Research Report

Automated Tools for Subject Matter Expert Evaluation of Automated Scoring

David M. Williamson

Isaac I. Bejar

Anne Sax

Research & Development

March 2004
RR-04-14

Automated Tools for Subject Matter Expert Evaluation of Automated Scoring

David M. Williamson and Isaac I. Bejar

ETS, Princeton, NJ

Anne Sax[1]

Chauncey Group International, Princeton, NJ

March 2004

Research Reports provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 19-R
ETS
Princeton, NJ 08541

Abstract

As automated scoring of complex constructed-response examinations reaches operational status, the process of evaluating the quality of resultant scores, particularly in contrast to scores of expert human graders, becomes as complex as the data itself. Using a vignette from the Architectural Registration Examination (ARE), this paper explores the potential utility of classification and regression trees (CART) and Kohonen self-organizing maps (SOM) as tools to facilitate subject matter expert (SME) examination of the fine-grained (feature level) quality of automated scores for complex data, with implications for the validity of the resultant scores. The paper explores both supervised and unsupervised learning techniques, the former being represented by CART (Breiman, Friedman, Olshen, & Stone, 1984) and the latter by SOM (Kohonen, 1989). Three applications comprise this investigation, the first of which suggests that CART can facilitate efficient and economical identification of specific elements of complex solutions that contribute to automated and human score discrepancies. The second application builds on the first by exploring CART's value for efficiently and accurately automating case selection for human intervention to ensure score validity. The final application explores the potential for SOM to reduce the need for SMEs in evaluating automated scoring. While both supervised and unsupervised methodologies examined were found to be promising tools for facilitating SME roles in maintaining and improving the quality of automated scoring, such applications remain unproven and further studies are necessary to establish the reliability of these techniques.

Key words: Automated scoring, computerized simulations, classification and regression trees, neural networks, human scoring, human computer agreement, quality control

Acknowledgements

The authors would like to thank the editor and three anonymous reviewers for numerous helpful comments throughout the submission and review process. The authors would also like to thank Ron Stevens for sharing his perspectives and experiences with neural networks during revisions to this paper. Any deficiencies are, of course, the sole responsibility of the authors. Any opinions expressed are those of the authors and not necessarily those of ETS. This research was initiated while the first author was employed by The Chauncey Group International.

Table of Contents

List of Tables

Page

List of Figures

Page

The rapid pace of advances in computer processing speed, storage capacity, graphics, multimedia and Internet applications has provided the capability to research and field fully computerized delivery and scoring of complex constructed-response tasks. Emerging capabilities for automated scoring of such tasks target particular advantages over scoring with expert human graders, including improved granularity of assessment, assured reproducibility of evaluations, known consistency in application of scoring criteria, complete tractability of scoring rationale, improved task specification, assured objectivity, improved reliability, and greater scoring efficiency (Williamson, Bejar, & Hone, 1999). Automated scoring also extends the advantages offered by computerized testing using multiple-choice items, such as year-round testing, remote testing, Web-based assessment, and adaptive testing, among other benefits. Potential advantages of computerization have inspired a variety of investigations of automated scoring of complex constructed-response tasks such as free-text entry essays (Burstein, Kukich, Wolff, & Lu, 1998); medical practice simulations (Clauser, Subhiyah, Nungenster, Ripkey, Clyman, & McKinley, 1995; Clyman, Melnick, & Clauser, 1995), mathematics (Bennett & Sebrechts, 1996; Sebrechts, Bennett, & Rock, 1991; Bennett, Steffen, Singley, Morley, & Jacquemin, 1997); computer programming (Braun, Bennett, Frye, & Soloway, 1990); computer networking (Bauer, Williamson, Steinberg, Mislevy, & Behrens, 2001); and aircraft diagnostics (Mislevy, 1994; Mislevy, 1996). The National Council of Architectural Registration Boards (NCARB) has implemented simulations for several graphic divisions of the Architect Registration Examination (ARE) (Bejar, 1991; Bejar & Braun, 1994; Kenney, 1997), which has provided the data for this paper. As many of these examples are currently being implemented, or are soon to be implemented, in operational assessments in educational and professional settings, it is clear that automated scoring of complex constructed response is now within the realm of applied measurement with important implications for decision-making. This paper explores classification and regression trees (CART) and Kohonen self-organizing maps (SOM) as potential techniques for improving the efficiency and effectiveness of subject matter expert (SME) evaluation of automated scoring for assessments using complex constructed-response tasks.

On the Role of Subject Matter Experts in Automated Scoring

It has long been recognized (e.g., Greene, 1941, p. 131) that all scoring processes, whether human or automated, must be undertaken in two fundamental stages: evidence identification, in which work products are examined for evidence of ability; and evidence accumulation, in which summary conclusions are derived on the basis of this evidence (see Almond, Steinberg, & Mislevy, 2002 for a more complete description of the four-process model of test administration/scoring that includes discussion of evidence identification and evidence accumulation). Expert human graders are well suited for evidence identification within their domains of expertise. Yet, even experts struggle with evidence accumulation for complex situations—a situation for which automated (statistical) methods can offer demonstrable improvements (e.g., Meehl, 1954). The potential improvement in accuracy of evidence accumulation over expert human graders, coupled with other advantages of automated scoring (e.g., Williamson et al., 1999), has motivated efforts to reduce or eliminate human scoring in favor of automated scoring.

Despite the fallibility, cost, inconsistency, and inefficiency of expert human graders it is precisely these expert human graders who remain the best resource for development of automated scoring algorithms and who serve as the baseline for validation (e.g., Clauser, Margolis, Clyman, & Ross, 1997; Sebrechts et al., 1991; Williamson et al. 1999). However, the role of expert human graders in postimplementation evaluation of automated scoring has received less attention in the literature. One cannot assume that once automated scoring is validated for operational use, there will be no further need for experienced human graders. Ongoing evaluation of operational assessment is a fundamental aspect of validity (e.g., Messick, 1989; Cronbach, 1988), and a multitude of characteristics of assessment administration and scoring (even interface design and circumstances of administration) can have validity implications (Bennett, & Bejar, 1998). Therefore, responsible researchers are adhering to the principle of validity that argues, “especially when an instrument is young, inquiry that clarifies possibilities, anomalies, and boundary conditions—formative evaluation—is worth most” (Cronbach, 1988, p.4). As such, postadministration analyses must be conducted for complex tasks scored with automated scoring systems just as they are routinely conducted for their multiple-choice counterparts. Operational implementation provides an opportunity to evaluate automated scoring for the full range of responses submitted by examinees. This facilitates identification of rare and/or unanticipated circumstances that may cause automated scoring to fail. Therefore, a critical part of the body of evidence needed to support the validity of an operational assessment is the ongoing evaluation of automated scoring. The current challenge for applied researchers is to define and communicate best practices for maximizing the effective use of expert human graders in ongoing operational evaluation of automated scoring—in an environment where diminishing returns may be expected from such postvalidation efforts.

Subject Matter Experts in Operational Evaluation

In operational evaluation of automated scoring five fundamental steps must be undertaken for each submitted response:

1. Determine automated and human scores.

2. Evaluate agreement between automated and human scores and identify discrepancies.

3. Select a discrepant case and identify cause(s).

4. Determine which "perspective" (automated, human, or some compromise) should stand as the evaluation of the examinee's response.

5. If the decision implies a change to automated scoring, devise and apply an alteration to the algorithm and/or assessment procedures generalizable to all responses. For fairness, this would imply the need to reprocess and reevaluate all submitted responses. If the decision is in favor of automated scoring, record as an instance of demonstrated advantage.

SMEs are central to these processes. Initially, SMEs are needed to obtain operational quality scores (as if human evaluation was the sole means for scoring), whether evaluation/scoring is conducted onsite or through remote scoring (e.g., Zhang, Powers, Wright, & Morgan, 2003). The evaluation of agreement is a relatively straightforward comparison of scores on the common scale when checking summary evaluations. However, for more complex scoring rubrics, once the evaluation of agreement extends to the more fine-grained layers of scoring criteria hierarchy the opportunity for discrepancies can grow substantially. Simultaneously, the relevance and importance of the discrepancies can diminish to issues of minutiae that are unlikely to impact summary scores (Bennett & Sebrechts, 1996; Braun et al., 1990). The determination of which perspective on scoring is accepted as the score is a key process with a risk for self-preferential bias. This suggests a need for a committee of SMEs with membership that balances considerations of those who may be "invested" in either human or automated scoring. To the extent possible, it is advisable that committee judgments be conducted blind to the source (automated or human) of the discrepant scores.

A concern of applying a proposed resolution (other than the technical feasibility required for algorithmic changes) is the risk of satisfying the immediate source of one or more scoring discrepancies at the expense of creating discrepancies where none previously existed. To mitigate such risk, all scoring changes must be implemented experimentally to assess the impact on all responses (and potential new discrepancies) before making changes to an operational system. Another issue to be cognizant of with regard to these committee activities is the potential for the personal characteristics of committee members (and the human tendency to be less than perfectly consistent in applying evaluation criteria and producing summary scores) to threaten the generalizability of decisions to the universe of the domain. In the extreme, this can result in a "round robin" of successive committees chasing their algorithmic tails as each new committee composition produces new recommendations for minute changes in automated scoring that reflect its own particular perspectives. A potential solution is to require some additional controls (such as reaching a threshold of impact on solution scores and/or further independent committee confirmation) before alteration of the automated scoring algorithm.

Operational evaluation faces a core challenge of identifying causes of discrepancy between automated and human scores. This requires the direct comparison of the rationales of automated scoring (embodied in the algorithm) with that of human graders. For automated scoring the rationales are transparent, given knowledge of the algorithm, while for human graders the true rationale can be elusive and circumstantially variable, even when following rubrics and documenting rationales for decisions. The potentially elusive and variable nature of human score rationale contributes to the laborious and uncertain nature of determining causes of discrepant scores. Even when the original human grader is present, the elicitation of a rationale can be problematic as there are occasions when human graders are unable to produce a justification for a score that they themselves produced (e.g., Williamson et al., 1999).

