DOCUMENT RESUME

ED 442 848 TM 031 268

AUTHOR Williamson, David M.; Hone, Anne S.; Miller, Susan; Bejar, Isaac I.
TITLE Classification Trees for Quality Control Processes in Automated Constructed Response Scoring.
PUB DATE 1998-04-00
NOTE 60p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 12-16, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Architects; Automation; *Classification; *Constructed Response; *Quality Control; *Scoring; *Test Scoring Machines
IDENTIFIERS Monitoring

ABSTRACT

As the automated scoring of constructed responses reaches operational status, the issue of monitoring the scoring process becomes a primary concern, particularly when the goal is to have automated scoring operate completely unassisted by humans. Using a vignette from the Architectural Registration Examination and data for 326 cases with both human and computer scores available, this study reports on the usefulness of an approach based on classification trees (L. Breiman, J. Friedman, R. Olshen, and C. Stone, 1984) as a means of quality control. Five studies were carried out analyzing different aspects of the "training set" and making efforts to cross-validate the results of the analysis by applying the resulting classification trees to data that had not been used in the development of the tree. The application of classification trees led to valuable insights with implications for operational quality control processes. Furthermore, classification tree methods were shown to be able to select cases for future quality control processes accurately and efficiently, thereby suggesting that future quality control selection procedures may be completely automated. However, further analyses are needed to establish whether classification trees can be relied on to identify cases that are the most likely to require some adjustment without incurring the potentially costly error of ignoring solutions that are likely to require adjustment. (Contains 10 tables, 7 figures, and 13 references.) (Author/SLD)

Running head: CLASSIFICATION TREES FOR QUALITY CONTROL

Classification Trees for Quality Control Processes in Automated Constructed Response Scoring

David M. Williamson

Anne S. Hone

The Chauncey Group International

Susan Miller

Isaac I. Bejar

Educational Testing Service

Presented at the annual meeting of the National Council on Measurement in Education, San Diego, California, April, 1998.

Abstract

As the automated scoring of constructed responses reaches operational status (e.g. Kenney, 1997) the issue of monitoring the scoring process becomes a primary concern, particularly when the goal is to have automated scoring operate completely unassisted by humans. Using a vignette from the Architectural Registration Examination (ARE) this study reports on the utility of an approach based on classification trees (Breiman, Friedman, Oshen, & Stone, 1984) as a means of quality control. Five studies were carried out analyzing different aspects of the "training set" and making efforts to cross-validate the results of the analysis by applying the resulting classification trees to data that had not been used in the development of the tree. The application of classification trees led to valuable insights with implications for operational quality control processes. Furthermore, classification tree methods were shown to be able to accurately and efficiently select cases for future quality control processes, thereby suggesting that future quality control selection procedures may be completely automated. However, further analyses are needed to establish whether classification trees can be relied upon to identify cases that are the most likely to require some adjustment without incurring the potentially costly error of ignoring solutions that are likely to require adjustment.

Classification Trees for Quality Control Processes in Automated Constructed Response Scoring

As the automated scoring of constructed responses reaches operational status (e.g., Kenney, 1997 in architecture) the issue of monitoring the scoring process becomes a primary concern. Initially, as an automated scoring system becomes operational, experts closely monitor the scoring process, thus providing an opportunity to gather data upon which to base statistical processes that may automate aspects of the quality control process itself. For example, if experts have a tendency to judge the automated scores unsatisfactory for specific classes of solutions then by identifying those classes it may be possible to make the quality control process more effective and efficient. Of course, the aim of automated scoring is not to emulate human scores. Human scorers typically operate under a set of scoring rules that are tailored to the characteristics of humans as graders. The aim of automated scoring is to emulate the best aspects of human graders but also to make it possible to consistently and fairly evaluate aspects of performance that human graders would find difficult, time consuming or impossible to analyze. Nevertheless, during the transition to operational status certain aspects of the automated scoring process may not function entirely satisfactorily and experienced human graders can provide valuable information to contrast with automated scoring. That is, disagreements between experienced graders and automated scoring are to be expected and may be the source of valuable information about both automated and human scoring processes. A study by Williamson, Bejar and Hone (1997) analyzed such differences for the constructed response portions of the Architect Registration Examination (ARE)

(Kenney, 1997). That study concluded that while the scoring policies implemented in the automated scoring are consistent with the scoring practices of independent groups of experienced graders of ARE solutions, automated scoring was able to extract far more detail from performances and to score with greater consistency than human scoring. Moreover, in the majority of cases humans were willing to accept the computer score once the details of computer evaluation and the rationale behind the computer score were presented to them. The present study uses human and computer grading data for one vignette (ARE constructed response task) from the Williamson et al. (1997) study.

The present study investigates the operating characteristics of automated scoring at the feature level (the finest level of ARE solution evaluation) and the score level (the coarsest level of ARE solution evaluation), both with regard to the integrity of the automated scoring engines and with an emphasis on examining the scoring engines for the potential of future development. The emphasis on the immediate integrity of the automated scoring is referred to as *first-order* quality control. Processes of *first-order* quality control are focused on the immediate performance of the automated scoring procedures and the results they produce as compared to the intent of their design. A distinguishing feature of first-order quality control is that it concerns aspects of scoring that have the potential to adversely impact the accuracy or validity of resultant scores if some aspects of scoring are not operating in the *intended way*. By implication any impact on resultant scores could demand intervention in the form of adjustments or corrections to automated scoring procedures to make them consistent with the intent of their design. This priority makes the identification of any such malfunctions a primary concern of first-order quality control processes. Clearly, when a scoring feature that is not functioning in the intended way is identified it should be fixed as soon as possible. In practice, it may not be possible to immediately institute the correction for a variety of reasons. In such events, there is significant value in efficiently identifying cases that may be affected by any malfunction.

In contrast, the term *second-order* quality control processes indicates investigations whose focus is on the long-term precision and evolution of automated scoring of complex constructed responses. Issues identified in second-order quality control procedures are those in which automated scoring is performing as it was intended to perform but a particular group of experts may feel that some 'tweaks' would be appropriate to better reflect their opinions (or biases) on particular issues. Examples of these types of issues may include different recommended weightings of criteria, different tolerance for less-than-perfect implementations of criteria, and inclusion or exclusion of criteria that may be marginally or tangentially related to the purpose of the examination. Of course, any two groups of experts will disagree on certain points of practice so the findings from second-order quality control processes can only be considered as 'suggestions' rather than as 'problems' with automated scoring, which would be the domain of first-order quality control. The nature of constructed response problems (e.g. allowing the candidate the freedom to implement a variety of complex solutions, or complex errors) in an automated examination prevents the accommodation of every *possible* solution a candidate may create; though every *reasonable* solution may be accommodated. This process of second-order quality control can help assure that all reasonable criteria are included and are evaluated appropriately by the automated scoring as well as providing possibilities for the future evolution of the constructed response examination.

Overview of the method

The intent of the present study is to evaluate the utility of classification trees (Breiman, Friedman, Oshen, Stone, 1984) for performing first-order and second-order quality control processes. A specific goal is to automate the identification of cases where experienced graders and automated scoring can be expected to disagree as a result of automated scoring malfunction (first-order quality control). The availability of experienced graders makes it possible to train a classification tree system to identify such cases so that the system can then be used once the experienced graders are no longer available. Specifically, given a training set of solutions for which we have available a measure of the computer-human agreement the aim is to identify which solutions would exhibit a disagreement in order to accurately and efficiently identify future cases. An expert would, of course, need to review the solutions identified in this manner but there would be substantial savings of time, effort and cost by limiting this examination process to those cases that are most likely to have exhibited a scoring disagreement. Such a targeted selection of solutions to review would seem to be more effective than random sampling techniques commonly used in quality control procedures and more efficient than a 100% quality control review process.

