DOCUMENT RESUME

ED 382 667 TM 023 105

AUTHOR Sheehan. Kathleen; Mislevy, Robert J.
TITLE A Tree-Based Analysis of Items from an Assessment of Basic Mathematics Skills.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-14
PUB DATE Apr 94
NOTE 42p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Basic Skills; *Computer Assisted Testing; *Difficulty Level; Educational Assessment; Identification; Interaction; *Mathematics Skills; Multiple Choice Tests; Pretests Posttests; Problem Solving; *Test Construction; Test Items
IDENTIFIERS Asymptotic Distributions; Free Response Test Items; *Praxis I; Three Parameter Model; *Tree Based Analysis

ABSTRACT

The operating characteristics of 114 mathematics pretest items from the Praxis I: Computer Based Test were analyzed in terms of item attributes and test developers' judgments of item difficulty. Item operating characteristics were defined as the difficulty, discrimination, and asymptote parameters of a three parameter logistic item response theory (IRT) model. Three types of item attributes were considered: (1) surface features (for example, whether or not the item stem contained an equation); (2) aspects of the solution process (for example, whether or not the solution required application of a standard formula); and (3) response type (free-response or multiple-choice). Because the attribute set included large numbers of categorical variables, an approach based on binary regression trees (Breiman, Friedman, Olshen, and Stone, 1984) was implemented. The results were quite impressive for asymptote parameters (85% of variance explained), somewhat less so for difficulty parameters (36% of variance explained) and fairly unimpressive for discrimination parameters (only 12% of variance explained). In addition, the tree-based approach was found to be particularly useful for identifying important interaction effects and for developing graphical summaries of the modeling results. Six tables and eight figures support the analyses. (Contains 11 references.) (Author/SLD)

RESEARCH REPORT

A TREE-BASED ANALYSIS OF ITEMS FROM AN ASSESSMENT OF BASIC MATHEMATICS SKILLS

Kathleen Sheehan
Robert J. Mislevy

Educational Testing Service
Princeton, New Jersey
April 1994

A Tree-Based Analysis of Items From An Assessment Of Basic Mathematics Skills

Kathleen Sheehan and Robert J. Mislevy

Educational Testing Service

March, 1994

This work was supported by Educational Testing Service through the Program Research Planning Council. The analyses of the Praxis I: CBT Mathematics pretest items were supported by ETS development funds and were carried out in collaboration with Valerie Folk, Kathleen Martin, Mary Morley & Judy Smith.

A Tree-Based Analysis of Items From An Assessment of Basic Mathematics Skills

The operating characteristics of 114 Mathematics pretest items from the Praxis I: Computer Based Test were analyzed in terms of item attributes and test developers' judgements of item difficulty. Item operating characteristics were defined as the difficulty, discrimination and asymptote parameters of a three parameter logistic IRT model. Three types of item attributes were considered: surface features (for example, whether or not the item stem contained an equation); aspects of the solution process (for example, whether or not the solution required application of a standard formula); and response type (free-response or multiple-choice). Because the attribute set included large numbers of categorical variables, an approach based on binary regression trees (Breiman, Friedman, Olshen, and Stone, 1984) was implemented. The results were quite impressive for asymptote parameters (85% of variance explained), somewhat less so for difficulty parameters (36% of variance explained) and fairly unimpressive for discrimination parameters (only 12% of variance explained). In addition, the tree-based approach was found to be particularly useful for identifying important interaction effects and for developing graphical summaries of the modeling results.

A Tree-Based Analysis of Items From An Assessment of Basic Mathematics Skills

The goal of this study was to determine the degree to which the operating characteristics of basic mathematics achievement test items could be predicted from an analysis of item attributes and test developers' judgements of item difficulty. Items' operating characteristics were defined as the difficulty, discrimination and asymptote parameters of the three parameter logistic (3PL) IRT model. Three types of item attributes were considered: surface features of the items (for example, whether or not the item stem included an equation); aspects of the solution process (for example, whether or not the solution required application of a standard formula); and response format (free-response or multiple-choice). Studies of this type may be conducted for a variety of reasons including: (1) reducing sample size requirements for item calibration (Mislevy, Sheehan & Wingersky, 1993); (2) providing for more systematic test design and construct validation (Embretson & Wetzel, 1987; and Bejar, 1991); and (3) diagnosing students' misconceptions (Tatsuoka, 1987, 1990).

The analyses reported in this paper were conducted using a combination of least-squares regression analysis and binary regression trees (Breiman, Friedman, Olshen, and Stone, 1984). Regression analysis has been used in numerous studies of the components of item difficulty (see for example, Enright, Allen & Kim, 1993; Scheuneman, Gerritz & Embretson, 1991; Sheehan & Mislevy, 1990; and Tatsuoka, 1987). This paper introduces tree-based models as an exploratory technique for determining the structure of the regression equation and for developing graphical summaries of the modeling results.

The Tree-Based Approach

For problems involving a single numeric response (y) and a set of predictor variables (x) a binary regression tree is fit by successively splitting the data on the basis of the independent variables into binary subsets with similar values of the response variable. At each stage of model fitting, the splitting algorithm considers all possible splits of all possible predictor variables. When the potential predictor is a multi-level categorical variable, as was the case for several of the variables considered in this study, the splitting algorithm considers all collapsing strategies resulting in exactly two levels. When the potential predictor is a numeric variable, such as the item difficulty rating considered in this study, the splitting algorithm considers all possible cut points for grouping the observations into low and high subsets. Potential splits are evaluated in terms of deviance, a statistical measure of the dissimilarity in the response variable among the observations belonging to a single subset. At each stage of splitting, the original subset of observations is referred to as the parent node and the two outcome subsets are referred to as the left and right child nodes. The best split is the one that produces the largest decrease between the deviance of the parent node and the sum

of the deviances in the two child nodes. The deviance of the parent node is calculated as the sum of the deviances of all of its members,

$$D(y,\hat{y}) = \sum (y_i - \hat{y})^2$$

where $\hat{y}$ is the mean value of the response calculated from all of the obervations in the node. The deviance of a potential split is calculated as

$$D(y,\hat{y}_L,\hat{y}_R) = \sum_L D(y_i,\hat{y}_L) + \sum_R D(y_i,\hat{y}_R)$$

$$= \sum_L (y_i - \hat{y}_L)^2 + \sum_R (y_i - \hat{y}_R)^2$$

where $\hat{y}_L$ is the mean value of the response in the left child node and $\hat{y}_R$ is the mean value of the response in the right child node. The split that maximizes the change in deviance

$$\Delta D = D(y,\hat{y}) - D(y,\hat{y}_L,\hat{y}_R)$$

is the split chosen at a given node. In the final fitted model, the predicted value for each observation is the mean response calculated from only those observations belonging to the same terminal node.

Figure 1 provides a graphical representation of a tree model estimated for a hypothetical set of 20 observations. In this particular representation, the number of observations in each node is plotted as the node label and the variables used to define each split are indicated on the lines connecting parents to children. Node locations indicate the predicted value of the response variable (read from the horizontal axis) and the estimated deviance value (read from the vertical axis).

