Journal of Modern Applied Statistical Methods

Volume 7 | Issue 1 Article 19

5-1-2008

Using Connectionist Models to Evaluate Examinees' Response Patterns to Achievement Tests

Mark J. Gierl
University of Alberta, mark.gierl@ualberta.ca

Ying Cui
University of Alberta, yc@ualberta.ca

Steve Hunka
University of Alberta, Albertasteve.hunka@ualberta.ca

Recommended Citation
Gierl, Mark J.; Cui, Ying; and Hunka, Steve (2008) "Using Connectionist Models to Evaluate Examinees' Response Patterns to Achievement Tests," *Journal of Modern Applied Statistical Methods*: Vol. 7 : Iss. 1 , Article 19.
DOI: 10.22237/jmasm/1209615480

Journal of Modern Applied Statistical Methods
May, 2008, Vol. 7, No. 1, 234-245

1538 – 9472/08/$95.00

Using Connectionist Models to Evaluate Examinees' Response Patterns to Achievement Tests

Mark J. Gierl Ying Cui Steve Hunka
University of Alberta

The attribute hierarchy method (AHM) applied to assessment engineering is described. It is a psychometric method for classifying examinees' test item responses into a set of attribute mastery patterns associated with different components in a cognitive model of task performance. Attribute probabilities, computed using a neural network, can be estimated for each examinee thereby providing specific information about the examinee's attribute-mastery level. The pattern recognition approach described in this study relies on an explicit cognitive model to produce the expected response patterns. The expected response patterns serve as the input to the neural network. The model also yields the cognitive test specifications. These specifications identify the examinees' attribute patterns which are used as output for the neural network. The purpose of the statistical pattern recognition analysis is to estimate the probability that an examinee possess specific attribute combinations based on their observed item response patterns. Two examples using student response data from a sample of algebra items on the SAT illustrate our pattern recognition approach.

Keywords: Attribute hierarchy method, multilayer perceptron, neural network, educational measurement.

Introduction

Educational measurement is undergoing profound changes, as developments in cognitive science, mathematical statistics, computer technology, educational psychology, and computing science are permeating the testing field. In particular, the influence of cognitive psychology on educational measurement, which began almost 20 years ago (Snow & Lohman, 1989), has become a source of great activity contributing to many of the ideas and innovations in cognitive diagnostic assessment (Leighton & Gierl, 2007a). One consequence of these interdisciplinary influences is the emergence of a new area of research called assessment engineering (AE) (Luecht, 2006). AE is an innovative approach to measurement where engineering-like principles are used to direct the design as well as the analysis, scoring, and reporting of assessment results. With this approach, an assessment begins with specific, empirically-derived cognitive models of task performance. Next, assessment task templates are created using established frameworks derived from the cognitive model to produce replicable test items. Finally, psychometric models are applied to the examinee response data collected using the templates to produce scores that are both replicable and interpretable.

Mark J. Gierl is a Professor of Educational Psychology and Canada Research Chair in Educational Measurement, Centre for Research in Applied Measurement and Evaluation. His primary research interests include cognitive diagnostic assessment, and assessment engineering. E-mail: mark.gierl@ualberta.ca. Ying Cui is an Assistant Professor of Educational Psychology. Her primary research interests include investigating cognitive factors underlying student performance and using cognitive models to interpret student test performance. E-mail her at yc@ualberta.ca. Stephen Hunka is Professor Emeritus. His areas of specialization include statistical models and computation. E-mail him at: Albertasteve.hunka@ualberta.ca.

AE differs from more traditional approaches to test design and analysis in four fundamental ways. First, cognitive models guide task design and item development, rather than content-based test specifications. While the database "tags" associated with content specifications can be included in the task templates, the assessment principles used to develop items are much more specific allowing items to be created quickly and efficiently during the development cycle. Second, explicit data models and assessment task templates are created to control and manipulate both the content and cognitive attributes of the items. Item writers are required to use the templates during development thereby producing items that adhere to strict quality controls and that meet high psychometric standards. Third, automated test assembly procedures are employed to build assessments that function to exacting specifications, as outlined in the task templates. Hence, multiple test forms can be created from a bank of items very efficiently according to both content and statistical specifications. Fourth, pursuant to scoring and score-reporting, psychometric models are employed in a confirmatory—versus exploratory—manner to assess the model-data fit relative to the intended underlying structure of the constructs or traits the test is design to measure. The outcomes from these model-data fit analyses also provide developers with guidelines for specific modifications to the cognitive models and task templates, as needed, to facilitate the acquisition of data that supports the intended assessment inferences.

