assumed that students would perform better on an item after instruction than prior to instruction. In addition, students who receive high amounts of instruction in a given content area were expected to perform better on items from that content area than students who receive less instruction in that content area. Items that conformed to the two above assumptions were then selected to form a reactive, sensitive measure of classroom learning. Using this technique for test development (i.e., item selection), the BTES tests did show a significant relationship to time allocation by content area (Fischer, Filby, Marliave, Cahen, Dishaw, Moore, and Berliner, 1978).

In order to focus our attention on a manageable data set, it was decided to work only with the fraction items of the mathematics grade 5 test. This further reduced the data set since the fraction items were not given on occasion A (October, 1976). The fifteen items from the fractions subtest tested the student's ability to identify equivalent fractions. The skills tested included reducing fractions and finding the missing numerator or denominator in a fractional equation. The items are contained in Appendix B. There were 127 cases on occasion B (December, 1976), 123 cases on occasion C (May, 1977), and 89 cases on occasion D (September, 1977). The individual students were drawn from 21 classrooms.

Besides having the instructional variables, another difference between the BTES analyses and the IEA analyses was the use of a "pretest". The model for the BTES analysis was the same except that two indeoendent varaibles were used. The dependent variables were the item responses on occasion C. The independent variables were the item responses on occasion B along with:

1. Allocated Time - minutes allocated to learning fractions divided by 1000,
2. Easy Time - estimated time spent doing work that is easy for the student, divided by 100,
3. Hard Time - estimated time spent doing work that is difficult for the student.

The regression equations are the same as those used in the IEA analyses, except that there are now a pair of independent variables in each equation.

The basic multilevel item characteristics are given in Tables 10a, 10b, and 10c. Two features of the tables are especially prominent. First, students scored appreciably higher on occasion C than on occasion B, and slightly lower on occasion D than on occasion C. As was expected, performance increased after instruction and fell off over the summer vacation. The second feature of these tables is that the average n^2 followed the same pattern as the mean response. Apparently, the same students working together within a classroom and getting roughly the same level of instruction within a classroom result in large between-class effects, but after the class breaks up, the between-class effect began to diminish. The pattern of summer loss is unrelated to class membership.

In Table 11 the point biserial correlations are given. The majority of the items correlated fairly highly with the subtest at all levels of analysis. This meant that students who did well on an item also did well on the rest of the test, relative to the rest of the class. Also, a class that scored high on the subtest was likely to get the individual items right. Hence, it appears that the test is fairly reliable for measuring either within-class or between-class differences.

The regression analyses are contained in Tables 12, 13, and 14. Each table is based on the prediction of item scores from the same item on an earlier occasion and an instructional variable. The "pretest" seems to have positive impact on both the between-class and within-class analyses in all three tables. The positive within-class effect shows that students who do better than their classmates on occasion B will do better than their peers on occaion C. The positive between-class effect shows that classes that do well on the item on one occasion will also do well on the item on the second equation.

benefited more from instruction.

Instructional effects were also found to be related to item response. Table 12 shows that for item 9, there was a significant psoitive relationship between average classroom allocated time and item response. Classes which spend more time learning fractions got better results on this item. None of the within-class coefficients were significant. We interpret this along with an η^2 of .720 for allocated time to mean that there is not a great deal of variation of the allocated time of different students within a class. Students within a class will often work on a given content area at the same time. However, individualization and learning centers can differentiate the time allocated to different students within a class.

In the case of items 1, 2, 4, 5, 8-10, and 12-15, there was a confounding of effects. While neither the between-class nor the within-class coefficients are significant, the total coefficient is significant. This is a case where multilevel item analysis would have suggested a different conclusion than a total analysis. Apparently, the combination of the between-class effect and the within-class effect does suggest that students who are allocated more time will perform higher on the item, but the partitioning of the variance masks the effect. This suggests that the between-student analysis can also give useful information. While partitioning effects into between-class effects and within-class effects often may help to better understand the classroom effects, the between-student effects may also yeild useful and interesting information.

Hard time and easy time are peculiar variables in that they have different substantive meaning at the two levels of analysis (i.e., between-class and within-class). At the between-class level, the variables can be interpreted as measures of what level the class is taught at. However, within-class effects can be atrributed largely to ability. A student with high ability will spend a great deal of time going through exercises which are easy simply because he knows more about the tasks assigned to the whole class. In contrast, a low ability student will find very little to be easy.

Tables 11 and 12 are consistent with our interpretation of the variables easy time and hard time. At the between-class level, too much easy time has a negative effect for items 1 and 2, and too much hard time has a positive effect for item 11. Apparently, too much easy time for the class is detrimental to learning; whereas, more hard time may be beneficial to the students. When a classroom is taught below its level, the material covered is already known and no learning occurs. However, when a classroom is taught at or above its level, the class excels because of the challenge. The within-class results were also consistent with the above discussion. A positive effect within-class for items 1 and 2 on easy time suggests that students who had more time spent on easy activities were the higher achievers. Conversely, a negative effect for items 2 and 11 on hard time suggests that students who experienced more time on difficult activities were low achievers.

The BTES analyses suggests that there is much to be learned about the relationship of instructional variables and item responses from a multilevel perspective. Effects can occur both between and within classes. Furthermore, some possible different substantive meanings were given to between-class and within-class effects.

