International Journal of Testing, 17: 302–321, 2017
Copyright © International Test Commission
ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2017.1297816

Routledge
Taylor & Francis Group

Challenges to the Use of Artificial Neural Networks for Diagnostic Classifications with Student Test Data

Derek C. Briggs
School of Education, University of Colorado at Boulder, Colorado, USA

Ruhan Circi
Partnership for the Assessment of College and Career Readiness, USA

Artificial Neural Networks (ANNs) have been proposed as a promising approach for the classification of students into different levels of a psychological attribute hierarchy. Unfortunately, because such classifications typically rely upon internally produced item response patterns that have not been externally validated, the instability of ANN estimates of attribute probabilities may not be widely appreciated. The present study illustrates the problem with both empirical and simulated data. In particular, it is shown that when an ANN is “trained” multiple times using the same data, attribute probability estimates can vary, sometimes dramatically. Researchers hoping to apply ANNs in the context of diagnostic classification models with student test data need to be very deliberate in checking the sensitivity of their findings.

Keywords: diagnostic classification models, cognitive diagnostic model, artificial neural networks, attribute hierarchy method, learning progression, ordered multiple-choice items

There are many of psychometric models and statistical techniques that can be used to classify students into knowledge/skill categories as a function of their test item responses (see, e.g., Rupp, Templin, & Henson, 2010). Among these, the Attribute Hierarchy Method (AHM), based on Tatsuoka’s Rule Space Method (Tatsuoka, 1983, 2009), has been used in some instances to classify examinees’ response patterns on a test into a set of attribute patterns that correspond to different hierarchically defined levels of mastery (Gierl, Leighton, & Hunka, 2007; Leighton, Cui, & Cor, 2009). The ability

Correspondence should be sent to Derek C. Briggs, School of Education, University of Colorado at Boulder, UCB 249, Boulder, CO 80309, USA. E-mail: derek.briggs@colorado.edu

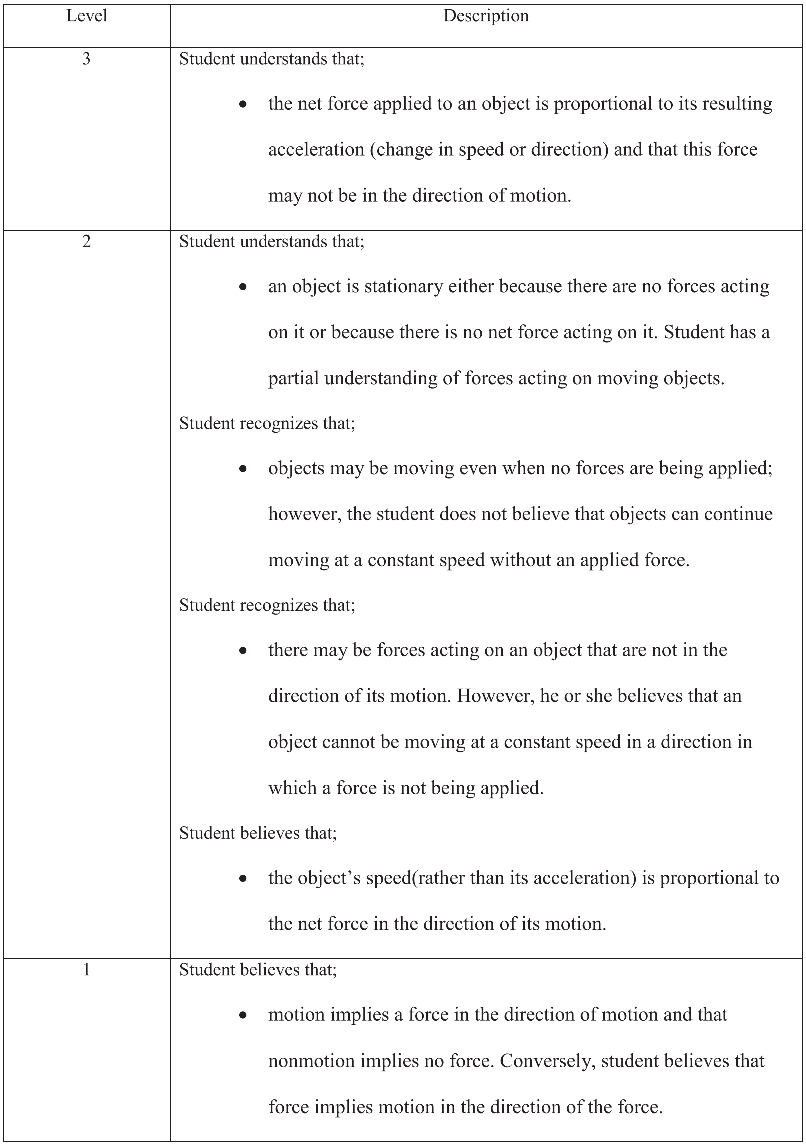
to classify students into levels using the AHM hinges on two key design elements. First, one must have some basis for a hypothesis about the cognitive attributes that students invoke when presented with test items that are all related to the same conceptual topic. Second, one must be able to write well-targeted items such that the responses to them can be used to make distinctions about the mastery states of these cognitive attributes. A methodological question of interest interacts with these design elements: how can one best convert a pattern of item responses into a defensible diagnostic classification? Artificial Neural Networks (ANNs) have been proposed as an attractive approach to this end (Gierl, Cui, & Hunka, 2008; Gierl, Wang, & Zhou, 2008). Unfortunately, use of an ANN in this data context can produce estimates of attribute probabilities that are very unstable, something that may not be well appreciated unless a great deal of care is taken to conduct sensitivity analyses. The purpose of this article is to establish this point directly using both empirical and simulated data. We conclude with some recommendations for best practices.