Overview of Attribute Hierarchy Model

Recently, Leighton, Gierl, and Hunka (2004; see also Gierl, Leighton, & Hunka, 2007) proposed the attribute hierarchy method (AHM). The AHM is a psychometric method used to classify examinees' test item responses into a set of structured attribute patterns associated with different components from a cognitive model of task performance (Leighton & Gierl, 2007b). Attributes include different procedures, skills, and/or processes that an examinee must possess to solve a test item. These attributes are structured using a hierarchy so the ordering of the cognitive skills is specified. As a result, the attribute hierarchy serves as an explicit cognitive model. This model, in turn, provides the structure for both developing test items and linking examinees' test performance to specific cognitive inferences about skill acquisition. The AHM was developed to address two specific problems associated with *feature creation* and *statistical pattern recognition* (Gierl, 2007). Our solutions to these problems are described in the next two sections.

Feature Creation with the AHM

To make specific inferences about problem solving, cognitive models are required to operationalize the construct of interest. A cognitive model in educational measurement refers to a simplified description of human problem solving on standardized tasks at some convenient grain size or level of analysis in order to facilitate explanation and prediction of students' performance. These models provide an interpretative framework that can guide item development so test performance can be linked to specific cognitive inferences about examinees' knowledge, processes, and strategies. These models also provide the means for connecting cognitive principles with measurement practices.

A cognitive model of task performance is specified at a small grain size because it magnifies the cognitive processes underlying test performance. Often, a cognitive model of task performance will also reflect a hierarchy of cognitive processes within a domain because cognitive processes share dependencies and function within a much larger network of inter-related processes, competencies, and skills. Assessments based on cognitive models of task performance should be developed so test items directly measure specific cognitive processes of increasing complexity in the examinees' understanding of a domain. The items can also be designed with this hierarchical order in mind, so that test performance is directly linked to information about students' cognitive strengths and weaknesses. Strong inferences about examinees' cognitive skills can be made because the small grain size in these models help illuminate the knowledge and skills required to perform competently on testing tasks. Specific diagnostic inferences can also be generated

when items are developed to measure different components and processes in the model.

To specify the relationships among the attributes in the hierarchy using the AHM, the adjacency and reachability matrices are defined. The direct relationship among attributes is specified by a binary adjacency matrix (A) of order (k, k), where k is the number of attributes, such that the ij^{th} element represents the absence (i.e., 0) or presence (i.e., 1) of a direct connection between two attributes. The adjacency matrix is of upper triangular form. The direct and indirect relationships among attributes are specified by the binary reachability matrix (R) of order (k, k), where k is the number of attributes. To obtain the R matrix from the A matrix, Boolean addition and multiplication operations are performed on the adjacency matrix, meaning $R = (A + I)^n$, where n is the integer required to reach invariance, $n = 1, 2, \ldots m$, and I is the identity matrix.

Next, the potential pool of items is generated. This pool is considered to be those items representing all combinations of attributes when the attributes are independent of one other. The size of the potential pool is $2^k - 1$, where k is the number of attributes. The attributes in the potential pool of items are described by the incidence matrix (Q) of order (k, p), where k is the number of attributes and p is the number of potential items. This matrix can be reduced to form the reduced Q matrix (Q_r) by imposing the constraints of the attribute hierarchy as defined in the R matrix. The Q_r matrix represents the items from the potential pool that fit the constraints defined in the attribute hierarchy. The Q_r matrix is formed using Boolean inclusion by determining which columns of the R matrix are logically included in each column of the Q matrix. The Q_r matrix is of order (k, i) where k is the number of attributes and i is the reduced number of items resulting from the constraints in the hierarchy.

Given a hierarchy of attributes, the expected response patterns for a group of examinees can then be generated. The expected response matrix (E) is created, again using Boolean inclusion, where the algorithm compares each row of the attribute pattern matrix (which is the transpose of the Q_r matrix) to the columns of the Q_r matrix. The expected response matrix, of order (j, i), is calculated, where j is the number of examinees and i is the reduced number of items resulting from the constraints imposed by the hierarchy.

