INSTRUCTIONALLY SENSITIVE PSYCHOMETRICS: APPLICATIONS TO THE SECOND INTERNATIONAL MATHEMATICS STUDY

CSE Technical Report 286

Bengt O. Muthen

UCLA Center for Research on Evaluation,
Standards, and Student Testing

September, 1988

The research reported herein was conducted with partial support from the U.S. Department of Education, Office of Educational Research and Improvement, pursuant to Grant No. G0086-003. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Introduction[1]

This paper discusses new psychometric analyses that improve capabilities for relating performance on achievement test items to instruction received by the examinees. The modeling discussion will be closely tied to data for U.S. eighth grade students provided by the Second International Mathematics Study (SIMS), comprising not only responses to a set of achievement items at the beginning and end of the eighth grade but also a relatively rich set of student background information, including opportunity-to-learn (OTL) information specific to each item (Crosswhite, Dossey, Swafford, McKnight, & Cooney, 1985).

Item Response Theory (IRT) is a standard psychometric approach for analyzing a set of dichotomously scored test items. Standard IRT modeling assumes that the items measure a unidimensional trait. This particular kind of latent trait model is used to assess the measurement qualities of each item and to give each examinee a latent trait score. As will be shown, however, IRT modeling is limited in ways that are a hindrance to properly relating achievement responses to instructional experiences. Taking IRT as a starting point, this paper summarizes the author's work on a set of new psychometric analysis techniques that give a richer description of achievement-instruction relations. Six topics that expand standard IRT and specifically deal with effects of varying instructional opportunities (OTL) will be discussed as outlined below:

1. Variation in latent trait measurement characteristics. This relates to the classic IRT concern of "item bias," here translated as the absence or presence of an added advantage due to OTL in getting an item right.

2. Multidimensional modeling. Inclusion of narrowly defined, specific factors closely related to instructional units in the presence of a general, dominant trait.

3. Modeling with heterogeneity in levels. Analyses that take into account that achievement data often are not sampled from a single student population but one with heterogeneity of performance levels.

4. Estimation of trait scores. Deriving scores based on both performance and background information for both general and specific traits.

5. Predicting achievement. Latent trait modeling that relates the trait to student background variables.

6. Analyzing change. Relating change in general and specific traits to OTL.

The SIMS data will be used throughout to illustrate the new methods. All analyses will be carried out within the modeling framework of the LISCOMP computer program (Muthen, 1984, 1987).

The second section describes the SIMS data to be analyzed. The third section describes general features of the psychometric problem. The fourth section presents a descriptive analysis of the achievement - instruction relation for the SIMS data and sets the stage for later modeling. The remaining sections discuss methods topics 1 - 6 listed above.

[1] I would like to acknowledge the helpful assistance of Chih-Fen Kao, Suk-Woo Kim, and Lynn Short.

The SIMS data

The Second International Mathematics Study (Crosswhite, Dossey, Swafford, McKnight, & Cooney, 1985) was conducted in order to study variations in mathematics knowledge for eighth and twelfth graders within and across several countries including the United States, Japan, France, etc. To this aim, multiple-choice mathematics achievement responses were collected on items in the areas of arithmetic, algebra, geometry, measurement, and statistics. The test was administered both in the Fall and in the Spring of each grade. The achievement test consisted of 180 items distributed among four test forms. Each student responded to a core test of 40 items and one of four randomly assigned rotated forms with about 35 items. For the part of the sample that we will be concerned with, the core test was administered both during the Fall and the Spring while the rotated forms varied. It is well known that particularly eighth grade math curricula vary widely, certainly for students in the U.S. To be able to better describe the variation in student math achievement, information related to these curricular differences was therefore also collected. A detailed part of this information was student opportunity-to-learn (OTL) for the topics covered by each test item. For U.S. eighth grade math students, information was also collected in order to make a distinction between "tracks" or class type, yielding a categorization into Remedial, Typical, Enriched, and Algebra classes. This classification was based on teacher questionnaire data and on information on textbooks used. A variety of other teacher-related information was also collected, such as topic emphasis, textbooks, and teaching style. Student background information on family, career interests, and attitudes was also collected. We will concentrate our analyses on U.S. eighth graders for whom there are about 4,000 observations from both Fall and Spring randomly sampled from about 200 randomly sampled classrooms, varying in size from about 5 to 35 students. We will be particularly concerned with analyses of the 40 core items, but will also report on analyses of the four rotated forms which, when combined with the core items, consist of about 75 items administered to about 1,000 students taking each form. The rotated form analyses will be presented as a cross-validation of findings for the core items. In this way, the SIMS data provide a uniquely rich set of data with which to study instructionally sensitive psychometrics.

In the analyses that follow, a key piece of instructional information was obtained from the teacher questionnaire. For each item teachers were asked two questions regarding student opportunity to learn.

Question 1:

"During this school year did you teach or review the mathematics needed to answer the item correctly?"

1. No
2. Yes
3. No response

Question 2:

"If in this school year you did not teach or review the mathematics needed to answer this item correctly, was it mainly because?"

1. It had been taught prior to this school year
2. It will be taught later (this year or later)
3. It is not in the school curriculum at all
4. For other reasons
5. No response

Given these responses, opportunity-to-learn (OTL) level will be defined as:

No OTL: Question 1 = 1, questions 2 = 2, 3, 4, or 5
Prior OTL: Question 1 = 1, or 3 and question 2 = 1
This Year OTL: Question 1 = 2, question 2 = 9 (other response combinations had zero frequencies)

In most analyses to follow, Prior OTL and This Year OTL will be combined into a single OTL category.

The General Problem

In general, psychometric modeling assumes independent and identically distributed observations from some relevant population. This assumption is also made in IRT. The assumption of identically distributed observations is not realistic, however, using data of the SIMS kind to describe either relationships between what is measured (achievement responses) and what the measurements are attempting to capture (the traits), or how traits vary with relevant covariates such as instructional exposure and student background. This is because of the instructional heterogeneity of the students analyzed. The distribution of responses conditional on various trait values cannot be expected to be identical for a student who has had no specific instruction on the item topic and a student who has. The trait distribution cannot be expected to be the same for students in enriched classes as for students in typical classes. The students are naturally sampled from heterogeneous populations. It is true that increased homogeneity can be obtained by dividing the students into groups based on instructional experiences. However, such groupings may have to be very detailed to achieve their purpose and any simple grouping may be quite arbitrary. A more satisfactory approach is to use modeling that allows for heterogeneity, using parameters that vary for varying instructional experiences. Such modeling also accomplishes the goal of instructionally sensitive psychometrics, namely explicitly describing the achievement response-instructional experiences relations.

Descriptive Analyses

It is informative to consider descriptively how the achievement responses vary with instructional exposure. This forms a basis for our subsequent modeling efforts. We will study this in terms of both univariate and bivariate achievement distributions using the posttest core items administered to the U.S. eighth graders. We will also study the change in univariate responses from pretest to posttest.

Univariate Response

Consider first the univariate responses for the posttest. The wording of the core items is given in Appendix B. The proportion correct for each item is described in Table 1 (see Appendix A), broken down by the class type categories Remedial, Typical, Enriched, and Algebra and by the OTL categories No OTL, This Year OTL, and Prior OTL. From the totals it is seen that both class type and OTL have a strong effect on proportion correct.

For most items the proportion correct is higher for Enriched and Algebra classes than for Remedial and Typical classes. For almost all items the proportion correct increases when moving from No OTL to This Year OTL to Prior OTL. The reason why Prior OTL gives higher proportion correct than This Year OTL is partly because Prior OTL is more common for Enriched and Algebra classes to which we presume students of higher achievement levels have been selected. OTL appears to have an overall positive effect on proportion correct also when controlling for class type, at least for typical classes. Also, when controlling for OTL, class type seems to still have a strong effect.

The univariate relationships are informative but confound effects of instructional exposure with effects of student achievement level. For example, the higher proportion correct for a certain item for students with Prior OTL may be solely due to such students having a higher achievement level on the whole test. It would be of interest to know if students with the same achievement level perform differently on a certain item for different instructional exposure. To this aim we may consider the total score on the posttest as the general mathematics achievement level of each student and study the variation of proportion correct for each item as a function of instructional exposure conditionally on the general achievement level. We have carried this out using the dichotomous version of OTL, combining Prior OTL with This Year OTL into a single OTL category. For each value of the achievement variable we then have a proportion correct for a No OTL and an OTL group and can study whether OTL makes a difference. Conversely, for each of the two OTL categories we will present the distribution of the achievement variable in order to study whether having OTL for an item implies that these students have a higher general achievement level. These plots are given in Figures 1-9 (see Appendix A).

Figure 1 describes items 1, 2, and 3. The left-most panel shows the total score distribution given No OTL and OTL, respectively. We note that the score distributions have different locations with the OTL distribution having somewhat higher mean, supporting the notion that students who receive OTL perform better as measured by this test. We also note that the variances of the two distributions are about the same. The score distributions shown are representative of all core items.

The right-most part of Figure 1 and Figures 2-9 contain curves showing the proportion correct for given total score for the two OTL categories. For each item and both OTL categories, proportion correct increases with total score indicating that for both OTL categories the item is a good measurement indicator of the general achievement variable which the total score represents. It is particularly noteworthy that this is true also for the No OTL category and that the No OTL and OTL curves most often are very close. Students who, according to their teachers, have not been taught the mathematics needed to answer the item correctly still appear to have a high probability of answering the item correctly and this probability increases with increasing total score. This may indicate that students can to a large degree draw on related knowledge to solve the item. It may also indicate unreliability in the teachers' OTL responses. However, the differences in score distributions for the core items show that the OTL measures have consistent and strong relations to the total score. Instead of unreliability there may be a component of invalidity involved in the teachers' responses, where OTL may to some extent be confounded with average achievement level in the class and/or the item's difficulty.

The score distributions show that OTL is correlated with performance. Our hypothesis is that OTL helps to induce an increased level of the general achievement variable and that in general it is this increased level that increases the probability of a correct answer, not OTL directly. In this way, moving from the No OTL status to the OTL status implies a move upwards to the right along the common curve for No OTL and OTL.

There are some exceptions to the general finding of common curves for the No OTL and OTL categories. For example, items 3, 17, and 39 show a large positive effect of having OTL. Several other items with sizeable numbers of students in the two OTL categories also show positive effects. This means that for these items, the added advantage of having OTL is not fully explained by a corresponding increase in total score. OTL directly affects the success in solving the item correctly. From Table 1 we find that for the three items listed, the proportion correct increases strongly when moving from the No OTL category to the OTL categories. However, Table 1 cannot be counted on for finding items with direct OTL effects of this kind, since several other items also show strong increases in proportion correct due to OTL. We will return to the interpretation of this type of effect later in this report. Note also that with the exception of item 3 any OTL effect appears to be such that the two curves are approximately parallel, implying that the OTL effect is constant across achievement levels. For item 3 the OTL advantage increases with increasing achievement level, perhaps because it is a difficult item.

Bivariate Responses

The various descriptive analyses carried out for the univariate responses can be carried over to bivariate responses. A common measure for studying relationships among dichotomous items is that of the tetrachoric correlation coefficient (Lord & Novick, 1968). In line with the previous section, we may study the strength of association between each pair of achievement items by computing three sets of correlations, using all students, students with No OTL on either of the pairs of variables, and students with OTL on both of the pair of variables. For each of the sets, the average correlation across all pairs gives an indication of the degree of homogeneity of the items in their measurement of achievement. It is of interest to study if this homogeneity is affected by OTL. Further, in line with the previous section, the corresponding three sets of correlations may be computed as conditional on the total test score viewed as a general achievement variable. For lack of space these analyses will not be presented here, except to note that the homogeneity of correlations does not seem to be affected by OTL.

Change in Univariate Responses

The SIMS core items also provide the opportunity to study changes in proportion correct for each item from the Fall testing to the Spring testing. This change can be related to OTL. For each item we may distinguish between three groups of students, those who did not have OTL before the pretest or before the posttest (the No OTL group), those who had OTL before the pretest (Prior OTL), and those who did not have OTL before the pretest but did have OTL before the posttest (This Year OTL). The change for the No OTL group gives an indication of change due to learning on related topics. The change for the Prior OTL group gives an indication of effects related to practice, review, and perhaps, forgetting. The change for the group having This Year OTL reflects the direct exposure to the topic represented by the item. These changes can be studied in Table 1. Table 1 shows that, where changes occur, they are largely positive for each OTL category with the largest changes occurring for students in the category of This Year OTL as expected. This may be taken to support the dependability of the teacher-reported OTL measure.

Variations in Latent Trait Measurement Characteristics

The study of the univariate achievement responses showed that the set of core test items served as good indicators of the total test score. We may hypothesize that this test score is a proxy for a general mathematics achievement variable as measured by the combined content of the set of core items. However, the total test score is a fallible measure and what we are interested in are the relationships between the items and the true score and estimates of the true scores. This is a situation for which Item Response Theory (IRT) has been proposed as a solution used (see for example, Lord, 1980). The curves of Figures 1-9 are in IRT language called empirical item characteristic curves, which as theoretical counterparts have the conditional probability curves describing the probability correct on an item given a latent trait score. We will now describe the IRT model and how it can be extended to take into account instructional heterogeneity in its measurement characteristic.

In formulas the IRT model may be briefly described as follows. Let y* be a p vector of continuous latent response variables that correspond to specific skills needed to solve each item correctly. for item j,

$$(1)\qquad y_j = 0,\ \text{if}\ y^*_j \le \tau_j$$
$$\qquad\qquad 1,\ \text{otherwise}$$

where 0 denotes the incorrect answer, 1 denotes the correct answer, and τ_j is a threshold parameter for item j corresponding to its difficulty. Assume also that the latent response variable y^*_j is a function of a single continuous latent trait v and a residual ε_j,

$$(2)\qquad y^*_j = \lambda_j \eta + \varepsilon_j .$$

where λ_j is a slope parameter for item j, interpretable as a factor loading. With proper assumptions on the right-hand-side variables, this gives rise to the two-parameter normal ogive IRT model. For each item there are two parameters τ_j and λ_j. The conditional probability of a correct response on item j is

$$(3)\qquad P\ (y_j = 1\ I\ \eta) = \phi[(-\tau_j + \lambda_j \eta)\,\theta^{-1/2}]$$

where θ is the variance of ε_j. This means that the threshold τ_j determines the item's difficulty, that is the horizontal location of the probability curve, and the loading λ_j determines the slope of the probability curve.