The use of classification and regression trees is an increasingly popular method in psychometric applications. Sheehan (1996) describes the application of tree-based methods for proficiency scaling and diagnostic assessment. Bejar, Yepes-Baraya and Miller (1997) discuss an application for modeling rater cognition. Holland, Ponte, Crane, Malberg (1998) discuss an application in computerized adaptive testing. Although firmly grounded in statistical theory (Breiman et al, 1984), classification trees share elements of techniques related to machine learning emerging from the artificial intelligence literature (e.g., Quinlan, 1979; Hunt, Marine, Stone, 1966). As a classification methodology it is a competitor of classical statistical methods, such as discriminant analysis, as well as more recent methods, such as neural networks. When compared with these techniques (Michie, Spigelhaulter, Taylor, 1994) classification trees were found to perform well with specific data sets. The methodology is claimed to possess many advantages, including the following:

- It is a nonparametric technique, and as such does not require distributional assumptions.
- It is suitable for both exploratory and confirmatory analyses.
- The method excels with data sets that are complex in nature.
- It is robust with respect to outliers and can handle cases with missing independent variables.

Several commercial implementations of classification and regression trees are available, including those by Salford Systems, SPSS, and S-Plus. The analyses in this paper were conducted using the program CART (Classification and Regression Trees) published by Salford Systems.

Description of the Method

Before considering the application of CART to the quality control of automated scoring it is useful to illustrate the method in the context of a small and familiar data set.

As in linear regression and discriminant function analyses, the analysis requires data (often called a training dataset) on the attributes (or independent variables) and the classification outcome (or dependent variable). Unlike linear regression analysis, where the outcome is a prediction equation, the outcome of CART is a tree, specifically a binary tree. A binary tree consists of a set of sequential binary decisions, applied to each case, that lead to further binary decisions or to a final classification of a that case. The independent variables can be numeric or nominal variables, which provides great flexibility for possible analyses.

Figure 1 shows a classification tree from the CART manual (Steinberg and Colla, 1992), based on a classic data set (Iris flower species) used by R.A. Fisher to illustrate discriminant analysis[1]. The same data were analyzed with CART, yielding a classification tree shown in Figure 1.

The CART procedure actually computes many competing trees and then selects an optimal one as the final tree. This is done, optionally, in the context of a "10-fold cross-validation" procedure (see Breiman et al. 1984, Chapter 11) whereby 1/10 of the data is held back and a classification tree grown. The procedure is repeated nine times and the final tree obtained by taking into consideration the ten different trees. The fit of the tree to the data, that is, how well it classifies cases, is measured by a misclassification table for the chosen tree.

A resultant tree can be used to classify new cases where the dependent variable is not available. Given a classification tree, new cases are “filtered down” the tree to a final classification. In this example using Iris data, there are 3 classes of final classification (Iris species), represented by the rectangles, and two classification decision nodes, represented by the diamonds. Decisions about which direction the data goes within the tree structure are based upon whether cases meet the specific criterion of the node. The first decision at Node 1 is based upon petal length (PETALLEN). The question, "Is petal length less than or equal to 24.5?" is posed. Those cases with a PETALLEN value of 24.5 or less (a "yes" answer) are deposited into Terminal Node 1, that is, they are classified in class 1 (Setosa species), while cases with a PETALLEN value greater than 24.5 (a "no" answer) continue through the decision tree. The Node 2question, "Is petal width less than or equal to 17.5?" is asked of those, as yet, unclassified cases. Cases where petal width (PETALWID) is less than or equal to 17.5 (a "yes" answer) end up at Terminal Node 2, with a classification of 2 (Versicolor species). Cases where PETALWID is greater than 17.5 (a "no" answer) end up at Terminal Node 3, with a classification of 3 (Verginica species). These terminal classification nodes may be characterized in table format by decision vectors that represent the decision sequence and outcome of the classification tree. The decision vectors corresponding to the Iris classification tree in Figure 1 are presented as Table 1. The fit of the model may be evaluated by examining the cross-validated misclassification table (which is different from, and typically less accurate than, the learning sample classification to prevent overfitting), which is included as Table 2 for the Iris data example. The table shows the joint occurrence of actual and predicted classification and probability. In this example the classification accuracy is high with 140 out 150 cases correctly classified.

The production of classification trees requires intense computations. The process can be conceptualized as splitting the data matrix into contiguous sets of rows that have been sorted on the variable that is being considered as the splitting variable (decision node variable). The two sets of rows that result are then dealt with recursively in the same fashion. If one of the dependent variable sets achieves a sufficiently high classification rate, those rows are not analyzed further. The remaining set is recursively analyzed until all rows are classified. A key aspect of this process is the selection of a splitting value. Several criteria are possible (see Ripley 1996, p. 217). The general idea

$$Entropy = \sum_j p_j \log p_j^2$$

is to compare whether the two sets resulting from a given split are "purer" than the parent set. A possible measure of purity is entropy and is given by

where p_j is the proportion of cases in category j.

However, in this study the Gini index, as suggested by Breiman et al. (1984), was used as the measure of purity and is given by

$$\text{Gini} = 1 - \sum_j p_j^2$$

The Gini index is 0 when the set contains all cases in a single dependent variable category and is largest when the set contains the same number of cases in each dependent variable category.

Figure 2 is a graphical representation of the Iris data set illustrating the concepts described above. The figure shows the cases (by dependent variable) on the x-axis and their independent variable measurements on the y-axis, and are sorted on petal length (PETALLEN) as can be seen by the monotonically increasing plot corresponding to that variable. The chart also displays the actual classification (variable Speno) of each case, which have been arbitrarily coded as 1 (Setosa), 2 (Versicolor), and 3 (Verginica).

Notice that the cases to the left of a split on PETALLEN at around 24 are all category 1. This is why in Figure 1 above those cases appear in a "terminal node" without further decision nodes. The remaining cases, to the right of the split, are then analyzed and all variables are considered as the next splitting variable. The process is repeated recursively until all cases have been classified.

A useful aspect of CART is that it characterizes variables in terms of their importance. Importance refers to the contribution a variable can make in classification accuracy , based on how well it can split the data as measured by the purity of the resulting sets. A variable's importance is based on potential and actual splitting behavior. Thus, a variable may be highly important even if it never appears as a primary node splitter in a specific tree. To allow comparison of the importance of different variables importance is normalized relative to the variable with highest importance. Thus the most important variable in given tree always has importance of 100. In the Iris data set, for example, the most important variable is PETALWID, followed by PETALLEN. Thus, order of appearance in the tree and importance are not necessarily the same.

Overview of the Studies

We present five separate studies. Study 1 is concerned with an analysis of the importance of evaluated automated scoring features in predicting or classifying cases according to the level and direction of human-computer disagreement. The difference between human and computer scores is regressed on the feature scores that are extracted as part of the automated scoring process. The human scores were obtained as part of a previous study (see Williamson, Bejar, & Hone, 1997). The present study focuses on a single ARE vignette. There were 326 cases for which both human and computer scores were available, which we refer to as the training set. The purpose of Study 1 is to see if the importance of the features in predicting differences from the CART analysis corresponds to what was previously known as a source of disagreement from the actual 100% quality control process that took place with these data. Because the focus is on the identification of features that may not be functioning correctly and which may require intervention this study is an instance of first-order quality control analysis, though additional second-order quality control elements were also identified. The second study is based on the same data and tree as in Study 1 but the focus of analysis is specifically on the second-order quality control process. The third study regresses the human scores on the feature scores and aims to determine if human graders are scoring on the basis of criteria other than those represented in the automated scoring features. The fourth study extends results from the first three studies as a means of determining whether, practically speaking, CART results can be relied upon to identify cases whose score may need to be adjusted as part of first-order quality control intervention. The fifth study examines the use of CART classification trees, regressing the adjudicated scores on features, for the specific purpose of identifying cases requiring first-order quality control intervention.

For a description of the procedures used to obtain the training dataset the reader is refered to Williamson, Bejar & Hone (1997). The human scores were produced by a "Grading Committee" (GC) consisting of six human graders experienced in the holistic grading of candidate submissions for the ARE. The committee was divided into two groups so that three graders examined each solution. Three hundred and twenty-six actual candidate solutions for an actual ARE vignette were considered in these studies. These solutions are evaluated on a feature by feature basis, with each feature receiving an evaluation of A (acceptable), I (indeterminate), or U (unacceptable). These feature evaluations are the independent variables in the CART analyses. These feature evaluations are aggregated to produce a final solution score of A, I, or U. It should be noted that the I evaluation represents a borderline implementation. For more information on the scoring of this examination see Bejar (1991), Bejar and Braun (1994), and Kenney (1997).