Assessment engineering principles are used explicitly with the AHM to design test items and analyze examinees' observed response patterns. To design test items, the Q_r matrix is used. Recall, the Q_r matrix is produced by determining which columns of the R matrix are logically included in columns of the Q matrix, using Boolean inclusion. The Q_r matrix can be interpreted as the cognitive test specification because it contains the attribute-by-item specification for each component of the cognitive model of task performance outlined in the A matrix. Hence, the results from the Q_r matrix can be used to develop items that measure each specific attribute combination defined in the hierarchy. Then, in the pattern recognition stage, as described in the next section, examinees' observed response patterns can be analyzed according to the cognitive characteristics probed by each item.

Pattern Recognition with the AHM

An examinee's observed response pattern is judged relative to expected response pattern with the AHM under the assumption that the cognitive model is true. Hence, the purpose of the statistical pattern recognition analysis is to estimate the probability that an examinee possess specific attribute combinations based on their response patterns. These probabilities provide examinees with specific information about their attribute-level mastery as part of the test reporting process. To estimate the probability that examinees possess specific attributes, given their observed item response pattern, an artificial neural network approach is used.

The input to train the neural network is the expected response vector derived from the cognitive model. The expected response vectors serve as the exemplars. For each expected

response vector, there is a specific combination of examinee attributes described in the transpose of the Q_r matrix. Recall, Q_r matrix is of order (k, i) where k is the number of attributes and i is the reduced number of items resulting from the constraints specified by the hierarchy. The transpose of this matrix is of order (j, k) where j is the number of examinees and k is the number of attributes. In other words, the transpose of the reduced incidence matrix has a distinct row and column interpretation—the rows serve as the examinees and the columns serve as the items. The examinee attribute patterns, like the expected response vectors, are derived from the cognitive model and, thus, specify the attribute pattern that should be associated with each expected response pattern. The relationship between the expected response vectors with their associated attribute vectors is established by presenting each pattern to the network repeatedly until it learns each association. The final result is a set of weight matrices that can be used to transform any observed response vector to its associate attribute vector. The transformed result can be interpreted as the attribute probability, scaled from 0 to 1, where a higher value indicates that the examinee has a higher probability of possessing a specific attribute (McClelland, 1998).

A multilayer perceptron is the parallel-processing architecture used in the neural network. This network transforms the stimulus received by the input unit to a signal for the output unit through the hidden units. The contribution of each input unit i to hidden unit j is determined by weight, w_{ji}. Similarly, the contribution of each hidden unit j to output unit k is determined by weight, v_{kj}. The input layer contains the exemplars (i.e., expected response patterns) the network is designed to learn. Learning is deemed to occur when the output layer, containing the desired response output (i.e., the attribute patterns), is correctly associated with the exemplars, as indicated by the value of the root mean square error. That is, the connection weights in the hidden layer transform the input stimuli into a weighted sum defined as

$$S_j = \sum_{i=1}^{p} w_{ji} x_i \qquad (1)$$

where S_j is the weighted sum for node j in the hidden layer, w_{ji} is the weight used by node j for input x_i, and x_i is the input from node i of the input layer with i ranging from 1 to p for the input node and j ranging from 1 to q for the hidden layer node. S_j is then transformed by the logistic function,

$$S_j^* = \frac{1}{1 + e^{-S_j}} \qquad (2)$$

Similarly, the hidden layer produces a weighted linear combination of their inputs which are transformed to non-linear weighted sums that are passed to every output layer unit to produce the final attribute-level responses. The output, S_j^*, from every hidden layer unit is passed to every output layer unit where a linearly weighted sum, T_k, is formed using the weights v_{kj}, and the result transformed for output T_k^* using a nonlinear function. In other words,

$$T_k = \sum_{j=1}^{q} v_{kj} S_j^* \qquad (3)$$

where T_k is the weighted sum for each of k output nodes using weights v_{kj}, with j ranging from 1 to q for the hidden layer nodes. T_k, like S_j, is transformed by the logistic function to T_k^*. Because the correct activation function is scaled using the logistic transformation, the output values range from 0 to 1. The result can be interpreted as the probability the correct or target value for each output will have a value of 1.