DATA CONTEXT: FORCE AND MOTION LEARNING PROGRESSION AND ORDERED MULTIPLE CHOICE ITEMS

The empirical context and motivation for this paper comes from a research project that sought to apply the AHM to polytomously scored Ordered Multiple-Choice (OMC) items (Briggs, Alonzo, Schwab, & Wilson, 2006; Briggs & Alonzo, 2012). OMC items differ from more traditional multiple-choice items in the sense that they have been written such that each of their response options can be linked to a level of a previously defined learning progression (c.f., Alonzo & Gotwals, 2012). Since a learning progression can be expressed as a hierarchy of discrete attributes (i.e., knowledge, skills, and abilities expressed at a small grain size), it follows that the AHM can be extended to accommodate OMC items, and thereby to facilitate diagnostic inferences that link student responses back to the hypothesized levels of the LP.[1] In the present context, we make use of empirical data from seven OMC items written to correspond to an LP on the physical science topic of Force and Motion. These items were administered to a sample of 1008 high school students at six schools in rural and suburban Iowa during the 2008–2009 school year. Figure 1 presents a simplified version of the Force and Motion LP that was the basis for the design of the OMC item options[2] (for details on the development of this LP, see Alonzo and Steedle, 2009).

[1] Throughout this article we use the terms “learning progression” and “attribute hierarchy” interchangeably.

[2] The learning progression has been simplified primarily by removing the lowest level “Student may understand force as a push or pull that may or may not involve motion.”



Level	Description
3	Student understands that; • the net force applied to an object is proportional to its resulting acceleration (change in speed or direction) and that this force may not be in the direction of motion.

2	Student understands that;
	• an object is stationary either because there are no forces acting on it or because there is no net force acting on it. Student has a partial understanding of forces acting on moving objects.
	Student recognizes that;
	• objects may be moving even when no forces are being applied; however, the student does not believe that objects can continue moving at a constant speed without an applied force.
	Student recognizes that;
	• there may be forces acting on an object that are not in the direction of its motion. However, he or she believes that an object cannot be moving at a constant speed in a direction in which a force is not being applied.
	Student believes that;
	• the object's speed(rather than its acceleration) is proportional to the net force in the direction of its motion.

1	Student believes that; • motion implies a force in the direction of motion and that nonmotion implies no force. Conversely, student believes that force implies motion in the direction of the force.

FIGURE 1

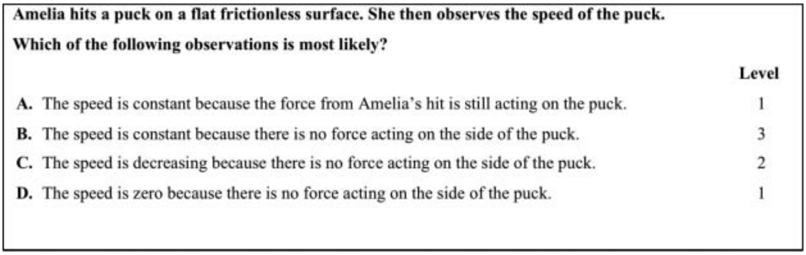
Simplified version of force and motion learning progression. From "Developing and Assessing a Force and Motion Learning Progression" by A. C. Alonzo and J. T. Steedle, 2009, *Science Education, 93*(3), 403–405. (c) 2008 Wiley Periodicals, Inc. Adapted with permission.

The levels of the Force and Motion LP can be defined with respect to three attributes:

A1 = motion implies force
A2 = net force is associated with speed
A3 = net force is associated with acceleration

This LP implies a simple linear conjunctive model such that A1→ A2 → A3. A student at level 1 of the learning progression typically thinks that motion implies force (A1); a student at level 2 typically believes that the speed of motion is associated with net force (A2); and a student at level 3 understands that the acceleration of motion is associated with net force (A3). The model is conjunctive not in the sense that each level requires a student to have mastered the preceding attribute, but in the sense that to master an attribute associated with a higher level of the progression (i.e., A3), a student must understand the context in which conceptions rooted in A1 and/or A2 would be insufficient to explain a relationship between force and motion in a given scenario.

Figure 2 provides an example of an OMC item that was written to distinguish between students at different levels of the Force and Motion LP. Notice that although option B is the most correct response in that it is linked to level 3 of the progression, level 2 is not entirely incorrect, and even responses A and D, although they represent a level 1 conception, provide some insights into the ways that students might be thinking about the relationship between force and motion. Table 1 shows the OMC items and the percent of student responses for each item option linked to each LP level.



Amelia hits a puck on a flat frictionless surface. She then observes the speed of the puck.

Which of the following observations is most likely?

		Level
A.	The speed is constant because the force from Amelia's hit is still acting on the puck.	1
B.	The speed is constant because there is no force acting on the side of the puck.	3
C.	The speed is decreasing because there is no force acting on the side of the puck.	2
D.	The speed is zero because there is no force acting on the side of the puck.	1

FIGURE 2
An example of an ordered multiple-choice item. From "Developing and Assessing a Force and Motion Learning Progression" by A. C. Alonzo and J. T. Steedle, 2009, *Science Education, 93*(3), 403–405. (c) 2008 Wiley Periodicals, Inc. Reprinted with permission.

TABLE 1
Response Frequencies for Nine OMC Items Associated with Simplified Forces and Motion Learning Progression

	Item 1	Item 2	Item 3	Item 4	Item 5
Level 3	35%	18%	38%	17%	38%
Level 2	24%	53% + 8%	40%	39% + 37%	38%
Level 1	39% + 2%	20%	23%	6%	36% + 9%
	Item 6	Item 7			
Level 3	20%	25%			
Level 2	58%	11% + 36%			
Level 1	4% + 18%	29%			

Note. $N = 1008$ Students. Multiple percentages added together reflect multiple response options linked to the same level of the learning progression.