The goal of operational evaluation is twofold: to identify differences in scoring rationale between human and automated scoring and to assess the potential impact of any discrepancies on scoring outcomes, and thus validity. The challenge for review committees in meeting this goal is to use observed discrepancies to identify differences in human and automated scoring rationale then determine whether such differences pose a threat to the veracity of automated scoring outcomes, thus implying a need for resolution. This process demands more cognitively of the review committee than scoring demands of the human graders. The review committee must not only evaluate responses under the human scoring rubric (that is, replicate the process of human graders), but it must also:

- *infer* the human graders' situation-specific rationale in applying the rubric,
- *understand* the rationale embodied in the automated scoring algorithm,
- *contrast* the automated rationale with the inferred human rationale,
- *determine* whether a difference in rationale poses a threat to automated scoring validity, and
- *devise* a generalizable resolution to the difference in rationales.

The same advantages of accuracy and efficiency in dealing with complexity promised by automated scoring may also apply for the complex process of evaluation in review committee. That is, automated algorithms may be able to compare, classify, and propose hypotheses about the nature of differences in automated and human scoring rationale for review committees. Such algorithms may also provide a form of sensitivity analysis by providing some sense of the extent to which any differences in rationale result in score discrepancies.

Automated Tools Assisting Subject Matter Experts

The challenges SMEs face in dealing with complexity are well known and the impetus for automated scoring is based on overcoming such inherently human limitations, as well as increasing scoring efficiency and convenience. The same methods used to deal with the complexities of automated scoring may also be potential tools for review committees dealing with complexities in resolving discrepancies between automated and human scoring. Specifically, it would be advantageous if automated tools existed to assist SMEs in:

1. gaining insight into their own scoring criteria (i.e., discovery of implicit criteria);
2. gaining insight about the way they draw summary conclusions from features of responses;
3. identifying inconsistencies and/or undesirable tendencies in SME scoring;

4. identifying needed changes to automated scoring, both immediate and evolutionary, including the inclusion/exclusion of criteria from automated scoring, differential weighting of criteria, and differential consideration of approximations to correct implementation of responses;

5. determining causes of scoring discrepancies between human and automated scores;

6. locating and correcting the automated scores for responses that require human intervention; and

7. distinguishing between important and inconsequential differences in scoring rationale between human and automated scoring with respect to impact on resultant scores.

A means by which automated tools may facilitate these processes is to filter responses by score discrepancy, classify responses by the apparent indicators of such discrepancy, and use this interaction between nature of score discrepancy and patterns present in responses to propose hypotheses about differences between automated and human scoring rationale. Such automated tools would promote efficient and cost-effective SME effort in review committees and therefore have implications for the feasibility and practicality of continued evaluation of operational automated scoring.

This paper explores CART (Breiman et al., 1984) and Kohonen SOMs (Kohonen, 1989) as potential automated tools for facilitating the accuracy, efficiency, and cost-effective use of human experts in automated scoring development and implementation. These methods are employed as automated filters for responses. Such filters would ease the tasks of SMEs by presenting the nature of discrepancies, the key indicators of discrepancy as hypotheses about differences in rationale, and frequencies of occurrence at the outset of committee processes. Beyond these fundamental steps, these methods may also serve to predict whether responses require human intervention to produce a valid score.

Automated tools for classifying responses in optimal groups for the work of SMEs would be of substantial benefit. The initial classification could be based on the nature and direction of score discrepancy. Subclassification could be based on patterns within responses associated with the discrepancy. Resultant sets of responses could be presented to SMEs as sets with a common tendency to elicit a particular type of discrepant scores as a function of differences in automated and human scoring rationale. Indicators of the characteristics of the response could be used by the committee for initial hypotheses about the nature of differences in rationale. The number of responses comprising such a set also suggests the prevalence, and therefore the potential impact, of any discrepancy. The utility of such a presentation is in facilitating review committee work by informing members of systematic discrepancy, proposing indicators of cause, identifying frequency of occurrence of discrepancy as a result of such causes.

Effective automated tools for SME evaluations would facilitate identification and resolution of two classes of automated and human scoring discrepancy. *First-order* findings have direct and immediate implications for score validity. This is typically due to a substantial failing of automated scoring or other critical elements of the assessment that directly impact scoring and may require human intervention. In practice, a variety of circumstances (e.g., programming limitations, interface limitations, software packaging, and distribution, etc.) may delay correcting such malfunctioning aspects of automated scoring. Therefore, it would be useful for an automated tool to also be capable of efficiently identifying responses that may be affected by the malfunction for intervention prior to release of scores. *Second-order* findings are automated scoring or related issues that are not a threat to score validity but provide opportunity for future modifications. In these instances, automated scoring is performing in a valid fashion, but a committee of experts may have suggestions for scoring adjustments. Examples include different recommended weightings of criteria, different tolerance for less-than-perfect implementations of criteria, and inclusion or exclusion of marginally important criteria. Of course, any two groups of experts may disagree on some aspects of practice, so such findings should typically be considered as recommendations rather than flaws in automated scoring. The consideration of second-order issues can enhance possibilities for the future evolution of the assessment and the automated scoring.

Unlike automated scoring, the purpose of automated tools facilitating SME evaluations is to facilitate rather than supplant SMEs' contribution to automated scoring development and operational evaluation. Automated classification tools can be applied to facilitate SME responsibilities for many aspects of assessment development and implementation (e.g., for standard setting see Bejar & Whalen, 1997). This paper briefly discusses both CART and SOM, explores how they may facilitate the role of SMEs in development and operational evaluation of automated scoring, and provides an application to a vignette from the ARE.

Classification and Regression Trees

CART, or classification and regression trees[2] (Breiman et al., 1984; Winston, 1992) are a tool for classification problems in which the target of inference is a discrete classification or prediction (dependent variable) produced on the basis of other variables (independent variables). CART has been successfully applied to classification problems in psychometrics (e.g., Sheehan, 1997, for proficiency scaling and diagnostic assessment; Bejar, Yepes-Baraya, & Miller, 1997, for modeling rater cognition; and Holland, Ponte, Crane, & Malberg, 1998, in computerized adaptive testing) and has some similarities to linear regression. Both CART and linear regression begin with data consisting of a set of observations for which both the independent variables and the dependent variable are known called the training set. This data forms the basis for "training" the model for prediction of the dependent variable from the independent variables. Such methods, which require training data for which observations on both the independent variables and the dependent variable of interest are available, are referred to as supervised learning techniques. In linear regression these training data are used to derive a linear equation. This equation is used to predict the value of the dependent variable on the basis of the independent variable values for subsequent cases. In this prediction, linear regression uses the value of each independent variable in the equation simultaneously to predict the dependent variable. CART also uses training data to derive a classification tree (tree-based vocabulary is pervasive in CART, from growing the tree to the establishment of pruning criteria to reduce the size of a tree). This tree predicts the dependent variable on the basis of the independent variables for subsequent cases, so the method and output of CART differs from linear regression. Figure 1 provides an example of a classification tree from a CART analysis (Steinberg, & Colla, 1992) of a classic data set (iris flower species) used by R. A. Fisher (1936) to illustrate discriminant analysis. The training data consists of 150 flowers; each flower has a dependent variable of species (50 flowers each of setosa, versicolor, and virginica) and four independent variables: petal length, petal width, sepal length, and sepal width.

The CART analysis uses a binary recursive partitioning algorithm to produce, or *grow*, the classification tree. This algorithm operates by conducting a series of *binary partitions* of the training data in which the data is divided into two subsets on the basis of the dichotomized value of a single independent variable. This variable is called the *branching variable* or splitting variable (e.g., all cases with a value greater than X on variable A are placed into one subset and

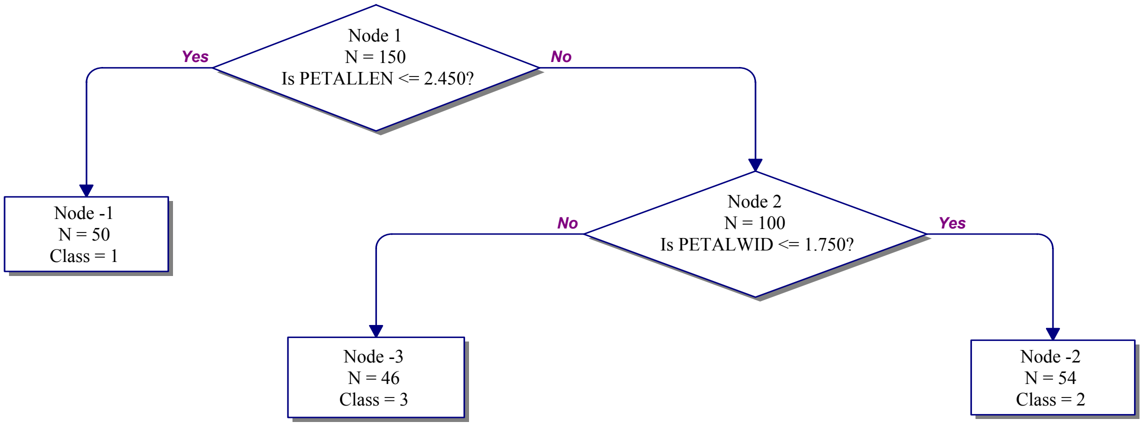


Figure 1. **Sample CART analysis classification tree for the iris data.**

all cases with a value less than or equal to X on variable A in the other). The partitioning algorithm selects the branching variable using a brute force method in which all possible splits for all variables in the analysis are considered for each binary partitioning. For example, if the data consisted of 200 cases with 10 independent variables CART considers up to 2,000 (200 x 10 = 2,000) splits to conduct the binary partitioning. Any problem has a finite number of possible partitions, and CART searches through all possible splits when selecting the partition to implement for each binary partitioning.