The attribute-based targets in the output units are compared to the pattern associated with the exemplars, which are the expected response patterns. However, the solution produced initially is likely to be discrepant resulting in a relatively large root mean square error. This discrepancy can be used to modify the connection weights leading to a more accurate solution and a smaller error term. With the AHM, the weights are approximated so the error term is minimized using the well-known learning algorithm called the generalized delta rule that is incorporated in the back propagation of error training procedure (Rumelhart, Hinton, & Williams, 1986a, 1986b). The final result is a set of weight matrices, one for cells in the hidden layer and one for the cells in the output layer, that can be used to transform any examinee response vector to its associate attribute vector. The functional relationship for mapping the examinees' observed response pattern onto the expected response patterns so their attribute probabilities can be computed is given as follows. Let

$$F(z) = \frac{1}{1+e^{-z}} \qquad (4)$$

and

$$a_k = \sum_{j=1}^{q} v_{kj} F(\sum_{i=1}^{p} w_{ji} x_i) \qquad (5)$$

then the output for unit k, M_k^*, is given as

$$M_k^* = F(a_k) \qquad (6)$$

where q is the total number of hidden units, v_{kj} is the weight of hidden unit j for output unit k, p is the total number of input units, w_{ji} is the weight of input unit i for hidden unit j, and x_i is the input received from input unit i. Using this transformation, attribute probabilities can be computed for each observed response pattern thereby providing examinees with specific information about their attribute-mastery level.

Two Examples Using SAT Algebra Items

To illustrate how a multilayer perceptron can be used to estimate the attribute probabilities in an actual testing situation, two examples are provided. Each example is based on the observed response data from a random sample of 5000 students who wrote the algebra items on the March 2005 administration of the SAT. The SAT is a college admissions test developed, analyzed, and scored by the College Board. The Mathematics section contains items in the content areas of Number and Operations; Algebra I, II, and Functions; Geometry; and Statistics, Probability, and Data Analysis. For our analysis, only a subset of items in Algebra I and II were evaluated. Sample algebra items from the SAT Mathematics section are available from the College Board website at www.collegeboard.com.

Note that cognitive models of task performance guide diagnostic inferences because they are specified at a small grain size and they magnify the cognitive processes that underlie performance. Ideally, a theory of task performance would direct the development of a cognitive model of task performance. But, in the absence of such a theory, a cognitive model must still be specified to create the attribute hierarchy. Another starting point is to develop a cognitive model from a task analysis of the items in the domain when a theory or model of task performance is unavailable. In conducting the task analysis of the SAT algebra items we, first, solved each test item and attempted to identify the mathematical concepts, operations, procedures, and strategies used to solve each item (see Gierl, Wang, & Zhou, 2006; Gierl, Leighton, Wang, Zhou, Gokiert, & Tan, 2006). These cognitive attributes were categorized so they could be ordered in a logical, hierarchical sequence to summarize problem-solving performance. The cognitive model used to characterize examinee performance on the items is presented in Figure 1. Each attribute is denoted with an A (e.g., A1, A2, etc.). Each attribute was measured by one test item. The cognitive model in Figure 1 was used to created the Q_r matrix.

This hierarchy presents a cognitive model of task performance for skills in the areas of ratio, factoring, function, and substitution. The hierarchy contains two independent

branches which share a common prerequisite, attribute A1. Aside from attribute A1, the first branch includes two additional attributes, A2 and A3, and the second branch includes a self-contained sub-hierarchy which includes attributes A4 through A9. Three independent branches compose the sub-hierarchy: attributes A4, A5, A6; attributes A4, A7, A8; and attributes A4, A9.

As a prerequisite attribute, attribute A1 includes the most basic arithmetic operation skills, such as addition, subtraction, multiplication, and division of numbers. Attributes A2 and A3 both deal with factors. In attribute A2, the examinee simply needs to have knowledge about the property of factors. In attribute A3, the examinee not only requires knowledge of factoring (i.e., attribute A2), but also the application of factoring. Therefore, attribute A3 is considered a more advanced attribute than A2. The self-contained sub-hierarchy contains six attributes. Among these attributes, attribute A4 is the prerequisite for all other attributes in the sub-hierarchy. Attribute A4 has attribute A1 as a prerequisite because A4 not only represents basic skills in arithmetic operations (i.e., attribute A1), but it also involves the substitution of values into algebraic expressions which is more abstract and, therefore, more difficult than attribute A1. The first branch in the sub-hierarchy deals, mainly, with functional graph reading. For attribute A5, the examinee must be able to map the graph of a familiar function (e.g., a parabola) with its corresponding function. Attribute A6 deals with the abstract properties of functions, such as recognizing the graphical representation of the relationship between independent and dependent variables. The second branch in the sub-hierarchy considers the skills associated with advanced substitution. Attribute A7 requires the examinee to substitute numbers into algebraic expressions. The complexity of attribute A7 relative to attribute A4 lies in the concurrent management of multiple pairs of numbers and \ multiple equations. Attribute A8 also represents the skills of substitution. However, what makes attribute A8 more difficult than attribute A7 is that algebraic expressions, rather than numbers, need to be substituted into another algebraic expression. The last branch in the sub-hierarchy contains only one additional attribute, A9, related to skills associated with rule substitution. It is the rule, rather than the numeric value or the algebraic expression, that needs to be substituted in the item to reach a solution.