ADAPTING THE AHM FOR USE WITH POLYTOMOUSLY SCORED ITEMS

A premise of the AHM is that when students encounter test items, the success or failure of a student when solving these items can be distinguished with respect to the psychological attributes the examinee must possess in order to solve them. These structured attributes are hypothesized to form a hierarchy that defines the *order* of attributes a student must invoke to solve a given test item. The main steps in applying the AHM are (a) the (confirmatory) specification of an attribute hierarchy which defines the ordering of attributes that must be mastered in order to solve items, (b) the specification of expected response patterns on the basis of a Q_r matrix (where the r subscript stands for "reduced") that shows the item-attribute combinations that are possible if the specified hierarchy is accurate, and (c) the estimation of attribute probabilities for each student on the basis of their observed item responses. A point we shall emphasize repeatedly in this article is that steps (a) and (b) do not invoke empirical student data, a fact that can have significant consequences for the use of an ANN to classify students on the basis of their estimated attribute probabilities.

TABLE 2
Q_r Matrix Excerpt for First Two Items: Item Option Level Approach

Item Option	1A	1B	1C	1D	2A	2B	2C	2D
A1	1	1	1	1	1	1	1	1
A2	0	1	1	0	1	0	1	1
A3	0	1	0	0	1	0	0	0
LP Level	1	3	2	1	3	1	2	2

TABLE 3
Expected Response Patterns for First Two OMC Items: Item-Option Level

Hypothetical student	Expected response	Attributes
1	[1/2 0 0 1/2] [0 1 0 0]	100
2	[00100] [0 0 1/2 1/2]	110
3	[01000][1000]	111

Since OMC items are scored polytomously, the Q_r matrix must be specified at the item *option* level. This is illustrated for the first two Force and Motion OMC items in Table 2 (the full Q_r matrix is too long to fit on a page).

A Q_r matrix provides a roadmap that indicates which item option students would be expected to select in a conditional sense—*if* their attribute profiles were known, and *if* the specified attribute hierarchy is accurate. It follows from this that a Q_r matrix can be used to simulate a matrix of expected response patterns for hypothetical students at each level of the LP—again, assuming that the hierarchy of attributes specified within the LP is accurate. The expected response matrix at the item option level for the OMC Force and Motion items 1 and 2 is represented in Table 3.

Model-Data Fit

In the context of a set of dichotomously scored items, the fit of observed student response patterns to the attribute hierarchy is calculated using the Hierarchical Classification Index (HCI; Cui, Leighton, Gierl, & Hunka, 2006). The HCI provides a student-specific statistic ranging between −1 and 1, with a value of 1 indicating perfect fit. According to the developers of the HCI, values greater than 0.8 suggest a pattern of item responses with an excellent fit to the hierarchy, values between 0.6 and 0.8 indicate a moderate fit, and values less than 0.6 suggest a poor fit (Cui, 2007; Cui et al., 2006). Because OMC items are polytomously scored, the HCI cannot be applied and interpreted in the same way. The strategy we take to evaluate the fit of OMC-based response patterns to our attribute hierarchy is similar to the one used for the HCI but also differs in key ways. The biggest distinction is that we specify an attribute hierarchy *within* an item (between item response options) rather than between items. As a result, given the expected response patterns found in our Q_r matrix, when a student selects an item option corresponding to an attribute combination at the high end of the Force and Motion attribute hierarchy (e.g., level 3), we assume that the student has mastered all these attributes and we expect the student to select the similar (higher level) response option when it is present in another item. Hence, the conception of fit requires *consistency* among the *options* selected by students who have mastered the same attributes. To this end, we compute a “Response Consistency Index” (RCI) as

$$RCI_i = 1 - \frac{M_i}{J(J-1)}. \tag{1}$$

The *RCI* is computed for each student, indexed by *i*. Let M_i indicates the total number of times a given student chooses a response option that is inconsistent with the response option chosen in the reference item, and *J* represents the total number of OMC items. It follows that the denominator in Equation (1) represents the number of times that any single item can be compared to all remaining items.

To give a simple example of how the RCI can be computed consider one of the observed response patterns we actually observed from our sample of 1,008 students: BDCADBC. The LP levels (scores) associated with the response pattern are 3232123. For item 1, a response of B is associated with a level 3 response, indicative of a student who has mastered attributes 1, 2, and 3 of the Force and Motion LP. We then compare this item to the remaining six items and count the number of times where this student chose a response option other than level 3. This was the case for items 2, 4, 5, and 6, so the number of misfits relative the first item response is four. The same process is repeated for the other six items. The sum of all misfits for this example is 30. The number of possible comparisons was 42. Therefore, the RCI for this student is $1 - \frac{30}{42} = 0.29$.

The RCI provides an index of consistency, but what is the threshold for acceptable fit? To provide a normative answer this question specific to the current

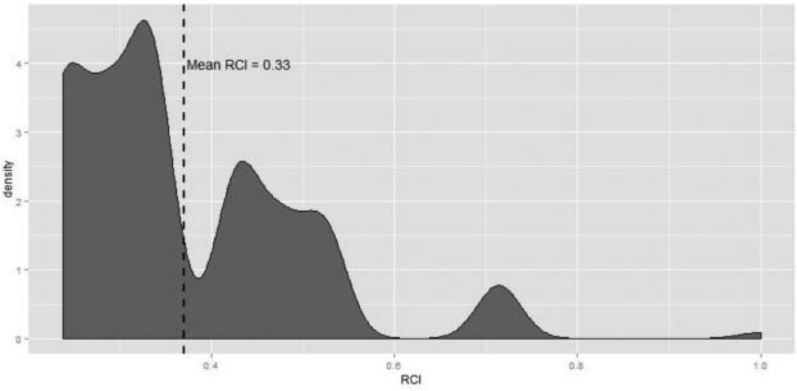


FIGURE 3
Observed distribution of the RCI for seven FM OMC items.

data context, we generated 1000 responses strings in which values between 1 and 3 were selected at random with equal probabilities. The mean RCI for this data was 0.33, and can be interpreted as the average response consistency one would expect to observe by chance. Figure 3 plots the distribution of the observed RCI values, and the value of 0.33 is shown with a vertical line. Because a good proportion of the observed RCI values overlap with values we would expect to observe by chance, it seems clear that many students in our sample did not respond to these OMC items in the way that we would have expected if the underlying learning progression hypothesis was accurate. Later, we explore the impact of this by simulating data with better fit to an underlying hierarchy.