Study 1

Method

Design and Procedure

The initial study investigated the utility of CART for first-order quality control processes. Specifically, this study focused on the identification of features evaluated by the automated scoring engines that may be sources of disagreement between human holistic evaluations and automated scoring evaluations of candidate submissions. The primary purpose of this investigation is to provide an additional method for ensuring that the evaluation of features in automated scoring is functioning as intended.

In this evaluation each of the 326 actual candidate solutions for an ARE vignette were scored holistically by the GC in addition to the scores provided by the automated scoring engine. The resultant scores of A (acceptable), I (indeterminate) and U (unacceptable) were then converted into numeric representations of 3, 2 and 1, respectively. A difference score was computed by subtracting the numeric value of the automated score from the numeric value of the human holistic score. The possible resultant values of this procedure for configurations of human and automated scores are presented in Table 3. These resultant difference scores were used as the dependent variable for the CART procedure. This CART procedure assigned relative importance values to the features used in the automated scoring according to each feature's ability to predict the resultant difference score. The variations in relative importance values were evaluated with regard to their ability to suggest specific automated features likely to be a source of disagreement between human holistic and automated scoring.

Results

In order to permit a detailed discussion of the findings of this research the results section of this and subsequent studies will reference a hypothetical "exemplar" vignette which has instructions, features, and characteristics which have been altered considerably and is which not actually used in the ARE. This exemplar vignette, and the architectural program requirements associated with it, are constructed to permit a faithful representation of the characteristics of relevant features and requirements of the actual ARE vignette which was the subject of these studies. For this exemplar vignette the candidate would be given a floor plan for an office and would be required to make modifications according to specific requirements from a hypothetical client.

A line graph of the relative importance of the features, ordered from most important to least important, is presented as Figure 3. The relative importance values suggest that feature F2 (skylight location) is the major contributing factor to discrepancies between human and automated scoring. Other features that may be contributing to discrepancies include F3 (flashing), F9 (eave height), F15 (water flow), and F1 (gutters).

These results prompted an architectural review of vignette solutions for which there were discrepant scores (those for which the difference score was not equal to zero) with particular attention to the features identified as possible sources of discrepancy. This review observed a high frequency of solutions with an additional skylight (represented by a square with an X) indicated by the arrow in Figure 4.

In this exemplar vignette the candidate would begin with the floor plan showing an open office area, a cubicle within the open office area, and toilet facilities. The candidate would then be required to complete the Floor and Roof plans according to specified client requirements. The building section portion of Figure 4 was not available to the candidate or the GC but is included here for the benefit of the reader.

One of the requirements of this vignette is that "all rooms must receive natural light", the intention of which is to have the candidate place a skylight in the roof over the toilet facilities, as this is the only room without windows. An examination of feature F2 (skylight location) for the solutions identified as receiving discrepant scores revealed that in these cases there were actually two skylights; one in the required location for the toilet and the other placed over the cubicle area (indicated by the arrow in Figure 4). For each skylight the candidate would typically place flashing (F3) around the skylight and a cricket to prevent water from leaking into the building (F15). The placement of an additional skylight over the cubicle area, and the accompanying flashing and cricket would be considered excessive use of skylights and flashing, and inappropriate water flow control and would cause automated scoring to provide an unfavorable evaluation of these features.

From this observation and the fact that human holistic evaluations tend to give credit to candidates providing the extra skylight over the cubicle (but not for placement over other areas of the room) it is possible to infer that the GC made allowances in scoring for the possibility that candidates were interpreting the partitioned cubicle in the floor plan as a room (keeping in mind that neither the candidate nor the GC had the building section view in Figure 4). While this discovery is not a deficiency in the automated evaluation of particular features, it did reveal a potential ambiguity for candidates in fulfilling the requirements of the vignette. On the basis of this possibility steps were taken to eliminate this potential for misinterpretation. Specifically, as shown in Figure 5, the floor plan was changed to include pre-existing windows for the cubicle (indicated by the arrow) so that there would be no confusion about the correct implementation of skylights.

Architectural examination of eave height (feature F9) in the solutions with discrepant scores revealed that the GC was at times overlooking this element in their evaluation process, despite the fact that it was included in their written criteria. An example of the type of situations in which the automated scoring was providing an unfavorable evaluation of eave height while the GC was considering solutions to be acceptable is presented in Figure 6. Since the GC would often rely on "eyeballing" to judge the correctness of the roof heights at various points, they at times missed the fact that given a specific ridge height and slope, the eave height would not be a practical solution. Figure 6 shows an exaggerated representation of the findings. In Figure 6 we have two roof plans, which are visible to the candidate and the GC, and their associated building section views, which are not available to the candidate or GC. Both plans in Figure 6 have a ridge height of 18'-0". The plan in Figure 6 (a) shows a slope ratio of 6:12 while the plan in Figure 6 (b) shows a ratio of 12:12. It is readily apparent from the building section views associated with the roof plans that given the different candidate-defined slopes and ridge heights, the two roof profiles would be quite different. Based on the requirements for the vignette, the solution in Figure 6 (a) would be a correct solution while the solution in Figure 6 (b) would be incorrect. Therefore, if the GC neglected to calculate the slopes in their holistic scoring they would have missed the fact that the solution in Figure 6 (b) was incorrect. Examination of solutions with discrepant scores revealed that in these cases the holistic scoring process failed to completely evaluate eave height (F9).

The examination of discrepant solutions with emphasis on the gutter (F1) feature revealed an apparent difference in the relative tolerance of less-than-perfect implementation and weighting of this particular feature as it is aggregated with other features to produce the final vignette score. Specifically, the GC appeared to have less tolerance of less-than-perfect implementation than was implemented in the automated scoring and the GC appeared to weight this feature more heavily than the automated scoring in the determination of overall score. The differences attributable to this feature were found to be relatively minor and were documented as second-order quality control issues for future consideration.

Discussion

Initial examination of the relative importance of features evaluated in the automated scoring suggested that feature F2 (skylight location) is the primary contributor to the discrepancies between human holistic scoring and automated scoring. Investigation of this issue led to the understanding of features F3 (flashing) and F15 (water flow) as factors related to the primary cause. This approach demonstrated the ability of this method to identify first-order quality control cases where there may be a problem with the scoring implementation or other vignette characteristics. The identification of this potential ambiguity resulted in a policy of performing an architectural review of 100% of candidate solutions until the new base floor plan could be implemented.

Investigation of features with relative importance values similar to those of F3 (flashing) and F15 (water flow) also revealed one of the advantages of automated scoring in its ability to precisely evaluate every aspect of a candidate solution, as exemplified in feature F9 (eave height). This CART procedure, then, seems capable not only of first-order quality control processes but also of documenting situations in which one scoring methodology may be more precise than another, thus helping to evaluate competing scoring procedures.

An unanticipated result of this investigation is the ability of relative importance output of the CART procedure to identify issues of second-order quality control processes. Specifically, this procedure was able to identify feature F1 (gutter) as a feature for which the GC utilized a somewhat different standard of tolerance for less-than-perfect implementations or somewhat different weighting in aggregation to the final solution score. As a result this investigation also identified a second-order quality control issue of overall criteria and content which can be examined by architectural test development committees in the continued evolution of ARE automated scoring.

Study 2

Method

Design and Procedure.

The participants and materials for Study 2 are identical to those for Study 1. This second study investigates the utility of CART specifically for second-order quality control processes. The investigation of features identified through Study 1 was shown to be a fruitful process. The results of Study 1, however, do not address the question of whether the holistic scoring of the GC might implicitly include criteria which are not currently evaluated by the automated scoring but which would improve the quality of the scoring if these features were included.