SAT Example 1: Training without Extra Output

In the first example, training was conducted without extra output. That is, the input to train the network is the expected response vectors produced from the AHM feature creation analyses and the output is the specific combination of examinee attributes derived from the transpose of the Q_r matrix for each expected response vector. The relationship between the expected response vectors with their associated attribute vectors was established by presenting each pattern to the network repeatedly.

Using nine hidden units, the network converged using a model with 9 input, 9 hidden, and 9 output units. The value for the root mean square was 0.00082 after 500 epochs. The probabilities associated with each attribute across the nine expected response patterns was used to define the functional relationship for mapping the examinees' observed response pattern onto the expected response pattern so their attribute mastery levels could be determined.

Seven examples are presented in Table 1. The first three include attribute probabilities for observed response patterns that are consistent with the cognitive model in Figure 1. Take, for instance, an examinee who possesses the first three attributes, A1 to A3, thereby producing the response pattern 111000000 (i.e., example 1). This observed response pattern is consistent with one of the 58 expected response patterns. The attribute probabilities for this response pattern are 0.91, 1.00, 1.00, 0.08, 0.02, 0.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively. Examples 2 and 3 illustrate the attribute probabilities associated with observed response patterns that are also consistent with the hierarchy in Figure 1.

Alternatively, examples 4 to 7 illustrate attribute probabilities for observe response patterns that are *inconsistent* with the attribute hierarchy. In other words, these response patterns are not one of the 58 patterns in

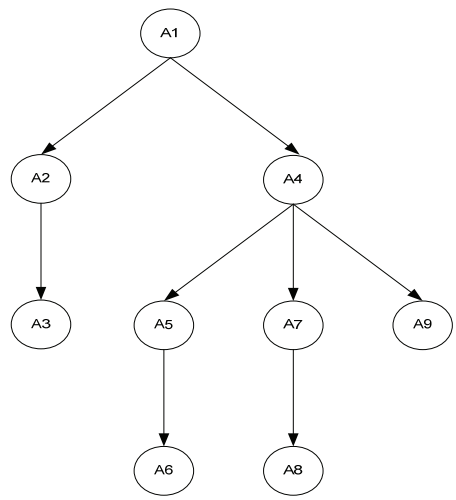

Figure 1.

expected response matrix. These inconsistency can be addressed using the network because its purpose is to define the functional relationship for mapping the examinees' observed response pattern onto the expected response pattern using $M_k^* = F(a_k)$.

The first inconsistent pattern, example 4, includes examinees who correctly solve the items associated with attributes A1 and A3, but then incorrectly solve the item associated with attribute A2. According to the cognitive model in Figure 1, this response patterns is not expected because A3 requires A1 and A2. Yet, we have an observed response pattern where A3 is solved correctly while A2 is not. This inconsistency or slip means that the examinee's item response is unexpected because the attributes probed by the item are assumed to be mastered by the examinee, given the cognitive model of task performance. The attribute probabilities for this observed response pattern are 0.92, 0.99, 1.00, 0.16, 0.04, 0.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively, indicating that it is very unlikely that an examinee who possesses attribute A3 would not also possess attribute A2, if the cognitive model in Figure 1 is true. The attribute probability level is also unusually high, in this example, because we only have one item measuring each attribute and this branch (A1 to A3) has only three attributes, in total. However, when a larger number of items are used to measure the attributes across a larger number of branches, the attribute probabilities decrease, as illustrated in examples 5 to 7.