ESTIMATING ATTRIBUTE PROBABILITIES WITH AN ARTIFICIAL NEURAL NETWORK

The purpose of an ANN is prediction: given some set of input covariates that can be linked to known outcomes, what weighted combination of these covariates maximizes the likelihood of observing these outcomes? The simplest version of an ANN collapses to a logistic, log-linear or linear regression model. For example, to predict whether a student has mastered the attributes associated with the Force and Motion LP, one could simulate data by replicating the three expected response strings associated with each of our three LP levels 100 times. Then two logistic regressions would be specified. In the first, the log odds of mastering A2 and A3 relative to A1 could be modeled as a function of scores on OMC items 1

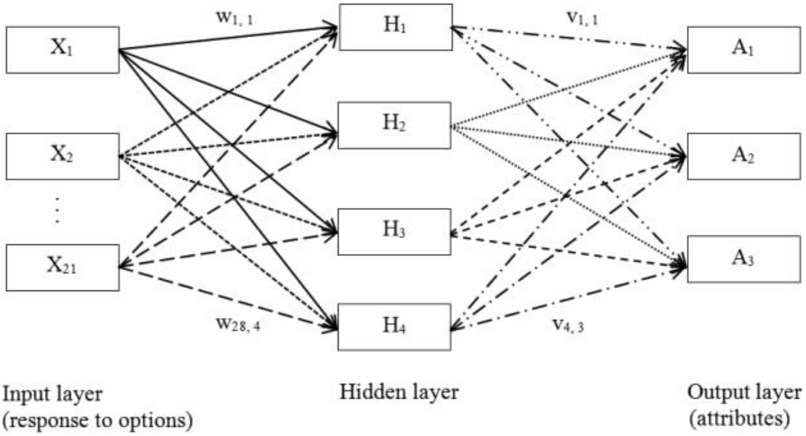


FIGURE 4
Neural network schema for force and motion LP.

through 9. A second logistic regression would use the log odds of mastering A3 relative to A2 and A1 as an outcome of interest.

An ANN approach elaborates this by making it straightforward to model interactions among covariates, and by incorporating multivariate outcomes. Figure 4 illustrates the typical components of a neural network. The neural network in Figure 4 includes three layers: one input layer, one hidden layer, and an output layer. Each layer consists of "neurons" (shown as rectangles), which have different interpretations depending upon the layer. In our context, for a given student, each neuron in the input layer represents a scored response to a test item. For polytomously scored items, the number of neurons depends on whether an expected response matrix has been specified at the item-option or item level. The neurons in the output layer are fixed to correspond to the different attributes hypothesized for the attribute hierarchy (e.g., in the present context there would be only three neurons in the output layer). A hidden layer in a neural network makes it possible to examine the impact of input neuron interactions on output neurons; if the hidden layer in Figure 4 where excluded, one would only be modeling the main effects of each input neuron on each output neuron. The arrows connecting the neurons between layers represent weights. The idea is to predict the output neurons given the input neurons as a function of these weights. The weights are estimated iteratively such that they collectively minimize the difference between the known value of attributes for an expected response pattern, and the predicted value. Because the estimation process is iterative, all weights are usually initialized with random values drawn from a standardized normal distribution. This is the default approach when using the R package `neuralnet` as we do in what follows. There are many different approaches that can be taken to converge upon estimates for these weights, including backpropagation and resilient backpropagation (Günther & Fritsch, 2010). A more detailed description of the feature and differences in ANN approaches is outside the scope of the present article (but see Anastasiadis, Magoulas & Vrahatis, 2005; Cui, Gierl, & Leighton, 2010; Günther & Fritsch, 2010; Intrator & Intrator, 2001).

The process of estimating weights in an ANN is referred to as "training" the ANN by providing a set of inputs for which outputs are already known. Once training is complete, a new data set for which inputs are known but outputs are unknown can be inserted, and the ANN can be used to generate a predicted value for each output. In the context of student test data, once an analyst has specified an attribute hierarchy, Q_r matrix and expected response matrix, it can be deceivingly easy to train an ANN and then generate attribute probabilities for observed item response patterns. However, to reiterate, one must keep in mind that no empirical data is necessary to estimate the parameters of an ANN in this context—one only requires an expected response matrix, and this is generated from theory. In other words, in a diagnostic assessment context using test data, an ANN is always trained using simulated data.

This is the most likely the reason that Cui and Leighton (2009) place considerable emphasis on establishing that an attribute hierarchy has adequate fit before it is used as a basis for training an ANN. However, even when a hierarchy appears to display acceptable fit for most respondents, there are still important reasons to be cautious about the results. As discussed by Intrator and Intrator (2001), there are three main reasons that caution is necessary:

1. The solution to any given ANN with some fixed number of hidden layers and units within these layers is not uniquely determined because ANN models are overidentified.
2. Because the cost functions in an ANN are nonconvex, estimates of weight parameters will often get stuck at local minima, so the choice of starting values can have a big impact on where each weight will give the appearance of converging.
3. The choice of ANN architecture (number of hidden layers and units within layers) is often done on a trial and error basis, and the optimal choice is never known in advance, making the estimation endeavor largely atheoretical.