==========

Insert Figure 1 Here

==========

As can be seen, the model has two splits yielding a total of three terminal nodes. The first split divides the data into subsets based on values of the categorical variable x_1. Observations with values of x_1 equal to A or B (denoted x_1=AB) are classified into the left child node. Observations with values of x_1 equal to C or D (denoted x_1=CD) are classified into the right child node. The horizontal distance between the left child node and the right child node is the amount by which the predicted response for observations of type A or B differs from the predicted response for observations of type C or D. The second split divides the set of ten observations in the x_1=AB node into subsets based on values of the second independent variable x_2. The six observations with x_1=AB and x_2<=10 are classified into one

subset; the four observations with x_1=AB and x_2>10 are classified into a second subset. There are no further splits of the x_1=CD node indicating that x_2 was only helpful at predicting the value of y for observations with x_1=AB. This type of interaction is common in problems involving several independent predictors. The final fitted model is specified in terms of the following three prediction rules (corresponding to the three terminal nodes in Figure 1):

$$\textit{if } x_1 = AB \textit{ and } x_2 \le 10 \textit{ then } \hat{y} = \frac{1}{6}\sum_{i=1}^{6} y_i$$

$$\textit{if } x_1 = AB \textit{ and } x_2 > 10 \textit{ then } \hat{y} = \frac{1}{4}\sum_{i=7}^{10} y_i$$

$$\textit{if } x_1 = CD \textit{ then } \hat{y} = \frac{1}{10}\sum_{i=11}^{20} y_i$$

The various splits shown in Figure 1 represent the optimal sequence of splits determined from a consideration of all possible remaining splits at each stage of fitting. Splits of binary variables require a single evaluation. Splits of multi-level categorical variables require 2^{k-1}-1 evaluations, where k is the number of levels. For example, a categorical variable with 3 levels (A, B, and C) would be evaluated at three possible binary cuts (A vs. BC, AB vs. C, and B vs. AC). Splits of numeric variables must be evaluated between each successive pair of ordered observations, a total of n-1 evaluations, where n is the number of observations in the node (excluding ties).

The thoroughness of this approach to model selection can be appreciated by noting that, even for the simple example presented above, which included only two variables and 20 observations, as many as 46 separate evaluations were required to determine the optimal model structure. The determination of the best initial split required 26 evaluations: 7 for the categorical variable x_1, and 19 for the numeric variable x_2. The determination of the best subsequent split of the x_1=AB node required 10 additional evaluations: one to determine whether or not to continue splitting based on x_1 (potentially yielding an x_1=A node and an x_1=B node), and nine to evaluate potential splits based on x_2. And finally, although the final model did not include any subsequent splits of the x_1=CD node, the decision to leave this node intact required 10 additional evaluations: one to evaluate a subsequent split based on x_1 and nine to evaluate a subsequent split based on x_2.

An Example: The Praxis I Mathematics Item Pool

The Praxis I: CBT measures the mathematics, reading and writing skills of prospective teachers during their college years. Our example concerns a pool of 510 mathematics items which were pretested in the Fall and Winter of 1992. The field test was structured so that examinees were administered overlapping subsets of items. This was accomplished by dividing the original item pool into representative 17-item blocks and

administering three blocks of items to each examinee. Under this design, each examinee received 51 items, and each item was administered to approximately 900 examinees. The entire pool of 510 items was then calibrated using a 3PL model, fit by means of Mislevy and Bock's (1983) BILOG program. A representative subset of 114 items was subsequently selected for use in the analysis of items' operating characteristics. This subset included 48 free-response items and 66 multiple-choice items.

Item Attributes

Item attributes were developed by asking members of the ETS Test Development staff to list surface features of the items and aspects of the solution process which would be expected to contribute to item difficulty. The resulting attribute list included 13 item feature variables and 13 solution process variables. Two members of the staff whose duties included the writing of similar types of items were then asked to rate each of the items on each of the item feature variables and each of the solution process variables. Raters were also asked to provide overall ratings of item difficulty expressed on a 1 to 5 scale. Except where noted, subsequent analyses are based on the average of the two sets of ratings obtained.

Information about item content was also available. In particular, each item was classified as belonging to one of five content areas:

A. Number Sense and Operations
B. Mathematical Relationships
C. Data Interpretation
D. Geometry and Measurement
E. Reasoning

The item feature variables and the solution process variables are listed in Tables 1 and 2 along with frequency statistics, rater agreement statistics and correlations with item parameters. (Attribute abbreviations are given in parentheses.) Rater agreement was fairly high, greater that 90% for all but one of the surface feature variables and averaging about 82% for the solution process variables. Correlations are reported for all of the items combined (n=114) and for subsets defined by content area and response format. These subsets were suggested by the tree analyses reported below.

The global judgements of item difficulty provided by the two raters are summarized in Table 3. For 92% of the items, the difference between the ratings provided by the two raters was less than or equal to one point (on a five point scale). Table 3 also provides correlations with item parameters calculated for all of the items combined and for items grouped by response format. For the set of all items combined, item difficulty was more highly correlated with the average difficulty rating than with either of the two individual ratings (.47 vs .40 or .43). The individual correlations show that the two raters were differentially adept at rating items with different response formats. In particular, Rater 1 was more adept at

rating the free-response items and Rater 2 was more adept at rating the multiple-choice items.

Insert Tables 1-3 About Here

Analysis of Item Difficulty

Our investigation into the components of item difficulty was conducted using a combination of tree-based modeling and regression analysis. Tree-based modeling can be considered as an exploratory technique for uncovering structure in data (Clark & Pregibon, 1992). In this study, tree models are used to identify important interaction effects, to select subsets of variables for consideration in subsequent regression analyses, and to provide graphic displays of the modeling results.

The tree-based analysis of item difficulty was conducted in stages. The set of predictor variables considered in the initial stage of the analysis consisted of all of the item attributes described above except for the item difficulty ratings provided by the two raters. The difficulty rating data was intentionally excluded to avoid swamping the information available from the other item attributes. This strategy allowed several interesting features of the data to be revealed.

Most of the attributes considered in this study were originally scored on a binary scale. Consequently, the average attribute scores considered in the tree-based analyses were specified on a three-point scale: 1 = both raters agreed that the feature was present; 0 = both raters agreed that the feature was not present; and 0.5 = the two raters disagreed on whether or not the feature was present. Potential splits of these variables were evaluated twice: once with disagreements grouped with 1s (feature present); and once with disagreements grouped with 0s (feature not present). As will be seen, the optimal grouping varied from one attribute to another and from one analysis to another.

Figure 2 provides a graphical representation of a tree model developed to predict item difficulty from the surface feature variables and the solution process variables. The predicted difficulty value associated with each node can be read from the horizontal axis. The item attributes used to define each split are indicated on the lines connecting parents to children. Split definitions also indicate the optimal treatment of rater disagreement. Split definitions of the form "attribute<1" and "attribute=1" indicate that, for that attribute, the disagreement items were grouped with the items coded as "feature not present". Split definitions of the form "attribute=0" and "attribute>0" indicate that, for that attribute, the disagreement items were grouped with the items coded as "feature present". In each case, the grouping selected was the one which provided the best prediction.

Insert Figure 2 Here

As can be seen, the first split divides the items into subsets based on values of the content area variable: the 61 items classified as content area A (Number Sense and Operations) or C (Data Interpretation) are assigned to the left child node; the 53 items classified as content area B (Mathematical Relationships), D (Geometry and Measurement), or E (Reasoning) are assigned to the right child node. The AC node is subsequently split into the 55 items rated as routine applications (NONROU<1) and the 6 items rated as nonroutine applications (NONROU=1) indicating that, although AC items are generally easier than BDE items, those AC items rated as nonroutine applications are among the most difficult items in the pool. Although the routine/nonroutine variable is highly predictive of the difficulty of AC items, the fact that it does not appear among the variables selected to define further splits of the BDE node indicates that it provides minimal information about gradations of difficulty among BDE items. As a matter of fact, Figure 2 shows that there is no overlap whatsoever between the subset of variables selected for the prediction model for AC items and the subset of variables selected for the prediction model for BDE items! Confirmation of this unexpected result can be found in Tables 1 and 2 which provide correlation coefficients calculated separately for the AC items and the BDE items. In all but one case, variables which are significantly correlated with the difficulty of AC items are not significantly correlated with the difficulty of BDE items and conversely, variables which are significantly correlated with the difficulty of BDE items are not significantly correlated with the difficulty of AC items. In addition, the magnitude of the correlations calculated from the appropriate subset (either AC items or BDE items) are greater than those calculated from the combined set of 114 items. The one exception noted concerns the solution process variable "Apply standard algorithm in a nonstandard manner". This variable is significantly correlated with both types of items but only appears in the tree model estimated for the BDE items. This discrepancy can be explained by the correlation of this variable with several of the surface feature variables.