For these three examples, attribute A4, which is the prerequisite attribute in each case, is missing. In example 5, the examinee correctly solves the items measuring A1, A5 and A6, but incorrectly solves the item measuring A4. The attribute probabilities for this observed response pattern are 0.69, 0.01, 0.00, 0.31, 1.00, 1.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1, A5, and A6, but likely not A4. A value of 0.50 is used in our example to interpret the probabilities, meaning that if the probability is greater that 50%, the examinee is believed to possess the attribute. In example 5, however, it is difficult to evaluate A4 because the examinee only solves two items correctly that required A4. In example 6, on the other hand, the examinee

correctly solves the items measuring A1 and A5 to A8. In this case, four items that require A4 are correctly solved. The attribute probabilities for this observed response pattern are 0.43, 0.00, 0.00, 0.95, 1.00, 1.00, 0.99, 0.78, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A4 to A8. The examinees may also possess A1, but the probability is low (the result for A1 in this example is unusual because the examinee must possess A1 to solve the remaining items). Notice that when all four items requiring the prerequisite attribute are correctly solved (i.e, A5 to A8), but the prerequisite attribute is incorrectly solved (i.e., A4), the probability is high that the examinee, in fact, possesses the prerequisite A4. Or, stated differently, it is unlikely that the examinee could solve the items associated with A5 to A8 without possessing A4, if the cognitive model in Figure 1 is accurate. When the final attribute is included, A9, in example 7, the attribute probabilities are 0.87, 0.01, 0.00, 0.96, 1.00, 0.99, 0.97, 0.62, 0.98 indicating that the examinee possesses A1, A4 to A9. The results across the seven examples are consistent with our expectations based on the cognitive model, for the most part. The only unusual results occurred in example 5 where the probability for A4 was unexpectedly low and example 6 where the probability for A1 was also low.

SAT Example 2: Training with Extra Output

In the second example, training was conducted with extra output (Gällmo & Carlström, 1995). That is, the input to train the network is the expected response vectors produced from the AHM feature creation analyses, as in example 1, but the target output is the specific combination of examinee attributes derived from the transpose of the Q_r matrix as well as the ability estimate for each expected response vector.

With a cognitive diagnostic model like the AHM, expected item and ability parameters can be estimated. The expected item parameters can be produced using an item response theory (IRT) model. For example 2, the two-parameter (2PL) logistic IRT model is used. This model is given by

Table 1. Attribute Probabilities for Seven Observed Examinee Response Patterns using the SAT Algebra Hierarchy in Figure 1 with No Extra Output

Pattern	*Attribute Probability*								
	1	2	3	4	5	6	7	8	9
Consistent									
1. A1 to A3	0.91	1.00	1.00	0.08	0.02	0.00	0.00	0.00	0.00
2. A1, A4 to A6	0.94	0.01	0.00	0.96	1.00	0.97	0.01	0.00	0.00
3. A1, A4 to A8	0.96	0.00	0.00	1.00	1.00	0.97	1.00	0.98	0.02
Inconsistent									
4. A1, A3 (Missing A2)	0.92	0.99	1.00	0.16	0.04	0.00	0.00	0.00	0.00
5. A1, A5, A6 (Missing A4)	0.69	0.01	0.00	0.31	1.00	1.00	0.00	0.00	0.00
6. A1, A5 to A8 (Missing A4)	0.43	0.00	0.00	0.95	1.00	1.00	0.99	0.78	0.00
7. A1, A5 to A9 (Missing A4)	0.87	0.01	0.00	0.96	1.00	0.99	0.97	0.62	0.98

$$P(u=1|\Theta)=\frac{1}{1+e^{-1.7a_i(\Theta-b_i)}}, \quad (7)$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, and Θ is the ability parameter. Using the 2PL logistic IRT function, the a and b parameters can be determined for each item using the expected item response patterns given by the columns of the expected response matrix. The expected ability parameters are then produced by locating the maximum of the likelihood function defined by

$$L\left(\underset{\sim}{\mathbf{u}}\middle|\theta_j\right)=\prod_{i=1}^{n}P_{ij}^{u_{ij}}Q_{ij}^{1-u_{ij}} \quad (8)$$

where $P_{ij}^{u_{ij}}$ is the probability, based on the 2PL logistic function, for a correct response to item i and $Q_{ij}^{1-u_{ij}}$ is $1\text{-}P_{ij}^{u_{ij}}$. The likelihood function is typically placed on a unidimensional scale with a mean of 0 and a standard deviation of 1.