Once all possible partitions are completed, the algorithm conducts a rank ordering of the possible partitions on the basis of a *quality-of-split* criterion. Consistent with Breiman et al. (1984) this study uses the *Gini* index as the quality-of-split criterion, which is given by

$$Gini = 1 - \sum_{j} p_j^2 \qquad (1)$$

where p_j is the proportion of cases in class j. The Gini index is 0 when the data consists of 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.[3] The Gini index partitions the data by targeting the extraction of cases from the largest class (the dependent variable class with the highest *n*), if this is possible given the data and the values for the independent variables, and isolating them from the remainder of cases in the data. Then, in the next partitioning iteration it attempts the extraction of the next largest dependent variable class from the remaining data. This process continues with each recursive binary partition until all classes are exhausted or some other termination criterion is met (e.g., the lower limit of permitted *n* per subset is met). The final subset of data is then characterized as a particular type with respect to the dependent variable. Typically, the class assignment mirrors the dependent variable comprising the greatest proportion of cases in the subset. This rule can be modified to a loss function to account for the costs of making a mistake in classification and to adjust for over- or under-sampling from certain classes.

In Figure 1 the diamond-shaped node labeled "Node 1" represents the first binary partitioning of the 150 flowers in the training set. The CART algorithm selected the value of 2.45 for the petal length variable as the best choice for the first binary partition. If a flower has a petal length less than or equal to 2.45 (*N* = 50), then the flower is placed in the subset of data *branching* to the left of the decision node in Figure 1 (the "Yes" direction) where the termination criterion is met and no further partitioning is possible (homogeneous set of data with respect to the dependent variable). This partition terminates in the rectangle labeled "Node-1," which is the coded classification for setosa species. If at the first partition a flower has petal length greater than 2.45 (*N* = 100), it is placed in the subset of data branching to the right of the decision node (the "No" direction) to the second partition labeled "Node 2."

The second partition uses petal width of 1.75 as the branching variable value, resulting in those cases with a petal width greater than 1.75 (*N* = 46) branching to the left to be classified as virginica (coded "Class-3"). Those with petal width less than or equal to 1.75 (*N* = 54) branch to the right to be classified as versicolor (coded "Class-2"). Note the classification error of four virginica flowers misclassified as versicolor in the "Class-2" node and that only two of the four independent variables were used in the classification tree.

This classification tree of binary partitions and their outcomes represents the optimal use of independent variables for dependent variable classification. The classification tree then serves as a means of prediction for new cases without dependent variables. It subjects each case to the series of sequential decisions represented by the classification tree, from the first branch to the final classification, which is the predicted value of the dependent variable. As an illustration, suppose we obtained an iris flower for which the petal measurements (independent variables) are available but the species (dependent variable) is unknown. We can classify the flower by subjecting it to the tests indicated by the branching variables in the classification tree. Suppose we observe a petal length (the first branching variable) >2.45 and so we branch to petal width, which is > 1.75 and so we classify the unknown flower as versicolor (Class = 2). This sequence of decisions about an individual case resulting in a final classification can be represented in table format by decision vectors representing the decision sequence and outcome of the classification tree. An example of such a set of decision vectors that correspond to Figure 1 is presented in Table 1.

Table 1

Decision Vectors Corresponding to the Iris Classification Tree

Classification	Node 1 PETAL LEN.	Node 2 PETAL WID.
1	<= 2.45	
2	> 2.45	<= 1.75
3	> 2.45	> 1.75

Several CART procedures help ensure a reliable classification tree. One such technique is successive growing and pruning of the tree. This technique involves initially growing a *maximal tree*, one that ignores termination criteria and continues partitioning until further partitioning is impossible. From this maximal tree, a number of alternative trees are produced by enforcing termination criteria to *prune* this tree to varying degrees. When the data set is sufficiently large the original sample is divided into a growing sample and a test sample. The growing sample is used to generate both the maximal tree and the pruned versions of this tree. The test sample is used to estimate each generated tree's respective misclassification rates (adjusted by any specified misclassification costs) as a measure of fit, and the tree with the lowest misclassification cost is taken as the optimal tree. The reader may recognize the similarity between this procedure and procedures used to reduce the potential for overfitting in linear regression.

Another procedure, suitable when there is insufficient data for split-half growing and pruning, is 10-fold cross validation. This proceeds by dividing the data into 10 equal-*n* subsets with similar distributions on the dependent variable. Nine of these subsets are used to generate the maximal tree and the pruned versions of the tree, while the remaining section is used to obtain initial estimates of the error rates of the generated trees (again with the potential for incorporating misclassification costs). This process is repeated nine more times, with each iteration using a different one of the 10 sections as the section for determining error rates. The remaining nine sections are used to generate the trees. The results of the 10 iterations are combined into overall error rates for each of the generated trees, which are then used to determine the optimal tree.

It should be noted that CART is appropriate for both discrete and continuous variables and that the use of a variable as a branching variable at one partition does not preclude the use of the same variable again at a subsequent partition. Furthermore, it is not uncommon to have a classification tree with multiple terminal classification nodes of the same classification (value of the dependent variable). Classification tree techniques are nonparametric and make no distributional assumptions, applicable to both exploratory and confirmatory analyses, and robust to outliers and missing data. They excel with complex data, and have been found to outperform certain competitive methodologies, such as discriminant analysis, (Michie, Spiegelhalter, & Taylor, 1994) for some data.

In contrast with regression, the CART goal of developing simple tree structure can result in an independent variable being an important variable for prediction, even if it doesn't appear in the tree as a partition variable. Therefore, CART also characterizes independent variables in terms of their importance to the series of recursive partitioning in the tree. *Importance* is calculated based on the quality-of-split measure attributable to each variable when treated as a *surrogate* (a substitute variable) for the primary splitting variable at each partition. These quality-of-split measures are summed across each partition for each variable and the values are then scaled relative to the best performing variable, which receives an importance value of 100. In the iris data set, for example, the most important variable is petal width, followed by petal length.

Kohonen Self-organizing Maps

Kohonen SOMs (Kohonen, 1989) are a neural network methodology akin to such dimensionality reductions methods as cluster analysis. Like CART, neural networks may be used for supervised learning on the basis of a criterion variable. However, this study applies neural networks as an *unsupervised learning* technique, that is, without reference to the criterion value of interest (e.g., the dependent variable required in CART). As such, SOM (as applied in this study) requires no external variable or training set to classify data. Neural network nomenclature stems from methodological development based on the reaction of brain neurons to sensory stimulus. For example, Hubel and Weisel (1962) presented a set of alternating white and black lines to a particular neuron in the brain. They found that the neuron reacted most strongly when these lines were of a particular orientation and less strongly as the lines were changed from the neuron-preferred orientation to some other orientation. Neural networks evolved from findings that these neuron patterns of greater or lesser excitation to specific stimuli may be mapped. Various sensory inputs can be described as being "close" or "distant" from each other on the basis of metric properties of the sensory input, which is reflected in the patterns of neuron excitation and spatial location in the brain. Neural network methods simulate this brain behavior and regional specificity by establishing an artificial network, or grid, of simulated neurons. Through repeated exposure to patterns in data, case by case, each individual neuron is conditioned to be "excited" by specific patterns of a case in the data and "inhibited" by patterns that do not conform to the those "excitement" patterns. The end result is that each of the cases in the data become associated with the neuron most receptive to the pattern appearing in the particular case. That is, associations are made between neurons and cases with greatest similarity to the "desired" vector of the particular neuron. An artificial *neuron* in neural networks is a classification cell in which member cases have patterns of data that are similar to patterns expected in the cell, thus causing the excitation or activation of the artificial neuron and inhibition of other neurons in the network. These classifications of cases by neuron excitation can then be interpreted, using the established neural network metric, on the basis of their proximity to the other neurons in the network that contain other cases.

To illustrate neural networks, specifically the Kohonen SOM explored in this paper, Fisher's iris data from the CART illustration is useful again. Because SOM is being applied as an unsupervised method, the input data, still referred to as the training set, for this example consists of the same set of iris flower measurements of petal length, petal width, sepal length, and sepal width, but without data on the species of flower. The SOM produces an array of artificial neurons, with each neuron having some initial weight (different for each neuron) attributed to each of the possible input variables, resulting in a weight vector for the neuron. An initial case vector, in this case one of the flower measurement vectors, is presented to the network and the artificial neurons all "compete" for this case. This competition is conducted by computing the Euclidean distance between the points represented by each neuron's weight vector and the input vector, called the neuron's *activation* level, for the case. As a result of this competition, the neuron with weights representing the closest match (the smallest Euclidean distance) to the case is the "winner" of the competition and the case becomes associated, or clustered, with the winning neuron. In turn, the weights of the "winning" neuron are permitted to adjust somewhat (the degree of adjustment is referred to as the *learning rate*) towards the case that was just associated with it. The neurons near the "winning" neuron, referred to as the *neighborhood*, are also permitted to adjust weights toward the case, but to a lesser extent than that of the winning neuron. This process of neuron competition for cases, case association, and updating of neuron vector weights in response to neuron association is continued for each case in the data set. For the iris data, each flower's vector of values for petal length, petal width, sepal length, and sepal width acts as the input. The Euclidean distance is computed between the flower vector and the weights for the competing neurons, with the winning neuron becoming associated with the flower in question. This process is continued for each of the flowers in the data.

Once the entire data set has been processed and each case associated with an artificial neuron, the entire training set is reprocessed in the same fashion. The weights for each artificial neuron that resulted from the initial processing serve as the initial weights for the reprocessing of the data. Although in the initial analysis of the data the neighborhood consists of the entire neural network, during the reprocessing of the data this neighborhood becomes more restricted around the winning neuron. In addition, the extent to which the weights for each neuron are altered after each case classification is also reduced from the initial processing of the training set. This iterative process of "learning" from the training set is repeated multiple times until, ultimately, the neighborhood is restricted to the single winning neuron, the adjustment to the winning neuron's weight becomes small, and termination criteria are met. Once neural network learning is complete, the association of each case with a particular neuron becomes the final classification of cases, with the regions of neighboring neurons in the network comprising regions of similarity among case classifications. Thus, a grid of artificial neurons is produced with each neuron associated more or less strongly with particular input data patterns. This grid of neurons and the cases comprise a pattern that might be conceptualized as a contour map of the data. Groups of similar cases cluster together in specific neurons or SOM regions and the number of cases associated with a particular neuron determines the "height" of various SOM neurons. An example of the output classification of Fisher's iris data from a 1 x 3 SOM (a one-dimensional neural network) for the iris data is provided as Figure 2. In Figure 2 each box represents the location of a particular neuron (labeled Neuron 1 through Neuron 3) in the 1 x 3 SOM. Immediately beneath the neuron number is the number of iris flowers associated with that neuron, and thus the number of flowers classified together with the neuron. Beneath these neuron sample sizes are the iris species (setosa, versicolor, and virginica) corresponding to the neuron, with the percentage of flowers of that species associated with the neuron in question immediately following in parentheses.