A regression analysis was conducted to evaluate the predictive capability of the solution process variables (SPVs) and the item feature variables (IFVs). Thirty variables were considered in the analysis: (1) all of the SPVs and IFVs with average frequencies of at least five (11 SPVs and 9 IFVs); (2) a dummy variable used to distinguish AC items from BDE items; and (3) a set of 9 interaction terms. The interaction terms were defined by crossing Type=AC with five other variables (Word Problem, Order & Match, Histogram, Nonroutine Application, and Recall or Recognize Facts) and by crossing Type=BDE with four other variables (Quantitative Comparison, Apply Standard Algorithm, Apply Standard Algorithm in Nonstandard Manner, and Apply Multistep Thinking). The results are presented as Model #1 in Table 4. Of the 30 variables originally considered, 8 were significant at an alpha level of 0.15, including four of the interaction terms suggested by the tree-based analysis. The estimated eight-variable model accounted for 28% of the variance in item difficulty.

Insert Table 4 Here

The analyses described above did not consider the information about item difficulty available from the global judgements of item difficulty provided by the two raters. Figure 3 presents the tree model obtained by adding the average difficulty rating (DR) to the set of variables considered previously. As can be seen, the average difficulty rating is now the most important predictor, accounting for the first several splits. Note that the average difficulty rating has divided the items into three distinct groups: the low group consists of items with average ratings between 1 and 2.5 inclusive, the medium group consists of items with average ratings between 3 and 4 inclusive, and the high group consists of items with average ratings between 4.5 and 5 inclusive. This grouping is highly correlated with item difficulty: low rated items tend to have estimated difficulties below -1.0; medium rated items tend to have estimated difficulties between -1.0 and 0.0; and high rated items tend to have estimated difficulties greater than 0.0. The most notable exception to this rule occurred for items involving a Quantitative Comparison (QC). Figure 3 shows that the difficulty of the QC items was consistently underrated. In particular, QC items with estimated difficulties in the medium range were given low ratings and QC items with estimated difficulties in the high range were given medium ratings.

Insert Figure 3 Here

Additional evidence of the raters' tendency to underrate the difficulty of QC items is provided in Figure 4, which depicts the least squares regression line estimated from the entire set of 114 items, along with points representing individual items. (QC items are plotted as circles, non-QC items are plotted as dots.) As can be seen, almost all of the QC items are underpredicted. The amount of variation in item difficulty accounted for by the regression on average difficulty rating was 21% (Model #2 in Table 4). When the regression was rerun with the QC variable included, the amount of variation accounted for increased to 29% (Model #3 in Table 4).

Insert Figure 4 Here

An additional analysis was conducted to determine whether any of the other item attributes provided improved prediction over and above that provided by the average difficulty rating and the QC variable. Using a stepwise procedure, four additional variables were selected. The additional variables included two SPVs (Apply Standard Algorithm and Translate Words to Symbols); one item feature variable (Histogram); and one interaction term

(type=AC crossed with Order and Match). Estimated coefficients are given as Model #4 in Table 4. The enhanced model accounted for 36% of the variation in item difficulty.

For practical applications requiring maximum predictive power, the enhanced model (Model # 4 in Table 4) is preferable, since it explains the most amount of variation in item difficulty. The other models provide useful information about what makes items easy or hard. In particular, residuals from the analytical model (Model #1 in Table 4) can be consulted for clues as to why some items are unexpectedly easy or hard, given the identifiable factors that are usually associated with item difficulty.

Analysis of Item Discrimination

The tree-based analysis of item discrimination considered all of the item attributes described previously. The fitted model is plotted in Figure 5. The first split shows that items containing equations (EQUA>0) tend to be more discriminating than those without, although this is not always the case. The most prominent exception occurs for multiple-choice items (MC=1) formulated as word problems (WORDP>0) which can not be solved through application of a standard algorithm (STDALG<1). The 15 items with this combination of attributes were among the most highly discriminating in the pool. The plot also shows that the least discriminating items were those which did not involve equations (EQUA=0) and could be solved through application of a standard algorithm (STDALG=1).

Insert Figure 5 Here

A linear prediction model for item discrimination was estimated using a stepwise regression procedure. The variables considered in the analysis included all of the item attributes listed in Tables 1 and 2 with average frequencies of at least five, plus three interaction terms suggested by the tree model. The interactions were defined as follows:

(1) MC*WORDP=1 if (MC=1 & WORDP>0), MC*WORDP=0 otherwise;
(2) NE*WORDP=1 if (EQUA=0 & WORDP>0), NE*WORDP=0 otherwise;
(3) NE*STDALG=1 if (EQUA=0 & STDALG=1), NE*STDALG=0 otherwise.

The estimated regression model included one of the item feature variables (EQUA), and two of the interaction terms (MC*WORDP and NE*STDALG). As shown in Table 5, EQUA and MC*WORDP have positive coefficents and NE*STDALG has a negative coefficient. Together, these variables account for 12% of the variance in item discrimination.

Insert Table 5 Here

Analysis of Item Asymptotes

The 3PL asymptote parameter measures the likelihood of responding correctly to an item through random guessing. Since the chances of guessing the correct response to a free-response item are extremely small, we followed the common practice of setting the asymptote parameter equal to zero for all of the free-response items in this study. Consequently, our analysis of item asymptotes was confined to the 66 items classified as multiple-choice (MC=1). This subset included 41 standard multiple-choice items with 3 or 4 options, and 25 nonstandard multiple-choice items with varying numbers of options, from eight to more than twenty. (The exact number of options was not tallied for items with more than twenty options.) The tree model estimated from this data is plotted in Figure 6. As can be seen, the number of choices is the single most important predictor. Items with five or more choices have low predicted asymptote values ($\hat{c}<0.15$); items with fewer than five choices have high predicted asymptote values ($\hat{c}>0.15$).

Insert Figure 6 Here

The linear regression models estimated to predict item asymptotes are listed in Table 6. A model including the single variable, "Number of Choices" accounts for 59% of the variance. A model including "Number of Choices" and four additional variables accounts for 85% of the variance. The additional variables include: a dummy variable coded as 1 for items with twenty or more options, and zero otherwise; the square of the number of choices variable; and two solution process variables "Apply standard algorithm in a nonstandard manner" and "Interpret mathematical vocabulary". The dummy variable was included to account for the fact that the No. of Choices variable was truncated at twenty.

Insert Table 6 Here

Past efforts to develop prediction models for item asymptotes were significantly less successful than the current effort. In the analysis of verbal items reported in Mislevy et al. (1993), for example, the prediction model for item asymptotes only accounted for 5% of the variance. The success of the current effort can be attributed to the many different types of items included in the Praxis I pool. Whereas most previous analyses have considered similarly formatted items (e.g. all four choice multiple-choice items) the Praxis pool includes items with many different formats, from standard 3- or 4-choice multiple-choice items, to items requiring the examinee to select a response from a table of more than 20 numbers. This variation in item format resulted in the large variation in the Number of Choices variable which accounted for the high value of R^2 obtained.

Evaluation of Model Fit

Predicted values of discrimination parameters, difficulty parameters and asymptotes are plotted vs. "true" values in Figure 7. Predicted values were obtained by applying the prediction equations with the highest values of R^2, as reported in Tables 4, 5 and 6. "True" values are the parameter estimates obtained in the original calibration of the entire pool of 510 items. Free-response items are plotted with a circle; multiple-choice items are plotted with a dot. Although considerable variation remains for discrimination parameters, much of the variation in difficulty parameters and asymptotes has been accounted for. In addition, the plots show no unusual outliers.