Earlier in this report, we investigated descriptively whether the conditional proportion correct given total test score varied across OTL groups. In IRT language this is referred to as investigating item bias or using a more neutral term, differential item functioning. Standard IRT assumes invariant item functioning across different groups of individuals. A variety of bias detection schemes related to IRT have been discussed in the literature. Concerns about item bias due to instructional heterogeneity have recently been raised in the educational measurement literature. Conflicting results have been found in empirical studies. For example, Mehrens and Phillips (1986, 1987) found little differences in measurement characteristics of standardized tests due to varying curricula in schools, while Miller and Linn (1988), using the SIMS data, found large differences related to opportunity to learn although these differences were not always interpretable. Muthen (1989) pointed out methodological problems in assessing differential item functioning when many items may be biased. He suggested a new approach based on a model which extends the standard IRT. The analysis is carried out by the LISCOMP program (Muthen, 1987). This approach is particularly suitable to the SIMS data situation with its item specific OTL information and it will be briefly reviewed here.

Let x be a vector of p OTL variables, one for achievement item. The x variables may be continuous, but assume for simplicity that x_j is dichotomous with $x_j = 0$ for No OTL and $x_j = 1$ for OTL. Consider the modification of equation (2)

$$(4)\qquad y^* = \lambda\eta + Bx + \varepsilon$$

where in general we restrict B to be a diagonal p x p matrix. The diagonal element for item j is denoted β_j. The OTL variables are also seen as influencing the trait η,

$$(5)\qquad \eta = \gamma x + \varsigma$$

where γ is a p-vector of regression parameter slopes and ς is a residual.

It follows that

$$(6) \quad P\ (\ y_j = 1 \mid \eta,\ x_j) = \Phi\ [\ (-\ \tau_j + \beta_j\ x_j + \lambda_j \eta)\ V(y^*_j \mid \eta)^{-1/2}\ ]$$

In effect, then, the β_j coefficient indicates the added or reduced difficulty in the item due to OTL. Equivalently, using equation (4), we may see this effect as increasing y^*_j, the specific skill needed to solve item j.

We note that this model allows for differential item functioning in terms of difficulty but not in terms of the slope related parameter λ_j. This is in line with the data analysis findings repeated previously where little difference in slopes of the conditional proportion correct curves was found across OTL groups (item 3 was an exception; we assume that this item will be reasonably well fitted by a varying difficulty model). More general modeling is in principle possible, but the data features do not seem to warrant such an extra effort.

This model disentangles the effects of OTL in an interesting way. Equation (5) states that OTL has an effect on the general achievement trait as measured by the γ coefficients. Here we are interested in finding positive effects of instruction. Through the expected increase in η, such effects also have an indirect positive effect on the probability of a correct item response. The strength of η's effect on item j is measured by the coefficient λ_j; see equations (4) and (6). In addition to the indirect effect of OTL for item j determined by γ and λ_j, there is also the possibility of a direct OTL effect on item j, which is determined by the β_j coefficient; see equations (4) and (6). Any direct effect indicates that the specific skill needed to solve item j draws not only on the general achievement trait but also on OTL. The size of the γ effect indicates the extent to which the achievement trait is sensitive to instruction. The size of the β_j effect indicates the amount of exposure sensitivity or instructional "over-sensitivity" in item j. While positive γ effects correspond to a positive educational outcome, positive β_j effects are of less educational interest in that they demonstrate effects of teaching that influences very narrow content domains. From a test construction point of view items that show such exposure sensitivity are less suitable for inclusion in standardized tests, since they are prone to "item bias" in groups of examinees with varying instructional history. If such item bias goes undetected, IRT analysis distorted. In the modeling presented above, however, exposure, sensitivity is allowed for and the analysis does not suffer from the presence of such effects.

Muthen, Kao, and Burstein (1988) presents examples of analysis of exposure sensitivity using the dichotomous OTL groupings. However, we will first consider an example from an earlier draft of this paper, where the OTL categories No OTL, This Year OTL, Prior OTL were used. Figure 10 shows the estimated item characteristic curves for item 17 having to do with acute angles. Since there are three OTL categories, there are three curves corresponding to three difficulty values. Since the curves for both This Year OTL and Prior OTL are above the No OTL curve, the β effects are positive for these two OTL groups. Exposure to the concept of acute angles produces a specific skill, which has the same effect as a reduced item difficulty, and this skill is not included in the general achievement trait. It is interesting to relate this finding to the percentage correct on item 17 broken down by OTL group as given in Table 1. Percentage correct increases dramatically from the No OTL category to the OTL categories, but the percentage correct is slightly higher for Prior OTL than for This Year OTL. For item 17 the Prior OTL students may do better than This Year OTL students, but Figure 10 shows that the recency of OTL gives an advantage for students at the same achievement trait level. Comparing the estimated item characteristic curves of Figure 10 with the empirical curves of Figure 5 we find a large degree of similarity but also differences. The estimated curves represent more correct and precise estimates of these curves.

Muthen, Kao, and Burstein (1988) found substantial exposure sensitivity in items 3, 16, 17, 38, and 39, corresponding to solving for x, the product of negative integers, acute angles, percentages, and the coordinate system (see Appendix B). While items 3, 17 and 39 provided rather poor measurements of the achievement trait as indicated by their estimated λ values, that was not the case for the other two. The authors hypothesized that the exposure sensitivity corresponded to early learning of a definitional nature. Further analyses of the rotated form items, carried out by Kao (1989), supported this hypothesis. For example, the rotated forms showed exposure sensitivity for items covering square root problems. Overall, about 15-30 % of the items exhibit mild exposure sensitivity, while only about 10-15 % exhibit strong exposure sensitivity. We may note that these percentages are considerably lower than the Miller and Linn (1988) findings using related parts of the SIMS data and standard IRT methodology. The effects of OTL on the achievement trait will be discussed in later sections.

Multidimensional Modeling

Standard IRT modeling assumes a unidimensional trait as was also done in the previous section. For a carefully selected set of test items, this is often a good approximation. However, in many achievement applications, it is reasonable to assume that sets of items draw on more than one achievement trait.

Muthen (1978) presented a method for the factor analysis of dichotomous items, where the model is

$$(7) \qquad y^* = \Lambda \eta + \varepsilon$$

$$(8) \qquad V\ (y^*) = \Lambda\ \Psi\ \Lambda' + \Theta$$

where Λ is a p x m factor loading matrix, Ψ is a factor covariance matrix, and Θ is a diagonal matrix of residual variances. In line with item analysis tradition (see Lord and Novick, 1968), Muthen fitted the model to a matrix of sample tetrachorics. For an overview of factor analysis with dichotomous items, see Mislevy (1986).

Although of great substantive interest, models with many minor factors are very hard to identify by usual means of analysis. For instance, assume as we will for the SIMS data that a general achievement factor is the dominant factor in that it influences the responses to all items.

Assume that, in addition to this general factor there are several specific factors, orthogonal to the general factor, that influence small sets of items of common, narrow content. It is well known that such models with continuous data cannot be easily recovered by ordinary exploratory factor analysis techniques involving rotations. This problem carries over directly to dimensionality analysis of dichotomous items using tetrachoric correlations.

Consider as an illustration of the problem an artificial model for forty dichotomous items. Assume that one general factor influences all items and eight specific factors each influence a set of five items. Let the general factor loadings be 0.5 and 0.6 while the specific factor loadings are 0.3 and 0.4. Let the factors be standardized to unit variances and let the factors be uncorrelated. The eigenvalues of the corresponding artificial correlation matrix are shown in Figure 11. Such a "scree plot" is used for determining the number of factors in an item set. The number of factors is taken to correspond to the first brake point in the plot where the eigenvalues level off. If the first eigenvalue is considerably larger than the others and the others are approximately equal, this is usually taken as a strong indication of unidimensionality.

Figure 11 clearly indicates unidimensionality despite the existence of the eight specific factors. There would be no reason to consider solutions of higher dimensionality.

As a comparison, Figure 12 shows the eigenvalues for the tetrachoric correlation matrix for the 39 core items of the SIMS data. The two eigenvalue plots are rather similar.

Models similar to the artificial one considered above have been studied by Schmid and Leiman (1957), where it was pointed out that the above hypothesized nine-factor model can also be represented as an eight-factor model with correlated factors. Each of the eight factors may be viewed as a function of both a general, second-order factor and the corresponding specific factor of the nine-factor model. The specific factor is then viewed as a residual contribution, orthogonal to the second order factor. hence, Schmid and Leiman used the term hierarchical factor analysis. Using exploratory factor analysis on the artificial correlation matrix, an oblique rotation of the eight factor solution did indeed identify the eight correlated factors of such a hierarchical reformulation of the model. Schmid and Leiman (1957) gave formulas for transforming such a solution back to the original model with a general factor and eight specific factors, all factors being uncorrelated. However, without knowing the correct number of factors, there would have been no guide to choosing this eight-factor solution.

The usefulness of hierarchical factor analysis has recently been pointed out by Gustafsson (1988a, b). He proposed to circumvent the difficulties of using exploratory factor analysis by formulating confirmatory factor analysis models. Hypothesizing a certain specific factor structure in addition to a general factor, the confirmatory model enables the estimation of factors with very narrow content. Applications of this type of modeling to the SIMS data are being considered by the author in collaboration with Burstein, Gustafsson, Webb, Kim, Novak, and Short. In line with our previous modeling, we may write a simple version of this model as

$$(9) \quad y^*_j = \lambda_{Gj}\, \eta_G + \lambda_{Sj}\, \eta_{S_K} + \varepsilon_j$$

where y^*_j is the latent response variable for item j (cf. earlier model), η_G is the general achievement factor, η_{S_K} is the specific factor for item j, and ε_j is a residual. The three right hand side variables are taken to be uncorrelated. This means that the items belonging to a certain specific factor correlate not only due to the general factor but also due to this specific factor.

In this simplified version of the model, it assumed that each item measures only one specific factor. For identification purposes we assume that each specific factor η_{S_K} is measured by at least two items. Also for identification purposes, our baseline model will set $\lambda_{Sj} = 1$ for all j's, although this can be relaxed as a need arises as will be discussed below. In this way, the general factor is assumed to influence each item to a different degree, while the specific factor has the same influence on all items in the corresponding set.

This multidimensional confirmatory factor analysis model allows an interesting variance component model interpretation. Standardizing the general factor variance to unity, while letting the specific factor variances be free parameters, the model implies a decomposition of the latent response variable variances into a general factor component, a specific factor component, and an error component:

$$(10) \quad V\ (y^*_j) = \lambda_{Gj}^2 + \Psi_{S_k} + \theta_j$$

where Ψ_{S_k} is the variance of the specific factor k. Since the items are dichotomous, the variances of the y*'s are standardized to one by restrictions on the θ_j's. The relative sizes of the first two terms on the right hand side of (10), the general and the specific components, are of particular interest. The specific component can also be interpreted as the average correlation remaining between items belonging to specific factor k when holding the general factor constant.

The model can be estimated by confirmatory factor analysis techniques for dichotomous items using the LISCOMP computer program, see Muthen (1978, 1987).

The SIMS items of the core and the rotated forms were classified into subsets corresponding to specific factors defined both by content and procedure. Examples of the narrow item domains that were considered are: Arithmetic with signed numbers (core items 3, 16, 25), percent calculations (core items 2, 3, 36, 38), estimations skills (size, distance; core items 6, 8, 9), and angular measurements (core items 17, 19, 21, 22).

The analysis steps are as follows. For a given hypothesized set of specific factors, a confirmatory factor analysis run can be performed. The initial model may then be refined in several steps. An inappropriate combination of items for a specific factor gives rise to a low or negative variance component estimate for this specific factor. Modifications may be assisted by inspection of model misfit indices. For this model a useful index is related to the loadings of the specific factors, λs_j, which are fixed to unity in the baseline model. The sign and size of the derivatives of these loadings are of interest. A positive value for a certain item indicates that if the loading is free to be estimated, the estimated value will be smaller than one. In effect, this allows the estimate of the variance component for the specific factor at hand to increase. This is because the specific variance component is related to the average correlation of the specific factor items, conditional on the general factor, where the decrease in the factor loading for a certain item means that the contribution from this item is weighted down. Thus modifying the initial analysis, items that obtain very low or negative specific factor loadings are candidates for exclusion from the set assigned to this specific factor. This modification process may be performed in several iterations. In the analyses performed for the SIMS data, this procedure appeared to produce substantively meaningful results in that the items that were singled out clearly had features that distinguished them form the others in the set.

Table 2 gives the estimated variance components for core items corresponding to three of the specific factors.

It is seen that the variance contribution from the specific factors can be as large as 50% of that of the general factor and are therefore of great practical significance. This is particularly so since the sets of items for a specific factor correspond closely to instructional units. Analyses of the rotated forms replicated most of the specific factors found for the core.

The confirmatory factor analysis procedure described is a cumbersome one involving many iterations and many subjective decisions. An attempt was therefore made to find an approach which would involve fewer steps and a more objective analysis. It was reasoned that if the influence of the general factor could be removed from the item correlations, the remaining correlations would be due to the specific factors alone. Such residual correlations could then be factor analyzed by regular exploratory techniques, at least if nesting of specific factors within each other was ignored. Given a proxy for the general factor, the residual correlations could be obtained by bivariate probit regressions of all pairs of items on the proxy, using the LISCOMP program.

An attempt was first made to approximate the general factor for the posttest core items with the posttest total score. However, this produced almost zero residual correlations. Instead, the pretest total score was used for the posttest items. An exploratory factor analysis of these residual correlations, using an orthogonal rotation by

Varimax, resulted in eleven factors with eigenvalues greater than one. The interpretation of these factors showed an extraordinary high degree of agreement with the specific factors previously obtained. The best agreement was obtained for factors that had obtained the largest variance component estimates. The exploratory analysis also suggested a few items to be added to the specific factors as defined earlier. The agreement of these two very different approaches is remarkable and it is interesting that the pretest score appears to be a better proxy for the general factor at the posttest occasion than the posttest score. This may indicate that the general factor is a relatively stable trait related to the achievement level before eighth grade instruction; we note from Table 1 that This Year OTL is the most prevalent category. Controlling for posttest score may in contrast control for a combination of the general factor and specific factors.