Neuron 1	Neuron 2	Neuron 3
n = 31	*n = 29*	*n = 30*
versicolor (74%)	versicolor (24%)	setosa (100%)
virginica (26%)	virginica (76%)	

Figure 2. **A 1 x 3 Kohonen self-organizing map result classification for Fisher's iris data.**

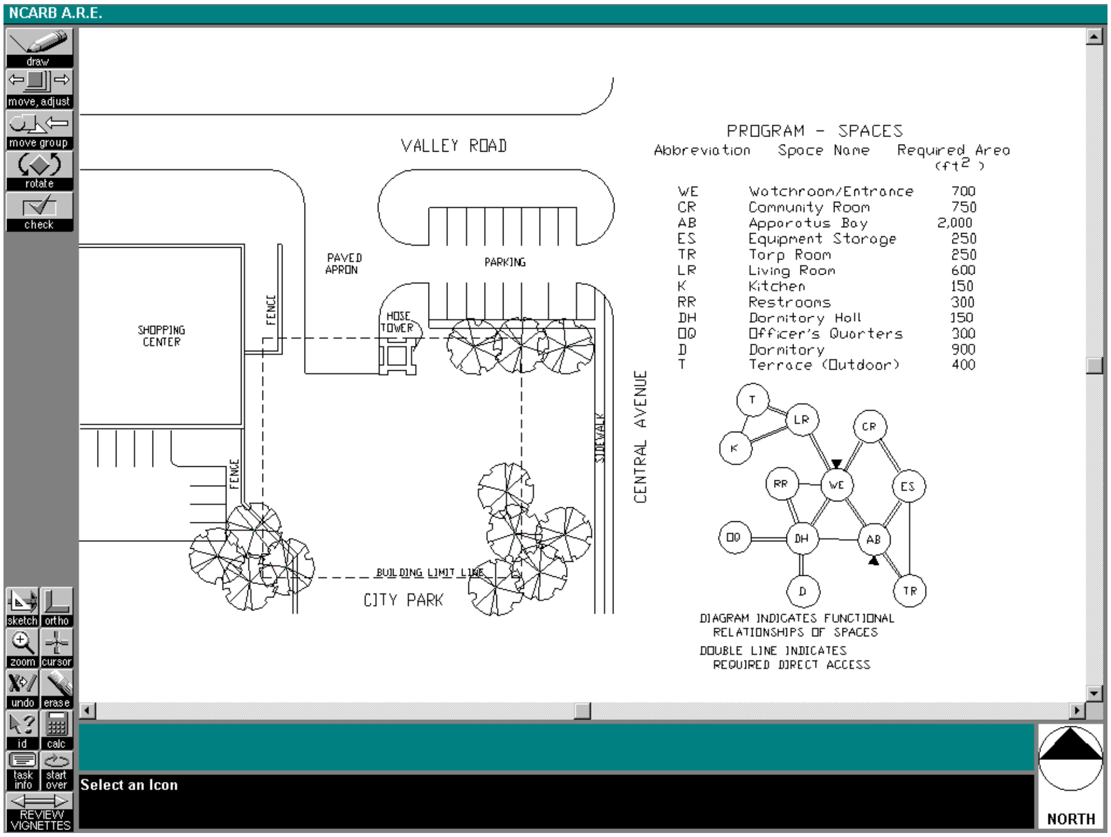
To the extent that various regions of the SOM correspond to certain aspects of interest in the data, such techniques permit rapid and efficient organization of complex data for a variety of purposes, including as proposed methods for automated scoring (e.g., Stevens, Ikeda, Casillas, Palacio-Cayetano, & Clyman, 1999). In addition, the representation of the data in a SOM provides a means for swiftly visualizing trends and tendencies within complex data. (For more about the Kohonen SOM, see Balakrishan, Cooper, Jacob, & Lewis, 1994; Murtagh & Hernandez-Pajares, 1995; Ripley, 1996; or Waller, Kaiser, Illian, & Manry, 1998).

Overview of the ARE

The assessment data used to explore the utility of CART and SOM as automated tools for SME evaluation comes from the ARE. The ARE consists of nine divisions; six of which are multiple-choice CBT examinations and three of which are fully computerized graphic simulations of architectural design tasks. The candidate receives either a “Pass” or “Fail” for each division and must pass all divisions to become registered. Each of the three graphic divisions is made up of a series of small, focused design problems called “vignettes.” A total of 15 different vignettes comprise the three graphic divisions of the ARE. An example of a base diagram representing the examinee starting point for an architectural design vignette not currently used in the ARE is provided as Figure 3. The candidate would complete the design of structures and their functional relationship on the given site using the CAD interface provided. (More information on the administration and scoring of the ARE may be found in Bejar & Braun, 1999; Kenney, 1997; and Williamson et al., 1999).

The scoring of the ARE vignettes, referred to as a “mental model” approach (see Williamson et al., 1999), is the culmination of studies of the criteria, judgments, weighting, and considerations of expert human graders during holistic grading processes for operational paper-and-pencil vignettes (Bejar, 1991; Oltman, Bejar, & Kim, 1993; Bejar & Braun, 1994). Scoring is automated and modeled after aspects of the human scoring processes (see Williamson et al., 1999). One stage of automated scoring is the representation of elements of each candidate’s response to a vignette (called a *solution*) as *features*, or specific and distinct scoring components. Individual features may represent any one of a multitude of solution aspects. They could range from simple assessments of completeness (e.g., did the candidate provide all nine rooms in the building?) or measurements of important elements (e.g., distances, ratios, or slopes of and between objects) to more complex assessments of relations among various aspects of a solution (e.g., sunlight and shadows or the ability for building users to enter each space in a building from the lobby). For a more detailed account of how automated processes transform graphic design elements into these scoring features refer to Bejar (1991) and Oltman et al. (1993).

In the automated scoring of each candidate’s vignette solution each feature is scored as an acceptable (A), indeterminate (I), or unacceptable (U) implementation of the required vignette feature. In this scoring the indeterminate (I) level represents a borderline implementation as judged by expert human graders who may disagree or be uncertain whether the implementation



PROGRAM - SPACES

Abbreviation	Space Name	Required Area (ft^2)
WE	Watchroom/Entrance	700
CR	Community Room	750
AB	Apparatus Bay	2,000
ES	Equipment Storage	250
TR	Tarp Room	250
LR	Living Room	600
K	Kitchen	150
RR	Restrooms	300
DH	Dormitory Hall	150
DQ	Officer's Quarters	300
D	Dormitory	900
T	Terrace (Outdoor)	400

DIAGRAM INDICATES FUNCTIONAL RELATIONSHIPS OF SPACES

DOUBLE LINE INDICATES REQUIRED DIRECT ACCESS

Select an Icon

Figure 3. **Computer screen showing the base block diagram drawing upon which the candidate would construct buildings according to the vignette program requirements.**

is acceptable or unacceptable. Like the individual features, the total vignette is scored as an acceptable (A), indeterminate (I), or unacceptable (U) demonstration of competence in the targeted domain content. This total vignette score is determined through an aggregation process that mimics the "chunking" process used by expert human graders, who scored the individual elements, grouped the element scores into blocks of related elements for which they made summary judgment, and scored several such sets of elements to determine the final vignette score. This expert human grading process for dealing with complexity was reproduced in the structure of the "mental model" for automated scoring.

CART and SOM as Automated Tools for SMEs

As previously discussed, in a typical evaluation of automated scoring SMEs must: identify scoring discrepancies, infer aspects of human scoring processes, understand the automated algorithm, contrast rationales of automated and human scoring, determine any threat to veracity of automated scoring, and devise a generalizable resolution. The complexity of solutions, the large volume of solutions, and the inherent limitations of human judgment processes make the identification of fine-grained flaws in automated scoring a difficult process.[4] The use of automated tools for classification of solutions prior to review are intended to improve the ability of SMEs to perform by conducting the first step (adequacy evaluation) and facilitating the second step (hypothesis development) by producing classifications that imply meaningful hypotheses about the nature of scoring discrepancies between automated and human scores. These tools might also be used to leverage preexisting solutions for which human scores are available (e.g., from operational administration prior to automated scoring, developmental efforts, validation studies, or through ongoing evaluation of operational implementation), subject the solutions to automated scoring, apply the automated tools to facilitate SME review, and conduct the final review and resolution of classification groups.

There are two potential applications of these automated tools: explanatory, to help articulate the causes of discrepancies between automated scores and those of expert human graders; and predictive, for automated identification of solutions of specific interest, particularly those that require human intervention to maintain valid scores. This paper explores both potential applications.

In the explanatory capacity, these procedures may make SME review more fruitful and efficient by eliminating discussion on issues that do not result in a human-automated score discrepancy (e.g., over minutia that can have no impact on the outcome of scoring). They help elicit more accurate representations of the criteria experts use during grading (by revealing implicit criteria or differential criteria weighting), or even eliminate the expense of convening a committee and instead permit expert review remotely over the Internet.

In the predictive capacity, a targeted selection of cases expected to have human-automated score discrepancy that would require a change to the automated score in order to maintain score validity would be more effective than random sampling techniques and more efficient than reviewing the entire body of examinee responses. This would be particularly true when the responses requiring human intervention are the result of rare combinations of response elements, which is a more likely situation than pervasive instances once an automated scoring system becomes operational.