All three of these issues can lead to large prediction variance, which is why Intrator and Intrator recommended that multiple "robustification" steps be taken when training an ANN. These include weight decay regularization, adding noise to the input neurons during training, and ensemble averaging (training the same ANN multiple times and then taking the mean of output probability estimates). Unfortunately, there is little evidence that such steps have been taken in past applications of the ANN for diagnostic classifications using test data. In the next section, we use empirical data from our OMC project to focus on the impact that the simplest of these steps, ensemble averaging, can have on subsequent interpretations. In the following section, we use simulated data to explore whether our findings generalize beyond the present data context.

THE VARIANCE OF ATTRIBUTE PROBABILITY ESTIMATES

We estimated attribute probabilities for the student sample responding to the Force and Motion OMC items based on a neural net specification with one hidden layer, four hidden units, backpropagation and a learning rate of 0.01. The response strings associated with the three attribute combinations in our expected response matrix were each replicated 100 times and this simulated data was then

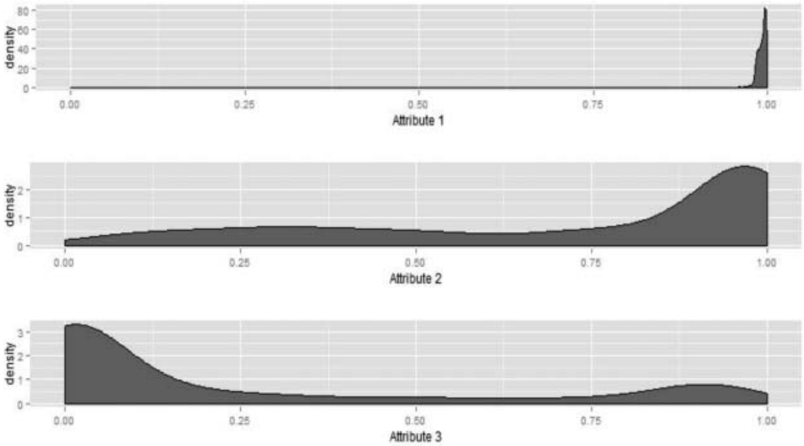


FIGURE 5
Distribution of attribute probabilities from single ANN training.

used to train an ANN.[3] For example, the attribute hierarchy associated with the Force and Motion OMC items converged in 9764 steps with a sum of square error of 0.018. As is common practice, we examined the probability estimates for the attributes of the expected response patterns with which we trained the network. The results showed that the ANN precisely mapped the input-output relationship that was specified by the expected response matrix.

Next we consider the contrast of interest: the distribution of estimated attribute probabilities for our sample of 1008 students when the estimates are based on a single ANN training as described, and the distribution that results when we conduct 100 different ANN trainings, produce attribute probability estimates for each training, and then average the results. Figures 5 and 6 allow for a comparison of the distribution of probability estimates for these two approaches. As can be seen, the distributions of the probability estimates for A2 and A3 are noticeably different when based on a single ANN training relative to the average across 100 trainings. This is because attribute probability estimates for the same student will vary depending upon the ANN training upon which it was based. Our results also do not appear to be sensitive to choices regarding the number of

[3] With OMC items, it is not really meaningful to speak of estimating the probability of a student having mastered the lowest attribute in the hierarchy. This is because each item response option is always assumed to encompass the lowest attribute in the hierarchy. That is, there is never a single item measuring just the lowest attribute (i.e., A1) where it would be possible to answer incorrectly. So in what follows, while we present attribute estimates for A1, these are generally between .98 and 1.00 and provide no unique information.

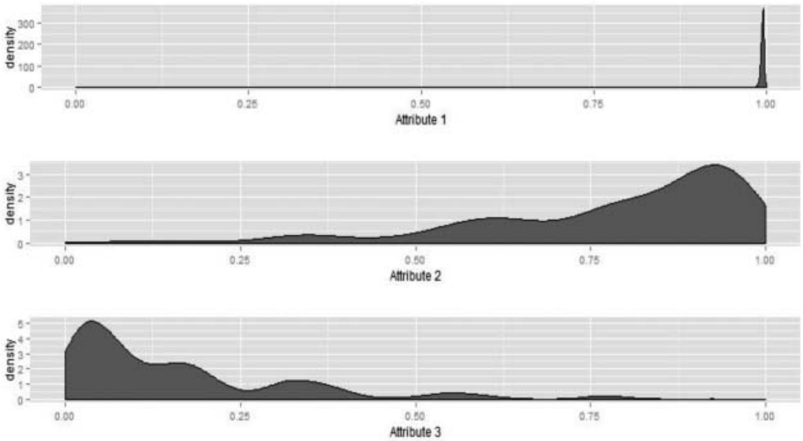


FIGURE 6
Distribution of attribute probabilities from average across 100 ANN trainings.

hidden layers and neurons within these layers. Adding or reducing layers and/or neurons did not change the basic patterns of our findings with this data.

One plausible explanation for the variability of these probability estimates across trainings of the ANN is that our distribution of RCI values is not much different than we would expect to observe by chance. However, even for the subset of students with relatively high RCI values the probability estimates for A2 and A3 vary dramatically, as is evident by the variability of attribute probability estimates shown in Table 4 for seven student response vectors drawn from our sample with RCI values that range from 0.286 to 1.