It is interesting to note that analyses of the core items administered at the pretest gave very similar results in terms of specific factors identified by the confirmatory approach. This indicates stability of the specific factors over the eighth grade. Attempting to compute residual correlations for exploratory factor analysis again gave near zero values when controlling for the total score, the pretest in this case, and this approach had to be abandoned.

Modeling with Heterogeneity in Levels

The factor analysis of the previous section was performed under the regular assumption of identically distributed observations, that is all students are assumed to be sampled from the same population with one set of parameters. However, we have already noted that the students have widely varying instructional histories and that the homogeneity of student populations is not a realistic assumption. This is a common problem in educational data analysis which has been given rather little attention. We may ask how this heterogeneity affects our analysis and if it can be taken into account in our modeling.

Muthen (1988a) considers covariance structure modeling in populations with heterogeneous mean levels. This research considers both the effect of incorrectly ignoring the heterogeneity and proposes a method to build the heterogeneity into the model. The method is directly applicable to the multidimensional factor analysis model considered in the previous section and can also be carried out within the LISCOMP framework. Consider the model of equation (7)

(11) $y^* = \Lambda\eta + \varepsilon$

In the previous section we made the usual standardization of $E\ (\eta_i) = 0$ for all observations i and assumed $V\ (\eta_i) = \Psi$. However, we know that it is unrealistic to assume that for example students from different class types have the same factor means levels and we may instead want to assume that the means vary with class type such that for student i in class c we have $E\ (\eta_{ic}) = a_c$. As pointed out in Muthen (1988a) this may be accomplished by considering in addition to (11) the equations

(12) $\eta_{ic} = \Gamma\ x_c + \varsigma_{ic}$

where x_c represents a vector of class type dummy variable values for class c, Γ is a parameter matrix, and ς_{ic} is a residual vector for student i in class c. We assume that conditional on class type membership the factor means vary while the factor covariance matrix remains constant,

(13) $E\ (\eta_{ic}\ I x_c) = \Gamma x_c$

(14) $V\ (\eta_{ic}\ I x_c) = \Psi$

The modeling also assumes that the matrices Λ and Θ are constant across class types, so that

(15) $E\ (y^*Ix_C) = \Lambda\Gamma xc$

(16) $V\ (y^*Ixc) = \Lambda\Psi\Lambda' + \Theta$

It is interesting to note that the assumption of constancy of the conditional covariance matrix $V(Y^*Ix_C)$ is in line with the findings of constancy of the homogeneity of correlations presented previously.

The structure imposed on the parameter matrices of (15) and (16) may correspond to an exploratory or a confirmatory factor analysis model. Muthen (1988a) points out that the conditional covariance matrix of (16) is not in general the same as the marginal covariance matrix $V\ (y^*)$. In our context this means that even when we have the same factor analysis structure in the different class types this covariance structure does not hold in the total group of students. The approach outlined here essentially provides a mean-adjusted analysis of pooled covariance matrices assumed to be equal in the population. In our situation the analysis effectively is carried out on pooled tetrachoric correlation matrices. This modeling has two important outcomes. The dimensionality analysis can be carried out without distortion due to the differences in factor mean levels across class types and the factor mean levels can be estimated.

The above mean-adjusted analysis was carried out on the SIMS core items using the multidimensional factor model from Table 2 of the previous section. Factor mean differences were allowed for class type using three dummy variables and also gender We will concentrate our discussion of the results on the factor structure. Despite large mean differences across class type for the general achievement factor, a factor structure very similar to the previous one emerged. The same specific factors showed large and small variances, respectively. Hence, the potential for a distorted structure is not realized in these data. The results are presented in parentheses in Table 2. It is seen that the variance contributions of the general factor are considerably reduced as compared to the first approach.

The reduction in variance contribution from the general factor is natural since holding class type constant reduces the individual differences in the general achievement trait due to selection of students. If the inference is to the mix of students encountered in the SIMS data the unreduced variation in the trait is the correct one, but this variation is not representative for a student from any given class type. It is also interesting to note that the specific factor variances are not similarly reduced by holding class type constant, presumably indicating that these specific skills are largely unrelated to the student differences represented by class type.

Estimation of Trait Scores

The sections above have considered various factor analysis models for the achievement responses. Assuming known or well-estimated parameter values for these models it is of interest to estimate each student's score on the factors of these models. For the standard, unidimensional IRT model estimation of the trait values is a standard task which may be carried out by maximum likelihood, Bayes' modal (maximum a posteriori), or expected a posteriori estimators (see for example Bock & Mislevy, 1986). The instructionally sensitive models we have considered for the SIMS data have however brought us outside this standard situation in the following three respects:

1. We want to consider factor score estimation that takes into account that certain items have different difficulty level depending on the students' OTL level.

2. We want to consider factor scores for both the general achievement factor and the specific factors in the multidimensional model.

3. We want to consider factor scores estimation that takes into account differences in student achievement level.

We note that (i) and (iii) are quite controversial since these points raise the issue of estimating achievement scores based not only on the student's test responses but also his/her instructional background. For example Bock (1972) has argued that prior information on groups should not be used in comparisons of individuals across groups. Nevertheless, it would seem that students who have had very limited OTL on a set of test items will be unfairly disadvantaged in comparison with students with different instructional exposure. The aim may instead be to obtain achievement scores for given instructional experiences.

Point (ii) is of considerable interest. While a rough proxy for the general achievement score is easily obtainable as the total test score, the adding of items corresponding to specific factors would involve only a few items resulting in a very unreliable score. As a contrast, estimating the specific factor scores draws on the correlated responses from all other items.

The following estimation procedure was discussed in Muthen and Short (1988) and handles all three cases above. For various density and probability functions g, consider the a posteriori distribution of the factors of η,

$$(14) \quad g\ (\eta\ I\ y,\ x) = \phi(\eta I x)\ g\ (y\ I\ \eta,\ x)\ /\ g\ (y\ I\ x)$$

Here, the first term on the right hand side represents a normal prior distribution for η conditional on x, where as before x represents instructional background variables such as OTL and class type. The factor covariance matrix may be taken as constant given x, while the factor means may vary with x. The second item on the right hand side represents the product of the item characteristic curves, which may vary in difficulty across OTL levels as discussed previously.

Muthen and Short (1988) considered an example of the situation of (i) and (iii). They generated a random sample of 1,000 observations from a model with forty items measuring a unidimensional trait. Observations were also generated from forty OTL variables and five other background variables. All background variables were assumed to influence the trait while the first twenty OTL variables had direct effects on their corresponding items, giving rise to exposure sensitivity in these items. Among other results, Muthen and Short considered differences in factor score estimates using the above method and the traditional IRT method. In Table 3 comparisons of the two corresponding score distributions are presented by quartiles, broken down in two parts - students with a high total sum of OTL and students of the low sum. The table demonstrates that for students of the low OTL group, estimated scores are on the whole higher with the new method, corresponding to an adjustment for having had less exposure, while for the high OTL group the estimated scores are on the whole lower for the new method.

Ongoing work by Muthen and Short investigates situation (ii) and the precision with which scores for specific factors can be estimated. Once the estimated factor scores have been calculated they may conveniently be related to various instructional variables and may also studied for change from pretest to posttest.

Predicting Achievement

Given the explorations of the previous sections, we may attempt to formulate a more comprehensive model for the data. Muthen (1988b) proposed the use of structural equation modeling for this task. He discussed a model which extends ordinary structural modeling to dichotomous response variables while at the same time extending ordinary IRT to include predictors of the trait. He studied part of the SIMS data using a model which attempted to predict a unidimensional algebra trait at the posttest occasion using a set of instructional and student background variables from the pretest. The set of predictors used and their standardized effects are given in Table 4. While pretest scores have strong expected effects, class type, being female, father being in the high occupational category, and finding mathematics useful to future needs also had strong effects. The OTL variables had very small effects overall, perhaps due to the fact that each item's OTL variable has rather little power in predicting this general trait.

Given the analysis results of the previous sections, this modeling approach can be extended to include a multidimensional model for both the set of pretest and posttest items, predicting posttest factors from pretest factors, using instructional and student background variables as covariates, and allowing for differential item functioning in terms of exposure sensitivity. This work is in progress.

Analyzing Change

The structural modeling discussed in the previous section is also suitable for modeling of change from pretest to posttest. We pointed out that in terms of change the SIMS data again exemplified complex population heterogeneity. For each item a student may belong to either of three OTL groups, corresponding to two types of no new learning and learning during the year. To again reach the goal of instructionally sensitive psychometrics for this new situation, we should explicitly model this heterogeneity. However, to properly model such complex heterogeneity is a very challenging task and this work has merely begun.

A basic assumption is that change is different for groups of students of different class types and OTL patterns. In a structural model where posttest factors are regressed on pretest factors the slopes may be viewed as varying across such student groups, where students groups for which a large degree of learning during the year has taken place, as measured by the set of OTL variables, are assumed to have steeper slopes than the other students. This methods area shows a very large degree of scarcity of psychometric work.

References

Bock, R.D. (1972). [Review of *The dependability of behavioral measurements*]. *Science, 178,* 1275-1275A.

Bock, R. D., Mislevy, R.J. (1986). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

Crosswhite, F.J., Dossey, J.A., Swafford, J.O., McKnight, C.C., & Cooney, T.J. (1985). *Second International Mathematics Study summary report for the United States.* Champaign, IL: Stipes.

Gustafsson, J-E. (1988a). Hierarchical models of individual differences in cognitive abilities. In R.J. Sternberg (Ed.), *Advances in the psychology of human intelligence, vol. 4.* Hillsdale, NJ: Lawrence Erlbaum.

Gustafsson, J-E. (1988b). *Broad and narrow abilities in research on learning and instruction. Learning and individual differences: Abilities, motivation, methodology.* Symposium conducted at the Minnesota Symposium, Minneapolis.

Kao, C-F. (in preparation). *An investigation of instructional sensitivity in mathematics achievement test items for U.S. eighth grade students.* Unpublished doctoral dissertation, University of California, Los Angeles.

Lord, F.M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Addison-Wesley: Reading, MA.

Mehrens, W.A., & Phillips, S.E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23,* 185-196.

Mehrens, W.A., & Phillips, S.E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24,* 357-370.

Miller, M.D., & Linn, R.L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement, 25,* 205-219.

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3-31.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,*551-560.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthen, B. (1987). *LISCOMP. Analysis of linear structural equations with a comprehensive measurement model: User's guide.* Mooresville, IN: Scientific Software, Inc.

Muthen, B. (1988a). *Covariance structure modeling in heterogeneous populations: Mean-adjusted analysis.* Los Angeles: UCLA Graduate School of Education.

Muthen, B. (1988b). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Erlbaum.

Muthen, B. (in press). Using item-specific instructional information in achievement modeling. *Psychometrika.*

Muthen, B., Kao, C-F, & Burstein, L. (1988). *Instructional sensitivity in mathematics achievement test items: Application of a new IRT-based detection technique.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Muthen, B. & Short, L.M. (1988). Estimation of ability by IRT models allowing for heterogeneous instructional background. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Schmid, J., & Leiman, J.M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53-61.

TABLE 1

Percentage Students and Percentages Correct for Core Items by OTL and Class Type

Item	Total*		No OTL			This Year OTL			Prior OTL		
	FR	FO	ST	FR	FO	ST	FR	FO	ST	FR	FO
ME01											
TOT	35	43	21	22	26	59	36	47	20	44	48
REM	11	18	33	7	8	60	12	23	7	21	21
TYP	30	38	24	21	27	64	34	43	12	28	34
ENR	42	52	17	25	24	71	48	63	12	29	29
ALG	61	64	6	64	64	5	39	50	89	62	65
AR02											
TOT	47	60	3	34	53	89	45	59	8	74	78
REM	12	21	9	17	33	91	11	20	0	0	0
TYP	42	57	3	34	40	97	42	57	0	0	0
ENR	58	74	4	46	86	90	57	73	6	74	81
ALG	74	75	0	0	0	43	73	71	57	74	78
AL03											
TOT	9	21	38	8	9	61	10	28	1	3	19
REM	15	9	78	15	8	22	13	13	0	0	0
TYP	8	14	49	7	9	50	8	18	2	3	19
ENR	8	21	16	12	11	84	7	23	0	0	0
ALG	16	64	7	0	19	94	17	68	0	0	0
AR04											
TOT	27	33	13	23	26	75	26	31	12	44	50
REM	16	13	40	16	9	60	16	15	0	0	0
TYP	24	29	11	16	20	87	25	30	2	30	30
ENR	29	38	15	39	48	70	25	34	15	37	45
ALG	47	54	0	0	0	33	41	50	67	50	56
ME05											
TOT	32	44	7	32	30	86	31	45	6	46	55
REM	17	18	6	27	27	85	17	18	9	17	8
TYP	27	40	8	20	17	90	27	42	2	22	43
ENR	37	55	5	49	60	95	37	54	0	0	0
ALG	56	63	8	75	66	48	53	62	44	55	64
ME06											
TOT	49	55	28	48	54	59	48	55	13	52	59
REM	20	31	41	23	35	45	21	31	14	11	22
TYP	47	52	27	48	53	65	48	52	8	42	47
ENR	52	61	32	51	60	65	52	62	2	82	68
ALG	66	73	10	83	80	28	68	75	62	63	72
	FR	FO	ST	FR	FO	ST	FR	FO	ST	FR	FO
GE07											
TOT	56	66	69	55	66	23	56	66	8	63	73
REM	26	39	75	25	36	25	27	46	0	0	0
TYP	54	64	66	54	64	24	55	63	10	56	67
ENR	58	72	71	56	71	27	64	75	3	62	77
ALG	77	85	75	76	84	6	82	95	19	83	87
ME08											
TOT	89	89	17	89	88	58	88	88	25	93	92
REM	67	61	34	62	55	58	69	64	8	76	67
TYP	89	89	17	94	93	66	88	89	18	89	88
ENR	93	93	16	90	91	59	93	93	26	96	94
ALG	98	97	14	96	100	12	96	98	74	99	97
ME09											
TOT	42	52	14	41	48	56	38	50	30	50	59
REM	16	18	27	18	19	58	15	18	15	21	15