In addition to the relative importance values for each feature CART produces a classification tree as described above. The classification accuracy rate for the classification tree produced using these difference scores is presented as Table 4. This second study seeks to determine whether this classification tree can be a useful tool in the identification of specific differences in criteria or tolerances and weighting between the GC and the automated scoring as part of second-order quality control. This was investigated by identifying feature vectors leading to the terminal nodes (final nodes indicating the resultant difference score). These feature vectors (labeled A through N) and their resultant difference score are presented in Table 5.

Feature vectors A, B, and C are all associated with the terminal node value of –2, in which the automated scoring result was A (acceptable) and the human holistic scoring result was U (unacceptable). These feature vectors are suggestive of solutions for which the GC is using additional criteria not assessed by the automated procedure, allowing less tolerance for less-than-perfect feature implementation, or utilizing greater feature weighting for inadequate features in the solution. Solutions with feature vectors of A, B, and C were selected and examined for any coherent architectural trends among the selected solutions which would suggest a difference in tolerance, weighting, or criteria implemented by the GC.

At the opposite pole of the difference score spectrum feature vector M is associated with difference scores indicating that the human holistic scoring provides a higher overall score than the automated scoring. Since the only feature with a U in this vector is the eave height feature (F9), and based on the knowledge gleaned from Study 1 it expected that feature vector M is indicative of cases where the human holistic scoring is overlooking the eave height feature (F9) as discussed above. Solutions with this vector of feature scores were selected to examine this hypothesis.

Feature vector N is also associated with difference scores that indicate the human holistic scoring provides a slightly higher overall score than the automated scoring. Since the two critical negative features in this vector are skylight location (F2) and flashing (F3), and based on the knowledge gleaned from Study 1 it is expected that feature vector N is indicative of cases where the GC made exceptions regarding the skylight location as discussed above. This possibility was evaluated through architectural examination of solutions with this vector of feature scores.

Results

Thirty of the 326 solutions were found to have feature vector A, of which 13 have human holistic scores identical to the automated scores (due to classification error in the tree). Architectural examination of these 30 solutions led to the identification of two features which may be criteria that were not specified by the GC in their documented criteria but were implicitly used in evaluating the solutions. Remaining consistent with the hypothetical ARE vignette discussed previously, for discussion purposes these criteria will be termed roof material and parapet walls.. The implicit feature of inappropriate roof material was observed in 11 of the 30 solutions (4 of the 17 with discrepant scores) and inappropriate use of parapet walls was observed in 16 of the 30 solutions (7 of the 17 with discrepant scores). Neither roof material nor parapet are evaluated in the automated scoring routines of this vignette.

Additionally, the architectural review identified 14 of the 30 cases (10 of the 17 with discrepant scores) for which the GC appeared to be weighting feature F2 (skylight location) more heavily than the automated procedures. A noteworthy aspect of this finding is that while this feature is the same feature which was the focus of attention for Study 1, the relevant aspects of this feature receive a different interpretation when examined in the context of solutions with feature vector A. It would appear that this distinction in interpretation of the F2 (skylight location) feature from Study 1 to Study 2 is a result of the restricted body of solutions being examined and the criterion value being considered. The architectural review of the large number of solutions in Study 1 identified the candidate interpretation issue as the primary conclusion based on the fact that it was a curiosity and it occurred with some frequency in the general set of solutions. By restricting the focus of architectural review through the selection of feature vector A solutions, the viewing of a subset of 30 solutions identified a trend which was masked in the Study 1 review of solutions. This identification was facilitated by the fact that the feature vector A solutions are solutions for which the criterion is that GC scores are *lower* than automated scores—a criterion different from expectations based on Study 1, in which GC exceptions for candidate interpretation would result in *higher* scores than the automated score.

It initially seemed curious that the use of CART methodology is capable of simultaneously identifying two potential points of investigation for a single automated feature. In an effort to obtain additional empirical support for the belief that these architectural observations of feature vector A solutions were not imaginary trends, an additional analysis was conducted controlling for the effects of candidate misinterpretation. This was conducted by examining each of the 326 solutions and correcting for instances of candidate misinterpretation described in Study 1 by altering the feature scores of candidates to accept the skylight implementation resulting from this misinterpretation. A new CART analysis was run using as the dependent variable the difference score between the human holistic score and this adjudicated automated score. The classification rate resulting from this analysis is presented as Table 6. A line graph of the resultant values of relative importance for each of the features, ordered from most important to least important, is presented as Figure 7. This analysis identified feature F9 (eave height) as the most important feature, which is consistent with the findings of Study 1 regarding this feature. The second most important feature is F2 (skylight location) despite the fact that the candidate interpretation of requirements is controlled. This provides some additional support for the conclusion about the GC weighting of F2 (skylight location) contributing to discrepant scores in the feature vector A solutions as well as offering some additional explanation for the dramatic difference between the relative importance of F2 (skylight location) and F3 (flashing) in Figure 3, despite the fact that these two features are conceptually and architecturally related.

Seven of the 326 solutions were found to have feature vector B, one of which had a GC holistic score identical to the automated score. Architectural examination of these solutions revealed that all are the result of a single difference in feature evaluation. In each case the GC was weighting a single feature, F5 (downspout/portal conflict), more heavily than the automated scoring engine.

Twenty-five solutions were identified as having feature vector C, all but 8 of which have GC holistic scores identical to the automated scores due to a higher rate of classification error for this particular terminal node. Architectural examination of these solutions identified feature F3 (flashing) as a feature that the GC was weighting more heavily than the automated routine. This feature was identified as a factor in 25 of the 30 solutions and in all 8 of the solutions for which resultant scores were discrepant.

Architectural examination of the 15 solutions with feature vector M, 14 of which are discrepant scores, supported the hypothesis that this vector was a representation of cases in which the GC appeared to overlook the measurement of the eave height feature (F9). This provides an additional corroborating source of evidence about the significance of this feature from Study 1.

Thirty solutions were identified as having feature vector N, 22 of which are discrepant scores. Architectural examination revealed that 17 of the 22 discrepant solutions are cases in which the candidate appeared to misinterpret the floor plan as described above. This result supports the hypothesis that feature vector N is a representation of cases where candidates are likely to be misinterpreting the floor plan.

Discussion

The architectural examination of solutions with feature vector A was successful in the identification of two features which the GC appeared to consider in their holistic scoring but which are not evaluated as part of the automated scoring. As a result these two features were documented for future consideration. In addition, the review of feature vector A solutions was able to identify an additional nuance of difference in scoring by the GC and automated procedures for a feature (F2) already identified as an important feature to be reviewed, but on a very different basis. The feature vectors were also able to contribute to the identification of two additional features whose weightings are worthy of review by architectural test development committees, though from the number of solutions selected these appear to occur infrequently. These results suggest that classification trees can be effective tools for second-order quality control processes.

The architectural examination of solutions with response vectors M and N confirmed that these vectors are indicative of cases for which issues identified in Study 1 are relevant. In this respect this constitutes additional evidence concerning the utility of CART procedures for first-order quality control processes as the results of Study 2 support conclusions from Study 1.

Study 3

Method

Design and Procedure.

Whereas the second study examined the utility of feature vectors using difference scores as the dependent variable, this study uses only the human holistic scores as the dependent variable. Thus, a classification tree was grown by regressing the human holistic scores onto the automated feature scores. The intent is to determine if the utilization of human holistic scores as the dependent variable results in any of the classification tree vectors being architecturally illogical. If a feature vector follows a pattern of entirely, or predominantly, acceptable automated features but results in a terminal node of unacceptable (as the human holistic score) this suggests that the GC is evaluating some additional features or implementing different tolerances or feature weighting. Subsequently, it may be fruitful to review these solutions as part of the second-order quality control process.

The feature vectors (designated O through Z) for the CART procedure using human holistic score as the dependent variable are presented as Table 7. The feature vectors Y and Z are architecturally surprising feature vectors for the overall score of U on the vignette. Feature vector Z contains predominantly A's as feature evaluations with one feature (F6) as I or U and resulting in a final GC vignette score of U. Since the feature F6 is a relatively minor feature it is curious that this would have enough influence to result in a human holistic score of U, particularly when feature vector Z is so similar to feature vector P, for which the GC holistic score is A for the vignette.