Insert Figure 7 Here

Analysis of Difficulty Rating Data

The test developers' global ratings of item difficulty was the single most important predictor of item difficulty among all those considered in this study. Because this variable turned out to be so important, an additional analysis was conducted to determine what could be learned about the "mental model" raters used to judge item difficulty. Figure 8 presents a tree model developed to predict the difficulty rating score from the other item attributes. Unlike the tree models presented previously, this model was built from the raw (unaveraged) rating data provided by the two raters. The item attributes considered in the analysis included all of the item attributes listed in Tables 1 and 2 with observed frequencies of at least five, a variable indicating whether the item was classified as free-response or multiple-choice, a variable indicating the source of the observation (Rater 1 or Rater 2), and a variable indicating the content area covered by the item (A,B,C,D or E). As shown in Figure 8, neither the rater identification variable nor the content area variable were selected for the tree-based prediction model. The tree also shows that low rated items and high rated items are easily identified: low rated items are those that do not involve multistep thinking (MTHINK=0) and can be solved by recalling or recognizing facts (RECALL=1). High rated items are those that do involve multi-step thinking (MTHINK=1). For items in the middle of these two extremes, the picture is more complicated, involving several other item attributes. This information may prove useful for future studies designed to refine the attribute scoring procedures.

Insert Figure 8 Here

Discussion

The tree-based approach described in this paper enabled us to develop a set of linear models for predicting the difficulty, discrimination and asymptote parameters of the Praxis I mathematics items. Using easily obtainable information about item features and test developers' ratings of item difficulty, we were able to explain 36% of the variation in item difficulty parameters, 12% of the variation in item discrimination parameters and 85% of the variation in item asymptote parameters. This is enough predictive power to be practically useful since, as was shown in Mislevy et al. (1993), similar models explaining even less variation, when used as prior distributions for item parameters, provided the information equivalent of approximately 250 additional pretest calibration subjects.

The tree-based approach employed in this study contributed to the success of the modeling effort in two ways: (1) it helped us to identify several important interaction effects which might not otherwise have been identified; and (2) it provided graphical displays of the modeling results which helped us to understand and discuss the models. We expect that the feedback provided by the tree-based displays will also prove useful in future efforts to refine the attribute scoring procedures.

Due to the limited number of items available, the models developed in this study could not be cross-validated. Additional research is needed to validate the model structure and to investigate the stability of the estimated parameters.

References

Bejar, I.I. (1991) A generative approach to psychological and educational measurement. (ETS Report No. RR-91-20). Princeton, NJ: Educational Testing Service.

Brieman, L., Friedman, J.H., Olshen, R., and Stone, C.J. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth International Group.

Clark, L.A and Pregibon, D. (1992) Tree-based models. In Chambers, J.M. and Hastie, T.J. (Eds.), *Statistical Models in S.* Belmont, CA: Wadsworth and Brooks/Cole.

Embretson, S.E. and Wetzel, C.D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175-193.

Enright, M.K., Allen, N. and Kim, M. (1993). A complexity anlaysis of items from a survey of academic achievement in the life sciences. (ETS Report No. RR-93-18). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. and Bock, R.D. (1983). *BILOG: Item analysis and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Mislevy, R.J., Sheehan, K.M. and Wingersky, M.S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30(1),* 55-78.

Scheuneman, J., Gerritz, K. and Embretson, S. (1991). *Effects of prose complexity on achievement test item difficulty.* (ETS Research Report No. RR-91-43). Princeton, NJ: Educational Testing Service.

Sheehan, K.M. and Mislevy, R.J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 20(4),* 345-354.

Tatsuoka, K.K (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement, 24,* 233-245.

Tatsuoka, K.K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredericksen, R. Glaser, A. Lesgold, and Shafto, M.G. (Eds.), *Diagnostic monitoring of skill and knowledge acquisition.* Hillsdale, NJ: Erlbaum.

Table 1

Correlation[a] of Item Parameters With Surface Feature Variables

Surface Features	Avg.[b] Freq.	% Agree ment	Correlation With Item Difficulty: AC[c] Items (n=61)	BDE[d] Items (n=53)	All Items (n=114)	Correlation With Item Discrimination: Free Resp. Items (n=48)	Mult. Choice Items (n=66)	All Items (n=114)	Corr. With Asymp. Mult. Choice Items (n=66)
Word Problem (WORDP)	57.0	82	.	.	.	.	.	.	.
Contains numerical expressions (NUMEXP)	23.5	92	.	.	.	.	.	.	.
Quantitative Comparison (QC)	21.0	93	.22	.	.19	.	.	.14	.36
Data presentation (DAPRES)	20.0	93	.	.	-.19	.	.	.	-.41
Equation or formula (EQUA)	16.0	96	.	.	.	.	.23	.19	.33
Geometric figure (FIG)	15.0	96	.	.	.	.	.	.	.
Complete a Table (CTAB)	8.0	98	.	.	.	.	.	.	-.41
Order & match (ORMAT)	6.5	99	-.20	.	-.14	.	.	.	-.23
Histogram (HIST)	5.0	100	-.27	.	-.25	.	.	.	-.39
Highlight Grid (HGRID)	2.5	99	.	.	.	.	.	.	-.26
Scale (SCALE)	2.0	98	.	.	.	.	.	.	-.24
Negative Stem (NEGSTM)	0.5	99	.	.	.	.	.	.	.
Number of Choices (choices)	N/A	N/A	-.23	.	-.20	.	.	.	-.77

a) Correlations not significantly different from zero (alpha=.15) are not listed. b) No. of items coded as having each attribute, averaged over two coders.
c) AC Content Areas = Number Sense & Operations & Data Interpretation. d) BDE content areas = Relationships, Geometry, Measurement & Reasoning.

Table 2

Correlation[a] of Item Parameters With Solution Process Variables

Solution Process	Avg.[b] Freq.	% Agree ment	Correlation With Item Difficulty: AC[c] Items (n=61)	Correlation With Item Difficulty: BDE[d] Items (n=53)	Correlation With Item Difficulty: All Items (n=114)	Correlation With Item Discrimination: Free Resp. Items (n=48)	Correlation With Item Discrimination: Mult. Choice Items (n=66)	Correlation With Item Discrimination: All Items (n=114)	Corr. With Asymp.: Mult. Choice Items (n=66)
Apply standard algorithm/procedure (STDALG)	73.5	59	.	-.33	-.24	.	.	-.16	.
Apply common sense reasoning (CSENSE)	49.0	65	.	.	.	.	.	.15	.
Complete multiple problem steps (MULSTP)	36.0	61	.	.21	.	.	.	.	-.26
Nonroutine application (NONROU)	25.0	77	.30	.	.25	.22	.	.	.
Examine each option (EXOPTN)	20.0	86	.	.	.	-.30	.	.	.
Apply multistep thinking (MTHINK)	18.0	80	.	.31	.25	.	.	.	.
Translate words to symbols (TRANS)	15.0	86	.	-.20	.	.	.	.	.
Interpret Math vocabulary (VOCAB)	11.0	89	.28	.	.23	.	.	.	.
Recall or recognize facts only (RECALL)	8.0	88	-.20	.	.	.	.	.	.
Apply std. alg. in nonstd. manner (NONSTD)	6.5	89	.24	.34	.29	.28	.	.	.
Represent given inf. in table or graph (REPINF)	5.5	92	.	.	.	.	.	.	-.43
Ignore irrelevant information (IGNORE)	4.5	94	.	.	.	.	.	.	-.19
Complete messy/prolonged calculations (MESSY)	2.5	96	.32	.	.24	.	.	.	.

a) Correlations not significantly different from zero (alpha=.15) are not listed. b) Number of items coded as having each attribute, averaged over two coders.
c) AC content areas = Number Sense & Operations & Data Interpretation. d) BDE content areas = Relationships, Geometry, Measurement & Reasoning.