The first two applications in this paper explore the potential utility of CART in exploratory and predictive capacities. The third application investigates the potential of SOM, applied as an unsupervised learning technique, to reduce the need for human experts in review processes by comparing the application of SOM without a criterion value to the results of the CART techniques. If similar results are obtained, SOM may be applied to sets of solutions without the prior calibration of data for which expert human evaluations are available, thereby reducing the need for SME evaluation in automated scoring development and operational evaluation.

Application 1

Method

Data

This application uses a training set of 326 candidate solutions for an ARE vignette that received both automated and human scores. The human scores were produced by a committee of six human graders experienced in holistic scoring of ARE solutions. The committee was divided into two groups so that three graders examined each solution and scoring was conducted according to their own documented holistic criteria (see Williamson et al., 1999). The vignette in question was selected for this study on the basis of its complexity. While it had a moderate number of scoring features for evidence identification in comparison to other ARE vignettes, these features included a mixture of relatively independent features and features that were highly interrelated but of varying degree of weighting in the evidence accumulation process. As such, this vignette was expected to provide the types of complexities commonly encountered in automated scoring of complex constructed-response tasks.

Vignette scores, both automated and human, of A, I, and U were transformed to numeric representations of 3, 2, and 1, respectively and difference scores were computed by subtracting the numeric automated score from the numeric human score. These difference scores comprise the CART dependent variable (all variables were treated as categorical in CART analysis and the difference score was selected as a convenient means of identifying the degree and direction of categorical separation, if any, between the human score and the automated score). Automated scoring features, which comprised the CART independent variables, were also recorded as A, I, or U by the automated evidence identification process and used to compute the final automated score of A, I, or U for each solution (Bejar & Braun, 1994).

Design and Procedure

This application addressed two aspects of CART. The first phase explored CART importance values as a tool for identifying malfunctioning features in automated scoring, and the second phase explored sets of solutions in CART classifications (terminal nodes in the CART tree) of interest for set-specific trends indicative of human and automated score discrepancies. The use of such solution pools in the second phase permits the targeting of specific types of differences in criteria or tolerances and weighting between human graders and automated scoring. That is, the classification tree acts as a filtering process that produces groups of solutions with consistent trends based on explicit type and degree of human and computer score discrepancy.

The first step of Phase Two was to identify feature vectors representing the navigation of data through the classification tree to the terminal classification nodes (see the earlier CART discussion for an example of how feature vectors were derived on the basis of Fisher's iris data). These feature vectors (labeled A through N) and their difference score are presented in Table 2, which is a representation of the CART trees for the ARE data in the same way that Table 1 is an alternative representation of Figure 1. However, Table 2 consists of categorical distinctions rather than divisions of continuous variables as in Table 1. There are a total of 13 binary partitions in the classification tree, represented in the table as columns N1-N13. Each binary partition indicator in the table also specifies the feature code (e.g., F2, F9) for the feature being evaluated at that partition. As previously mentioned, features can be used as a partitioning variable in multiple nodes of a classification tree. Feature vectors A, B, and C are associated with Terminal Node –2, in which the automated score was A and the human score was U. These feature vectors suggest solutions for which the human graders are using additional criteria not addressed by automated scoring, allowing less tolerance for less than perfect feature implementation, or utilizing greater weight for inadequate features in the solution. The opposite pole finds feature vectors M and N associated with solutions in which the human scores are higher than automated scores. Each set of solutions, based on feature vector, were examined by a registered architect for coherent architectural trends that would suggest the nature of differences between automated and human scoring.

Table 2

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

Results

The discussion for this paper references a hypothetical "exemplar" vignette context that has characteristics altered from the actual vignette to permit discussion of findings while maintaining test security. Although all analyses and results are from the actual operational vignette (including the use of the same feature nomenclature), the vignette task and feature characteristics are altered for security purposes. However, in this alteration of context the hypothetical features retain the same purpose and functional relationships, including dependence characteristics very similar to the "live" features, to maintain the fidelity of discussion to the actual feature findings. (To remind the reader of fabricated nature of content discussion that mirrors real findings about operational features, invented vignette detail will be indicated with an asterisk). For this exemplar vignette the candidate would begin with a 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 specific requirements (an example of a completed solution is provided as Figure 4).

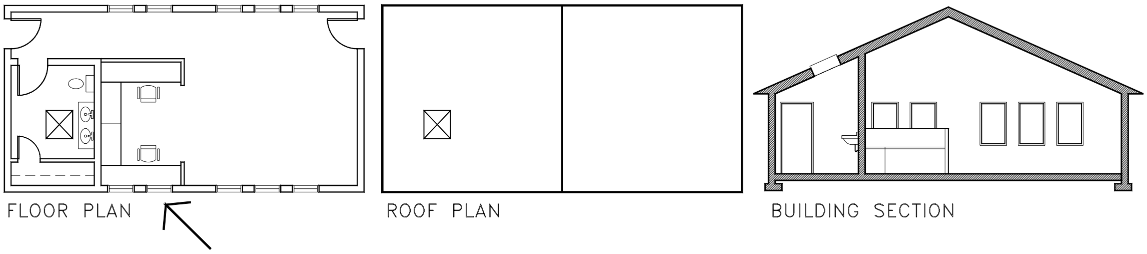


***Figure 4.* Floor plan, roof plan, and section view of the exemplar vignette showing the location of additional skylight as a result of candidate misinterpretation of the floor plan.**

A bar graph of CART feature importance is presented as Figure 5. The values suggest that feature F2 (skylight location*) is the major contributing factor to discrepancies between human and automated scoring, with F3 (flashing*), F9 (eave height*), F15 (water flow*), and F1 (gutter length*) being contributing factors. Architectural review of vignette solutions with score discrepancies was conducted with emphasis on these features. A vignette requirement 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. The architectural examination of the solutions with discrepant scores revealed that 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* (F15) to prevent water from leaking into the building. The placement of an extraneous skylight,* and the accompanying flashing* and cricket* would result in an unfavorable evaluation of these features by automated scoring.

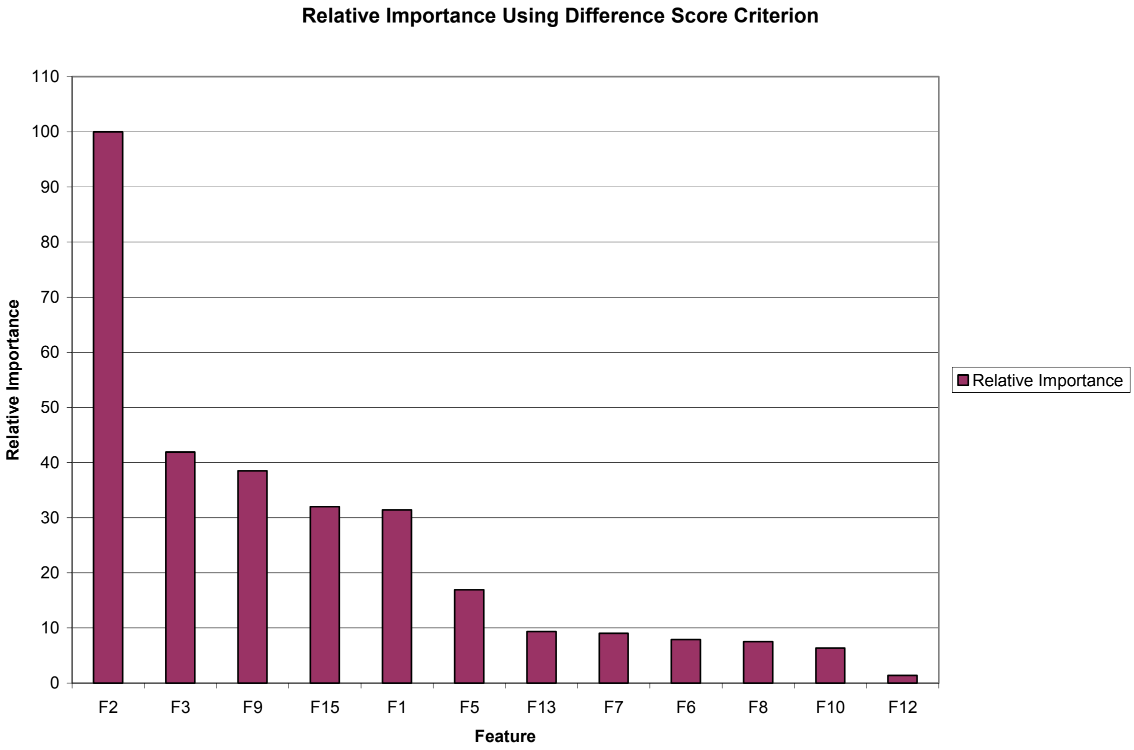


Figure 5. **Bar graph showing the relative importance of automated scoring features using the difference score as the dependent variable.**

From this observation and the fact that human holistic evaluations tended to give credit to these solutions (but not for placement of extra skylights* over other areas of the room) it was inferred that human graders made allowances for the previously unanticipated possibility that some candidates were interpreting the partitioned cubicle as a room. (The building section view of Figure 4 was not available to the candidate or the human graders but is included here for the benefit of the reader.) Although not a deficiency in the automated scoring itself, this issue is potentially l ambiguous for candidates trying to fulfill the requirements of the vignette and so examinees were given credit for such a misinterpretation of the base diagram.

Architectural examination of eave height* (F9) in the solutions with discrepant scores revealed that the human graders were occasionally overlooking this element in their evaluation process, despite the fact that it was included in their written criteria. Because the human graders would often rely on “eyeballing” to judge roof height* at various points, they occasionally missed cases in which the ridge height* and slope* were not consistent with the eave height*. An example presented in Figure 6* shows two roof plans (as seen by the candidate and human graders) and their associated building section views (created for this illustration).