TABLE 1

Percentage Students and Percentages Correct for Core Items by OTL and Class Type

Item	Total*		No OTL			This Year OTL			Prior OTL		
TYP	37	48	14	41	49	62	36	47	23	38	49
ENR	48	64	11	42	53	63	46	65	27	56	65
ALG	67	73	12	76	78	2	56	33	85	66	73
GE11											
TOT	26	31	40	20	26	56	29	34	4	33	39
REM	9	8	77	11	7	19	4	10	4	0	27
TYP	20	27	43	16	25	54	22	29	2	23	21
ENR	31	38	29	31	36	68	32	38	3	15	35
ALG	57	54	24	49	46	62	59	55	14	57	59
AR12											
TOT	34	44	10	32	40	85	34	44	5	41	48
REM	18	22	35	19	23	65	19	21	0	0	0
TYP	30	40	6	22	29	90	31	41	4	25	35
ENR	39	51	9	51	65	89	38	49	3	43	57
ALG	54	62	16	46	57	63	55	64	21	56	58
	PR	PO	ST	PR	PO	ST	PR	PO	ST	PR	PO
AL13											
TOT	58	71	12	46	59	85	59	73	2	74	85
REM	31	46	32	28	36	68	33	51	0	0	0
TYP	54	67	45	48	62	84	55	67	1	68	91
ENR	63	81	2	94	94	94	62	81	4	69	94
ALG	87	89	7	46	77	88	90	92	6	87	65
AR14											
TOT	56	61	15	49	53	78	56	61	7	66	76
REM	29	26	29	27	23	64	32	27	7	17	28
TYP	53	58	15	46	50	82	54	58	4	62	79
ENR	61	70	13	59	70	85	62	70	2	35	65
ALG	77	82	8	97	88	51	75	81	41	76	81
AR15											
TOT	22	32	10	20	28	77	20	30	14	34	45
REM	18	18	10	22	15	90	17	18	0	0	0
TYP	20	28	12	18	26	83	20	28	5	28	31
ENR	21	38	8	26	39	83	21	39	9	15	28
ALG	38	47	0	0	0	23	23	25	77	42	54
AL16											
TOT	23	58	6	9	16	92	24	60	2	37	88
REM	9	14	52	10	9	48	7	20	0	0	0
TYP	18	50	3	6	11	97	18	52	0	0	0
ENR	28	74	2	17	89	94	28	73	4	34	94
ALG	53	89	0	0	0	94	53	89	6	41	77
GE17											
TOT	47	59	13	39	38	72	46	62	15	59	63
REM	24	24	41	22	15	48	25	26	10	29	46
TYP	42	56	11	42	37	82	43	60	8	35	40
ENR	53	68	12	44	44	80	55	72	8	53	68
ALG	76	80	10	61	85	18	78	93	72	78	77
AL18											
TOT	43	51	20	32	29	78	46	56	2	58	60
REM	25	23	55	20	17	45	31	31	0	0	0
TYP	39	44	24	36	31	76	40	48	0	0	0
ENR	47	63	4	28	36	89	47	65	6	59	57
ALG	71	78	7	31	58	88	75	81	6	55	68
	PR	PO	ST	PR	PO	ST	PR	PO	ST	PR	PO

TABLE 1

Percentage Students and Percentages Correct for Core Items by OTL and Class Type

Item	Total*		No OTL			This Year OTL			Prior OTL		
GE19											
TOT	23	33	76	23	32	23	22	38	1	52	57
REM	10	19	0	10	19	0	0	0	0	0	0
TYP	22	30	72	22	29	28	21	33	0	0	0
ENR	25	39	71	25	35	29	25	49	0	0	0
ALG	39	49	89	38	48	0	0	0	11	52	57
AR20											
TOT	73	77	2	55	60	86	71	76	12	89	90
REM	31	37	7	28	33	93	31	37	0	0	0
TYP	71	75	3	64	69	93	71	75	5	78	85
ENR	80	87	6	0	88	94	80	87	0	84	0
ALG	94	94	0	0	0	29	93	96	71	94	93
GE21											
TOT	20	34	60	20	30	37	21	39	3	23	39
REM	16	16	97	16	17	3	25	13	0	0	0
TYP	18	30	60	17	29	39	20	33	1	22	11
ENR	20	39	46	20	34	52	20	44	2	6	33
ALG	34	50	65	33	45	18	44	71	17	28	49
GE22											
TOT	37	59	13	26	26	80	37	64	7	62	67
REM	21	18	79	23	19	17	9	11	4	30	40
TYP	33	55	8	28	26	90	33	58	2	29	37
ENR	40	71	6	20	15	92	40	75	2	59	59
ALG	70	81	9	47	82	44	70	85	47	73	78
ME23											
TOT	33	47	19	25	30	73	33	50	8	47	65
REM	17	18	52	17	18	48	18	17	0	0	0
TYP	29	41	19	25	30	80	31	44	2	16	29
ENR	33	58	15	29	41	79	34	62	6	23	53
ALG	59	74	7	38	35	43	62	78	51	60	76
AR24											
TOT	52	59	7	37	36	83	50	55	10	78	81
REM	23	18	15	33	18	85	21	18	0	0	0
TYP	47	53	10	38	40	89	48	55	1	50	58
ENR	60	66	0	0	0	95	60	65	5	61	75
ALG	80	82	0	0	0	21	71	76	79	82	83
	FR	FO	ST	FR	FO	ST	FR	FO	ST	FR	FO
AL25											
TOT	42	46	7	28	34	92	42	47	2	70	59
REM	12	15	28	8	13	72	13	16	0	0	0
TYP	38	42	7	36	40	92	37	43	2	68	44
ENR	48	55	3	40	60	97	49	55	0	0	0
ALG	69	67	0	0	0	94	69	66	6	73	86
AL27											
TOT	46	57	53	38	50	47	54	64	1	67	71
REM	27	30	91	26	30	9	36	24	0	0	0
TYP	42	52	58	37	48	41	49	59	1	67	71
ENR	50	63	49	49	64	51	50	62	0	0	0
ALG	69	82	7	50	65	93	71	83	0	0	0
AR28											
TOT	51	62	9	44	49	74	49	61	16	63	73
REM	20	29	20	19	19	76	21	33	4	0	18
TYP	47	57	11	47	49	80	47	58	9	44	59

TABLE 1

Percentage Students and Percentages Correct for Core Items by OTL and Class Type

Item	Total*		No OTL			This Year OTL			Prior OTL		
ENR	59	72	6	56	79	83	58	71	11	61	69
ALG	77	86	0	0	0	25	73	85	75	78	86
ME29											
TOT	77	75	10	63	60	64	76	75	25	83	81
REM	40	44	22	40	22	68	41	49	11	34	55
TYP	75	74	9	65	69	71	75	74	19	78	74
ENR	85	82	13	71	64	71	87	84	16	85	88
ALG	92	89	0	0	0	11	95	95	89	91	89
AL30											
TOT	31	40	52	28	36	45	34	48	3	34	43
REM	25	23	83	27	23	17	13	20	0	0	0
TYP	27	37	59	25	35	38	28	40	3	39	29
ENR	34	46	37	32	41	57	35	48	6	28	58
ALG	50	57	25	48	61	75	51	56	0	0	0
AR33											
TOT	45	50	5	34	33	87	44	49	8	62	66
REM	20	19	22	20	12	78	20	21	0	0	0
TYP	41	47	5	39	41	91	41	47	4	52	57
ENR	51	59	0	0	0	97	50	59	3	74	61
ALG	65	69	2	75	75	47	65	67	51	65	71
	FR	FO	ST	FR	FO	ST	FR	FO	ST	FR	FO
AR34											
TOT	24	39	4	16	19	90	22	39	7	45	53
REM	10	15	19	14	16	81	9	14	0	0	0
TYP	19	34	4	17	22	96	19	34	0	0	0
ENR	29	54	0	0	0	97	29	54	3	39	35
ALG	44	53	0	0	0	43	43	50	57	45	55
AL35											
TOT	51	59	29	39	44	70	55	65	1	54	92
REM	38	30	78	37	33	22	41	22	0	0	0
TYP	46	55	36	40	46	63	49	59	1	54	92
ENR	53	68	11	37	52	89	55	70	0	0	0
ALG	78	83	0	0	0	100	78	83	0	0	0
AR36											
TOT	47	56	7	44	38	86	46	56	7	64	73
REM	33	31	19	37	31	81	32	30	0	0	0
TYP	44	52	8	47	41	92	44	53	0	0	0
ENR	51	66	4	41	32	93	52	68	3	43	57
ALG	66	72	0	0	0	43	65	68	57	66	75
AR37											
TOT	31	37	15	21	23	65	29	36	21	44	52
REM	14	12	38	11	8	62	16	14	0	0	0
TYP	26	31	17	24	24	73	27	33	9	28	32
ENR	36	46	6	19	30	62	36	48	32	40	46
ALG	57	69	5	39	67	24	49	63	71	61	71
AR38											
TOT	36	51	3	26	23	91	34	51	7	61	72
REM	16	25	9	25	17	91	16	25	0	0	0
TYP	31	45	3	27	25	97	31	46	0	0	0
ENR	42	66	0	0	0	97	43	66	3	35	52
ALG	61	69	0	0	0	43	57	62	57	63	74
GE40											
TOT	35	47	47	33	41	50	37	52	3	52	56

TABLE 1

Percentage Students and Percentages Correct for Core Items by OTL and Class Type

Item	Total*		No OTL			This Year OTL			Prior OTL		
REM	24	31	93	24	31	7	21	21	0	0	0
TYP	32	43	46	30	38	54	34	46	0	0	0
ENR	39	56	32	35	44	66	42	63	2	22	50
ALG	52	60	56	53	59	19	44	68	26	57	57

TABLE 1

Percentage Students and Percentages Correct for Core Items by OTL and Class Type

* Percentages of students by class type are:
REM= Remedial: 7.1 (N=268), TYP= Typical: 57.6 (N=2148)
ENR= Enriched: 24.4 (N=909), ALG= Algebra: 10.7 (N=399)

ST= Percentage students
PR= Percentage correct for pretest
PO= Percentage correct for posttest

ME= measurement
AR= Arithmetic
AL= Algebra
GE= Geometry

Table 2

Variance Components for Selected Items from the Core*

Item	General Factor	Specific Factors		
		Percent	Estimate	Angular Measurement
AR02	33(24)	9(9)		
AR34	39(32)	9(9)		
AR36	32(27)	9(9)		
AR38	35(26)	9(9)		
ME06	20(14)		9(10)	
ME08	38(27)		9(10)	
ME09	38(29)		9(10)	
GE17	28(17)			11(12)
GE19	17(12)			11(12)
GE21	24(17)			11(12)
GE22	43(30)			11(12)

*Given in parenthesis is the estimate when controlling for mean level heterogeneity. (See section 7)

Table 3

TRAIT ESTIMATES BY TRADITIONAL AND NEW APPROACHES*

LOW OTL GROUP

	TRADITIONAL				
NEW	25%	50%	75%	100%	TOTAL
25%	136 -1.323 -1.255	6 -0.610 -0.724	0	0	142 -1.293 -1.233
50%	10 -0.783 -0.624	125 -0.361 -0.338	5 0.037 -0.119	0	140 -0.375 -0.351
75%	0	13 -0.094 0.058	111 0.309 0.316	7 0.827 0.691	131 0.297 0.311
100%	0	0	6 0.691 0.834	124 1.282 1.308	130 1.255 1.286
TOTAL	146 -1.286 -1.212	144 -0.347 -0.318	122 0.317 0.324	131 1.257 1.275	543

Table 3 (cont'd)

HIGH OTL GROUP

	TRADITIONAL				
NEW	25%	50%	75%	100%	TOTAL
25%	99 -1.306 -1.349	9 -0.578 -0.743	0	0	108 -1.245 -1.298
50%	5 -0.726 -0.581	94 -0.340 -0.366	12 0.049 -0.119	0	111 -0.315 -0.349
75%	0	3 -0.167 -0.022	110 0.345 0.322	5 0.870 0.640	118 0.355 0.327
100%	0	0	6 0.653 0.782	114 1.386 1.334	120 1.349 1.306
TOTAL	104 -1.278 -1.312	106 -0.355 -0.389	128 0.332 0.302	119 1.364 1.305	457

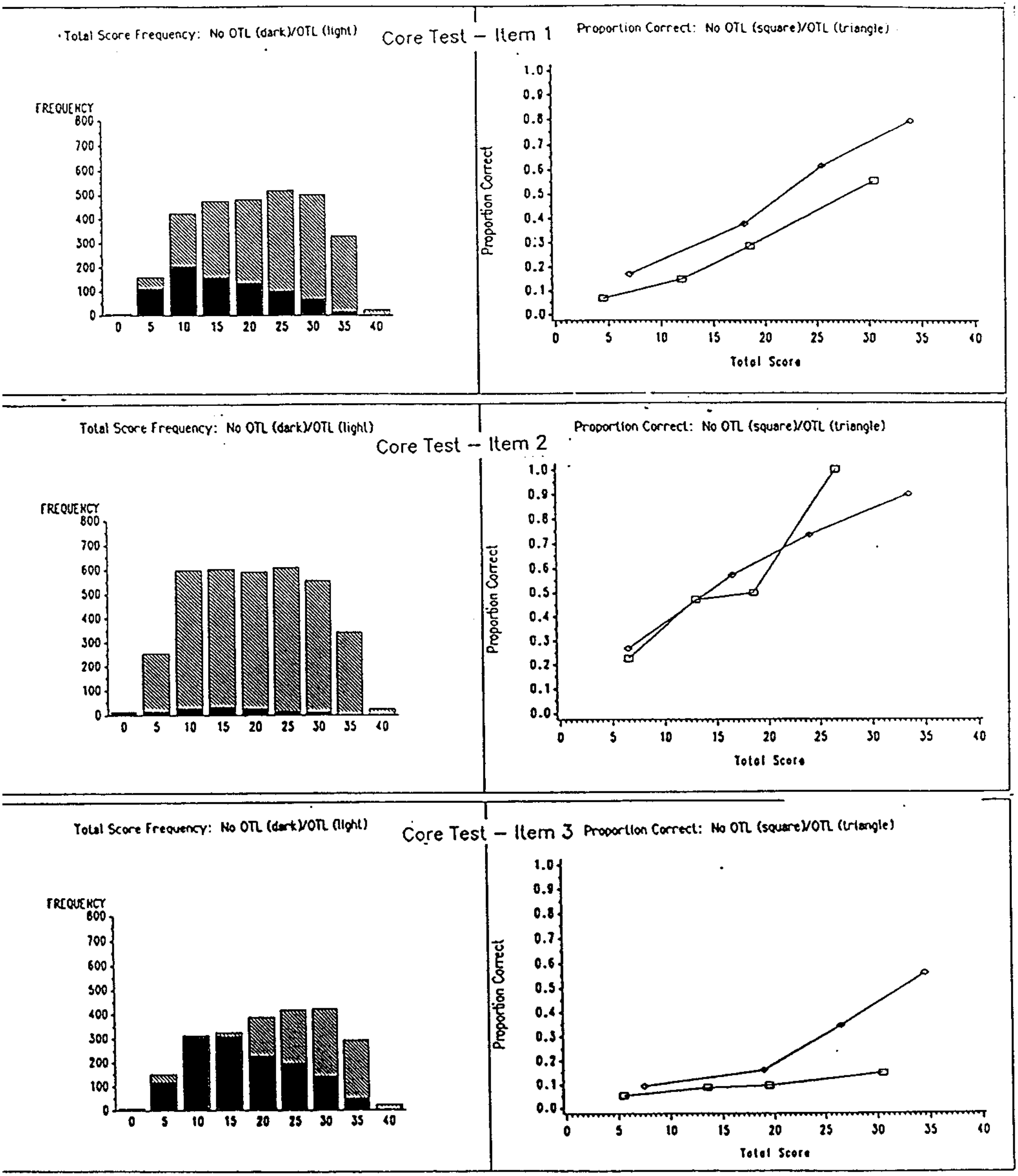
*Entries are
Frequency
mean value by the traditional approach
mean value by the new approach