Possible Utility of Multilevel Item Analysis

Major concerns about standardized norm-referenced tests have centered around their program relevance and instructional sensitivity. These concerns are generated by the weak evidence of program and instructional effects (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld, and York, 1966; Averch, Carroll, Donaldson, Kiesling, and Pincus, 1972) in conjunction with findings that test performance is higher when there is a substantial overlap between test content and instructional content (Armbruster, Steven, and Rosenshin, 1977; Jenkins and Pany, 1976; Madaus, Kellaghan, Rakkow, and King, 1979; Walker and Schaffarzik, 1974) and that even the most broadly based achievements vary substantially in content coverage (Porter, Schmidt, Floden, and Freeman, 1978).

Clearly, more effort is needed to develop instructionally sensitive measures. Efforts to develop instructionally sensitive and program relevant tests have followed two lines. First, there has been an effort to develop, by curricula and test analysis, tests such that program content and instructional content overlap with test content. Second, as in the BTES, investigators have attempted to develop insstructionally sensitive tests using logical empirical methods (e.g., as discussed on page 13). However, whether either of these test develoonment strategies would have a large impact on the quality of testing in schools is unclear. The majority of testing currently being conducted involves either standardized norm-referenced tests or state assessment and competency testing. Typically, the local school district has little input to the test development process and must rely on the publisher's and state educational agency (SEA) generated results.

While a large-scale development effort may not be possible, there does seem to be some virtue in developing test analysis strategies that district personnel can use to "customize" the standardized test and assessment data to their local needs. Such strategies should be within the technical and economic means of district research and evaluation staff.

One way of attacking the problem is to develop methods to improve instructional sensitivity that test publishers and SEA testing agencies would willingly employ in their test development activities. Such methods would have to both command the respect of the applied psychometric community and be viewed as economically and politically advantageous.

One possible step in the right direction in the development of instructionally sensitive tests and test use may be found from an examination of the multilevel characteristics of test item data. As has been seen, different items are sensitive to different background and class processes. Possibly, through the use of multilevel analyses of item data, subtests can be formed which are more sensitive to the between-class or within-class process of interest, or at least, items could be excluded from the test results which are insensitive to the variables of interest.

Conclusions

Items can be sensitive to background and instructional variables. They can be sensitive either within-groups and/or between-groups. That is, classrooms can have an effect on the student's response to an item. In addition, the relative rank of a student with respect to an instructional or a background variable can affect the item response.

Explaining multilevel effects on item response is still at a rudimentary stage and needs to be explored further. What is clear is that a between-student analysis fails to take into account the instructional context and its effect on student item response. This failure has two effects. First, the relationship between item response and other variables cannot be explained since it is a conglomerate of two different processes. That is, the between-class effect and within-class effect may have different substantive meanings which cannot be sorted out in a between-student analysis. Second, the between-student analysis may give a distorted view of whether an effect does or does not exist. That is, the combination of the between-class effect and within-class effect can work in opposite directions to obscure an effect that does exist, or they can work in the same way to produce a statistically significant effect when neither source is statistically significant by itself.

Clearly, more work is needed to better understand the multilevel characteristics of items. One possible avenue which may prove fruitful is the expansion of the present model. Items may relate to variables in more complex ways. A model might be built that takes into account socioeconomic status, verbal ability, a "pretest", and instructional variables simultaneously. Another approach might be to examine a variety of indices of grouping effects for their applicability to test item data. Finally, the properties of subtests which might be formed using multilevel item analysis should be examined.

References

Armbruster, L., Everston, C., and Brophy, J. The First Grade Reading Group Study: Technical Report of Experimental Effects and Process-Outcome Relationships. Austin, Texas: Research and Development Center for Teacher Education, 1978.

Averch, H., Carroll, S.J., Donaldson, T., Kiesling, H.J., and Pincus, J. How effective is schooling? A critical review and synthesis of research findings (R-956-PCSF/RC). Santa Monica, California: The Rand Corporation, 1972.

Burstein, L. Implications from the Beginning Teacher Evaluation Study for the IEA Second Mathematics Study. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 1978.

Burstein, L., Fischer, K., and Miller, M.D. Social policy and school effects: A cross-national comparison. Paper presented at the IX World Congress of Sociology Meeting, Uppsala, Sweden, August 1978.

Burstein, L., Linn, R.L., and Capell, F.J. Analyzing multilevel data in the presence of heterogeneous within-class regressions. Journal of Educational Statistics, 1978, 3(4), 347-383.

Burstein, L., and Miller, M.D. Alternative analytical models for identifying educational effects: Where are we? Paper presented at the Annual Meeting of the American Educational REsearch Association, Toronto, Canada, March 1978.

Carver, R.P. Two dimensions of tests: Psychometric and edumetric. American Psychologist, 1974, 29, 512-518.

Coleman, J.S. Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, S., Weinfeld, F.D., York, R.L. Equality of educational opprotunity (2 Vols.). Office of Education, U.S. Department of Health, Education and Welfare. Washington, D.C.: U.S. Government Printing Office, 1966.

Comber, L.C., and Keeves, J.P. Science education in nineteen countries, International studies in evaluation (Vol. 1). Stockholm: Almqvist & Wiksell; and New York: Wiley, 1973.