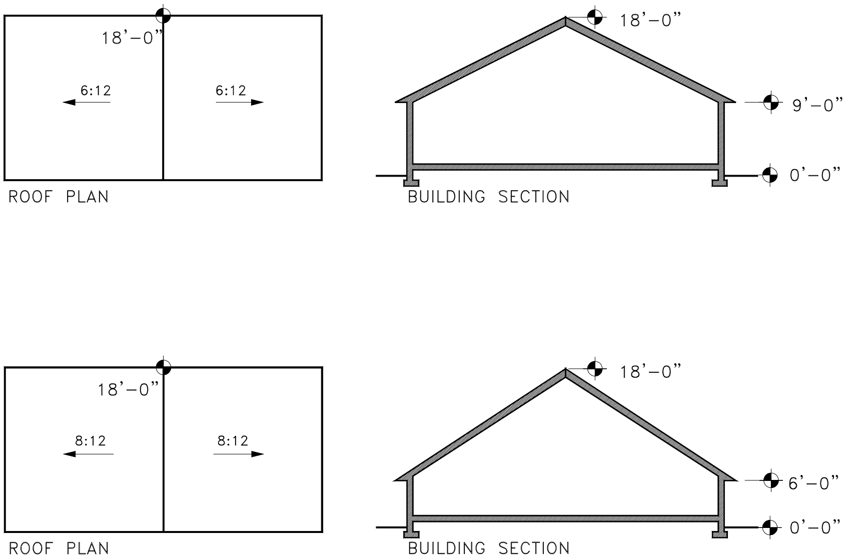


Figure 6. **Roof plans and section views of the exemplar vignette showing the correct implementation of eave height (top) and an incorrect implementation (bottom).**

Although many of these inappropriate implementations that human graders missed were of low to moderate severity, some were dramatically inappropriate. In this case CART techniques served to document a specific advantage of automated scoring.

Architectural examination that emphasized the gutter length* (F1) feature revealed an apparent minor difference in the relative tolerance of less-than-perfect implementation and weighting. Human graders appeared to have less tolerance of less-than-perfect implementation or greater weighting of the feature in aggregation than was implemented in the automated scoring.

In Phase Two, the CART classification tree pools of solutions (the terminal nodes of the tree) containing discrepancies between automated and human scores (pools with feature vectors A, B, C, M, and N) were evaluated with the intent of identifying coherent trends that impacted score discrepancy. Architectural review of these solutions revealed specific types of differences in automated and human scoring processes associated with each CART feature vector. These differences are summarized in Table 3 and include the human use of criteria not considered in automated scoring, heavier weighting of existing criteria by the human graders than by the automated scoring, human graders overlooking a feature utilized in automated scoring, and human graders providing for possible misinterpretation of the base diagram by examinees.

Table 3

Summary of Architectural Review Findings for CART Vector Pools

Feature vector	*N*	Findings
A	30	Two additional criteria used by experts and experts weighting one feature (F2) more heavily than automated scoring.
B	7	Heavier weighting of a particular feature (F5) by experts.
C	25	Heavier weighting of a particular feature (F3) by experts.
M	15	Experts overlooking a feature (F9).
N	30	Experts providing lenience for candidate misinterpretation of base diagram.

Two of these causes of discrepancies were also identified in Phase One as discussed above, providing some corroborating evidence of the previous findings. As such, the CART feature vectors were able to assist in a characterization of the vignette solutions by the specific issues that caused the discrepancy between human and automated scores. Of particular interest is the tendency for feature vector N to characterize cases where candidates are misinterpreting the floor plan in the base diagram, first-order instances in which human intervention is necessary to maintain the validity of automated scores.

Discussion

Phase One utilized CART relative importance values as indicators of sources of discrepancy between human graders and automated scoring. This resulted in the identification of a first-order validity issue of potential ambiguity in the base diagram that required human adjudication to maintain score validity. The CART relative importance values also led to the identification and documentation of a specific advantage of automated scoring in providing a more accurate determination of feature F9. The use of classification trees therefore appears capable of assisting in the identification of sources of scoring discrepancies, both for the improvement of scoring and in the documentation of relative merits of competing scoring procedures. This technique is not limited to human-automated score comparisons; a similar procedure could be used to compare competing methods of automated scoring or differences in human scoring (e.g., Bejar et al., 1997).

Phase Two used CART feature vectors to produce pools of related solutions with score discrepancies in order to more easily identify potential causes of discrepancies. This procedure identified the same issue from Phase One regarding candidate misinterpretation of a base drawing (from vector N). This examination also corroborated Phase One in identifying a specific feature for which automated scoring has an advantage over human scoring. The fact that this corroboration was accomplished with fewer solutions being examined suggests that examination of solution pools from the classification tree has some advantage over reviewing the entire pool on the basis of importance values. In practice, this suggests that the importance values would best serve as a guide for examining the resultant pools of solutions from the classification tree rather than as a guide for examination of the complete pool of solutions. Several other issues were identified from the classification tree pools, including the identification of two elements that human graders were considering but which were not incorporated into automated scoring (feature vector A). Also identified were several features for which automated scoring appears to implement different weighting and tolerances than human graders (feature vectors A, B, and C), although some of these discrepancies appear to occur infrequently.

The use of these techniques to classify solutions and conduct coherent reviews for sources of score discrepancy has obvious advantages over random selection and/or selection of 100% of solutions for review. Of course, conscientious reviewers without automated tools would likely also use discrepancy between human and automated scores as a basis for selecting solutions to review. However, using CART as an automated tool in the review process would appear to have specific advantages over even this best course for unaided review. Specifically, using only the degree and direction of score discrepancy (the full information available for classifying the solution discrepancies in the absence of more sophisticated automated methods emphasizing pattern recognition abilities) would result in solutions with feature vectors A, B, and C (all with discrepancy score of –2) being in a single, undifferentiated group of solutions. Similarly, solutions with feature vectors M and N (all with discrepancy score of 1) would be in another undifferentiated group. Yet it is evident from Table 3 that the application of CART techniques to classification is capable of yielding more fine-grained sets of solutions (e.g., sets with feature vectors A, B, and C), each containing solutions with consistent and coherent trends, but with different interpretations across sets. Most notably, these that feature vectors with the same value for the score discrepancy (feature vectors M and N) can have substantially different interpretations. One (vector M) supports the advantages of automated scoring over human scores, and the other (vector N) indicates of a situation in which there is a true problem with the assessment. Although it is certainly possible, if not likely, that the SMEs would identify these particular extremes from the unaided mixed set of vector M and N solutions, the ability to create more issue-specific data sets for analysis would nonetheless be advantageous. In particular, CART not only classified the solutions by nature of the causes of the scoring discrepancy, but also showed that these categorizations serve as immediate, automatically generated hypotheses about the cause of discrepancy between automated and human scoring. These hypotheses could then be used by the SME committee as starting points for their investigation of the scoring discrepancy. For example, feature vector A in Table 2 suggests to the reviewers that the source of discrepancy is strongly related to features F3, F2, and F9, and not other potential scoring features. The utility of this automatically generated hypothesis about the causes of scoring discrepancies between human and automated scores is by necessity focused only on the automated features. It cannot examine human feature scorings without those as input, yet it still serves several purposes for operational evaluation, including:

- focusing SME attention on most likely sources of the discrepancy (initial hypotheses about the source of error generated for the SMEs),

- creating a coherent set of solutions that exemplify this discrepancy and the causal pattern contributing to the discrepancy for examination of trends across these related solutions, and
- discouraging dwelling on "red herring" issues regarding minor scoring features or other issues that cannot impact the resultant summary solution scores.

Indeed, for more subtle differences, greater variety of differences, and larger solution pools that characterize operational scoring review this ability to provide automatically generated hypotheses about the source of scoring discrepancy may mean the difference between success and failure in identifying issues that impact score discrepancy. This finding illustrates the potential for such automated tools to facilitate SME evaluations of automated scoring.

Application 2

Method

Data

This application utilized the same training set utilized in Study 1 and an additional set of 1,117 operational candidate solutions for which no human scores were available.

Design and Procedure

This application builds on the first by exploring the predictive use of classification tree vectors as an automated solution selection algorithm to identify vignette solutions requiring human intervention to maintain score validity. Specifically, given that some candidates may be misinterpreting the cubicle as a room requiring a skylight, can classification tree vectors provide a means for identifying solutions where this misinterpretation would result in a different score? The benefit of automating the selection of solutions for both corrective and exploratory procedures is a potentially substantial increase in efficiency and corresponding reduction in effort and costs that avoids certain sampling risks. The first phase of this study explores a technique that maximizes overall accuracy and efficiency at the expense of omitting some solutions of interest (appropriate for exploratory reviews) whereas the second phase uses a technique that is more liberal in solution selection at the expense of efficiency (appropriate for score interventions to maintain validity).

Architectural examination of 100% of solutions (N = 1,117) that did not receive human scores was performed to identify and adjudicate scores for candidates who misinterpreted the cubicle as a room. This 100% examination and intervention process can be time consuming and costly. Because the first study suggests that feature vector N indicates cases for which candidates misinterpret the cubicle, the first phase investigates the direct use of this feature vector for identifying cases of candidate misinterpretation requiring score adjudication. Cases from the sample of 1,117 solutions with feature vectors matching vector N were identified, and the accuracy of solution selection for adjudication was evaluated on the basis of a 100% review.

As the classification tree producing feature vector N was trained to classify solutions based on the criterion discrepancy between human scores and automated scores, the feature vector may not be as appropriate as one trained specifically to identify cases where intervention is required for the particular issue of candidate misinterpretation. Therefore, Phase Two of this study examines the use of a CART tree produced on the training set using discrepancy between original automated score and adjudicated score as the criterion. The resultant feature vector is used to select solutions requiring score adjudication from the sample of 1,117 solutions.

Results

Outcomes of utilizing feature vector N to identify solutions requiring score adjudication are presented as Table 4. The overall predictive error rate is low, with only 69 (6%) misclassifications. The use of feature vector N for case selection would substantially reduce the human review burden as only 114 (10%) solutions would be selected. This substantial efficiency would come at the cost of 40 (32%) of the solutions requiring score adjudication remaining unidentified. For such first-order interventions, in which candidate scores are being adjusted as a result of the selection process to ensure the validity of scores, this error rate is problematic. Therefore, this highly efficient but less accurate technique may be appropriate for exploratory reviews. For such exploratory reviews, the intent is not to adjudicate scores but to investigate the rate of occurrence or tendencies toward certain actions, ranges of particular types of solution implementations, or score discrepancies.