Table 4

Structural Parameters with the Latent Construct as Dependent Variable

Regressor	Estimate	Estimate/S.E.
PREALG	0.68	11
PREMEAS	0.45	7
PREGEOM	0.33	5
PREARITH	2.09	16
FAED	0.07	1
MOED	0.02	0
MORED	0.18	3
USEFUL	0.45	7
ATTRACT	0.04	1
NONWHITE	-0.02	0
REMEDIAL	0.07	1
ENRICHED	0.22	3
ALGEBRA	0.56	4
FEMALE	0.14	6
LOWOCC	0.02	1
HIGHOCC	0.12	3
MISSOCC	0.05	2
NONW X REM	0.10	1
NONW X ENR	0.19	3
NONW X ALG	-0.18	-1
PREARITH X REM	-1.45	-3
PREARITH X ENR	-0.10	-1
PREARITH X ALG	-0.54	-2
NONW X PREARITH	-0.19	-1

FIGURE 1



Core Test – Item 1

Core Test – Item 2

Core Test – Item 3

FIGURE 2

Proportion Correct: No OTL (square)/OTL (triangle)

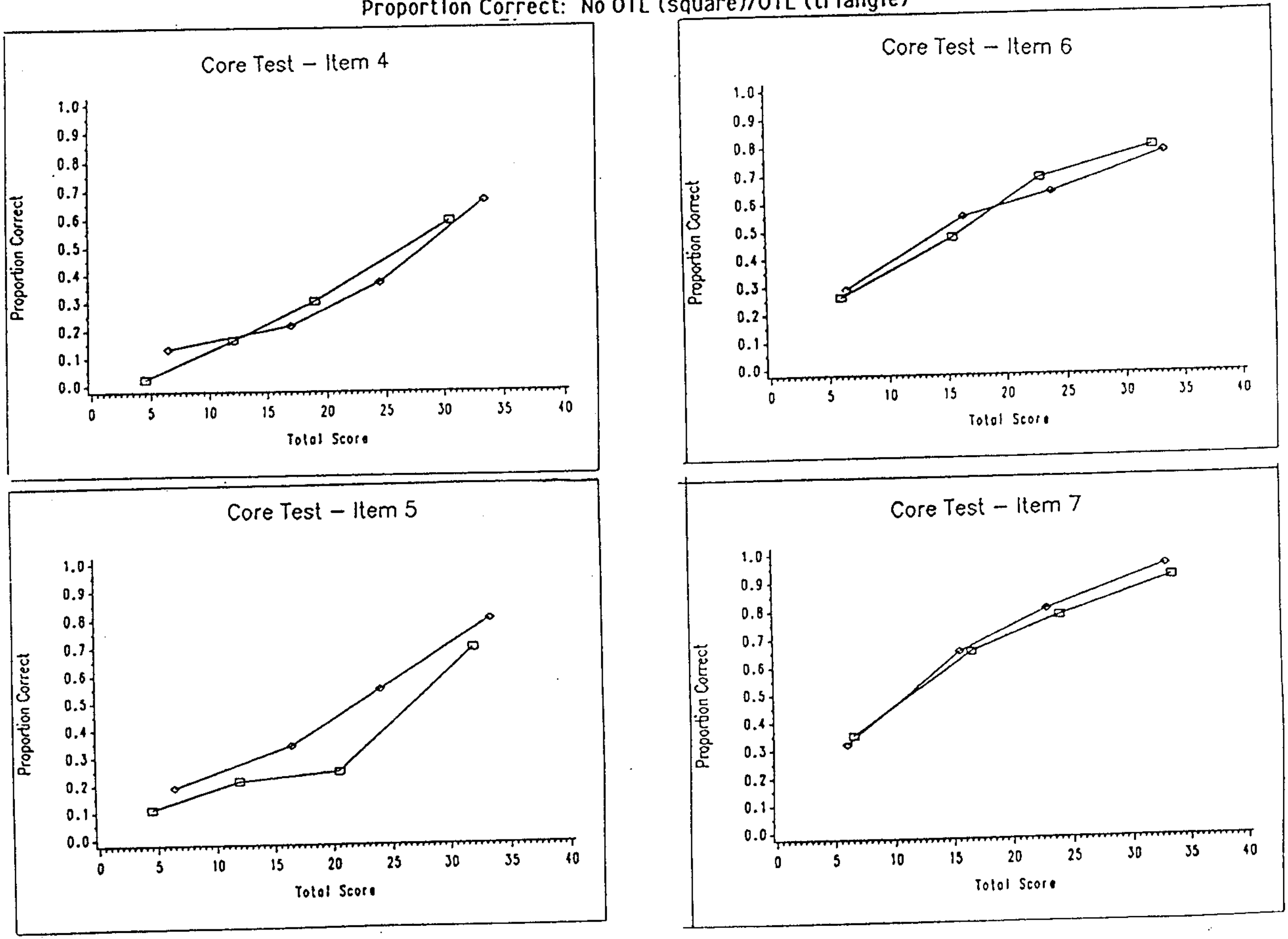


FIGURE 3

Proportion Correct: No OTL (square)/OTL (triangle)

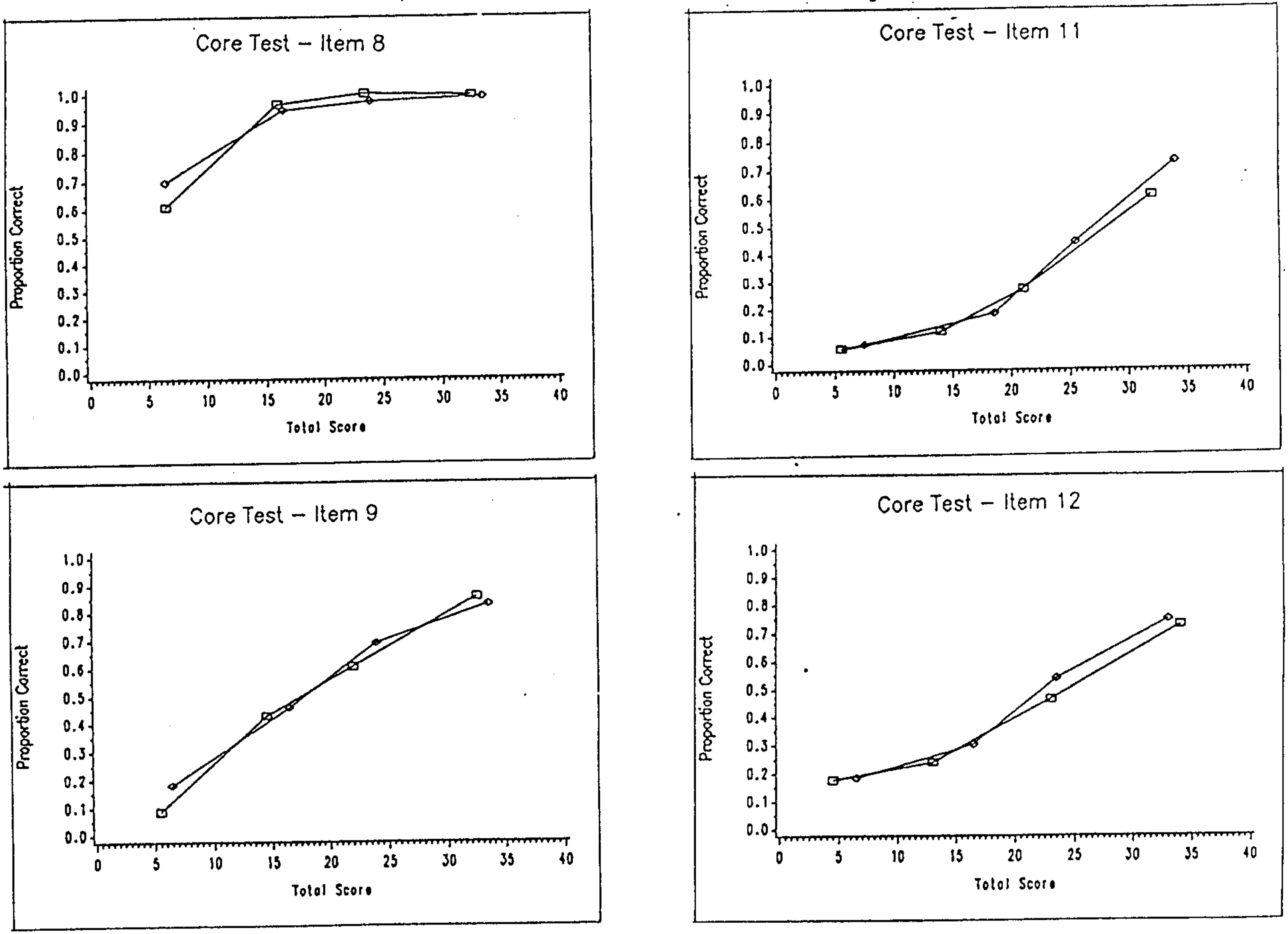
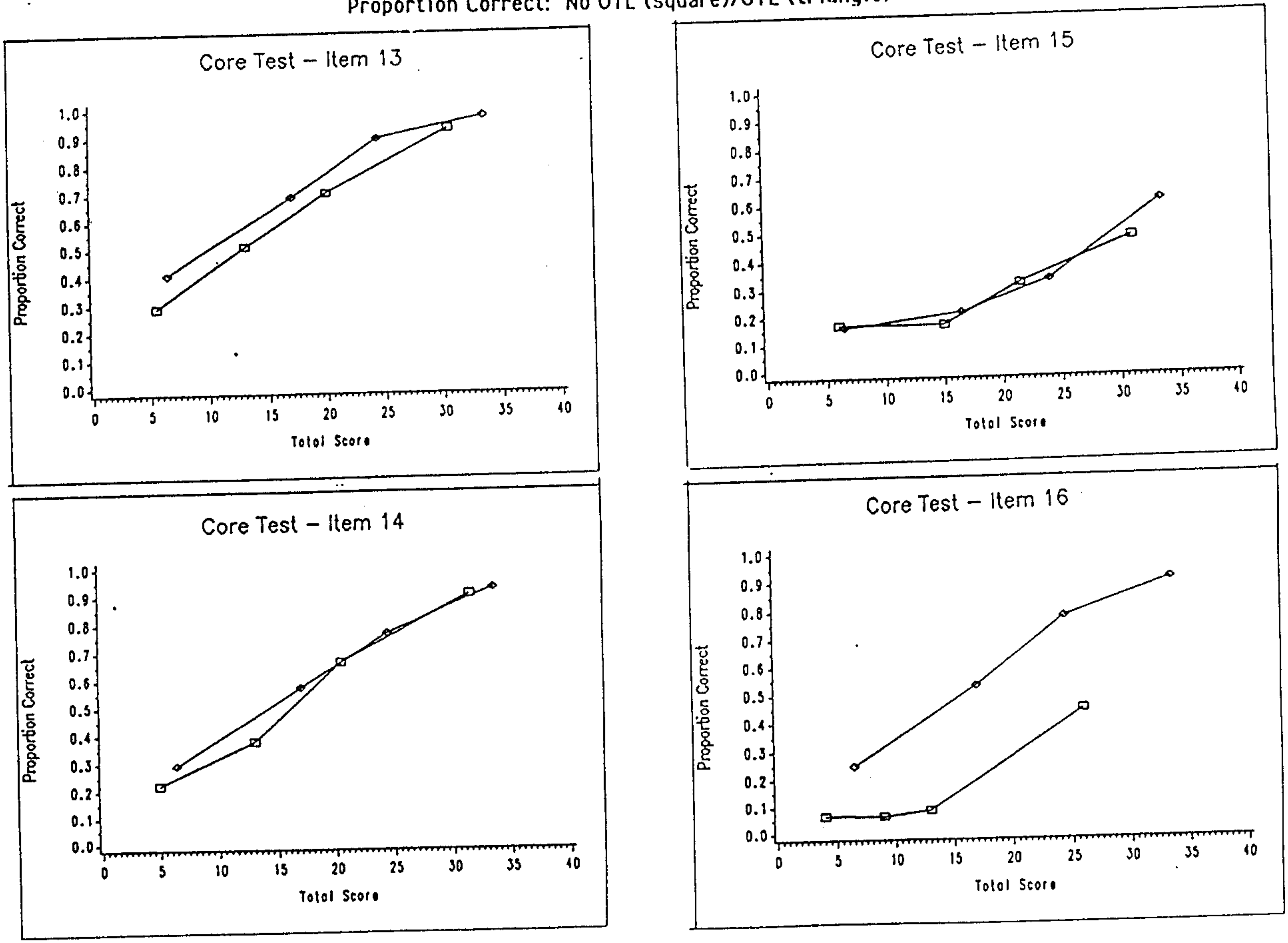