Table 3

Correlation[a] of Item Difficulty Ratings With Item Parameters

		Correlation With Item Difficulty			Correlation With Item Discrimination			Corr. With Asymp.
	% Agree-ment	Free Resp. Items (n=48)	Mult. Choice Items (n=66)	All Items (n=114)	Free Resp. Items (n=48)	Mult. Choice Items (n=66)	All Items (n=114)	Mult. Choice Items (n=66)
Difficulty Rating - Coder 1	N/A	.59	.27	.40	.33	.	.	-.21
Difficulty Rating - Coder 2	N/A	.28	.52	.43	.	.	.	.
Difficulty Rating - Average of 1 and 2	92[b]	.53	.44	.47	.29	.	.	.

a) Correlations not significantly different from zero (at the .15 level) are not listed.
b) Percent of times that the difference in codings was less than or equal to one.

Table 4

Summary of Item Difficulty Modeling Results: Estimated Regression Coefficients and R^2 Values

	Alternative Models			
Parameter[a]	1	2	3	4
Intercept	-.158	-2.146	-2.489	-1.899
Difficulty Rating		.482	.542	.497
Quantitative Comparison	.403		.709	.559
Apply Std. Algorithm	-.545			-.437
Histogram	-.971			-.844
Order & Match	1.185			
Translate Words to Symbols				-.405
BDE*(NonstdApplication)	.477			
BDE*(Apply Mul.Thinking)	.525			
AC*(Order & Match)	-1.668			-.601
AC*(Recall/Recog. Only)	-.685			
df	(8,105)	(1,112)	(2,111)	(6,107)
R^2	.33	.22	.30	.39
Adjusted R^2	.28	.21	.29	.36

a) All regression coefficients were significant at an alpha level of .15.
The adjusted R^2 is corrected for the number of variables in the model.
AC content areas = Number Sense & Operations & Data Interpretation.
BDE content areas = Mathematical Relationships, Geometry, Measurement & Reasoning.

Table 5

Summary of Item Discrimination Modeling Results: Estimated Regression Coefficients and R^2 Values

	Alternative Model	
Parameter[a]	1	2
Intercept	.928	.930
Equation	.146	.133
MC*(Word Problem)		.159
NE*(Apply Standard Alg.)		-.096
df	(1,112)	(3,110)
R^2	.04	.14
Adjusted R^2	.03	.12

a) All regression coefficients were significant at an alpha level of .15. The adjusted R^2 is corrected for the number of variables in the model. MC = Multiple Choice Item Format. NE = The item does not contain an equation or formula.

Table 6

Summary of Item Asymptote Modeling Results: Estimated Regression Coefficients and R^2 Values

	Alternative Models	
Parameter[a]	1	2
Intercept	.257	.553
No. of Choices	-.014	-.108
Choices>=20?		-.679
(No. of Choices)^2		.006
Apply Std.Alg. in Nonstd. Manner		-.063
Interpret Math. Vocabulary		.035
df	(1,64)	(5,60)
R^2	.60	.87
Adjusted R^2	.59	.85

a) All regression coefficients were significant at an alpha level of .15. The adjusted R^2 is corrected for the number of variables in the model.

Figure Captions

Figure 1. A sample tree model for 20 observations.

Figure 2. Prediction of item difficulty from solution process variables and item features.

Figure 3. Prediction of item difficulty from solution process variables, item features and difficulty rating.

Figure 4. Relationship of item difficulty to average difficulty rating.

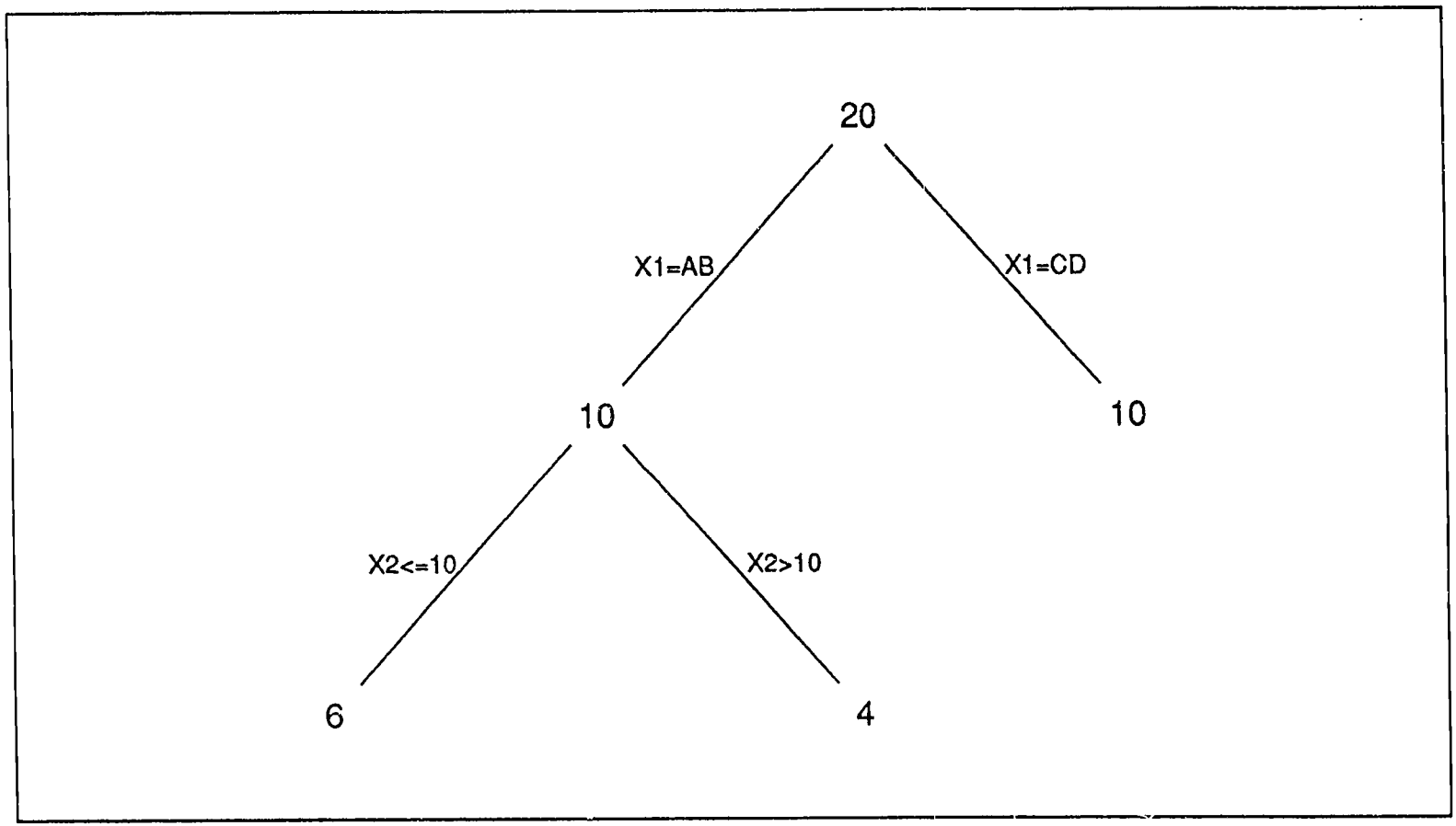
Figure 5. Prediction of item discrimination from solution process variables & item features.

Figure 6. Prediction of item asymptote from solution process variables & item features.

Figure 7. Evaluation of model fit.

Figure 8. Prediction of difficulty rating from solution process variables and item features.