Similarly, feature vector Y also has predominantly A's for the individual features with one relatively minor feature, F5 (downspout/portal conflict), receiving a U and resulting in an overall vignette GC evaluation of U. The minor feature, F5 (downspout/portal conflict), is the primary distinguishing feature between feature vector S, for which the GC typically evaluated the solution as an I, and feature vector Y. To investigate this use of classification trees solutions with feature score vectors Y and Z were selected and examined for architectural trends.

Results

Five of the 326 solutions were found to have response vector Y (two of which had been previously identified from feature vector A). Each of these had GC holistic scores that were discrepant from the automated scores. An architectural examination of these solutions concluded that the discrepancy in scores was the result of a consistent difference between the GC and the automated scoring in the weighting of two features; F2 (skylight location) and F5 (downspout/portal conflict). Each of the 5 solutions were inadequate implementations of both of these features. The feature F2 (skylight location) was previously identified as the cause of the discrepancies from feature vector B. The direction of score discrepancies from feature vector Y is consistent with the interpretation from Study 2. The feature F5 (downspout/portal conflict) was also previously identified in Study 2 as the feature weighting discrepancy from analysis of solutions with feature vector A. It is interesting that examination of feature vector A identified feature F2 (and additional GC criteria), feature vector B identified cases discrepant purely on the basis of feature F5, and feature vector Y isolated cases with discrepant scores resulting from the combination differential weightings of both features F2 and F5.

Thirteen solutions were identified as fitting response vector Z (of which 9 had been previously identified as part of feature vector A). Architectural examination of these solutions again revealed a difference in the weighting of the feature F2 (skylight location) originally identified from feature vector A in Study 2. This not unexpected when it is recognized that 9 of the 13 solutions were part of the feature vector A solution set. What is more relevant is that the evaluation of the set of 13 solutions for response vector Z resulted in the identification of an additional feature, F6, which appears to be receiving differential weighting between the GC and the automated procedures. This feature was identified in 8 of the 13 solutions as a potential source of differential weighting. It seems that this feature weighting difference was not apparent in the larger set of 30 solutions from vector A but when the restricted set of 13 solutions from vector Z was isolated the pattern of weighting feature F6 became more obvious.

Discussion

The identification of illogical feature vectors and the architectural examination of solutions with these feature vectors corroborated the results of previous studies in identification of two features that may be receiving different feature weighting between the GC and the automated scoring. This examination also identified a feature (F6) which appears to be receiving different weighting but which was not previously identified. However, since the number of occurrences of this feature as a factor in discrepant scores is relatively small it would appear to be less of a priority for examination by architectural test development committees responsible for continued test development.

Study 4

Method

Design and Procedure.

This study builds on the results of the past 3 studies and evaluates the utility of knowledge gleaned for the operational selection of cases for human intervention to resolve first-order quality control issues. Specifically, given the previous finding that some candidates may be misinterpreting the cubicle as a room requiring a skylight would the CART results provide a means for identification of instances where this misinterpretation would result in a different vignette score.

This interpretation issue was identified at the outset of operational testing through a policy of performing architectural examinations of 100% of solutions, with each solution examined by several architects. As a result it was determined that candidates who misinterpreted the cubicle as a separate room as described above would have the automated scoring evaluations adjudicated to accommodate this misinterpretation. Subsequently, there were a number of candidates whose overall vignette score was changed as a result of this adjudication. This process of examining 100% of solutions and making interventions where appropriate was relatively time consuming and expensive.

Since the results of Study 2 suggest that feature vector N indicates cases for which candidates misinterpret the cubicle, the possibility that use of this feature vector is a sufficient method for identifying cases of candidate misinterpretation which would result in a difference in vignette score was investigated. To evaluate this possibility an additional sample of 1117 candidate solutions which had been subjected to the process described above, but which did not receive scores from the GC, were analyzed. Cases which had feature vectors matching vector N were identified and the resultant accuracy of identification of cases resulting in a change in vignette score was evaluated.

Study 1 suggests a single feature, F2 (skylight location), is the primary feature that accounts for score discrepancies. Since this feature is related to the issue of candidate misinterpretation the possibility that selection on this single feature would be a sufficient technique for identification of cases of candidate misinterpretation which would result in a difference in vignette score was examined. This possibility was investigated through the selection of solutions from the extended sample of 1117 solutions described above for which this feature score, F2 (skylight location), was other than A (acceptable). The resultant accuracy of identification of cases resulting in a change in vignette score was evaluated.

Results

The results of utilizing feature vector N for the identification of cases for which intervention is required is presented in Table 8. The overall predictive error rate of using vector N for the identification of cases to receive a change in solution score is low, with only 69 (1%) misclassifications. The use of feature vector N for the selection of cases would certainly reduce the burden of reviewing solutions as only 114 (10%) of solutions would be selected for architectural examination. However, this reduction in solutions reviewed would have come at the cost of 40 (32%) of the solutions which required a change in solution score as a result of the candidate's misinterpretation remaining unidentified. For first-order quality control such as this, in which actions are being taken on candidate scores as a result of the selection process, this error rate is unacceptable.

For cases of second-order quality control, in which the intent is not to take actions on candidate scores but to investigate the occurrence or tendencies toward certain actions this may prove to be a useful technique of selecting cases for architectural examination.

The results of utilizing the single feature (F2) for the selection of solutions to be reviewed are presented in Table 9. The overall classification error rate of this technique is higher than for the feature vector N selection with 229 (21%) misclassifications. The use of the feature F2 as the selection criteria for solutions to be examined also reduces the burden and expense of the review process, though not to the extent of the feature vector N method, as 354 (32%) of all cases were selected for review. An advantage of this method for the example in question is that all of the solutions for which a change in score was warranted were selected for examination.

Discussion

These results suggest that selection of solutions for architectural examination based solely on the feature vectors resulting from the CART procedures (using the difference between human holistic scores and automated scores as the dependent variable) would not be a prudent method for first-order quality control interventions. This methodology however, may be a fruitful technique for second-order quality control processes of an investigative nature. The use of empirical and logical architectural knowledge gleaned from the previous studies, however, appears to be an effective method for selecting a reduced number of solutions for architectural examination with very little error. In such cases this methodology may make the quality control process more efficient and less expensive than the policy of reviewing 100% of cases.

Study 5

Method

Design and Procedure.

The results of Study 4 suggest that while the knowledge gleaned from classification tree quality control processes can inform effective selection procedures for case examination, the actual feature vectors (using the difference between human holistic scores and automated scores as the dependent variable) cannot be relied upon. However, the classification tree utilized in Study 4 was not produced for the purpose of identifying cases of score intervention; only for differences between human and automated scores. Therefore, it may be unrealistic to expect the resultant feature vector to be able to identify cases requiring a score change: a criterion for which the classification tree was not specifically trained. This study examines the question of whether an appropriately trained classification tree (using the criterion of interest—score interventions) is able to produce a feature vector which may be relied upon to select future cases for architectural examination and first-order quality control interventions.

The determination that score interventions would be implemented for candidates who misinterpreted the cubicle as a room resulted in 29 of the 326 solutions for which vignette scores were changed. From this training set of 326 solutions a classification tree was produced using as the dependent variable whether or not there was a difference in score between the automated score and the adjudicated score. The subsequent feature vector for classifying scores requiring an intervention was then used as a selection criterion for identifying cases for architectural review in the extended sample of 1117 solutions described above in Study 4. The resultant accuracy for identification of cases requiring a change in vignette score was evaluated.

Results

The CART analysis utilizing discrepancy between automated score and adjudicated score as the dependent variable identified a single feature, F2 (skylight location), as the predictive feature vector for changes in the automated score. Specifically, solutions with an A for F2 (skylight location) were classified as not predicting a change in score while solutions with an I or U for F2 (skylight location) were predictive of solutions with a score change. The resultant cross-validation results for the difference score between the automated and adjudicated scores are presented in Table 10. This procedure empirically identified the same feature and criterion for selection of cases requiring review that the architectural-logical procedures identified in Study 4. The resultant accuracy in the extended sample of 1117 solutions described above is identical to the results from Study 4 presented as Table 9. That is, this procedure resulted in the identification of 100% of the solutions that required a change in the automated score while requiring the review of only 32% of the solutions.