Table 4

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	1,003	114	1,117

In Phase Two, the CART analysis identified a single feature, F2 (skylight location*), as the feature vector. The accuracy of using this feature vector for case selection from the sample of 1,117 solutions is presented as Table 5. The overall classification error rate of this technique is higher than for feature vector N selection, with 229 (21%) misclassifications. The use of this feature vector reduces the burden and expense of the review process, though not to the extent of feature vector N, as 354 (32%) of all cases were selected for review. An advantage of using this feature vector for the example in question is that all of the solutions that required adjudication were selected with a substantial (68%) reduction in adjudication effort and expense. Therefore, although not as efficient as the feature vector N selection technique, the use of a feature vector specifically trained to identify the issue was successful in selecting all future cases requiring adjudication, demonstrating high accuracy and substantially greater efficiency than a 100% review process.

Table 5

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	1,117

Discussion

The results from Phase One suggest that for exploratory reviews the automation of case selection on the basis of classification tree vectors utilizing a criterion of difference score between human and automated scores can be a very efficient means for collecting solutions containing examples of scoring issues. However, this method would appear inadvisable for cases of first-order intervention as this efficiency comes at the cost of failing to select all cases that may require human intervention. That is, this method maximizes overall accuracy and efficiency, but may be too conservative in case selection for instances in which it is critical that 100% of relevant cases be selected.

The results of Phase Two suggest that classification tree vectors can be used to automate case selection for first-order intervention when the appropriate criterion is utilized in the training set (e.g., the difference score between automated score and adjudicated score). Such automation provided for a substantial increase in efficiency (68% reduction in solutions reviewed) while selecting 100% of the cases that required adjudication. The more expansive and complex the factors contributing to the discrepancy, the more advantageous the ability to automatically generate an algorithm for the automated selection of cases requiring first-order intervention.

The primary difference between the classification trees produced in Phase One and Phase Two is the use of a different, and more issue-specific, criterion for growing the classification tree in Phase Two. However, there are other options for controlling the accuracy and efficiency of automated case selection through specifying the relative cost of error types (errors of solution exclusion or inclusion in the automated selection) when producing the classification tree from the training set. Through this process differences in classification error severity can be controlled during tree production, and an examination of the resultant cross-validation classification accuracy can help the user determine if the classification tree is sufficiently accurate to be relied on for the selection of future cases.

Application 3

Method

Data

This study used the same training set from Application 1 and Application 2. As the SOM is applied as an unsupervised learning technique, neither the human holistic scores nor the automated scores were included in the data. Instead, only the automated scoring feature scores were used as the data.

Design and Procedure

This study explored the potential for SOM, applied as an unsupervised technique capable of performing in the absence of human evaluations, to provide advantages similar to those of CART, thereby providing for additional efficiency if used in conjunction with CART in SME review processes. The data were analyzed using Neural Connection (SPSS, Inc., & Recognition Systems Group, 1998). Given that there were 14 feature vectors in Table 2, a 5 x 5 neuron grid was specified for SOM, creating a total of 25 possible neurons in the network.[5] The composition of the resultant neuron clusters was compared to the previous CART feature vector results.

Results

The correspondence between results of SOM analysis and the original CART feature vectors is provided as Figure 7. In Figure 7 (which uses the same neuron representation as the iris data in Figure 2) each box represents the location of a particular numbered neuron in the 5 x 5 SOM. Immediately beneath the neuron number is the number of solutions associated with the neuron. Beneath these neuron sample sizes are the feature vector codes (A through N) from Table 2 for cases classified with the neuron. In parentheses following each feature vector code is the percentage of solutions in the neuron having that feature vector.

It is initially striking that a single neuron (Neuron 25) contains 45% of the total sample of cases (261) whereas the remaining neurons contain no more than 7% of the cases. However, the fact that all of the cases in Neuron 25 correspond to CART feature vector J, in which the human graders and the automated scoring were in perfect agreement, provides some reassurance that this peak in the neural net is an appropriate representation of the data. It is certainly not

Neuron 1 *n = 12* B (50%) D (17%) H (33%)	Neuron 2 *n = 0*	Neuron 3 *n = 10* C (30%) J (10%) N (60%)	Neuron 4 *n* *= 3* K (100%)	Neuron 5 *n = 19* C (11%) I (11%) K (11%) M (58%) N (11%)
Neuron 6 *n = 0*	Neuron 7 *n = 13* C (85%) G (15%)	Neuron 8 *n = 13* K (8%) L (8%) N (85%)	Neuron 9 *n = 0*	Neuron 10 *n = 1* C (100%)
Neuron 11 *n = 6* A (17%) J (50%) K (17%) M (17%)	Neuron 12 *n = 0*	Neuron 13 *n = 19* A (74%) E (26%)	Neuron 14 *n = 0*	Neuron 15 *n = 11* A (9%) F (45%) K (18%) L (9%) N (18%)
Neuron 16 *n = 4* I (25%) K (50%) L (25%)	Neuron 17 *n = 2* C (50%) E (50%)	Neuron 18 *n = 3* A (100%)	Neuron 19 *n = 0*	Neuron 20 *n = 3* J (100%)
Neuron 21 *n = 5* A (20%) I (40%) K (40%)	Neuron 22 *n = 3* E (33%) I (67%)	Neuron 23 *n = 17* A (12%) C (6%) J (76%) N (6%)	Neuron 24 *n = 0*	Neuron 25 *n = 117* J (100%)

***Figure 7.* A 5 x 5 Kohonen self-organizing map output for the ARE data showing CART/Kohonen comparisons.**

surprising to find a peak that corresponds to feature vectors indicative of perfect agreement between automated and human scoring when the validity of the automated scoring has already been evaluated through such comparisons (Williamson et al., 1999). With the contribution of Neuron 25, the lower right-hand region of the SOM (Neurons 20, 23, and 25) represents a region of agreement between human and automated scores. This region comprises 97% of all feature vector Js in the sample and 84% of all feature vectors representing complete agreement in the sample.

Neuron 5 represents the cases corresponding to feature vector M from the CART analyses, with 92% of all feature vector M cases associated with this neuron. Similarly, Neurons 3 and 8 are associated with feature vector N from the CART analyses, with 81% of all feature vector N cases associated with these two neurons. Taken together (and ignoring the three cases that appear in Neuron 4), these suggest that the upper right-hand region of the SOM (Neurons 3, 5, and 8) represents a region in which the human graders were somewhat more lenient than the automated scoring (i.e., automated score of I and human score of A or automated score of U and human score of I). Yet, even with this region of the network associated with this interpretation, the separation between the M and N vectors, and their separate interpretations from the CART study, remain distinct.

Neurons 13 and 18 are associated with feature vector A from the CART analyses, with 85% of all cases with this feature vector associated with these two neurons. Similarly, Neuron 7 is associated with feature vector C from the CART analyses (with 58% of all cases of feature vector C), and Neuron 1 is associated with feature vector B from the CART analyses (with 100% of all cases of feature vector B). Together these create a diagonal region in the network representing instances in which the human graders were more strict than the automated scoring (with human scores of U and automated scores of A). Yet, again within the larger region, the former CART feature vectors remain in distinct pockets of the SOM.

Neuron 15 is associated with feature vector F from the CART analyses. All, 100%, of cases with this feature vector associated with this neuron represent cases in which the automated scoring was slightly more lenient than that of the human graders. Finally, the lower left-hand region of the SOM (Neurons 11, 16, 17, 21, and 22) may be vaguely described as a region in which human and automated scores are in complete agreement, although the number of cases associated with these neurons is small.

Discussion

The results suggest that there is a reasonably high degree of concordance between the SOM and the classification tree terminal nodes from CART analyses. The neurons (or neighborhoods of neurons) in the SOM tend to be associated with cases parallel to the cases classified by the individual feature vectors from CART. Furthermore, the regions of the SOM tend to be consistent with particular classes of human and automated scoring agreement. The lower right-hand region is a region of agreement, the upper right-hand region represents cases where the human grading was more lenient than automated scoring, and the diagonal from the center to the top left-hand corner denotes a region of cases in which the automated scoring was more lenient than human grading.

Given this consistency between the neurons and regions of the SOM (an unsupervised method) and the classification trees from CART analysis (requiring a dependent variable), the use of SOM may provide additional efficiencies in the review process by prescreening data for sampling human scores. For example, if the SOM for this sample of solutions had been produced prior to any human evaluation, and thus any CART analyses, the resultant neuron clusters of solutions could have been used to provide a stratified sampling of solutions to receive human grades If, for example, 15 solutions from each neuron were selected to receive human grades, then 149 (57%) of the solutions would be human graded rather than the entire set of 261 solutions. Such a savings of 43% of human grading effort is not inconsequential because each human grading effort requires a committee of three graders and some degree of time for them to evaluate the solution, discuss their observations, and come to a determination of final score. Once the human scores were obtained on this stratified sample of solutions they could then be subjected to CART analyses on the basis of the smaller stratified sample to (presumably) produce a classification tree and feature vectors similar to those previously obtained with the larger data set. These CART results could then be used to the advantages suggested by the first two applications. Although this is insufficient evidence for operational use, these results suggest that SOM are worthy of further investigation with regard to their ability to reduce the burden of human evaluation in review processes in SME evaluation of automated scoring.[6]

Conclusion

This exploration of CART and SOM as automated tools for aiding SME evaluation of automated scoring of complex constructed responses has suggested potential for these methods to facilitate the development, validation, and operational evaluation of automated scoring. Classification trees appear to have promise as a means of facilitating the identification of sources of score discrepancies, in favor of either human graders or automated scoring. As such, CART may be a useful tool not only for facilitating the SME evaluation of automated scoring, but also for facilitating the comparison of multiple scoring systems. Those systems could be competing methods of automated scoring, or multiple human rubrics, or hybrids of the two. The results using classification tree feature vectors suggest CART as a potential technique for specific types of automated case selection, including first-order human intervention in support of score validity, and for the exploration and documentation of issues that may contribute to the future evolution of automated scoring. The apparent capacity of SOM to classify complex data into proximity regions consistent with specific trends, specifically with regard to the degree and direction of agreement between automated and human scores, suggests the potential for neural network techniques to become a useful tool in investigations of complex constructed-response data.