FIGURE 4

Proportion Correct: No OTL (square)/OTL (triangle)



Core Test – Item 13

Core Test – Item 15

Core Test – Item 14

Core Test – Item 16

FIGURE 5

Proportion Correct: No OTL (square)/OTL (triangle)

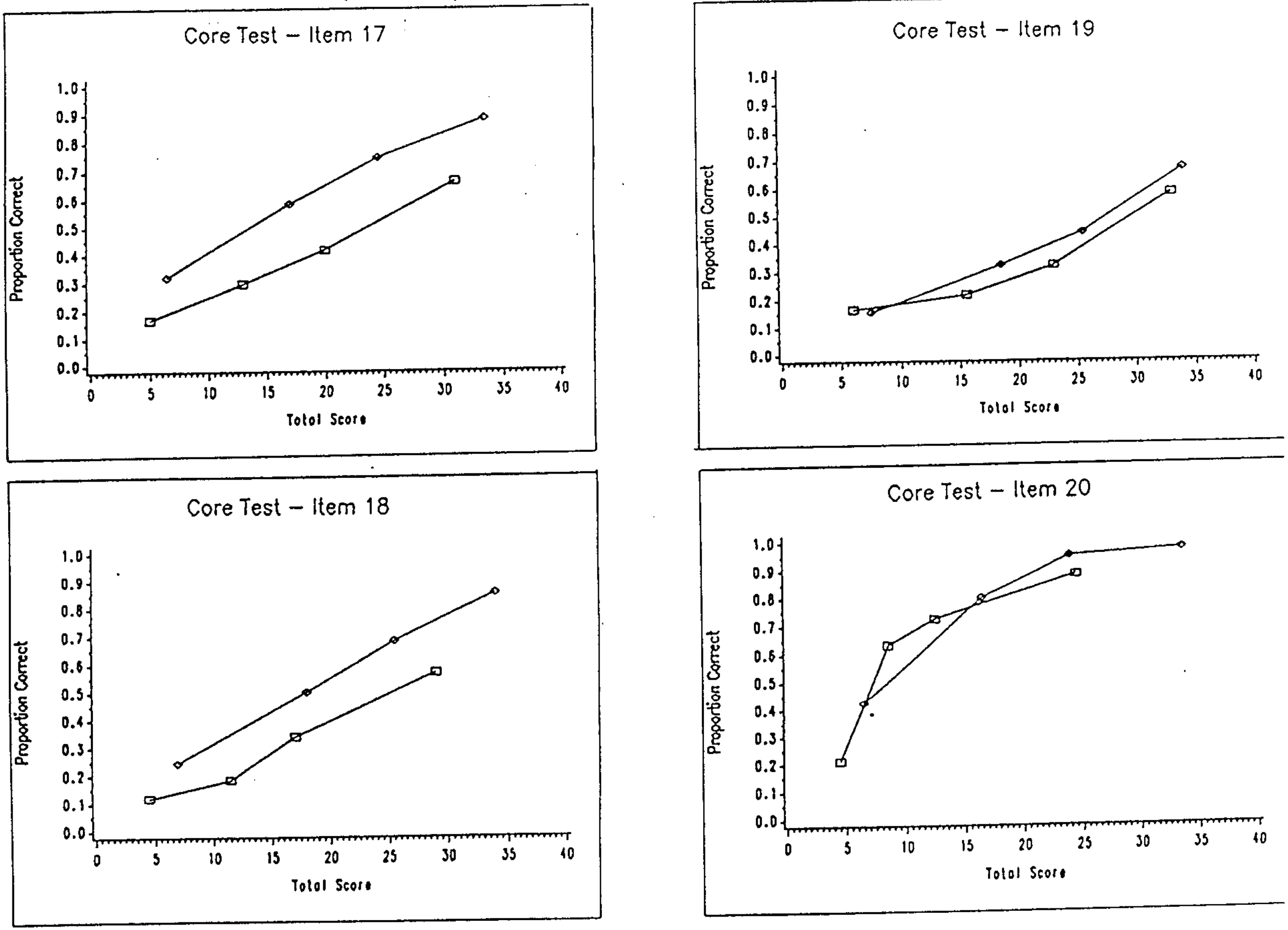


FIGURE 6

Proportion Correct: No OTL (square)/OTL (triangle)

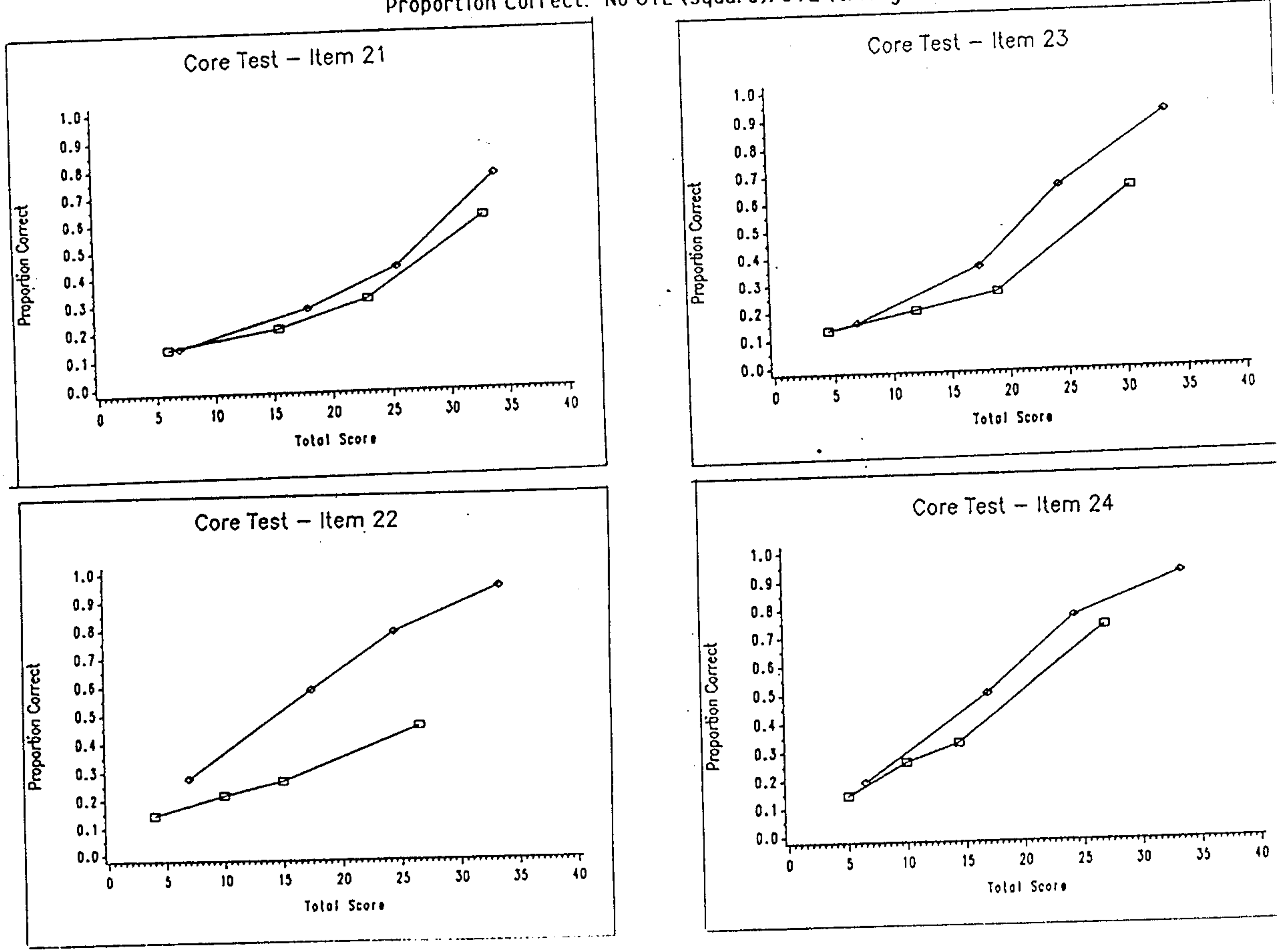


FIGURE 7

Proportion Correct: No OTL (square)/OTL (triangle)

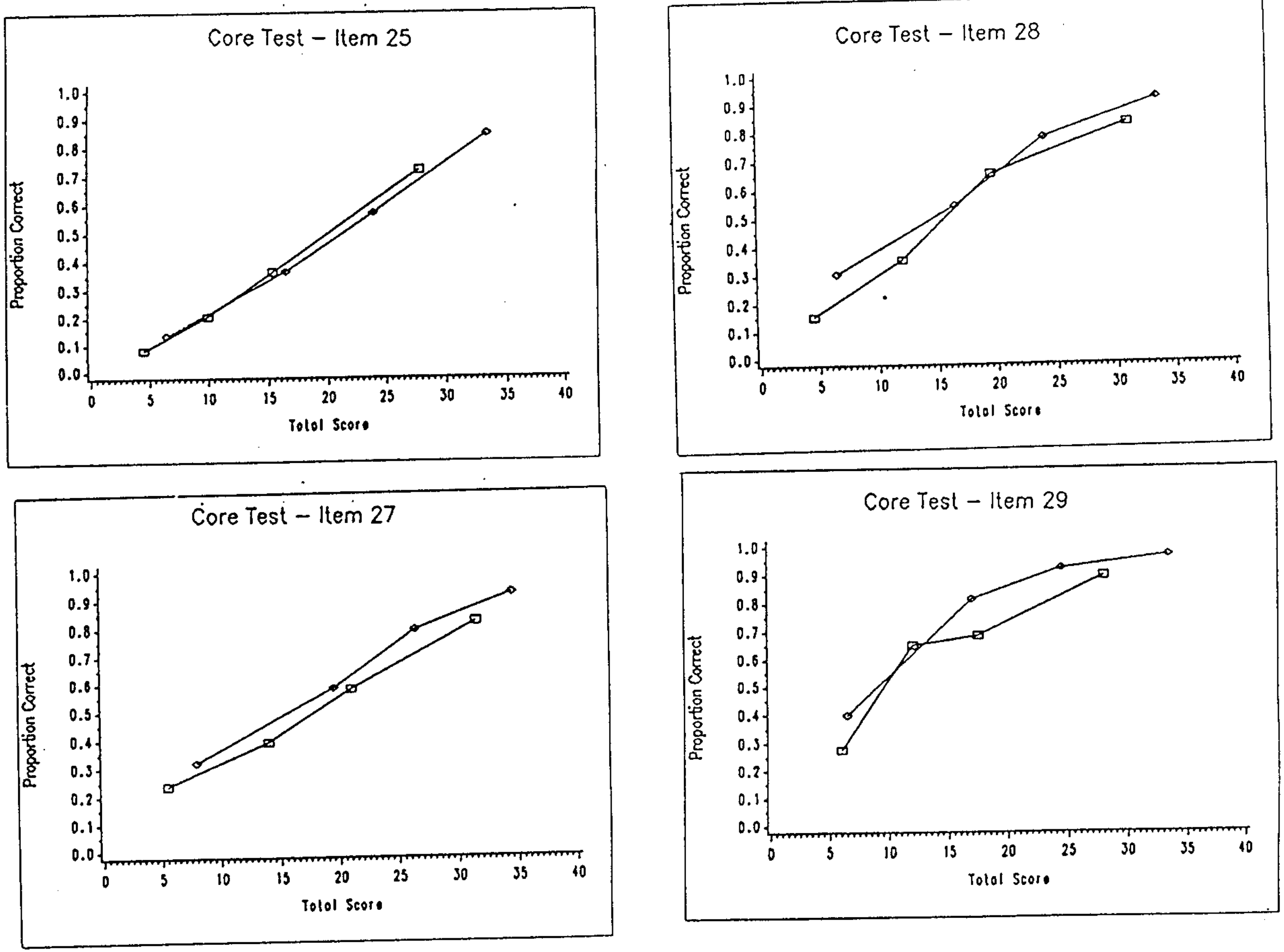


FIGURE 8

Proportion Correct: No OTL (square)/OTL (triangle)

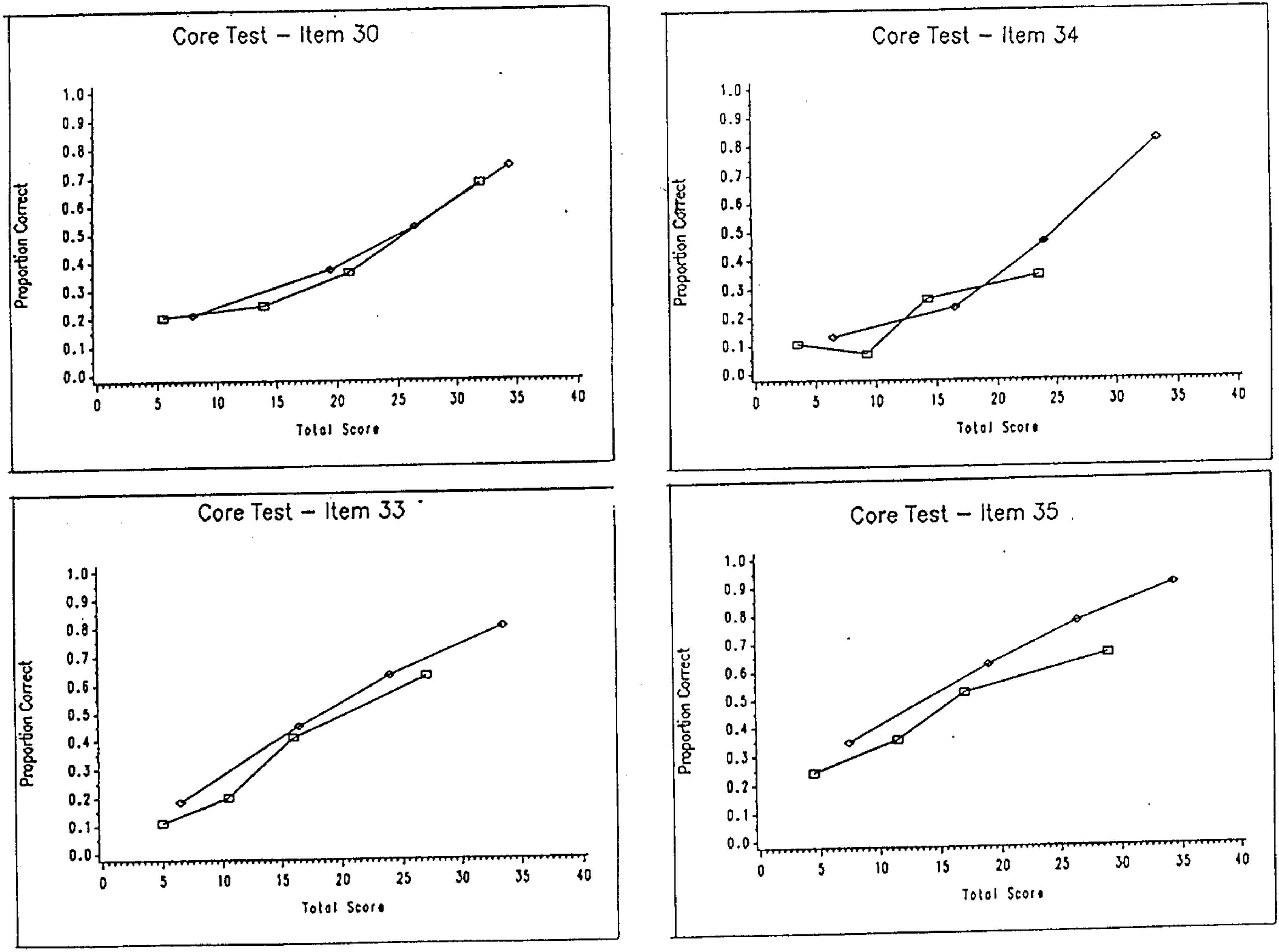


FIGURE 9

Proportion Correct: No OTL (square)/OTL (triangle)