Discussion

The results of Study 5 suggest that classification tree vectors can be utilized to accurately and efficiently identify cases requiring score intervention as part of first-order quality control processes when these classification trees are produced for this purpose. The accuracy of the cross-validation classification for the training set held for the extended set of additional solutions. As these results mirror the results from the architectural—logical analysis in Study 4 this suggests that both purely empirical and empirical—logical classification tree analyses can provide evidence about criteria for efficient and accurate future case selection. Since the relative cost of error types can be specified in producing a classification tree differences in importance of classification error can be controlled when the initial tree is produced from the training set. An examination of the resultant cross-validation classification accuracy can help the user determine if the classification tree is sufficiently accurate to rely on for the selection of future cases for review.

Conclusion

This series of investigations has examined the utility of classification trees for several aspects of quality control processes associated with automated scoring of open-ended responses. Generally these methods have proven to be fruitful approaches to both first-order and second-order quality control. In applications directed at first-order quality control these methods indicated specific features which required intervention and suggested others which upon investigation provided evidence about the advantage of specificity and thoroughness provided by automated scoring systems. Examinations with respect to second-order quality control processes revealed aspects which may be worthy of consideration for the continued evolution of automated scoring of constructed responses as well as giving some indication of the frequency and conditions for which these possibilities may be relevant. The use of feature vectors from classification trees for the selection of solutions for first-order quality control interventions was shown to be inadequate when the classification trees were not produced expressly for this purpose. When the classification trees were produced for this purpose, however, they were shown to be effective in the selection of future cases for first-order quality control intervention while reducing the burden of the review process by 68%. The architectural evaluation of solutions identified by feature vectors from human/automated classification trees was also shown to be fruitful for determining and/or confirming criteria for the selection of future cases for first-order quality control intervention. With further investigation and refinements of the fit parameters used to grow these classification trees these feature vectors may be proven to be an efficient and accurate way to completely automate the selection of solutions for quality control purposes. Further studies are needed to sufficiently evaluate and determine the extent to which the results of these analyses can be relied upon for such an automated quality control process.

References

Bejar I. I. (1991). A methodology for scoring open-ended architectural design problems. Journal of Applied Psychology, 76, (4), 522-532.

Bejar, I. I., & Braun, H. (1994). On the synergy between assessment and instruction: Early lessons from computer-based simulations. Machine-Mediated Learning, 4, 5-25.

Bejar, I. I., Yepes-Baraya, M., & Miller, S. (1997, March). Characterization of complex performance: From qualitative features to scores and decisions. Paper presented at the annual conference of the National Council on Measurement in Education, Chicago, IL.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, California: Wadsworth.

Holland, P. W. Ponte, E., Crane, E. Malberg, R. (1998) Treeing in CAT: Regression trees and adaptive testing. Paper presented at the annual conference of the National Council on Measurement in Education, San Diego, CA.

Hunt, E. B., Marin, J., & Stone, P. J. (1966). Experiments in induction. New York: Academic Press.

Kenney, J. F. (1997). New testing methodologies for the Architect Registration Examination. CLEAR Exam Review, 8, (2), 23-28.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (Ed.). (1994). Machine learning, neural and statistical classification. West Sussex, England: Ellis Horwood.Michie

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1, 81-106.

Hunt

Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge University Press.

Sheehan, K. M. (1996). A tree-based approach to proficiency scaling and diagnostic assessment. Journal of Educational Measurement, 34, (4), 333-352.

Steinberg, D. & Colla, P. (1992). CART, Evanston, IL: SYSTAT, Inc.

Williamson D. M., Bejar I. I., & Hone, A. S. (1997, November). Report on the empirical and qualitative analysis of RDS automated scoring: An examination of the agreement between RDS automated scoring and human holistic scoring with additional analysis of evaluative factors contributing to discrepant scores. Unpublished report: Princeton, NJ: The Chauncey Group International.

Footnotes

[1]This and other historical datasets can be found at

http://www.comcat.com/~hutch/DASL/overview.htm.

Table 1

Decision Vectors Corresponding to the Iris Classification Tree

	N1	N2
Classification	PETALLEN	PETALWID
1	<=2.45	
2	>2.45	<=1.75
3	>2.45	>1.75

Table 2

Cross-Validation for Iris Example

Actual Classification	Classification Probability Predicted			Classification Predicted		
	1	2	3	1	2	3
1	1.00	0.00	0.00	*50*	0	**0**
2	0.00	0.90	0.10	0	*45*	5
3	0.00	0.10	0.90	**0**	5	*45*

Table 3

Possible Difference Score Values (Human-Automated)

Human Score	Automated Score: A	I	U
A	0	1	2
I	-1	0	1
U	-2	-1	0

Table 4

Cross-Validation for Difference Score (Human Minus Automated)

CART Cross-Validation	Classification Probability Predicted				Classification Predicted			
Actual Class	-2	-1	0	1	-2	-1	0	1
-2	0.412	0.000	0.588	0.000	*21*	0	30	0
-1	0.186	0.237	0.288	0.288	11	*14*	17	17
0	0.205	0.031	0.697	0.067	40	6	*136*	13
1	0.000	0.238	0.000	0.762	0	5	0	*16*

Table 5

Feature Vectors Utilizing the Difference Scores as the Dependent Variable

Terminal Node (Difference Score)	Feature Vector	N1 (F3)	N2 (F9)	N3 (F2)	N4 (F2)	N5 (F13)	N6 (F5)	N7 (F2)	N8 (F5)	N9 (F2)	N10 (F2)	N11 (F1)	N12 (F5)	N13 (F1)
-2	A	I,A	I,A	I										
-2	B	I,A	I,A	U,A	I,A	U								
-2	C	U								U,A	I,A	U,A		
-1	D	I,A	U					A	U					
-1	E	I,A	I,A	U,A	U									
-1	F	I,A	I,A	U,A	I,A	U								
-1	G	U								U,A	I,A	I		
-1	H	U								I			U	
0	I	I,A	U					U,I						
0	J	I,A	I,A	U,A	I,A	I,A	I,A							
0	K	U								U,A	U			
0	L	U								I			I,A	U,I
1	M	I,A	U					A	I,A					
1	N	U								I			I,A	A

Table 6

Cross-Validation for Difference Score (Human Minus Adjudicated)

CART Cross-Validation	Classification Probability Predicted				Classification Predicted			
Actual Class	-2	-1	0	1	-2	-1	0	1
-2	0.386	0.088	0.509	0.018	*22*	5	29	1
-1	0.127	0.365	0.365	0.143	8	*23*	23	9
0	0.119	0.124	0.743	0.015	24	25	*150*	3
1	0.000	0.000	0.000	1.000	0	0	0	*4*

Table 7

Feature Vectors Utilizing the Human Holistic Score as the Dependent Variable

Terminal Node (Human Score)	Feature Vector	N1 (F9)	N2 (F2)	N3 (F1)	N4 (F13)	N5 (F5)	N6 (F10)	N7 (F2)	N8 (F13)	N9 (F15)	N10 (F6)	N11 (F5)
A	O	I,A	A	A	I,A	I,A	A					
A	P	I,A	I,U					I	A	U		
I	Q	I,A	A	A	I,A	U						
I	R	I,A	A	A	I,A	I,A	I					
I	S	I,A	I,U					I	A	I,A	A	A
U	T	U										
U	U	I,A	A	I,U								
U	V	I,A	A	A	U							
U	W	I,A	I,U					U				
U	X	I,A	I,U					I	U			
U	Y	I,A	I,U					I	A	A	A	U
U	Z	I,A	I,U					I	A	A	I,U	

Table 8

Solution Identification Accuracy of Feature Vector N

Solution Score	Not Vector N	Feature Vector N	Row Totals
Changed	40	85	125
Unchanged	963	29	992
Column Totals	1003	114	1117

Table 9

Solution Identification Accuracy of Feature F2

Solution Score	F2 of A	F2 of I or U	Row Totals
Changed	0	125	125
Unchanged	763	229	992
Column Totals	763	354	1117

Table 10

Cross-Validation for Difference Score (Automated and Adjudicated)

CART Cross-Validation	Classification Probability Predicted		Classification Predicted	
Actual Class	No Change	Change	No Change	Change
No Change	0.771	0.229	*229*	68
Change	0.000	1.000	0	*29*

Figure Captions

Figure 1. Sample CART analysis classification tree for the Iris data.