Cronbach, L.J. (with the assistance of J.E. Deken & N. Webb). Research on classrooms and schools: Formulation of questions, design, and analysis. Occasional Paper, Stanford Evaluation Consortium, Stanford, California, July, 1976.

Cronbach, L.J. and Webb, N. Between-class and within-class effects in a reported aptitude X treatment interaction: Reanalysis of a study by G.L. Anderson. Journal of Educational Psychology, 1975, 67, 717-724.

Duncan, O.D., Cuzzort, R.P., and Duncan, B.D. Statistical geography: Problems in analyzing areal data. Glencoe: Free Press, 1961.

Filby, N.N. and Dishaw, M. Development and refinement of reading and mathematics tests for grades 2 and 5. Technical Report III-1, Far West Laboratory for Educational Research and Development, Beginning Teacher Evaluation Study, August 1975.

Filby, N.N. and Dishaw, M. Refinement of reading and mathematics tests through an analysis of reactivity. Technical Report III-6, Far West Laboratory for Educational Research and Development, Beginning Teacher Evaluation Study, November 1976.

Fisher, C.W., Filby, N.N., Marliave, R.S., Cahen, L.S., Dishaw, M.M., Moore, J.W., and Berliner, D.C. Teaching behaviors, academic learning time and student achievement: Final Report of Phase III-B, Beginning Teacher Evaluation Study (Technical Report V-1). San Francisco: Far West Laboratory for Educational Research and Development, June 1978.

Haney, W. Units of analysis issues in the evluation of Project Follow Through. Unpublished report, Cambridge, Mass.: Huron Institute, 1974.

Jenkins, J.R. and Pany, D. Curriculum biases in reading achievement tests. Technical Report No. 16. Urbana, Illinois: University of Illinois, Center for the Study of Reading, November 1976.

Knapp, T.R. The unit-of-analysis problem in applications of simple correlation analysis to educational research. Journal of Educational Statistics, 1977, 2, 171-186.

Madaus, G.F., Kellaghan, T., Rakow, E.A., and King, D.J. The sensitivity of measures of school effectiveness. Harvard Educational Review, 1979, 49(2), 207-230.

McLarty, J.R. Multilevel item analysis. Paper presented at the Annual Conference of the California Society fo Educational Program Auditors and Evaluators, San Francisco, California, May 1979.

Porter, A.C., Schmidt, W.H., Floden, R.E., and Freeman, D.J. Practical significance in program evaluation. American Educational Research Journal, 1978, 15(4), 529-539.

Robinson, W.S. Ecological correlations and the behavior of individuals. American Sociological Review, 1950, 351-357.

Singer, J.D. and Goodrich, R.L. Aggregation and the unit of analysis in the National Day Care Study. Paper presented at the Annual Meeting of the American Educational Research Association , San Francisco, April 1979.

Walker, D.F., and Schaffarzick, J. Comparing curricula. Review of Educational Research, Winter 1974, 44(1), 83-111.

Wolf, R.M. Achievement in America: National report of the International Educational Achievement Project. New York: Teachers College Press, Columbia University, 1977.

Table 1:

Item Descriptive Data from IEA

(N=1210)

Item #	Mean	Standard Deviation			n^2
		Total	Between	Within	
2	.33	.47	.12	.45	.07
3	.65	.48	.15	.45	.10
4	.71	.46	.13	.44	.08
5	.19	.39	.11	.38	.08
6	.63	.48	.13	.47	.07
7	.49	.48	.15	.46	.10
8	.25	.43	.14	.41	.10
9	.32	.47	.14	.44	.09
10	.34	.45	.11	.43	.06
Total	4.00	1.82	.76	1.65	.18

SOURCE: McLarty, 1979.

Table 2:

IEA Item Intercorrelations

(N=1210 students, 50 schools)

Item #	2	3	4	5	6	7	8	9	10
2	(total) (between) (within)								
3	-.02 .03 -.03								
4	.05 .19 .04	.16 .66 .11							
5	.05 .26 .04	.06 .26 .04	.02 .19 .01						
6	.04 -.04 .04	.15 .53 .12	.11 .47 .09	.08 .40 .05					
7	.04 -.01 .05	.16 .58 .12	.20 .52 .17	.03 .29 .00	.25 .53 .23				
8	.10 -.36 .07	.14 .33 .12	.11 .40 .08	.08 .19 .07	.15 .25 .15	.14 .36 .11			
9	.13 .43 .10	.12 .49 .08	.10 .44 .07	.07 .29 .04	.07 .37 .04	.11 .41 .08	.18 .38 .16		
10	.05 .15 .05	.04 .40 .01	.03 .32 .01	.06 .52 .03	.09 .32 .08	.08 .28 .06	.10 .23 .09	.03 .16 .02	

SOURCE: McLarty, 1979.

Table 3. IEA corrected item - total correlation

Item	Total	Between	Within
2	.12	.52	.10
3	.23	.43	.17
4	.22	.45	.17
5	.12	.49	.08
6	.27	.47	.24
7	.29	.44	.24
8	.28	.47	.25
9	.22	.45	.17
10	.13	.50	.10

Table 4. Regression of IEA biology items on sex.