Figure 1
A Sample Tree Model for 20 Observations



Deviance

Predicted Response

Figure 2
Prediction of Item Difficulty
From Solution Process Variables & Item Features

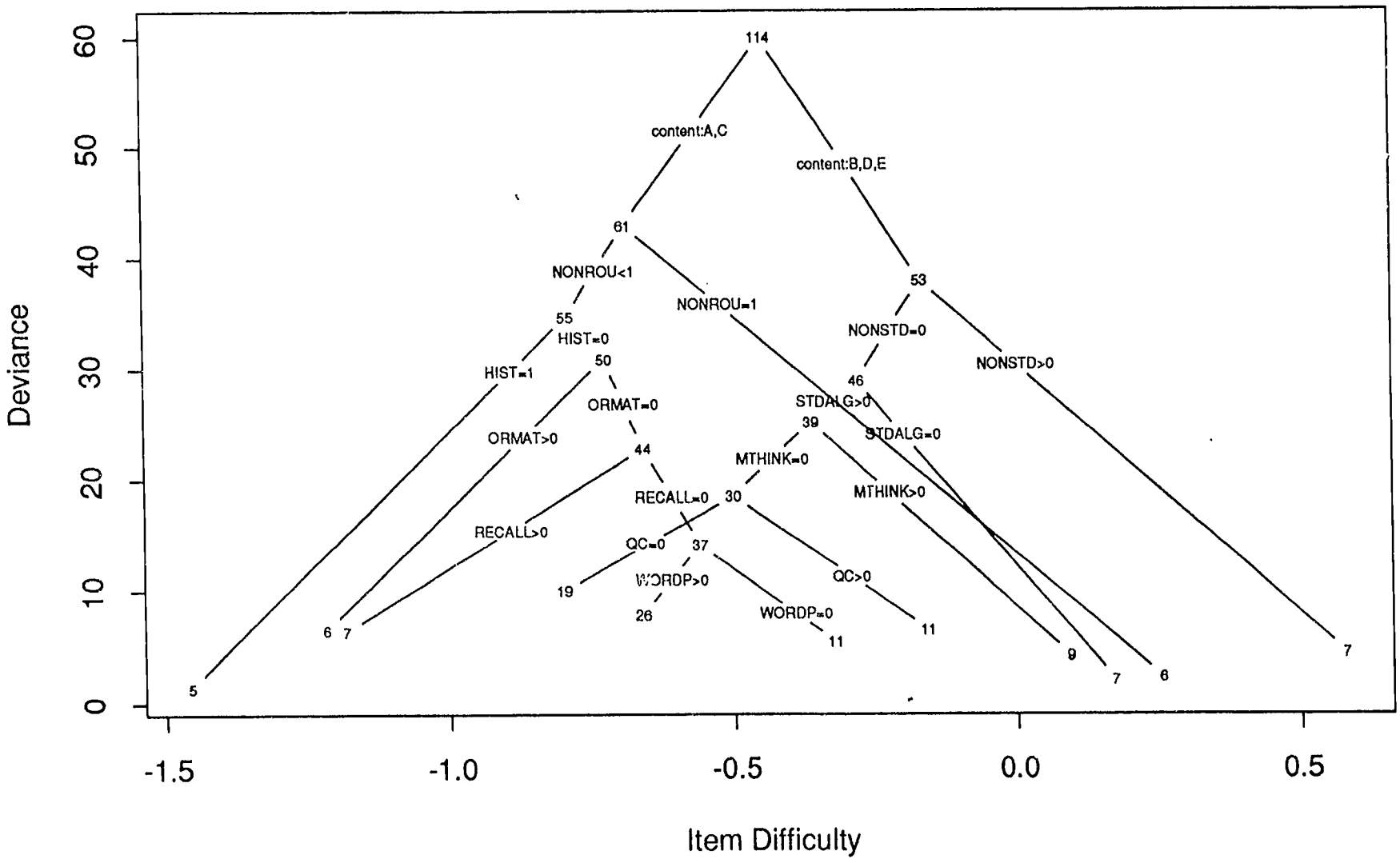


Figure 3
Prediction of Item Difficulty
From Solution Process Variables, Item Features & Difficulty Rating