Figure 2. Line graph of the four measurements representing independent variables and the resultant classification for the Iris data

Figure 3. Line graph of the relative importance of automated scoring features using the difference score as the dependent variable.

Figure 4. Floor plan, roof plan, and section view of the exemplar vignette showing the location of additional skylight as a result of candidate misinterpretation of the floor plan.

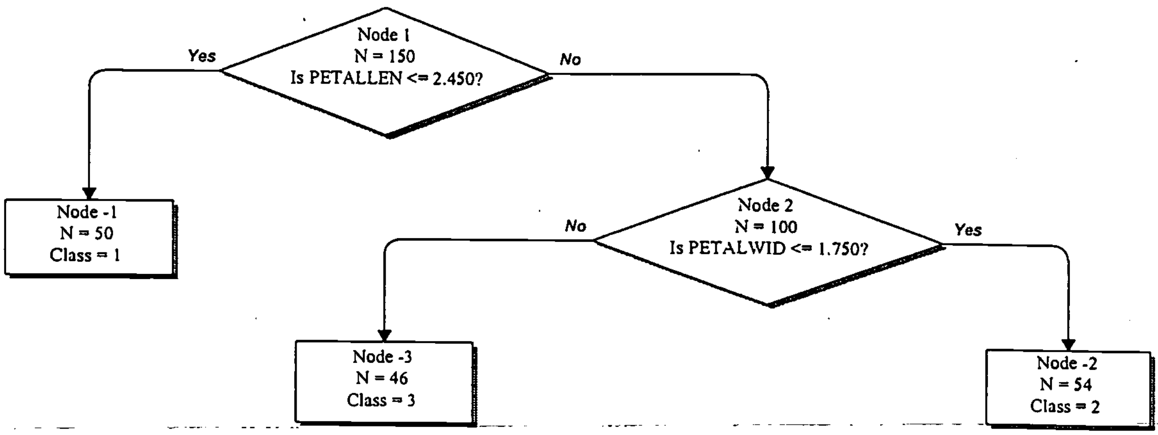
Figure 5. Floor plan, roof plan, and section view of the exemplar vignette showing the correct implementation of the skylight feature and the windows added to prevent candidate misinterpretation of the floor plan.

Figure 6a. Roof plan and section view of the exemplar vignette showing the correct implementation of eave height.

Figure 6b. Roof plan and section view of the exemplar vignette showing the incorrect implementation of eave height.

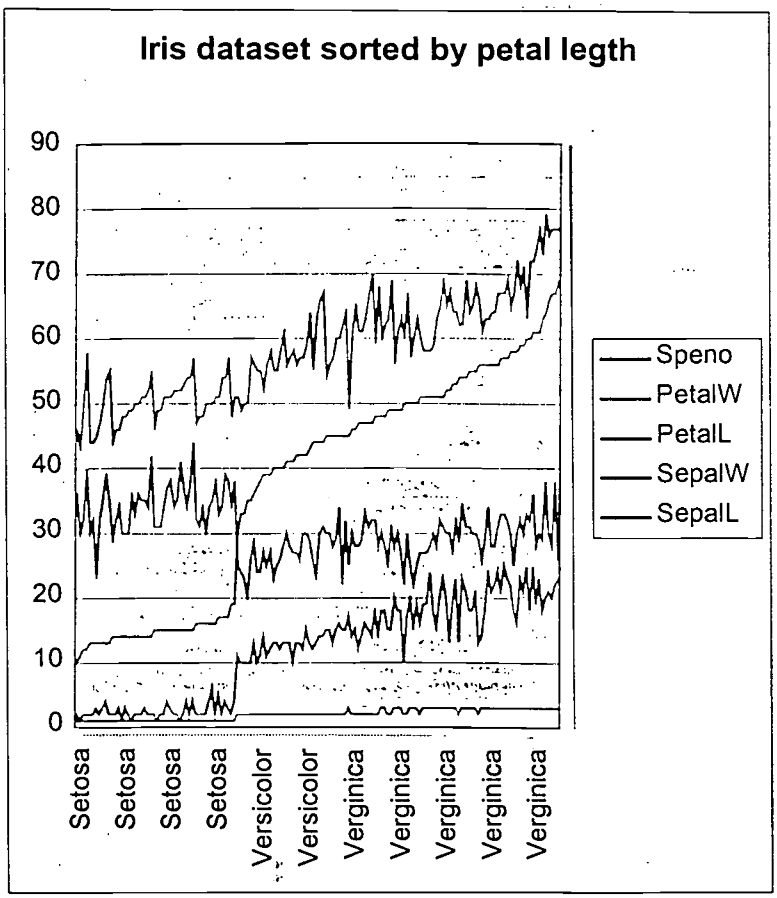
Figure 7. Line graph of the relative importance of automated scoring features using the difference score as the dependent variable and controlling for instances of candidate misinterpretation of skylight location.