	EFFECT ESTIMATES					
	Unstandardized			Standardized		
Item	Between	Within	Total	Between	Within	Total
2	-.186*	-.025	-.045	-.06	-.03	-.05
3	-.068	-.083*	-.090*	-.02	-.09	-.09
4	-.219*	-.074*	-.197*	-.08	-.08	-.11
5	-.006	.070*	.070*	-.00	.09	.09
6	-.047	-.035	-.040	-.02	-.04	-.04
7	-.198*	-.006	-.027	-.07	-.01	-.03
8	-.223*	-.104*	-.127*	-.08	-.12	-.15
9	-.071	-.070*	-.078*	-.02	-.08	-.08
10	.007	.053	.054*	.00	.06	.06

* Coefficient exceeds twice its standard error.

Table 5. Regression of IEA biology items on Raw ~~Word~~ Knowledge.

	EFFECT ESTIMATES					
	Unstandardized			Standardized		
Item	Between	Within	Total	Between	Within	Total
2	.009	.003	.005	.05	.03	.05
3	.018*	.018*	.022*	.09	.19	.24
4	.013*	.018*	.021*	.07	.22	.25
5	.008	.007*	.009*	.05	.09	.12
6	.014*	.012*	.015*	.07	.14	.17
7	.018*	.017*	.021*	.09	.19	.23
8	.005	.018*	.019*	.03	.22	.24
9	.022*	.010*	.014*	.12	.11	.17
10	.003	.007*	.008*	.02	.09	.09

* Coefficient exceeds twice its standard error.

Table 6. Regression of IEA biology items on students' liking of biology.

	EFFECT ESTIMATES					
	Unstandardized			Standardized		
Item	Between	Within	Total	Between	Within	Total
2	-.025	.024	.021	-.01	.04	.03
3	.163*	.066*	.085*	.09	.10	.13
4	.099	.045*	.056*	.05	.07	.09
5	.133*	-.010	.006	.08	-.02	.01
6	-.039	.084*	.079*	-.02	.13	.12
7	.039	.085*	.090*	.02	.13	.14
8	-.012	.108*	.106*	-.01	.18	.18
9	.044	.072*	.077*	.02	.11	.12
10	.013	.076*	.077*	.01	.12	.13

*Coefficient exceeds twice its standard error.

Table 7. Regression of IEA biology items on number of books in the home.

	EFFECT ESTIMATES					
	Unstandardized			Standardized		
Item	Between	Within	Total	Between	Within	Total
2	-.011	.006	.005	-.01	.01	.01
3	.193*	.051*	.073*	.11	.09	.13
4	.190*	.063*	.084*	.12	.11	.15
5	.097*	.026	.037*	.07	.06	.08
6	.118*	.067*	.080*	.07	.11	.14
7	.247*	.020	.048*	.14	.03	.08
8	.162*	.027	.045*	.11	.05	.09
9	.137*	.021	.037*	.08	.04	.07
10	.132*	.011	.026	.08	.02	.05

* Coefficient exceeds twice its standard error.

Table 8. Regression of IEA biology items on biology instruction.

	EFFECT ESTIMATES					
	Unstandardized			Standardized		
Item	Between	Within	Total	Between	Within	Total
2	-.086*	.019	-.002	-.08	.03	-.00
3	.032	.063*	.071*	.03	.11	.12
4	-.038	.038*	.029	-.04	.07	.05
5	.036	.003	.012	.04	.01	.02
6	-.046	.059*	.047*	-.04	.10	.08
7	-.014	.047*	.044*	-.01	.08	.08
8	.006	.043*	.044*	.01	.08	.09
9	.011	.012	.014	.01	.02	.03
10	-.023	.047*	.041*	-.02	.09	.08

* Coefficient exceeds twice its standard error.

Table 9. Regression of IEA biology items on biology homework.

	EFFECT ESTIMATES					
	Unstandardized			Standardized		
Item	Between	Within	Total	Between	Within	Total
2	-.066	.024	.009	-.05	.04	.02
3	.111*	.027	.051*	.08	.04	.08
4	-.007	.040*	.039*	-.01	.07	.06
5	.059	-.008	.005	.05	-.02	.01
6	-.036	.057*	.049*	-.03	.09	.08
7	-.004	.056*	.055*	-.00	.09	.09
8	.023	.038*	.043*	.02	.07	.07
9	.023	.032	.037*	.02	.05	.06
10	-.022	.057*	.052*	-.02	.10	.09

* Coefficient exceeds twic its standard error.

Table 10a. Descriptive statistics of BTES fractions subtest - occasion B.

	Mean	Standard Deviation			n^2
Item		Total	Between	Within	
1	.58	.50	.30	.39	.37
2	.54	.50	.30	.40	.36
3	.58	.50	.23	.44	.21
4	.54	.50	.26	.42	.28
5	.16	.37	.18	.32	.25
6	.50	.50	.26	.43	.27
7	.42	.50	.31	.31	.39
8	.13	.34	.14	.31	.18
9	.09	.28	.12	.26	.19
10	.47	.50	.25	.43	.25
11	.42	.50	.25	.42	.26
12	.31	.46	.24	.40	.26
13	.21	.41	.19	.36	.22
14	.41	.49	.22	.44	.20
15	.27	.45	.23	.39	.25
Total Test	5.63	3.47	2.40	2.51	.48

Table 10b. Descriptive statistics of BTES fractions subtest-occasion C.