The exploration of these techniques has provided a promising beginning to the development of potential automated tools capable of facilitating SME evaluation of automated scoring. The specific results of this investigation of classification trees are encouraging. The ability of classification tree vectors to be customized for the specific purpose of automated case selection processes, both by altering the criterion variable used to grow the trees and by adjusting the weights for classification error to reflect the cost of specific case classification and selection errors, suggests greater potential for this technique as such an automated tool. As an unsupervised procedure, SOM has promise as a method of reducing the need for expert human graders in order to train a model (such as a classification tree), to identify important issues in complex solutions, and to classify the solutions by specific trends. Together, these techniques offer a promising start for further investigations into automated tools intended to facilitate SME processes in automated scoring. Of particular interest is the potential for the tools to make the procedures in development, validation, and operational evaluation more efficient and cost-effective. More assessments are beginning to use fully automated administration and scoring of complex constructed-response tasks. Therefore, it will be necessary to direct more research toward such applied issues and tools, particularly with regard to efficiency and cost-effective implementation, in order to maximize the number of assessment domains that can explore the potential advantages of automated scoring for complex constructed-response tasks.

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four process architecture. *The Journal of Technology, Learning and Assessment, 1*(5). Retrieved January 5, 2003, from http://www.bc.edu/research/intasc/jtla/journal/pdf/v1n5_jtla.pdf

Balakrishan, P. V., Cooper, M. C., Jacob, V. S., & Lewis, P.A. (1994). A study of the classification capabilities of neural networks using unsupervised learning: A comparison with K-means clustering. *Psychometrika, 59,* 509-525.

Bauer, M., Williamson, D. M., Steinberg, L. S., Mislevy, R. J., & Behrens, J. T. (April, 2001). *How to create complex measurement models: A case study of principled assessment design*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

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., & Braun, H.. (March, 1999). *Architectural simulations: From research to implementation: Final report to the National Council of Architectural Registration Boards* (ETS RM-99-2). Princeton, NJ::ETS.

Bejar, I. I., & Whalen, S. J. (1997, March). *A system for interactive standard setting*. Paper presented at the annual conference of the National Council on Measurement in Education, Chicago, IL.

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.

Bennett, R. E., & Bejar I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9-17.

Bennett, R. E., & Sebrechts, M. M. (1996). The accuracy of expert-system diagnoses of mathematical problem solutions. *Applied Measurement in Education, 9,* 133-150.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M, & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computerized-adaptive tests. *Journal of Educational Measurement, 34,* 162-176.

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

Braun, H. I., Bennett, R. E., Frye, D, & Soloway, E. (1990). Scoring constructed responses using expert systems. *Journal of Educational Measurement, 27,* 93-108.

Burstein, J., Kukich, K., Wolff, S., & Lu, C. (1998). *Computer analysis of essay content for automated score prediction*. Presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Clauser, B. E., Margolis, M. J., Clyman, S. G., & Ross, L. P. (1997). Development of automated scoring algorithms for complex performance assessments: A comparison of two approaches. *Journal of Educational Measurement, 34,* 141-161.

Clauser, B. E., Subhiyah, R. G., Nungenster, R. J., Ripkey, D. R., Clyman, D. R., & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgement process of experts. *Journal of Educational Measurement, 32,* 397-415.

Clyman, S. G., Melnick, D. E., & Clauser, B. E. (1995). Computer-based simulations. In E. L. Mancall & E. G. Bashook (Eds.), *Assessing clinical reasoning: The oral examination and alternative methods* (pp. 139-149). Evanston, IL: American Board of Medical Specialties.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Fisher, R. A. (1936). The use of multiple measurements in axonomic problems. *Annals of Eugenics 7,* 179-188.

Greene, E. B. (1941). *Measurements of human behavior.* New York: The Odyssey Press.

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.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *Journal of Physics, 160,* 106-154.

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

Kohonen, T. (1989). *Self-organization and associative memory* (3rd ed.). New York: Springer-Verlag.

Meehl, P.E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (3rd ed.). New York: American Council on Education—Macmillan Publishing Company.

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

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometika, 59*(4), 439-483.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

Murtagh, F., & Hernandez-Pajares, M. (1995). The Kohonen self-organizing map method: An assessment. *Journal of Classification, 12,* 165-190.

Oltman, P. K., Bejar, I. I., & Kim, S. H. (1993). An approach to automated scoring of architectural designs. In U. Flemming & S. van Wyk (Eds.), *CAAD Futures* (vol. 93, pp. 215-224). Pittsburgh, PA: Elsevier Science Publishers B. V.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. England: Cambridge University Press.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Agreement between expert system and human raters' scores on complex constructed-response quantitative items. *Journal of Applied Psychology, 76,* 856-862.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*(4), 333-352.

SPSS, Inc., & Recognition Systems Group. (1998). Neural Connection 2.1 [Computer software]. Chicago, IL: Author.

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

Stevens, R., Ikeda, J., Casillas, A., Palacio-Cayetano, & J., Clyman, S. (1999). Artificial neural-network based performance assessments. *Computers in Human Behavior, 15*(3-4), 295-313.

Waller, N. G., Kaiser, H. A., Illian, J. B., & Manry, M. (1998). A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms. *Psychometrika, 63,* 5-22.

Williamson D. M., Bejar I. I., & Hone, A. S. (1999). "Mental model" comparison of automated and human scoring. *Journal of Educational Measurement, 36*(2), 158-184.

Winston, P. H. (1992). *Artificial intelligence* (3rd ed.). Boston: Addison-Wesley.

Zhang, Y., Powers, D. E., Wright, W & Morgan, R. (2003,April). *Applying the Online Scoring Network (OSN) to Advanced Placement Program (AP) tests* (ETS RR-03-12). Princeton, NJ: Educational Testing Service.

Notes

[1] This author previously published under the name Anne S. Hone.

[2] The authors used CART (classification and regression trees) software published by Salford Systems for the classification and regression tree analyses in this paper.

[3] The Gini index is named after Italian statistician and demographer Corrado Gini (1884-1965) and is a numerical representation of a geometric depiction known as a Lorenz Curve, an indicator of economic inequality. The Lorenz curve is a plot of cumulative percentage of population on the X-axis and cumulative percentage of income on the Y-axis, for which a 45-degree diagonal from the origin represents perfect equity of income. The Gini index is the ratio of the difference between the observed distribution of income and perfectly equitable distribution (area between the curves) to the perfectly equitable distribution of income (area under the 45-degree curve for perfect equity, which is always equal to .5). As a numeric representation of the Lorenz Curve, it varies from 0 to 1 with 0 indicating perfect equality of distribution of wealth across a population and 1 indicating the entirety of income concentrated in a single individual. (Note that Gini index computation in CART is 1 minus the ratio rather than the direct ratio, thereby resulting in the reversal of the meaning of values of 0 and 1 from the Gini index computation in economics.) Other methods for the partitioning in CART include entropy, which derives its name from the common definition of entropy as an indicator of the orderliness of a system, whereby a more highly ordered system (low entropy) can be described more concisely than a more disordered system. The Gini index targets the extraction of single classes of the dependent variable in partitioning, whereas the Entropy index operates by segmenting the dependent variable classes into two groups such that each of the two groups contains half the number of cases in the data. The procedure then searches for the partition that optimally separates these two groups of data. A similar method for establishing a quality-of-split criterion is called twoing. As one might imagine, there are a number of variations on these approaches to partitioning and the interested reader can refer to Ripley (1996, p. 217) for more information.

[4] For example, Williamson, Bejar, & Hone (1999) found that on occasion a committee of graders reviewing scores could not justify their own committee's holistic score even with reference to their original documented criteria. This suggests that committee processes

directed at more fine-grained features and specific aggregation processes would be even more likely to suffer difficulties.

[5] Selecting an optimal number of artificial neurons to specify in neural network applications remains an inexact process. For supervised neural networks, in which there are a targeted number of clusters expected a priori, the number to specify is straightforward since it should be no less than the number of desired clusters. When there is no a priori expectation, such as for unsupervised networks, then the training can occur with an excessive number of clusters. While excessive numbers of clusters provide a certain margin for ensuring separability, this comes at the expense of an increased risk of difficulty of interpretability of resultant clusters as more minor differences in variable vectors result in classification in different neurons. One approach that attempts to balance these concerns generates multiple neural networks with different numbers of artificial neurons (as well as layers and dimensions) *and to prefer* networks that produce a balance between the number of neurons with "peaks" (high *n*'s) and "valleys" (low *n*'s). The decision regarding the number of artificial neurons to specify is related to the ratio of number of observations to number of artificial neurons, the variance among observations, and the researcher's prior expectations.. As such, the decision regarding the number of artificial neurons to select is similar to the determination of the number of factors to extract in exploratory factor analysis, and in the future such factor analytic techniques as eigenvalues and scree plots may inspire similar techniques for neural network specification. In this research the neural networks were trained using unsupervised techniques, but with prior information from CART analysis that suggested there should be no less than 14 neurons. To ensure adequate separation among the patterns of neurons and to seek a mixture of "peaks" and "valleys" of observation, clusters across the neurons a 5 x 5 neuron grid was chosen.

[6] A reviewer suggested that this portion of the study could be beneficially extended through a comparison of the results from this sampling procedure to both the full CART sampling results and one or more smaller random samples. We agree that this is a desirable extension to the study and are regrettably unable to conduct further analyses of these data due to the expiration of our usage agreement with the owners of the data. It is hoped that future research along similar lines will be able to incorporate this recommendation.