To illustrate the extra output training method, a random sample of 5000 simulated examinees was generated for the 58 unique patterns in the expected response matrix with the constraint that the distribution of total score be normal in shape. Then, the simulated response data were fit to the 2PL logistic IRT model to estimate the item and ability parameters. Estimation was conducted with the computer software BILOG-MG (du Toit, 2003). Default settings in BILOG-MG were used, except the calibration option that was set to "float" indicating that the means of the priors on the item parameters were calculated using marginal maximum likelihood estimation, and both the means and the item parameters were updated after each iteration. The ability estimates provide a measure of the expected examinees' score on a (0, 1) unidimensional scale which typically ranges from -4 to +4. Thus, a higher score indicates a higher ability level.

These ability scores have an important role in the example 2 analysis: They serve as extra output or "hints" that provide prior knowledge to the neural network about a feature in each expected response pattern that may increase the accuracy of learning. The ability level extra output is only included to help the network learn, and once training is complete, the extra output is removed. The benefit of adding an extra output, like ability level, is that it can act as a side constraint thereby increasing the representational power of the network and potentially increase the accuracy and generalizability of the network solution.

Using nine hidden units, the network converged using a model with 9 input, 9 hidden, and 9 output units. The value for the root mean square was 0.00028 after 500 epochs. The probabilities associated with each attribute across the nine expected response patterns was used to define the functional relationship for mapping the examinees' observed response patterns from the SAT dataset onto the expected response patterns derived from the cognitive model so their attribute mastery levels can be determined. The attribute probabilities for the same seven response patterns in Table 1 are presented in Table 2.

The results between Tables 1 and 2 are similar, except for two important exceptions. Recall, for example 5 in Table 1, the examinee correctly solved the items measuring A1 and A5, but incorrectly solved the item measuring A4. The attribute probabilities for this observed response pattern was 0.69, 0.01, 0.00, 0.31, 1.00, 1.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1, A5, and A6, but not A4. The same example, but with extra output, shown in Table 2, yields a more interpretable result. The attribute probabilities are 0.95, 0.01, 0.00, 0.60, 1.00, 0.99, 0.01, 0.01, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1 and A5, and likely possesses A4, which is expected given that the examinee correctly solved the item measuring A5. In Table 1, example 6, the attribute probability for A1 was low, given that the examinee required this attribute to solve the items. But, in Table 2, example 6, the attribute probabilities are more consistent with the cognitive model at 0.97, 0.01, 0.00, 0.95, 1.00, 0.99, 1.00, 0.99, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1, A4 to A8. When A9 is added in example 7, the attribute probabilities are 0.98, 0.04, 0.00, 1.00, 1.00,

0.98, 1.00, 1.00, 0.98 indicating that the examinee possesses A1, A4 to A9. The probability for A8 in example 1 was reasonably high at 0.62. But, in example 2, the probability for A8 is much higher at 1.00 and, thus, easier to interpret. To summarize, the results across the seven examples in Table 2 are consistent with our expectations based on the cognitive model in Figure 1, particularly when compared to the results in Table 1. These outcomes also reveal that extra output learning improved the interpretability of the network solutions.

Discussion

Assessment engineering with the AHM relies on two stages. In the feature creation stage, principled test design procedures are used to develop items that systematically measure each component in the cognitive model. In the pattern recognition stage, the functional relationship between the examinees' expected response patterns and item attributes is established so the attribute probabilities for the examinees' observed response patterns can be estimated. The purpose of the present study was to describe the analytic procedures in the pattern recognition stage.

Using response data from a sample of examinees who wrote algebra items on the SAT, the results from two different examples were presented. In the first example, the attribute probabilities were computed by training the network without extra output. The value for the root mean square was small at 0.00082. The results across the seven examples were consistent with our expectations from the cognitive model, for the most part, as only two anomalous results were noted. In the second example, the attribute probabilities were computed by training the network with extra output associated with the ability estimates for each expected response pattern. The ability estimates served as an excellent source of extra learning output because they were derived from an IRT model fit to the expected response patterns to produce a single score for each unique pattern. The network yielded a smaller root mean square (0.00028) compared to the network without extra output, and the results across all seven examples were consistent with the cognitive model indicating that extra output training increased the interpretability of the network solution.

One limitation of the current study stems from the use of a post-hoc or retrofitting approach when identifying and applying the cognitive model of task performance to the algebra items on the SAT. In the current study, we generated a cognitive model of task performance by conducting a content review of the SAT algebra I and II items to identify the mathematical concepts, operations, procedures, and strategies used by students to solve items on the SAT. However, no new items were developed from the cognitive models of task performance used to produce the attribute hierarchies in Figure 3. This decision was made, in part, because the purpose of the study was to describe and illustrate the analytic procedures in the pattern recognition stage. However, in future applications of the AHM, researchers and practitioners implementing the AHM for AE should begin by specifying the cognitive model and use the attribute hierarchy to develop test items. These model-based test items can then be analyzed using the neural network procedures.