Item	Mean	Standard Deviation			n^2
		Total	Between	Within	
1	.50	.37	.21	.31	.32
2	.83	.38	.21	.32	.32
3	.60	.49	.22	.44	.20
4	.75	.44	.19	.39	.19
5	.39	.49	.27	.41	.31
6	.73	.45	.22	.39	.24
7	.67	.47	.26	.39	.31
8	.28	.45	.31	.32	.48
9	.26	.44	.26	.35	.36
10	.60	.49	.31	.39	.39
11	.62	.49	.27	.41	.29
12	.54	.50	.30	.40	.36
13	.36	.48	.25	.42	.27
14	.59	.49	.33	.37	.45
15	.36	.48	.27	.40	.32
Total Test	8.08	3.63	2.41	2.72	.44

Table 10c. Descriptive statistics of BTES fractions subtest-occasion D.

	Mean	Standard Deviation			n^2
Item		Total	Between	Within	
1	.74	.44	.22	.39	.24
2	.77	.42	.20	.37	.23
3	.58	.50	.16	.47	.10
4	.72	.45	.27	.36	.37
5	.24	.43	.25	.35	.33
6	.61	.49	.26	.42	.28
7	.53	.50	.27	.42	.29
8	.33	.47	.27	.39	.32
9	.31	.47	.23	.50	.23
10	.60	.49	.26	.42	.27
11	.54	.50	.21	.46	.17
12	.61	.49	.22	.44	.20
13	.36	.48	.23	.43	.21
14	.63	.49	.23	.43	.23
15	.49	.50	.28	.42	.30
Total Test	8.06	3.74	2.40	2.87	.41

Table 11. BTES fraction subtest - corrected part-whole correlations.

	Total			Between			Within		
OCCASION:									
Item	B	C	D	B	C	D	B	C	D
1	.552	.486	.485	.601	.549	.737	.444	.439	.354
2	.563	.343	.419	.725	.400	.701	.429	.320	.293
3	.169	.194	.267	.394	.078	.299	-.014	.257	.219
4	.415	.388	.261	.579	.313	.396	.287	.371	.203
5	.035	.260	.548	-.148	.589	.807	-.027	.121	.366
6	.523	.419	.353	.616	.492	.551	.404	.414	.269
7	.660	.410	.580	.634	.492	.685	.559	.420	.540
8	.549	.624	.632	.667	.769	.810	.449	.519	.577
9	.337	.595	.602	.606	.793	.804	.281	.454	.457
10	.457	.372	.542	.476	.476	.630	.372	.319	.466
11	.511	.521	.382	.288	.630	.191	.509	.438	.274
12	.375	.443	.320	.595	.622	.351	.221	.368	.256
13	.195	.370	.380	.389	.666	.475	.139	.242	.318
14	.303	.385	.320	.571	.428	.094	.137	.304	.342
15	.337	.536	.345	.491	.583	.359	.210	.395	.249

Table 12. Regression of BTES fraction items occasion C on fraction items occasion B (PRE) and allocated time (A.T.).

	UNSTANDARDIZED						STANDARDIZED					
	Between		Within		Total		Between		Within		Total	
Item	PRE	A.T.	PRE	A.T.	PRE	A.T.	PRE	A.T.	PRE	A.T.	PRE	A.T.
1	.012	.118	.048	.100	.125*	.115	.02	.23	.04	.13	.24	.15
2	.107	.030	-.128	.162	.113*	.114	.18	.06	-.10	.21	.21	.15
3	-.092	.077	-.151	.392*	.001	.363*	-.12	.11	-.07	.39	.00	.36
4	.047	.108	.412*	.098	.134*	.197*	.07	.19	.25	.11	.24	.23
5	-.004	.222	.319	.343*	.236*	.420*	-.01	.34	.12	.25	.36	.31
6	-.050	.140	.296	.005	.094	.094	-.07	.23	.17	.01	.15	.11
7	-.002	.063	.124	.177	.055	.255*	-.00	.10	.08	.18	.08	.24
8	.168	.051	.555	.284*	.170*	.374*	.23	.08	.17	.21	.27	.28
9	.296*	-.012	.243	.227	.198*	.251	.42	-.02	.07	.14	.33	.16
10	.030	.174	.482*	.124	.181*	.246*	.04	.27	.24	.13	.28	.25
11	-.172	.200	.441	.305*	.133	.387*	-.22	.29	.23	.31	.19	.40
12	.041	.162	.322	.257	.197*	.353*	.05	.23	.15	.24	.28	.33
13	.079	.162	.570	.168	.212*	.287*	.10	.24	.22	.15	.31	.25
14	.318	-.071	.549*	-.127	.181*	.012	.41	-.11	.25	-.13	.28	.01
15	.288	-.076	-.007	.342*	.134*	.385*	.38	-.12	-.00	.32	.21	.36

*Coefficient exceeds twice its standard error.