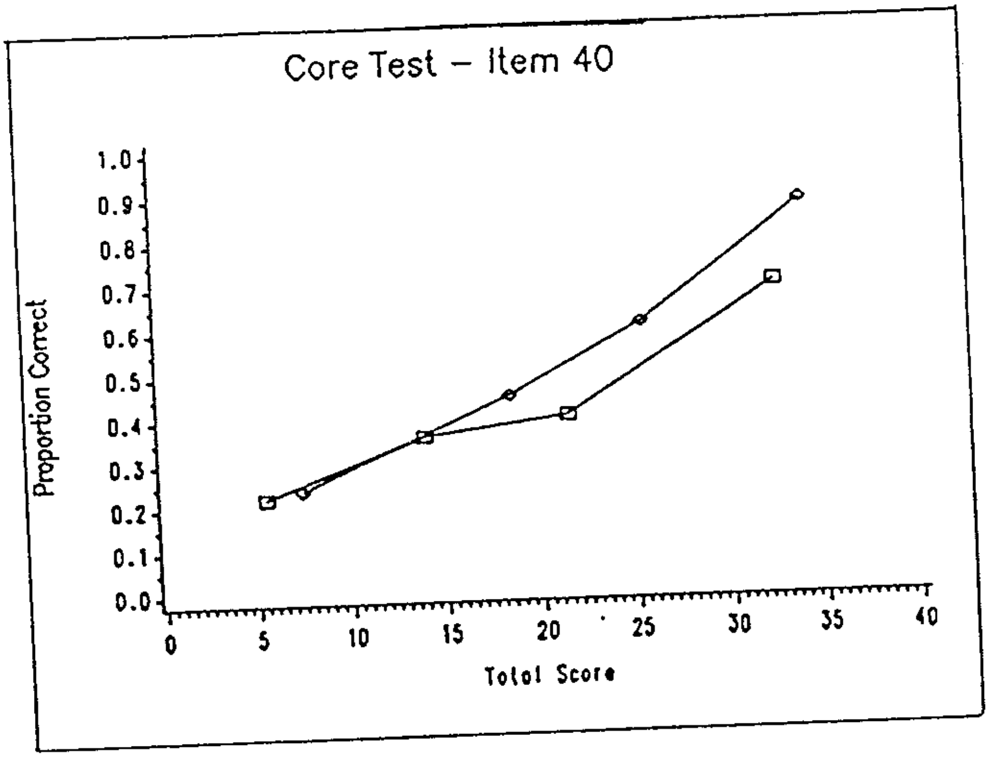


FIGURE 10

for Item 17

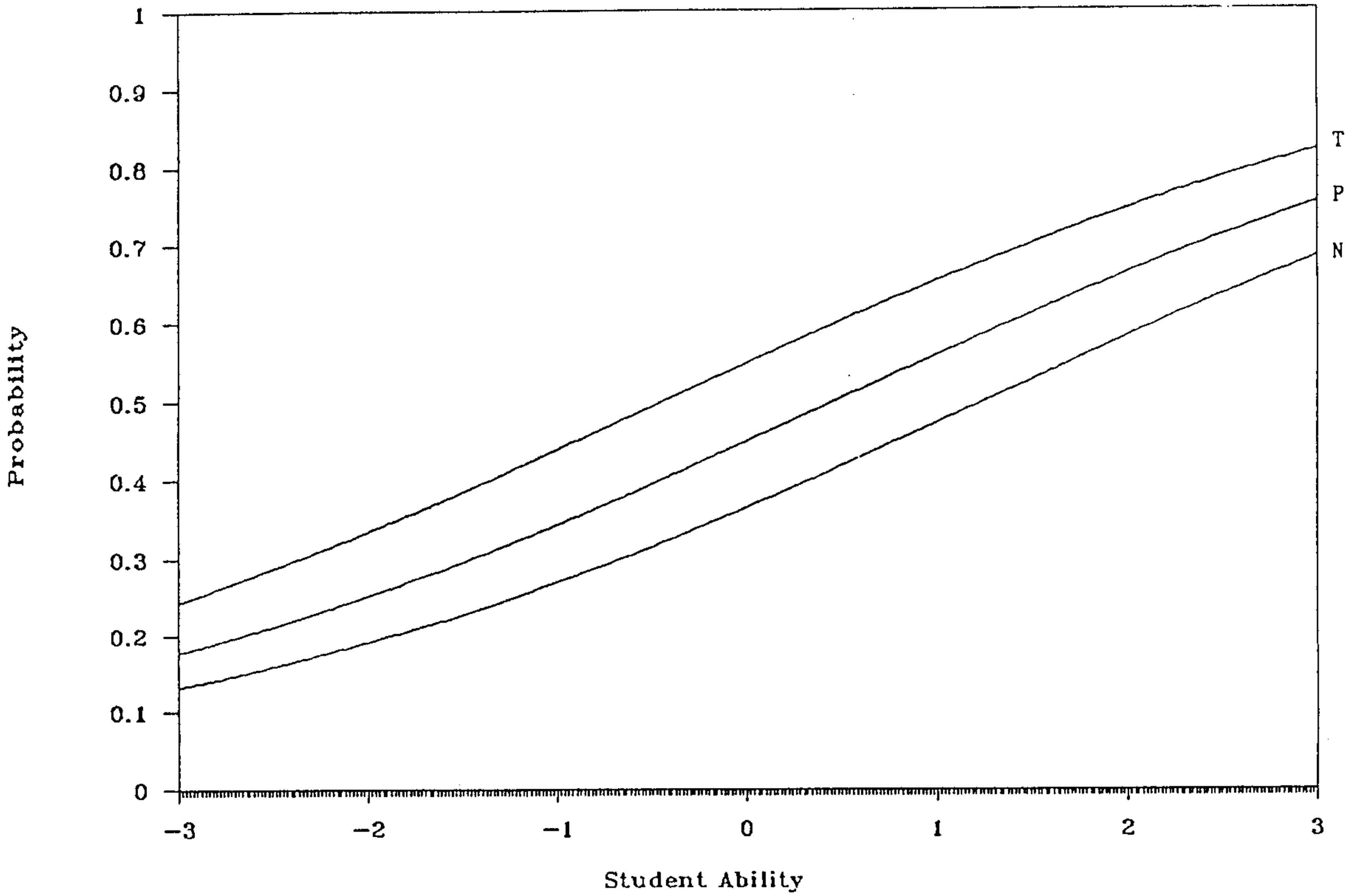


FIGURE 11

Scree Plot for Tetrachoric Correlations
with Artificial Model for 40 Items

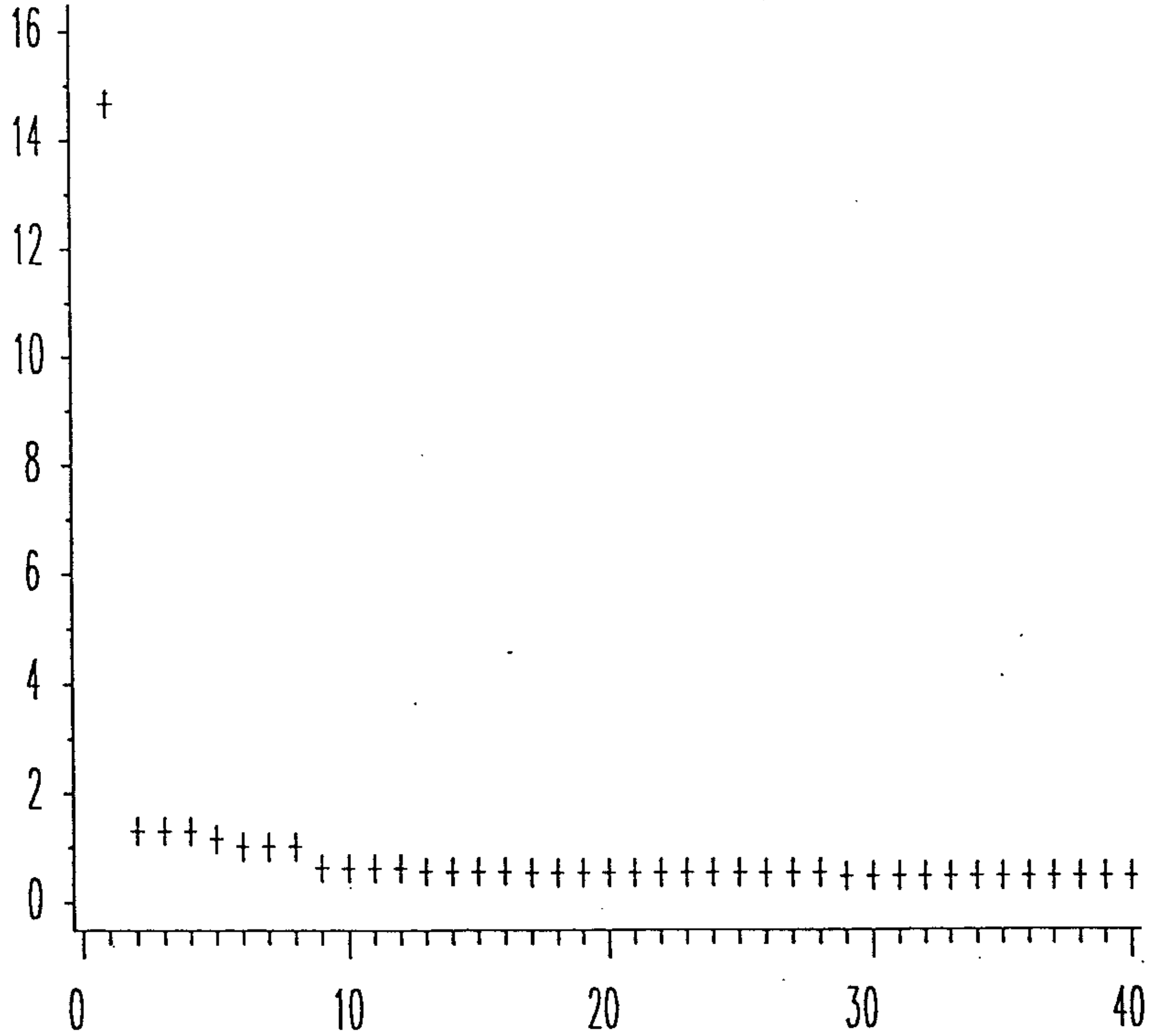
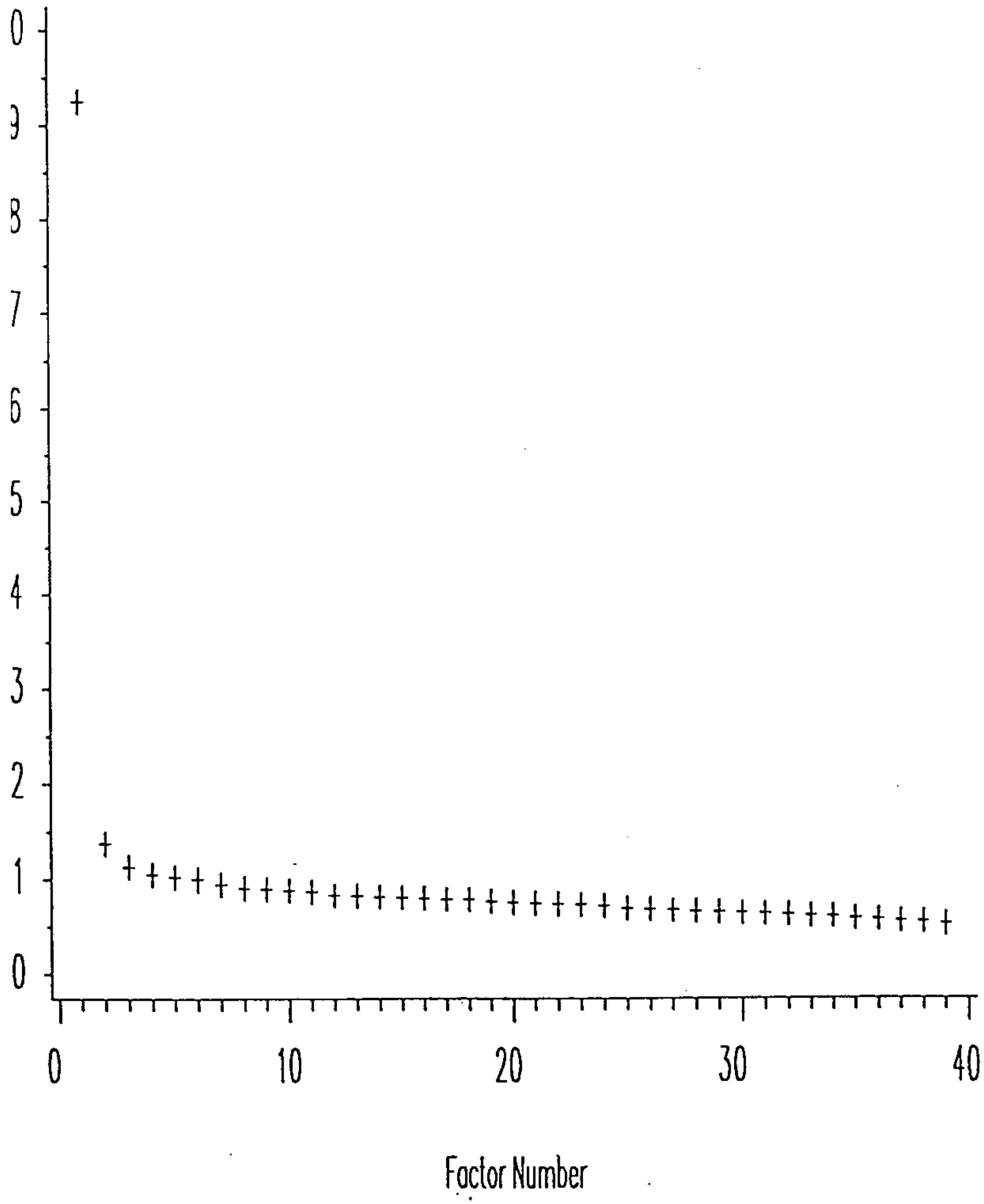


FIGURE 12

Scree Plot of Latent Roots for 39 Items Based on Tetrachorics



Appendix B

Wording of the 40 Core Items

1. 2 meters + 3 millimeters is equal to

A 2.0003 meters

B 2.003 meters

C 2.03 meters

D 2.3 meters

E 5 meters

2. $\frac{1}{5}$ is equal to

A 0.20%

B 2%

C 5%

D 20%

E 25%

3. If $5x + 4 = 4x - 31$, then x is equal to

A -35

B -27

C 3

D 27

E 35

4. Four 1-liter bowls of ice cream were set out at a party. After the party, 1 bowl was empty, 2 were half full, and 1 was three quarters full. How many liters of ice cream had been EATEN?