In closing, the role that pattern recognition procedures could one-day play in educational measurement is significant. In May 2006, Eduventures, a market research firm that specializes in educational products and applications, claimed that new applications of formative testing, like cognitive diagnostic assessment, may soon emerge to redefine the educational measurement practices in American classrooms. But they also noted that this emergence will only occur when several key objectives are met, including "the building of truly advanced analytic capabilities, relying on a neural network architecture to act as the engine to convert assessment inputs into prescriptive action" (Wiley, 2006). Our study provides one example of the "advanced analytic capabilities" that are possible when psychometric methods like the AHM incorporate pattern recognition procedures to classify examinees' response patterns on educational tests.

Table 2. Attribute Probabilities for Seven Observed Examinee Response Patterns using the SAT Algebra Hierarchy in Figure 1 with Ability as Extra Output

Pattern	*Attribute Probability*									*Ability*
	1	2	3	4	5	6	7	8	9	
Consistent										
8. A1 to A3	0.96	1.00	0.98	0.04	0.00	0.00	0.01	0.01	0.00	-2.408
9. A1, A4 to A6	0.99	0.01	0.00	0.98	1.00	0.98	0.01	0.00	0.01	-0.001
10. A1, A4 to A8	0.98	0.01	0.00	0.99	1.00	0.99	1.00	0.99	0.02	1.205
Inconsistent										
11. A1, A3 (Missing A2)	0.95	0.99	0.95	0.06	0.00	0.00	0.01	0.00	0.00	-2.408
12. A1, A5, A6 (Missing A4)	0.95	0.01	0.00	0.60	1.00	0.99	0.01	0.01	0.00	-0.369
13. A1, A5 to A8 (Missing A4)	0.97	0.01	0.00	0.95	1.00	0.99	1.00	0.99	0.00	1.072
14. A1, A5 to A9 (Missing A4)	0.98	0.04	0.00	1.00	1.00	0.98	1.00	1.00	0.98	1.429

Acknowledgement

The research reported in this study was conducted with funds provided to the first author by the College Board (Contract 2005-013) and by the Social Sciences and Humanities Research Council of Canada (SSHRC). We would like to thank the College Board and SSHRC for their support. The authors are solely responsible for the methods, procedures, and interpretations expressed in this study. Our views do not necessarily reflect those of the College Board or SSHRC.

References

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* [Computer Programs], Lincolnwood, IL : Scientific Software.

Gällmo, O, & Carlström, J. (1995). Some experiments using extra output learning to hint multi layer perceptrons. In L. F. Niklasson, M. B. Boden (Eds.), *Current trends in connections—Proceedings of the 1995 Swedish conference on connections* (pp. 179-190). Mahwah, NJ: Erlbaum.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement, 44,* 325-340. Invited Paper appearing in Special Issue of IRT-Based Cognitive Diagnostic Models and Related Methods, Lou DiBello & William Stout (Guest Editors).

Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and applications. (pp. 242-274).* Cambridge, UK: Cambridge University Press.

Gierl, M. J., Leighton, J. P., Wang, C., Zhou, J., Gokert, R., & Tan. A. (2006, December). *Validating cognitive models of task performance in algebra on the SAT©*. New York: College Examination Board.

Gierl, M. J., Wang, C., & Zhou, J. (2006, June). *Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT©*. New York: College Examination Board.

Leighton, J. P., & Gierl, M. J. (Eds.) (2007a). Cognitive diagnostic assessment for education: Theory and Applications. Cambridge, UK: Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. Educational Measurement: Issues and Practice, 26, 3-16.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41,* 205-236.

Luecht, R. M. (2006, September). *Assessment engineering: An emerging discipline*. Paper presented in the Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada.

McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition.* Oxford: Oxford University Press. 21-53.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning representations by back-propagating errors. *Nature, 323,* 533-536.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986b). *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd Edition., pp. 263-331). New York: American Council on Education, Macmillian.

Wiley, T. (2006, May). *Formative instruction and the quest for the killer application*, Eduventures, Boston, MA.