TABLE 4
Variability in Attribute Probability Estimates for Seven Illustrative Students Across 100 ANN Trainings

		Attribute 2			Attribute 3		
Case	RCI$_i$	Min	Max	SD	Min	Max	SD
1	0.286	0.10	0.99	0.32	0.10	0.92	0.21
2	0.333	0.10	0.99	0.33	0.10	0.91	0.27
3	0.429	0.10	0.99	0.31	0.10	0.98	0.32
4	0.476	0.10	0.99	0.21	0.10	0.99	0.14
5	0.524	0.10	0.98	0.32	0.10	0.20	0.18
6	0.714	0.40	0.99	0.02	0.10	0.95	0.04
7	1.000	0.99	1.00	< 0.01	0.97	0.98	< 0.01

A fundamental argument in favor of taking a probabilistic approach to classifying students for diagnostic purposes is that such an approach offers more nuanced insights into a student's strengths and weaknesses than taking a more straightforward approach, such as simply classifying a student as a function of his or her modal response. Of the 1008 students in our sample, 835 (83%) could be classified into a level of the Force and Motion LP on the basis of the OMC response level selected the most frequently (i.e., the modal response). We can contrast this with the classifications that would be made using a cutoff of 0.75 on each attribute probability estimate from the ANN that was based on either one training or the average of more than 100 trainings. This is not to suggest that the model classification represents the truth, but it does represent a fixed criterion that will not change as a function of ANN trainings. When the modal classifications are compared to probabilistic classifications from a single ANN training, the two methods have exact agreement for only about 50% of the students. When compared to probabilistic classifications from the mean of 100 ANN trainings, the agreement increases to about 68%. The upshot is that making diagnostic classifications based on a single ANN training tend to magnify differences in interpretation relative to the fixed modal approach, but a significant portion of these differences are unreliable—a combination of students with poor fit to the hypothesized hierarchy and ANN weight parameter estimates that are susceptible to local minima as a function of randomly generated starting values from a single trial.

DEMONSTRATİON WİTH DATA SİMULATED TO FİT THE ATTRİBUTE RESPONSE HİERARCHY

A reasonable objection to the results shown above is that they may be an artifact of our empirical data, since the OMC items in question appear to show poor fit to the hypothesized Force and Motion attribute hierarchy. To examine whether these results were unique to this particular hierarchy of attributes and its associated expected responses, we simulated data from two alternative attribute hierarchy structures: a five-attribute hierarchy with a simple conjunctive linear structure, and an eight-attribute nonlinear branched hierarchy. The second of these attribute hierarchies, pertaining to abilities of students to solve algebra problems, was previously illustrated using simulated data in Gierl, Leighton, and Hunka (2007).

Figure 7 depicts both hierarchies. The five-attribute hierarchy data differs from the three attribute hierarchy data in our FM LP example in that it features two additional attributes but just five items that are scored dichotomously, with each item linked to one and only one attribute. This simplifies the input layer of the ANN while adding more options to the output layer. The eight-attribute hierarchy differs in that it features a much more complex, nonlinear set of relationships between attributes and features a larger set of 21 items, dichotomously

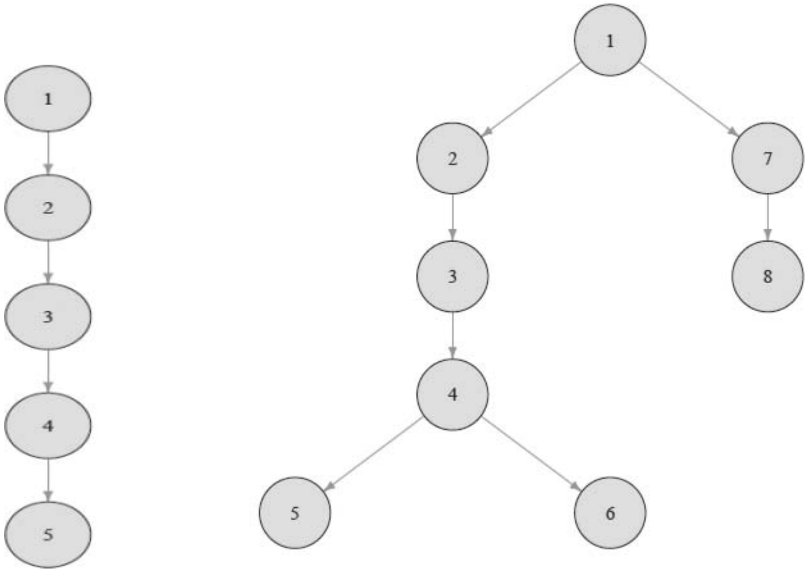


FIGURE 7
Five-attribute linear hierarchy (on the left) and eight-attribute branched hierarchy (on the right).

scored, as inputs. A key feature of the student response data that we simulate is that it is has been simulated to show greater average fit to each attribute hierarchy than was found in our empirical context. We accomplish this as follows:

1. For a given hierarchy, we specify a Q_r matrix for the minimum number of items needed to identify each attribute in the attribute hierarchy uniquely. For the conjunctive five-attribute case, this requires five items; for branched eight-attribute case, this requires 21 items.
2. Starting from the expected response patterns, we generate a larger set of possible observed response vectors by replacing some incorrectly answered items with “guesses” (correct responses when faced with an item that requires an attribute a student has not mastered) and some correctly answered items with “slips” (incorrect responses when faced with an item that requires an attribute a student has in fact mastered). For the five-attribute case we simply generate all 32 (5^2) possible five-item response vectors. For the branched eight-attribute case where there are 21 distinct expected response vectors, we generate an additional 84 response vectors by inserting four random switches (0 to 1 or 1 to 0) per each of the 21 expected response vectors. This results in 110 possible responses vectors.
3. For each response vector generated in step 2, we compute the HCI.

TABLE 5
Average Variability (Standard Deviation) in Attribute Estimates Across 100 ANN Trainings Using Simulated Responses to Conjunctive 5 Attribute Hierarchy

	HCI Bins for 300 Simulated Item Response Vectors			
Attribute	< 0.40	0.40–0.59	0.60–0.79	0.80–1.00
1	0.01	0.01	< 0.01	< 0.01
2	0.14	0.09	0.03	< 0.01
3	0.12	0.22	0.04	< 0.01
4	0.15	0.17	0.27	< 0.01
5	0.22	0.16	0.21	< 0.01
N	70	53	36	141

4. Randomly draw with replacement N = 300 from the set of possible observed item responses (N = 32 and N = 110 for each attribute hierarchy example) under the constraints that the mean HCI = .60 (i.e., the conventional cutoff criterion for “moderate” fit), and that all expected response patterns from the original Q_r matrix be included at least once.