A $3\frac{3}{4}$

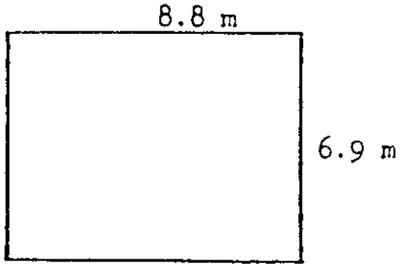
B $2\frac{3}{4}$

C $2\frac{1}{2}$

D $1\frac{3}{4}$

E None of these

5.



Which of the following is the closest approximation to the area of the rectangle with measurements given?

A $48\ m^2$

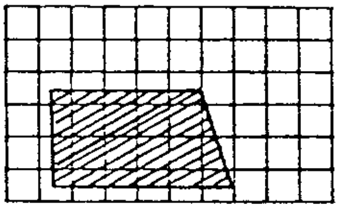
B $54\ m^2$

C $56\ m^2$

D $63\ m^2$

E $72\ m^2$

6.



□ 1 square unit

The area of the shaded figure, to the nearest square unit, is

A 23 square units

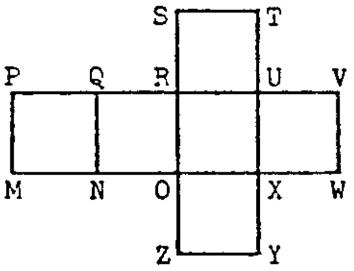
B 20 square units

C 18 square units

D 15 square units

E 12 square units

7.



The diagram shows a cardboard cube which has been cut along some edges and folded out flat. If it is folded to again make the cube, which two corners will touch corner P?

A corners Q and S

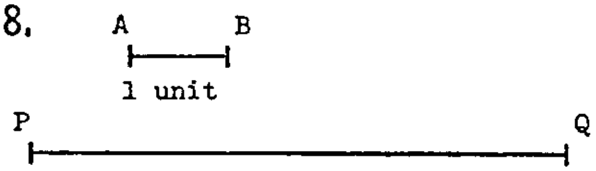
B corners T and Y

C corners W and Y

D corners T and V

E corners U and Y

8.



The length of $\overline{AB}$ is 1 unit. Which is the best estimate for the length of $\overline{PQ}$?

A 2 units

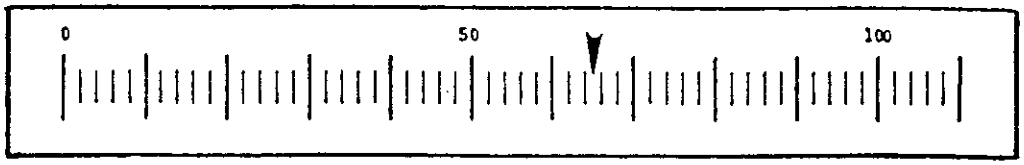
B 6 units

C 10 units

D 14 units

E 18 units

9.



On the above scale the reading indicated by the arrow is between

A 51 and 52

B 57 and 58

C 60 and 62

D 62 and 64

E 64 and 66

10. A solid plastic cube with edges 1 centimeter long weighs 1 gram. How much will a solid cube of the same plastic weigh if each edge is 2 centimeters long?

A 8 grams

B 4 grams

C 3 grams

D 2 grams

E 1 gram

11. On a number line two points *A* and *B* are given. The coordinate of *A* is -3 and the coordinate of *B* is +7. What is the coordinate of the point *C*, if *B* is the midpoint of the line segment $\overline{AC}$?

A -13

B $-\frac{1}{2}$

C +2

D +12

E +17

12. A painter is to mix green and yellow paint in the ratio of 4 to 7 to obtain the color he wants. If he has 28 liters of green paint, how many liters of yellow paint should be added?

A 11

B 16

C 28

D 49

E 196

13. If P = LW and if P = 12 and L = 3, then W is equal to

A $\frac{3}{4}$

B 3

C 4

D 12

E 36

14. A model boat is built to scale so that it is $\frac{1}{10}$ as long as the original boat. If the width of the original boat is 4 meters, the width of the model should be

A 0.1 meter

B 0.4 meter

C 1 meter

D 4 meters

E 40 meters

15. The value of 0.2131 × 0.02958 is approximately

A 0.6

B 0.06

C 0.006

D 0.0006

E 0.00006

16. $(-2) \times (-3)$ is equal to

A -6

B -5

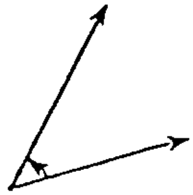
C -1

D 5

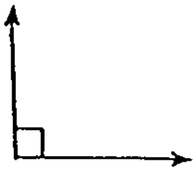
E 6

17. Which of the indicated angles is ACUTE ?

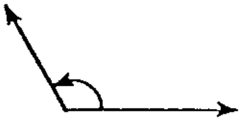
A



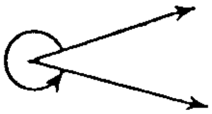
B



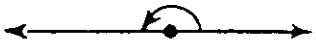
C



D



E



18. If $\frac{4x}{12} = 0$, then x is equal to

A 0

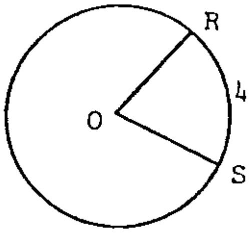
B 3

C 8

D 12

E 16

19.



The length of the circumference of the circle with center O is 24, and the length of arc RS is 4. What is the measure in degrees of the central angle ROS ?

A 24

B 30

C 45

D 60

E 90

20. In the discus-throwing competition, the winning throw was 61.60 meters. The second place throw was 59.72 meters. How much longer was the winning throw than the second place throw?

A 1.12 meters

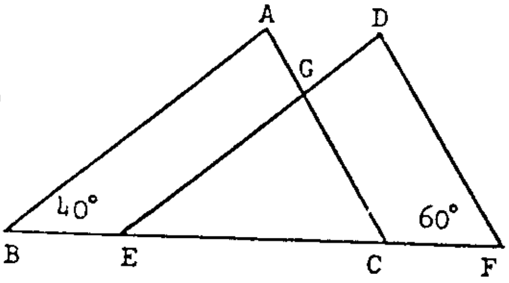
B 1.88 meters

C 1.92 meters

D 2.12 meters

E 121.32 meters

21.



In the above diagram, triangles ABC and DEF are congruent, with BC = EF. What is the measure of angle EGC ?

A 20°

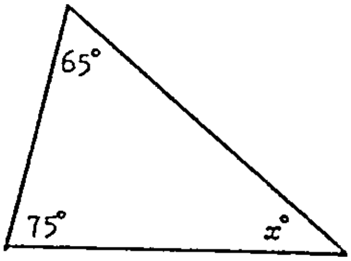
B 40°

C 60°

D 80°

E 100°

22.



x is equal to

A 75

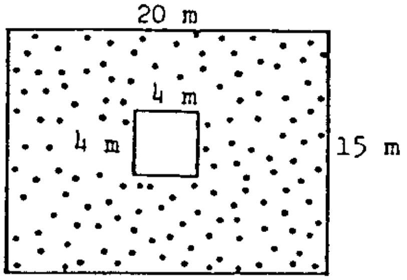
B 70

C 65

D 60

E 40

23.



A square is removed from the rectangle as shown. What is the area of the remaining part?

A 316 m^2

B 300 m^2

C 284 m^2

D 80 m^2

E 16 m^2

24.

Cloth is sold by the square meter. If 6 square meters of cloth cost $4.80, the cost of 16 square meters will be

A $12.80

B $14.40

C $28.80

D $52.80

E $128.00

25. The air temperature at the foot of a mountain is 31 degrees. On top of the mountain the temperature is -7 degrees. How much warmer is the air at the foot of the mountain?

A -38 degrees

B -24 degrees

C 7 degrees

D 24 degrees

E 38 degrees

26. 0.40 × 6.38 is equal to

A .2552

B 2.452

C 2.552

D 24.52

E 25.52

27. A shopkeeper has x kg of tea in stock. He sells 15 kg and then receives a new lot weighing $2y$ kg. What weight of tea does he now have?

A $x - 15 - 2y$

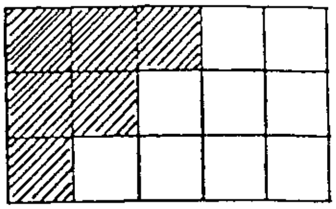
B $x + 15 + 2y$

C $x - 15 + 2y$

D $x + 15 - 2y$

E None of these

28.



In the figure the little squares are all the same size and the area of the whole rectangle is equal to 1. The area of the shaded part is equal to

A $\frac{2}{15}$

B $\frac{1}{3}$

C $\frac{2}{5}$

D $\frac{3}{8}$

E $\frac{1}{2}$

29. When using the metric system, the distance between two towns is usually measured in

A millimeters

B centimeters

C decimeters

D meters

E kilometers

30. The table below compares the height from which a ball is dropped (d) and the height to which it bounces (b).

d	50	80	100	150
b	25	40	50	75

Which formula describes this relationship?

A $b = d^2$

B $b = 2d$

C $b = \frac{d}{2}$

D $b = d + 25$

E $b = d - 25$

31. $\frac{2}{5} + \frac{3}{8}$ is equal to

A $\frac{5}{13}$

B $\frac{5}{40}$

C $\frac{6}{40}$

D $\frac{16}{15}$

E $\frac{31}{40}$

32. $7\,\frac{3}{20}$ is equal to

A 7.03

B 7.15

C 7.23

D 7.3

E 7.6

33. In a school of 800 pupils, 300 are boys. The ratio of the number of boys to the number of girls is

A 3 : 8

B 5 : 8

C 3 : 11

D 5 : 3

E 3 : 5

34. 20 is what percent of 80 ?

A 4%

B 20%

C 25%

D 40%

E None of these

35. The sentence *"a number x decreased by 6 is less than 12"* can be written as the inequality

A $x - 6 > 12$

B $x - 6 \geq 12$

C $x - 6 < 12$

D $6 - x \geq 12$

E $6 - x < 12$

36. 30 is 75% of what number?

A 40

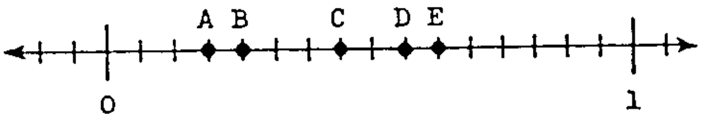
B 90

C 105

D 225

E 2250

37. Which of the points A, B, C, D, E on this number line corresponds to $\frac{5}{8}$?



A point A

B point B

C point C

D point D

E point E

38. 20% of 125 is equal to

A 6.25

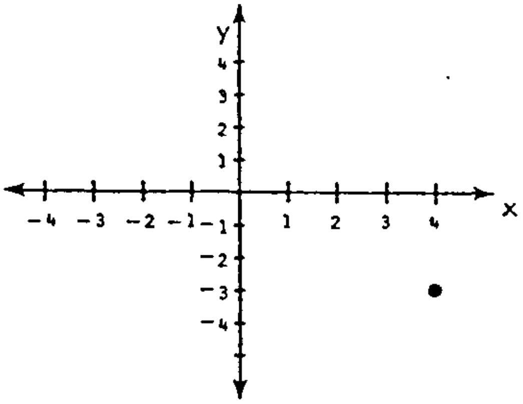
B 12.50

C 15

D 25

E 50

39.



What are the coordinates of point P ?

A (-3,4)

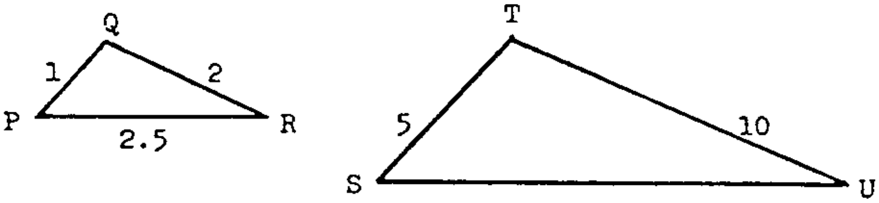
B (-4,-3)

C (3,4)

D (4,-3)

E (-4,3)

40.



Triangles PQR and STU are similar. How long is $\overline{SU}$?

A 5

B 10

C 12.5

D 15

E 25