Figure 1



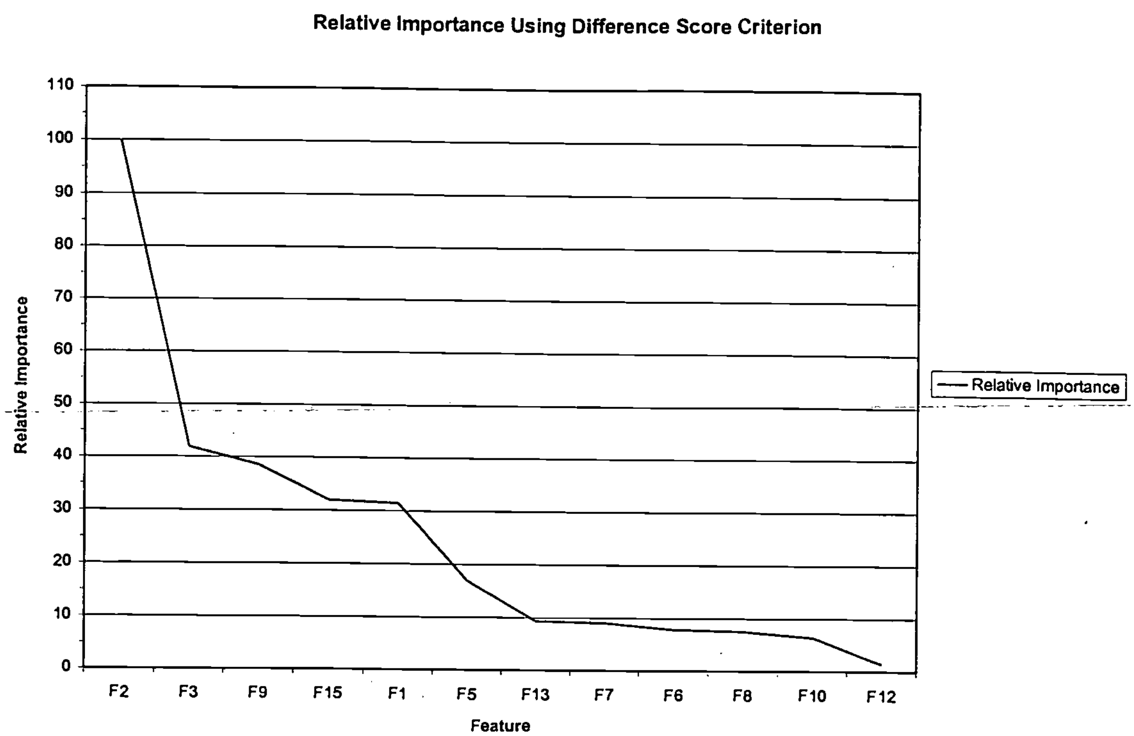
BEST COPY AVAILABLE

Figure 2



BEST COPY AVAILABLE

Figure 3



BEST COPY AVAILABLE

Figure 4

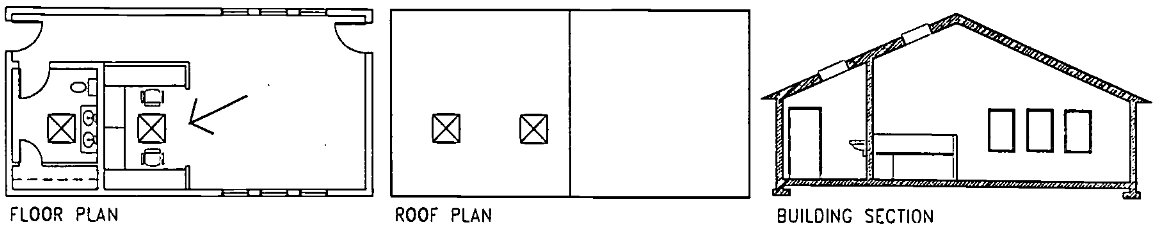


Figure 5

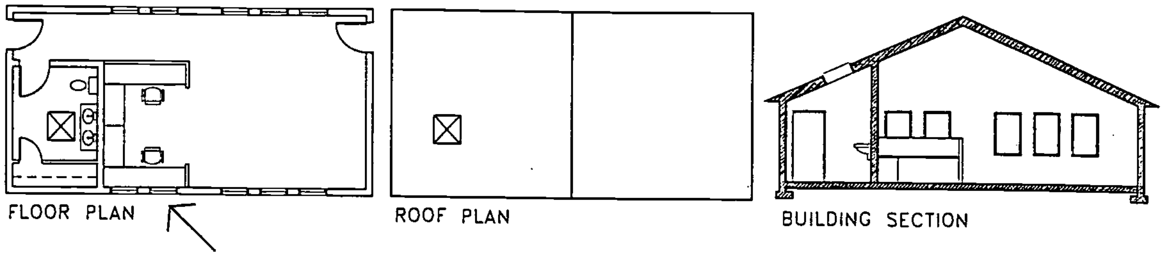
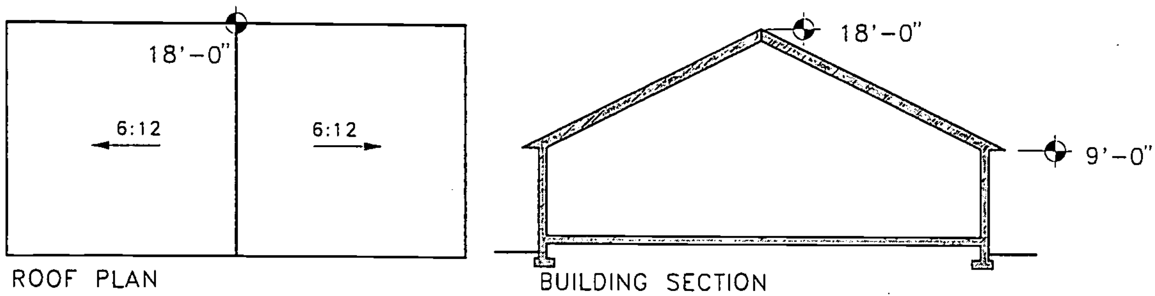


Figure 6 (a) and (b)



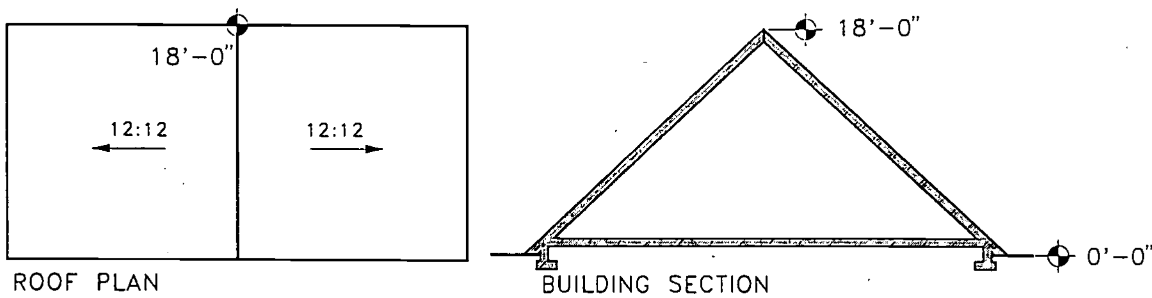
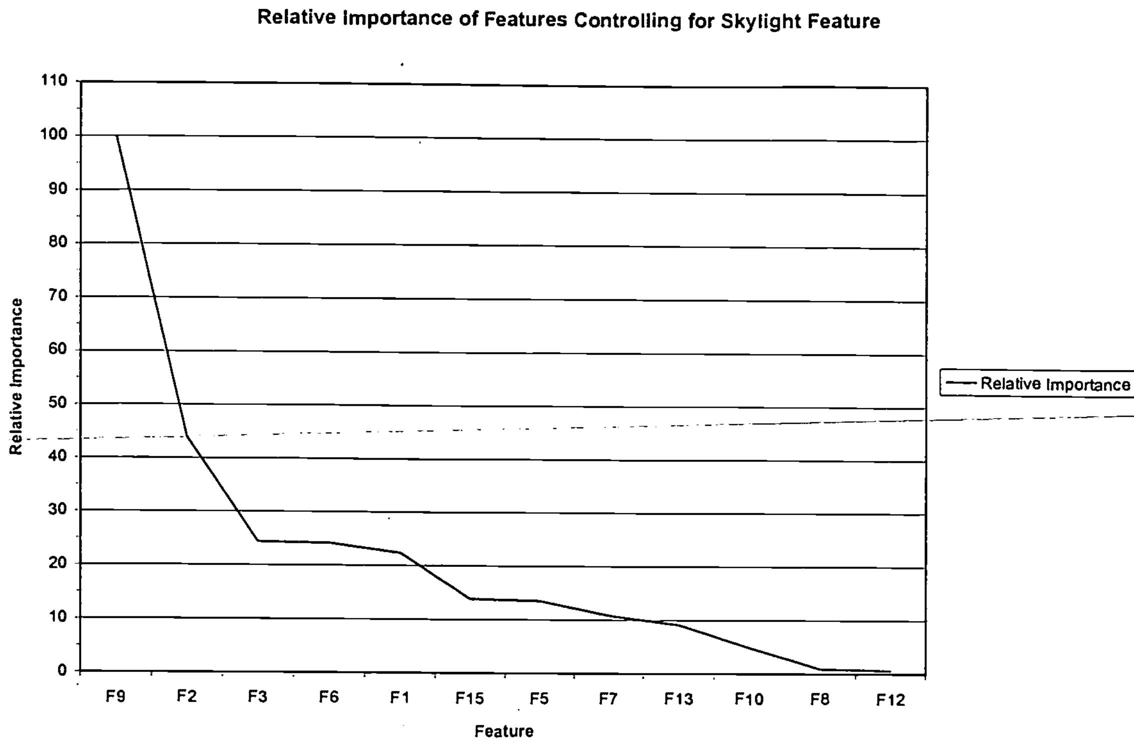


Figure 7


Relative Importance of Features Controlling for Skylight Feature
110
100
90
80
70
60
50
40
30
20
10
0
Relative Importance
F9
F2
F3
F6
F1
F15
F5
F7
F13
F10
F8
F12
Feature
Relative Importance

TM031268

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Classification Trees for Quality Control Processes in Automated Constructed Response Scoring	
Author(s): David M. Williamson; Anne S. Hone; Susan Miller; Isaac I. Bejar	
Corporate Source: Presented at NCME conference, 1998	Publication Date: April, 1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
[X]	[ ]	[ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: David M. Williamson	Printed Name/Position/Title: David M. Williamson	
Organization/Address: The Chauncey Group; 664 Rosedale Rd; Princeton, NJ 08540	Telephone: (609) 720-6608	FAX: (609) 720-6550
	E-Mail Address: dwilliamson@chauncey.org	Date: 5/8/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com