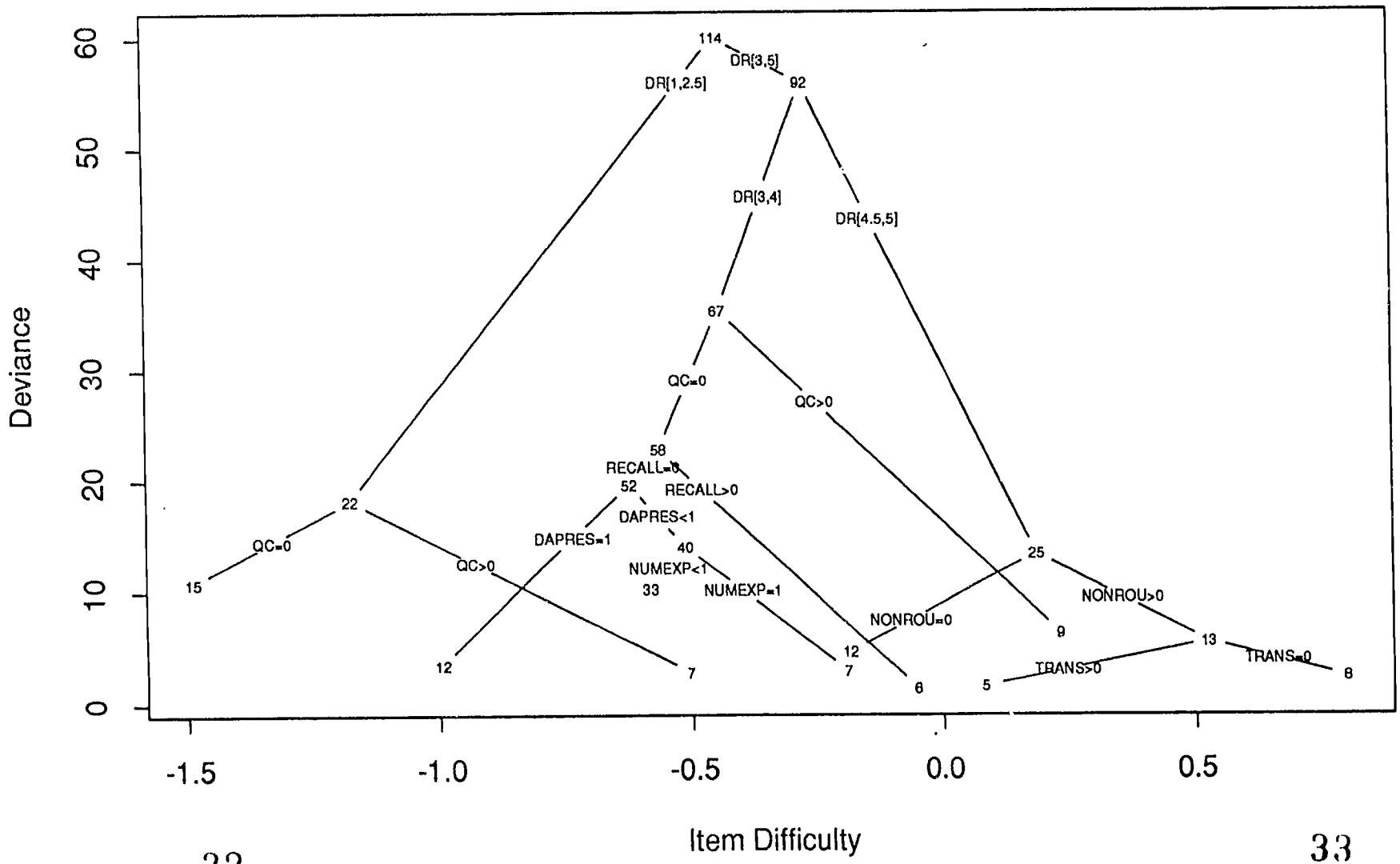


Figure 4
Relationship of Item Difficulty
To Average Difficulty Rating

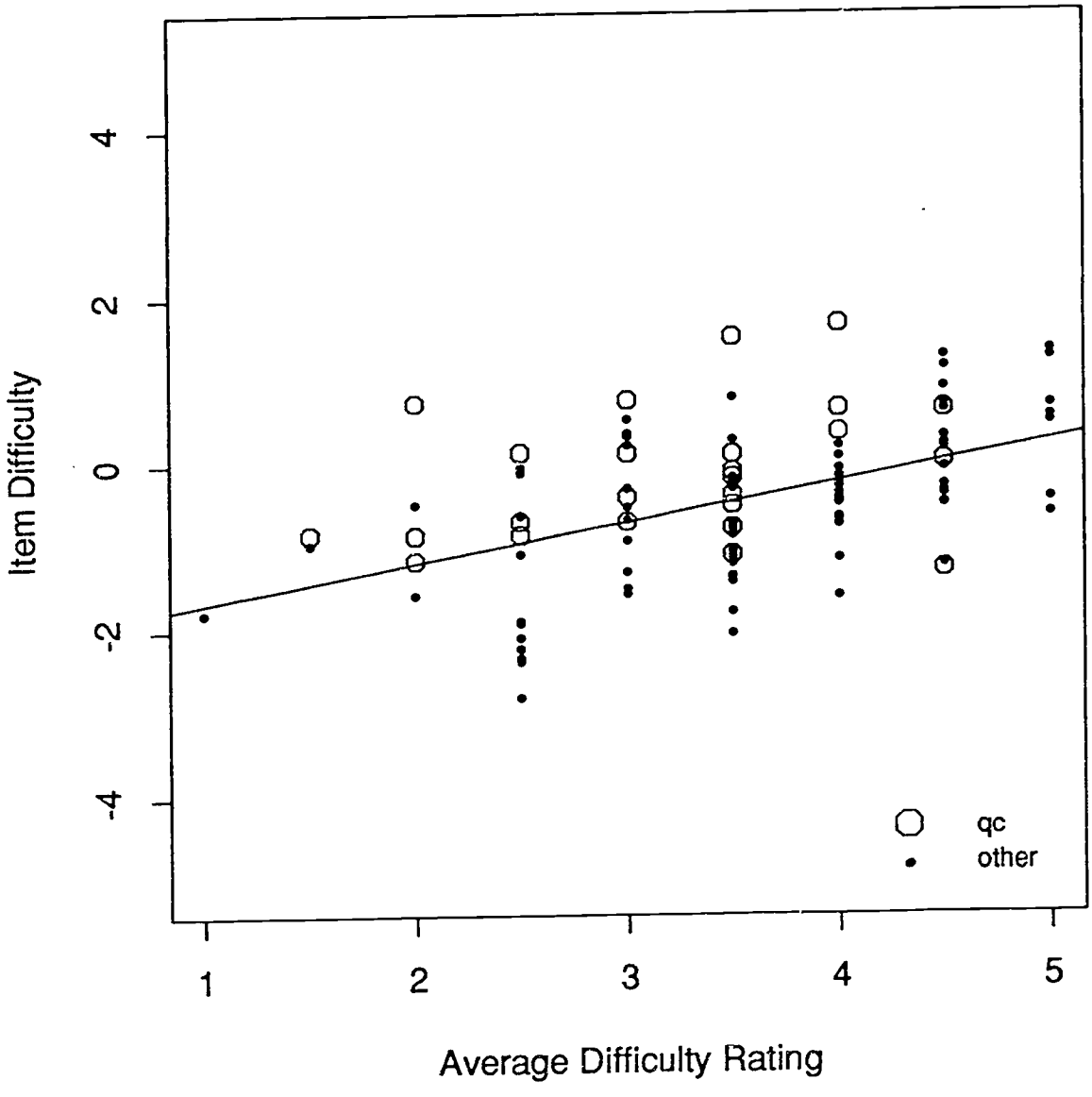


Figure 5
Prediction of Item Discrimination
From Solution Process Variables & Item Features

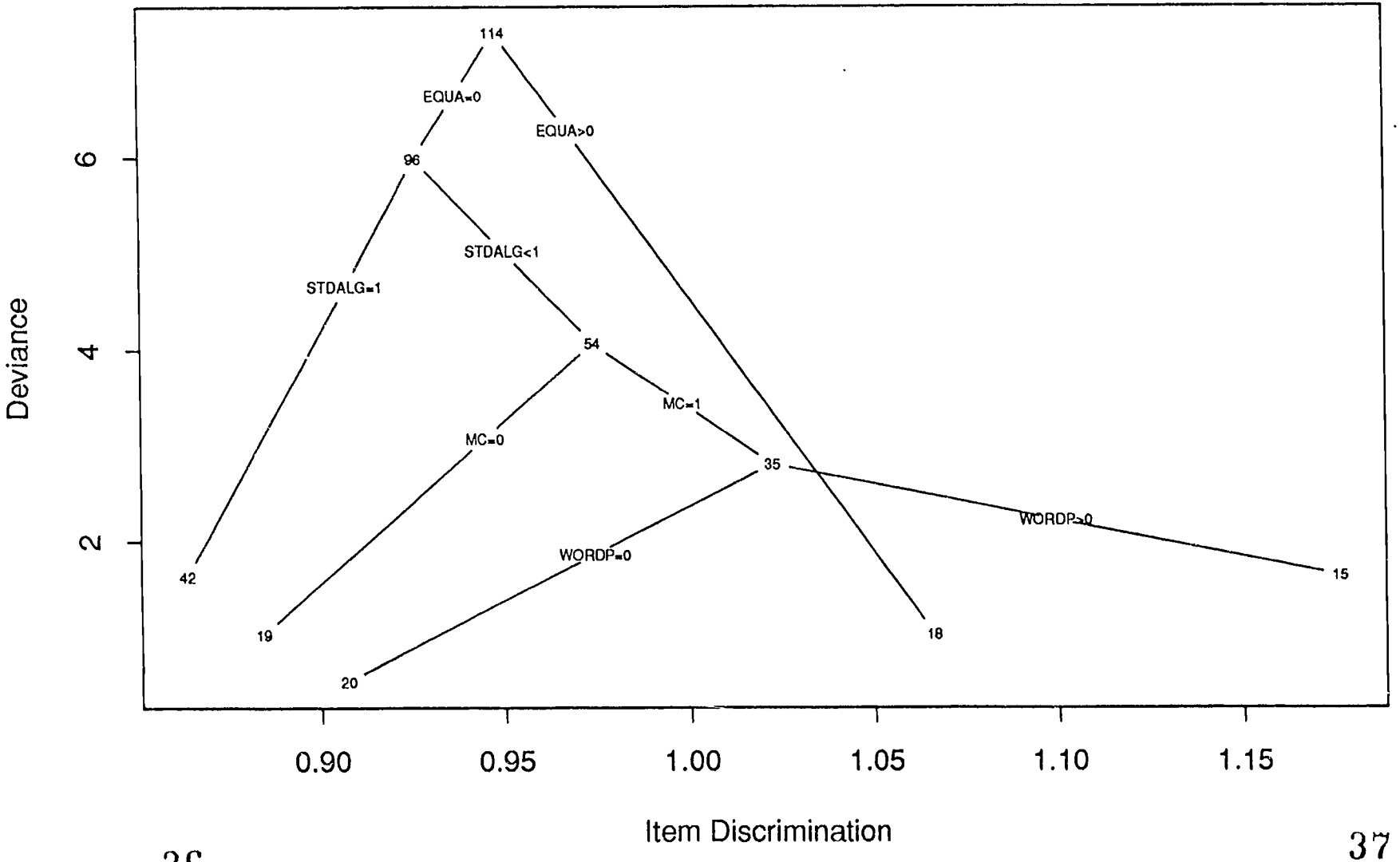


Figure 6
Prediction of Item Asymptote
From Solution Process Variables & Item Features