The result is two simulated data sets, one with 300 responses to five items, and another with 300 responses to 21 items. Both data sets had similar HCI distributions with means of 0.60, and medians of 0.70. Next, we repeated the same process applied to the empirical data from our FM LP. We trained an ANN 100 times, each time using the same expected response vectors for each attribute hierarchy. We then use the results of each of these 100 trainings to estimate attribute probabilities for the 300 simulated response patterns described.[4] Note that the outcome of interest in these simulations is not classification accuracy but precision. We operationalize this as the average variability in estimates of mastery (i.e., probabilities) for each attribute in a given hierarchy computed over each of 300 simulated respondents.

We summarize the results from these simulations in Tables 5 and 6. The main rows in each table represent each of the unique attributes for a given hierarchy; the columns bin the 300 simulated response vectors by their fit to the attribute hierarchy using the HCI. Each cell in the table shows the average of the SD of attribute probability estimates over 100 ANN trainings. In general, it is desirable

[4] We also implemented a version of this simulation exercise in which we repeated steps 2–4 above 100 times in order to get a sense for the empirical standard error around our statistic of interest, the mean standard deviation of a given attribute probability. We find that this standard error is never larger than .002. Because of this, we report only the results from the simpler version of our simulation. The results from the more elaborate version are available upon request.

TABLE 6
Average Variability (Standard Deviation) in Attribute Estimates Across 100 ANN Trainings Using Simulated Responses to Branched 8 Attribute Hierarchy

	HCI Bins for 300 Simulated Item Response Vectors			
Attribute	< 0.40	0.40–0.59	0.60–0.79	0.80–1.00
1	< 0.01	< 0.01	< 0.01	< 0.01
2	0.18	0.18	0.13	0.02
3	0.21	0.22	0.20	0.04
4	0.14	0.18	0.25	0.05
5	0.15	0.23	0.17	0.10
6	0.18	0.19	0.15	0.08
7	0.12	0.10	0.14	0.04
8	0.10	0.21	0.11	0.04
N	66	62	45	127

for these values to be small, indicating that attribute probability estimates are insensitive to factors such as ANN starting values and convergence to local minima. Holding the attribute constant one expects the average SD to get smaller across columns as fit to the hierarchy increases.

The results are largely consistent with this expectation, but there are some notable exceptions. With respect to the simple conjunctive five-attribute hierarchy, when response vectors show excellent fit to the hierarchy (HCI between 0.80 and 1), there will be almost no variability in attribute probability estimates across multiple trainings of the ANN. However, once fit to the hierarchy drops below 0.80, we begin to see some significant variability in attribute probability estimates. For the 36 response vectors with HCI values between 0.60 and 0.79 (indicative of moderate fit), the average SD jumps from less than 0.01 to 0.27 and 0.21 for attributes 4 and 5. For response vectors with HCI values below 0.60, we begin to see significant variability in four of the five attribute estimates. Interestingly, the pattern in average SD by attribute as a function of HCI is generally not monotonic with the exception of A3. Many of the same patterns found for the five-attribute hierarchy are present in the results of the eight attribute example, but here we see that even in the case of 127 response vectors with near perfect fit to the hierarchy, we see evidence of uncertainty in attribute probability estimates. Notably, the average variance is largest for the two attributes (5 and 6) that are not ordered relative to one another while still depending on mastery of four precursor attributes (recall Figure 7). For HCI values below 0.80, the magnitudes and the average SD by attribute is very similar to the magnitudes found for the five attribute example, suggesting that the amount of uncertainty caused by choice of ANN training has no clear relationship with the complexity of the attribute hierarchy being modeled.

DİSCUSSİON

Researchers need to be very deliberate in their application of ANNs for estimating attribute probabilities. The weight parameters in ANNs are not directly identifiable and are prone to converging upon local minima as a function of randomly selected starting values (c.f., Li, Alnuweiri, Wu, & Li, 1993). Good suggestions are available in the ANN literature to guard against such problems and to increase the interpretability of results (e.g., Intrator & Intrator, 2001; Panchal, Ganatra, Shah, & Panchal, 2011). We have illustrated the importance of just one of these approaches: averaging probability estimates across multiple ANN trainings.[5] Another sensitivity analysis to consider would be to use a fixed initial value for all weights, or to fix the seed used to generate weights randomly to examine the impact this has on probability estimates. However, this runs the risk of a solution that will be biased towards some particular set of weights. Two additional checks to deal with potential problems with estimated weights can be investigating the variability of the final weight estimates, and to employ several global search techniques (e.g., particle swarm optimization; Al-Shareef & Abbod, 2010). Finally, Cui, Gierl, and Guo (2016) suggested the potential exploratory use of an unsupervised ANN known as a "self-organizing map" (Kohonen, 2001) in diagnostic classification modeling. In contrast to the supervised ANN we have illustrated in this article, a self-organizing map allows the nodes in the output layer to be treated as unknown.

The supervised ANN is by far the most frequent ANN approach and has been applied for decades in medical research with health outcome data. In such contexts, the success of the approach can be readily evaluated because outcomes are eventually observed—for example, patients get cancer or they do not. Given the aim of successful prediction, the black box nature of an ANN is not problematic. If it leads to significantly better predictions than a simpler modeling approach, it is empirically useful. The conventional recommendations for good practice in ANN applications involve gathering data that can be split into three sets for training, testing and validation data. The training data is used to train the ANN architecture, and a test set is used to examine whether the established architecture generalizes. More specifically, a test set is used to examine whether the ANN has too many or too few hidden layers and neurons within these layers. Finally, a validation data set is used to evaluate the predictive accuracy of the ANN. Importantly, all three data sets are based on empirical data in which both inputs and outputs of interest are known in advance.