Table 13. Regression of BTES fraction items occation C on fraction items occasion B (PRE) an

	UNSTANDARDIZED						STANDARDIZED					
	Between		Within		Total		Between		Within		Total	
Item	PRE	E:T.	PRE	E.T.	PRE	E.T.	PRE	E.T.	PRE	E.T.	PRE	E.T.
1	-.108*	.097	-.008	.105	.022	.087	-.39	.42	-.01	.14	.09	.11
2	-.171*	.114*	-.131	.168	-.000	.102	-.55	.46	-.10	.22	-.00	.13
3	-.051	-.021	-.188	.395*	-.053	.371*	-.13	-.07	-.08	.39	-.17	.37
4	.010	.026	.351	.108	.037	.192*	.03	.10	.21	.13	.14	.22
5	-.036	.054	.561	.349*	.038	.496*	-.10	.18	.21	.26	.12	.36
6	-.033	.053	.247	.266	.032	.090	-.10	.19	.15	.03	.11	.10
7	-.018	.014	.093	.188	.002	.221*	-.05	.05	.06	.20	.01	.23
8	-.030	.032	.500	.279	.009	.358*	-.08	.09	.15	.20	.03	.26
9	-.034	.048	.169	.230	.020	.261	-.10	.14	.05	.14	.06	.16
10	.058	-.019	.445	.127	.025	.244*	.16	-.05	.23	.13	.07	.25
11	-.103	.090	.474*	.293*	.005	.442*	-.30	.30	.25	.30	.02	.45
12	-.126	.099	.431	.181	.001	.339*	-.37	.33	.21	.17	.00	.32
13	-.085	.090	.574	.138	.014	.293	-.26	.32	.22	.12	.05	.26
14	-.012	.028	.631*	-.114	.019	.013	-.03	.10	.29	-.11	.06	.01
15	-.123	.117	.268	.330*	.021	.416*	-.37	.40	.13	.31	.07	.39

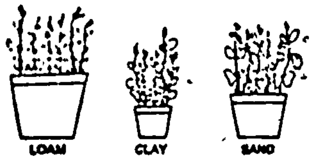
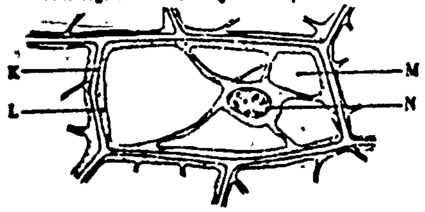
[a]Coefficient exceeds twice its standard error.

Table 14. Regression of BTES fraction items occasion C on fraction items occasion B (PRE) and hard time (H.T.).

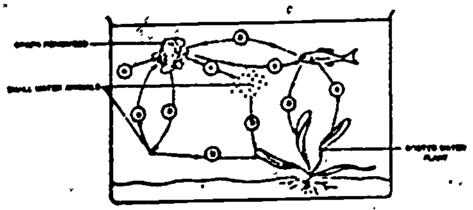
	UNSTANDARDIZED						STANDARDIZED					
	Between		Within		Total		Between		Within		Total	
Item	PRE	H.T.	PRE	H.T.	PRE	H.T.	PRE	H.T.	PRE	H.T.	PRE	H.T.
1	.017	-.030	.023	.082	-.020	.095	.11	-.25	.02	.11	-.17	.12
2	.022	-.047*	-.116	.139	-.036*	.103	.12	-.37	-.09	.18	-.28	.13
3	-.032	-.016	-.261	.376*	-.032*	.332*	-.14	-.10	-.12	.37	-.20	.33
4	.016	-.005	.373	.111	.006	.207*	.09	-.03	.22	.13	.04	.24
5	.042	-.029	.545	.358*	-.001	.503*	.20	-.19	.21	.26	-.01	.37
6	.003	-.009	.270	.019	-.005	.096	.02	-.06	.16	.02	-.04	.11
7	-.052	.021	.116	.210	-.008	.226	-.26	.14	.07	.22	-.05	.24
8	-.008	-.008	.480	.274	-.013	.348*	-.04	-.05	.15	.20	-.09	.26
9	.004	-.020	.167	.249	-.017	.282	.02	-.15	.05	.16	-.13	.18
10	-.013	.022	.440	.134	.010	.245*	-.06	.15	.22	.14	.07	.25
11	.056*	-.054*	.487*	.290*	-.021	.430*	.29	-.37	.26	.30	-.14	.44
12	.021	-.029	.395	.193	-.014	.327	.11	-.20	.19	.18	-.10	.31
13	.031	-.017	.577	.161	.001	.294*	.17	-.12	.22	.14	.00	.26
14	.014	-.006	.627*	-.119	.005	.010	.07	-.04	.29	-.12	.03	.01
15	.043	-.040	.285	.328*	-.015	.416*	.23	-.28	.13	.31	-.11	.39

*Coefficient exceeds twice its standard error.