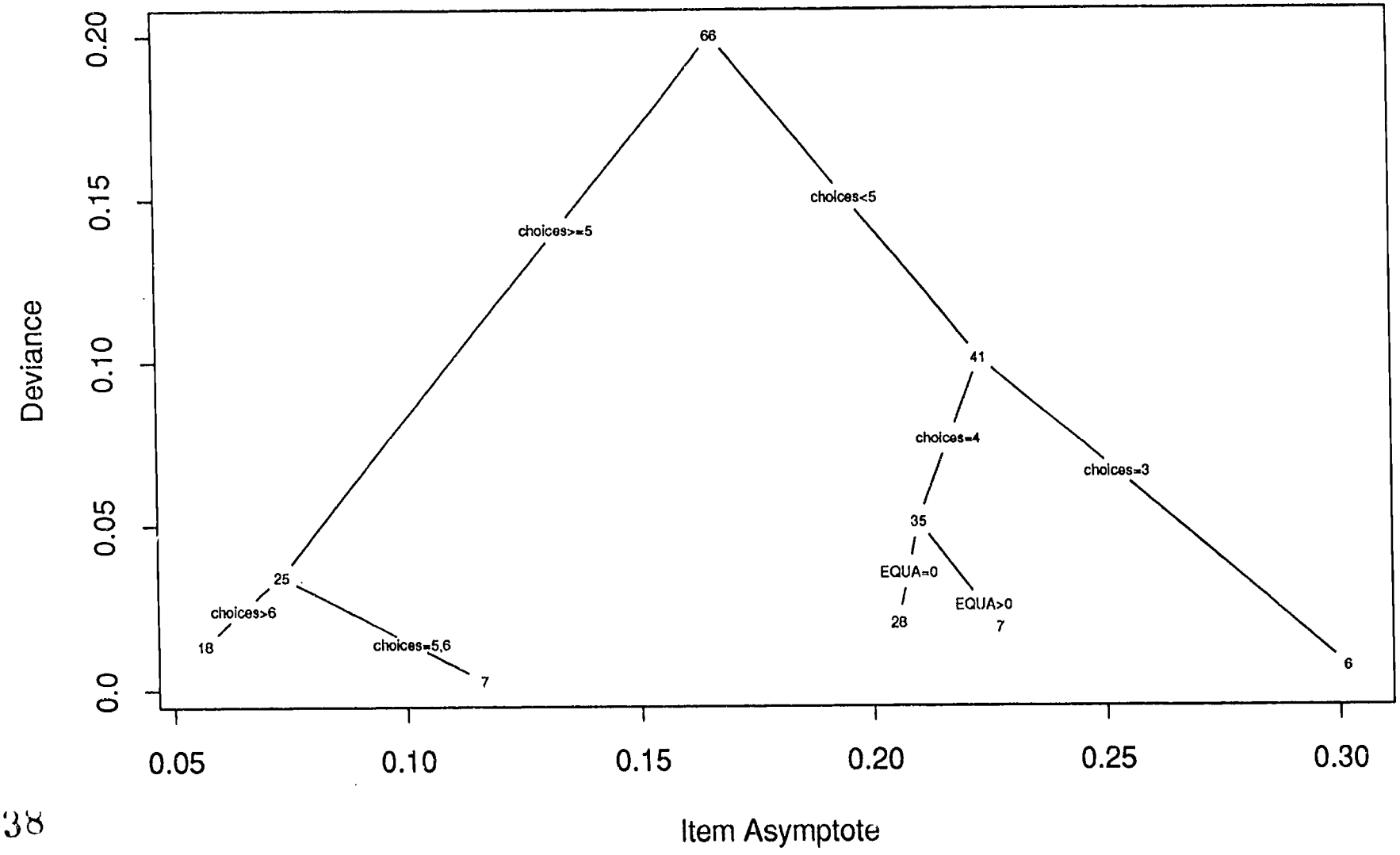
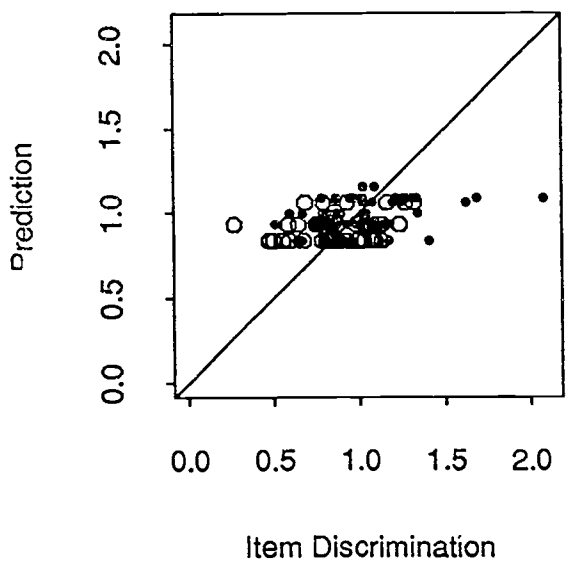
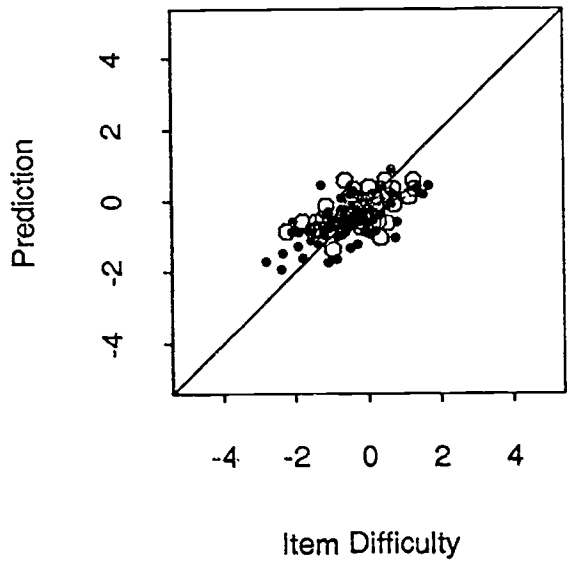
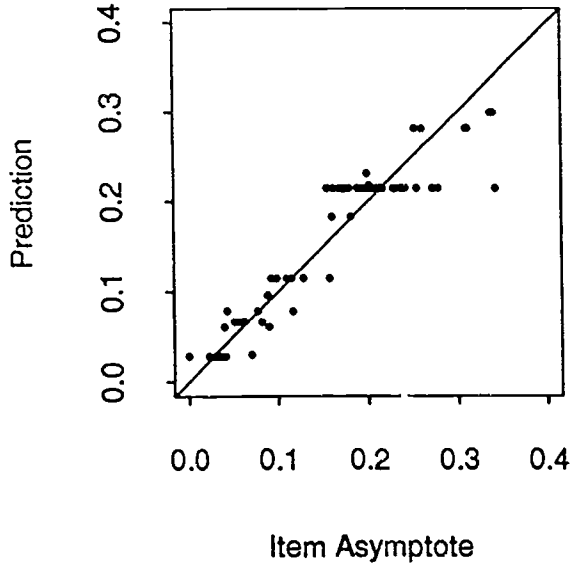


Figure 7
Evaluation of Model Fit







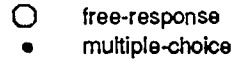


Figure 8
Prediction of Difficulty Ratings
from Item Attribute Data