In contrast, because an observable criterion is not available for latent classification models, it is not possible to evaluate an ANN using empirical data in these contexts. Instead, an ANN is evaluated using expected

[5] Similar advice has been offered in the past by Venaples & Ripley, 2002 (pp 342–344).

responses simulated based on the confirmatory theory used to generate a learning progression/attribute hierarchy and associated items. If the theory and/or items are somewhat flawed, the ANN will naturally incorporate those flaws. As we have demonstrated, in the absence of evidence of near perfect student fit to a hierarchy, prediction variance in estimated attribute probabilities can be quite large. This raises a bit of a paradox in that for students whose responses show perfect fit, there is really no need to produce attribute probability estimates. For such students one would only need to find the item with the score associated with the highest attribute and classify a student accordingly. It is precisely for students with response patterns that do not fit perfectly that one would hope a statistical model could provide some insights beyond that which is apparent to the eye. Yet these are the students for whom an ANN is least likely to produce reliable estimates, especially if based on just one training. At a minimum then, analysts interested in using ANNs should always check the sensitivity of their attribute probability estimates to different ANN trainings and consider averaging over multiple training when variability is evident.

Another alternative might be preferable. Instead of training an ANN using simulated data based on expected response patterns, it may be more sensible to gather empirical data with the most common observed student response patterns and then have a panel of content experts give each pattern a holistic score with respect to each attribute of interest. The scoring could be informed by an underlying Q_r matrix that maps each item (or item response option in the case of OMC items) to the attribute or attributes to which it was targeted by design, but the panel would have the flexibility to take into account ambiguous item wording, or slips and guesses within responses patterns they would be willing to overlook in their diagnosis of the student. If an ANN training were to proceed in this manner, it would now become possible to follow the conventional practices for ANNs used in other settings (i.e., splitting of data into training, testing, and validation sets). It seems likely this might reduce uncertainty due to choice of ANN training because the data being used to train the ANN would more closely resemble the type of data for which estimates of attribute probabilities are desired. This would make the use of ANNs for diagnostic assessment more like the use of ANNs for automated scoring of essays or constructed responses. The value behind such an approach is that it seems likely to not only improve the accuracy and precision of attribute probability estimates, but through the process of having panelists analyze and score observed student responses, it is likely to provide developers of learning progressions with actionable insights about the underlying cognitive model that motivated the learning progression in the first place.

ORCID

Derek C. Briggs http://orcid.org/0000-0003-1628-4661

REFERENCES

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education*, *93*, 389–421.

Alonzo, A. C., & Gotwals, A. W. (2012). *Learning progressions in science: Current challenges and future directions*. Rotterdam, The Netherlands: Sense Publishers.

Al-Shareef, A. J. & Abbod, M. F. (2010, March). *Neural networks initial weights optimization*. Paper presented at 12th International Conference on Modelling and Simulation. Retrieved February 2015 from IEEE Xplore. doi: 10.1109/UKSIM.2010.19

Anastasiadis, A, Magoulas, G, & Vrahatis, M. (2005). New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing*, *64*, 253–270.

Briggs, D. C., Alonzo, A. C., Schwab, S., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33–63.

Briggs, D. C. & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In A. Alonzo & A. Gotwals (Eds.), *Learning progressions in science* (pp. 293–316). Rotterdam, The Netherlands: Sense Publishers.

Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. (2006, April). *A person fit statistic for the attribute hierarchy method: The hierarchy consistency index*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cui, Y., Gierl, M., & Guo, Q. (2016). Statistical classification for cognitive diagnostic assessment: An artificial neural network approach. *Educational Psychology*, *36*(6), 1065–1082.

Cui, Y., Gierl, M., & Leighton, J. P. (2010). *Estimating the attribute hierarchy method with mathematica*. Retrieved October 2010 from http://www.ualberta.ca/~mgierl/files/conferences/Estimating%20the%20Attribute%20Hierarchy%20Method%20With%20Mathematica.pdf

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person-fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 429–449.

Gierl, M., Cui, Y., & Hunka, S. (2008). Using connectionist models to evaluate examinees' response patterns to achievement tests. *Journal of Modern Applied Statistical Methods*, *7*(1), 234–245, http://digitalcommons.wayne.edu/jmasm/vol7/iss1/19.

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

Gierl, M., Wang, C., & Zhou, J. (2008). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *The Journal of Technology, Learning and Assessment*, *6*(6). Retrieved June 8, 2009 from http://www.jtla.org

Günther, F. & Fritsch, S. (2010). Neuralnet: training of neural networks. Retrieved October 2012 from http://journal.r-project.org/archive/2010-1/RJournal_2010-1_Guenther+Fritsch.pdf

Intrator, O., & Intrator, N. (2001). Interpreting neural-network results: A simulation study. *Computational Statistics and Data Analysis*, *37*, 373–393.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Berlin, Germany: Springer-Verlag.

Leighton, J. P., Cui, Y., & Cor, M. K. (2009). Testing expert-based and student-based cognitive models: An application of the attribute hierarchy method and hierarchical consistency index. *Applied Measurement in Education*, *22*, 229–254.

Li, G., Alnuweiri, H., Wu, Y., & Li, H. (1993). *Acceleration of back propagation through initial weight pre-training with delta rule*. Paper presented at IEEE International Conference on Neural Networks. Retrieved July 2014 from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=298622.

Panchal, G., Ganatra, A., Shah, P., & Panchal, D. (2011). Determination of over-learning and over-fitting problem in back propagation neural network. *International Journal on Soft Computing*, *2* (2), 40–51.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the Rule Space Method*. New York: Routledge.

Venaples, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.