Biology Test Items Form B

	Target Popul.	Content Area	Behavior	Average Facility	Med. Discrim.	Popul. Discrim.	Effect. Distr.	Easier in	Harder in
Population II—Test 4 B									
2. In an experiment green leaves were put in a jar and the apparatus was kept in the dark. Lime water was turned cloudy by the gas that formed in the jar. Which of the following gives the best explanation of this result? A. O_2 was produced by photosynthesis B. O_2 was produced by respiration • C. CO_2 was produced by respiration D. O_2 was used up in respiration E. CO_2 was produced by photosynthesis	II	Biol. (13)	Under.	36.4	27	E, D	E	India Iran	FRG Finland Netherlands
3. John brought the skull of an animal to school. His teacher said she did not know what the animal was but she was sure that it was one that preyed on other animals for its food. Which clue, do you think, led her to this conclusion? A. The eye sockets faced sideways B. The skull was much longer than it was wide C. There was a projecting ridge along the top of the skull • D. Four of the teeth were long and pointed E. The jaws could work sideways as well as up and down	II (1B.6)	Biol. (12)	Applic.	63.3	33	–	–	Belgium (Fr) Finland Italy	Scotland Thailand
4. Tom wanted to learn which of three types of soil—clay, sand, or loam—would be best for growing beans. He found three flowerpots, put a different type of soil in each pot, and planted the same number of beans in each, as shown in the drawing. He placed them side by side on the window sill and gave each pot the same amount of water.  Why was Tom's experiment *NOT* a good one for his purpose? A. The plants in one pot got more sunlight than the plants in the other pots • B. The amount of soil in each pot was not the same C. One pot should have been placed in the dark D. Tom should have used different amounts of water E. The plants would get too hot on the window sill	II (1A.8)	Biol. (13)	Under.	54.8	29	–	–	Belgium (Fr) Netherlands USA	Belgium (Fl) FRG Hungary Japan
5. The drawing represents a plant cell. In which of the four regions marked might chloroplasts be found?  A. In K only • B. In L only C. In M only D. In N only E. In both K and L	II	Biol. (14)	Inform.	17.1	17	D	C	Finland	Hungary
6. The energy for photosynthesis is generally obtained from A. chlorophyll. B. chloroplasts. C. sunlight. D. carbohydrates. E. carbon dioxide.	II (11B.1)	Biol. (13)	Inform.	62.1	25	–	–	Chile India Iran New Zealand Thailand	Belgium (Fl) Belgium (Fr) FRG

Biology Test Items - Form B

	Target Popul.	Content Area	Behavior	Average Facility	Med. Discrim.	Popul. Discrim.	Effect. Distr.	Easier in	Harder in
7. The diagram below shows an example of interdependence among aquatic organisms. During the day the organisms either use up or give off ⓐ or ⓑ as shown by the arrows.  Choose the right answer for ⓐ and ⓑ from the alternatives given. • A. ⓐ is oxygen and ⓑ is carbon dioxide B. ⓐ is oxygen and ⓑ is carbohydrate C. ⓐ is nitrogen and ⓑ is carbon dioxide D. ⓐ is carbon dioxide and ⓑ is oxygen E. ⓐ is carbon dioxide and ⓑ is carbohydrate	II (11B.2)	Biol. (13)	Under.	48.7	36	–	–	Thailand	Netherlands
8. What does an active muscle, that is, a muscle which is doing work, give up to the blood? • A. Carbon dioxide B. Oxygen C. Nitrogen D. Vitamin B E. Glucose	II	Biol. (14)	Inform.	21.8	31	B, D, E	C	Hungary Italy	Japan
9. The Andes are high mountains in South America and their inhabitants live and work at high altitudes. These people have almost twice as many red corpuscles in their blood as do the people living in the valleys. Which one of the following is the best explanation of this? A. In the Andes there is less air pressure acting on the inhabitants' blood vessels and so new red corpuscles can be produced more quickly B. Because there is a smaller amount of oxygen in the air of the Andes the inhabitants breathe more deeply in order to increase the total amount of oxygen in their lungs • C. In the Andes there is less oxygen entering the lungs of the inhabitants so that an increase in the number of red corpuscles enables a larger proportion of this oxygen to be absorbed D. Inhabitants of the Andes need more red corpuscles to transport oxygen through the blood vessels because there is less oxygen in the air they breathe E. The lower air pressure in the Andes causes blood to circulate more quickly through the blood vessels and so more red corpuscles are needed to transport the oxygen	II	Biol. (17)	Higher	20.8	19	D, E	D	India Iran	Belgium (Fl) Sweden
10. All of the following are aspects of the reproductive process. Which one of them must occur before we can be certain that fertilisation has taken place? A. A male organism must find a mate B. Reproductive organs must be produced • C. The nucleus of a male gamete must fuse with that of a female gamete D. A spermatozoon must reach an egg cell E. A female gamete must have a store of food for the embryo	II (10A.7)	Biol. (16)	Inform.	33.9	18	D	D	Belgium (Fl) Iran	Japan

APPENDIX B

STANDARDIZED

Between		Within		Total	
PRE	H.T.	PRE	H.T.	PRE	H.T.
.11	-.25	.02	.11	-.17	.12
.12	-.37	-.09	.18	-.28	.13
-.14	-.10	-.12	.37	-.20	.33
.09	-.03	.22	.13	.04	.24
.20	-.19	.21	.26	-.01	.37
.02	-.06	.16	.02	-.04	.11
-.26	.14	.07	.22	-.05	.24
-.04	-.05	.15	.20	-.09	.26
.02	-.15	.05	.16	-.13	.18
-.06	.15	.22	.14	.07	.25
.29	-.37	.26	.30	-.14	.44
.11	-.20	.19	.18	-.10	.31
.17	-.12	.22	.14	.00	.26
.07	-.04	.29	-.12	.03	.01
.23	-.28	.13	.31	-.11	.39

STANDARDIZED

Between		Within		Total	
PRE	E.T.	PRE	E.T.	PRE	E.T.
-.39	.42	-.01	.14	.09	.11
-.55	.46	-.10	.22	-.00	.13
-.13	-.07	-.08	.39	-.17	.37
.03	.10	.21	.13	.14	.22
-.10	.18	.21	.26	.12	.36
-.10	.19	.15	.03	.11	.10
-.05	.05	.06	.20	.01	.23
-.08	.09	.15	.20	.03	.26
-.10	.14	.05	.14	.06	.16
.16	-.05	.23	.13	.07	.25
-.30	.30	.25	.30	.02	.45
-.37	.33	.21	.17	.00	.32
-.26	.32	.22	.12	.05	.26
-.03	.10	.29	-.11	.06	.01
-.37	.40	.13	.31	.07	.39

STANDARDIZED

Between		Within		Total	
PRE	A.T.	PRE	A.T.	PRE	A.T.
.02	.23	.04	.13	.24	.15
.18	.06	-.10	.21	.21	.15
-.12	.11	-.07	.39	.00	.36
.07	.19	.25	.11	.24	.23
-.01	.34	.12	.25	.36	.31
-.07	.23	.17	.01	.15	.11
-.00	.10	.08	.18	.08	.24
.23	.08	.17	.21	.27	.28
.42	-.02	.07	.14	.33	.16
.04	.27	.24	.13	.28	.25
-.22	.29	.23	.31	.19	.40
.05	.23	.15	-.24	.28	.33
.10	.24	.22	.15	.31	.25
.41	-.11	.25	-.13	.28	.01
.38	-.12	-.00	.32	.21	.36

Table 14. Regression of BTES fraction items occasion C on fraction items occasion B (PRE) and hard time (H.T.).

UNSTANDARDIZED

	Between		Within		Total	
Item	PRE	H.T.	PRE	H.T.	PRE	H.T.
1	.017	-.030	.023	.082	-.020	.095
2	.022	-.047*	-.116	.139	-.036*	.103
3	-.032	-.016	-.261	.376*	-.032*	.332*
4	.016	-.005	.373	.111	.006	.207*
5	.042	-.029	.545	.358*	-.001	.503*
6	.003	-.009	.270	.019	-.005	.096
7	-.052	.021	.116	.210	-.008	.226
8	-.008	-.008	.480	.274	-.013	.348*
9	.004	-.020	.167	.249	-.017	.282
10	-.013	.022	.440	.134	.010	.245*
11	.056*	-.054*	.487*	.290*	-.021	.430(
12	.021	-.029	.395	.193	-.014	.327
13	.031	-.017	.577	.161	.001	.294*
14	.014	-.006	.627*	-.119	.005	.010
15	.043	-.040	.285	.328*	-.015	.416*

*Coefficient exceeds twice its standard error.

Table 13. Regression of BTES fraction items occation C on fraction items occasion B (PRE) and easy time (E.T.).

UNSTANDARDIZED

	Between		Within		Total	
Item	PRE	E.T.	PRE	E.T.	PRE	E.T.
1	-.108*	.097	-.008	.105	.022	.087
2	-.171*	.114*	-.131	.168	-.000	.102
3	-.051	-.021	-.188	.395*	-.053	.371*
4	.010	.026	.351	.108	.037	.192*
5	-.036	.054	.561	.349*	.038	.496*
6	-.033	.053	.247	.266	.032	.090
7	-.018	.014	.093	.188	.002	.221*
8	-.030	.032	.500	.279	.009	.358*
9	-.034	.048	.169	.230	.020	.261
10	.058	-.019	.445	.127	.025	.244*
11	-.103	.090	.474*	.293*	.005	.442*
12	-.126	.099	.431	.181	.001	.339*
13	-.085	.090	.574	.138	.014	.293
14	-.012	.028	.631*	-.114	.019	.013
15	-.123	.117	.268	.330*	.021	.416*

[a] Coefficient exceeds twice its standard error.

Table 12. Regression of BTES fraction items occasion C on fraction items occasion B (PRE) and allocated time (A.T.).

	UNSTANDARDIZED					
	Between		Within		Total	
Item	PRE	A.T.	PRE	A.T.	PRE	A.T.
1	.012	.118	.048	.100	.125*	.115
2	.107	.030	-.128	.162	.113*	.114
3	-.092	.077	-.151	.392*	.001	.363*
4	.047	.108	.412*	.098	.134*	.197*
5	-.004	.222	.319	.343*	.236*	.420*
6	-.050	.140	.296	.005	.094	.094
7	-.002	.063	.124	.177	.055	.255*
8	.168	.051	.555	.284*	.170*	.374*
9	.296*	-.012	.243	.227	.198*	.251
10	.030	.174	.482*	.124	.181*	.246*
11	-.172	.200	.441	.305*	.133	.387*
12	.041	.162	.322	.257	.197*	.353*
13	.079	.162	.570	.168	.212*	.287*
14	.318	-.071	.549*	-.127	.181*	.012
15	.288	-.076	-.007	.342*	.134*	.385*

*Coefficient exceeds twice its standard error.