Artificial Neural Network-based Performance Assessments Using Simulations

Ron Stevens (Presenter)[1], Adrian Casillas & Terry Vendlinski
UCLA IMMEX Project, 5601 W. Slauson Ave. #255, Culver City, CA 90230
Tel: 310-649-6568, Fax: 310-649-6591

ABSTRACT

We have explored the ability of artificial neural network technologies to generate performance models of complex problem-solving tasks without detailed a priori knowledge of the nature of the task. To test the generalizability of this approach, we have applied this analysis to two diverse content domains – high school science and clinical patient management. In both domains, the neural networks, using only the sequence of actions taken while performing the task, generated multiple classification groups defining different levels of competence.

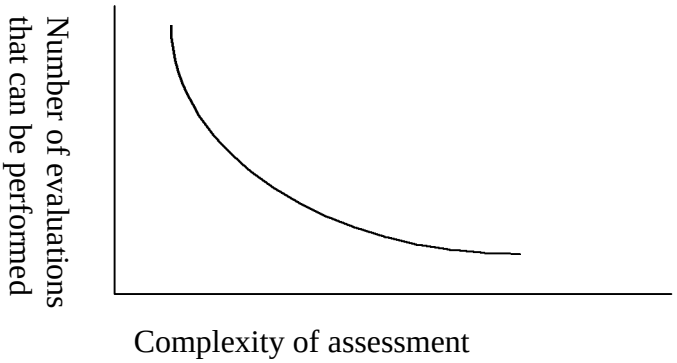
Introduction and Background

Most educators would agree that assessments in which students are actually required to apply what they have learned are better predictors of future success than tests that require mere recall of concepts (Maker, 1994), and these assessments can themselves become episodes of learning for test takers (Wolf, 1993). Examples of tasks that provide a rich reasoning environment include case-based and problem-based learning and these have been used as effective instructional strategies in professional schools for several decades (Elstein, 1993; Kolodner, 1993; Schank, 2000). These approaches have been attractive and effective not only for the motivational and realistic contexts they provide, but also for the balance of knowledge styles needed during the problem solving tasks and the different perspectives they reveal about student learning (Barrett and Depinet, 1991). Case-based reasoning is grounded in the belief that real-life reasoning and problem-solving behavior is almost never original, and that solutions to new problems are adaptations of previous problem solutions (Kolodner, 1997). Whether this is through the recall of representative or contradictory exemplar cases (Berry and Broadbent, 1988), by mental model generalizations (Johnson-Laird, 1983), or by applying scripts across a number of cases (Hudson, et al., 1992) is less clear. This is true because many aspects of case-based reasoning may involve the use of compiled knowledge or implicit memory that, for the most part, involves unconscious thought. The above theoretical considerations, combined with practical experiences of Shayer and Adey (1993), indicate the need to provide students with diverse, continual and prolonged problem-solving experiences. Wiggins (1993) and others (e.g. Schwartz, 1989) would also suggest the need to begin routinely assessing students in such formats.

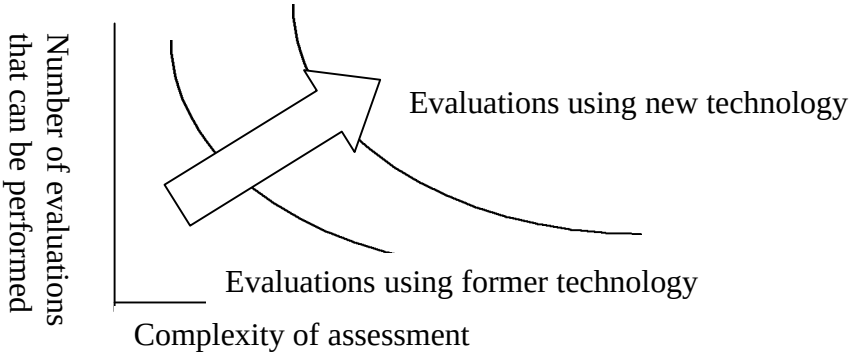
The benefits of alternative assessments, however, begin to encounter practical barriers as the transition is made from theory to practice. These include the difficulties of providing comprehensive coverage of a discipline, lengthy task development times of months to years, costs, logistics of implementation, etc. There are also human factors to consider.

For instance, as an assessment becomes more contextually sensitive and more holistic we are able to better understand student thinking, but the number of students we can evaluate decreases.

[1] Address correspondence to immex_ron1@hotmail.com



The curve above is similar to a production curve in manufacturing and can be thought of as an assessment production curve in educational settings. The curve represents the trade off between the number of evaluations that are possible and the complexity of each evaluation. Assessments become more complex as they allow or require students to consider the interaction among an increasing number of concepts. The function describing the relationship between complexity and number of evaluations, however, need not remain fixed. Experience suggests that new tools or technology can move the production curve up and to the right (de Neufville, 1990). As seen below, this allows a larger number of more complex evaluations and allows greater insights into student learning and understanding.



Computer technologies hold promise for accelerating the adoption of alternative learning and assessment modes in education, and, in particular, for extending, enhancing and scaling case-based reasoning. Extensions include the rapid creation of many cognitively complex cases, resulting in broad curricular coverage. Technology enhancements of case-based methods allow for better documentation and reporting of student performance, permit the development of cognitive and strategic profiles and can expand feedback mechanisms. Scale is accomplished by web-based delivery of problem-sets and aggregation of data across classrooms and schools. In these ways technology can make case-based reasoning not only possible in K-12 settings but also practical.

Parallel advances in cognitive science are also suggesting ways that technologies can be used to derive a richer understanding of student learning (see for instance, this symposium). Many

current systems take advantage of the observation that humans do not attempt to solve problems using completely random actions (Mann and Jepson, 1993), and the observation that novices and experts will often employ a different subset of actions when solving a particular problem. Even if the path to a solution is unknown and one merely searches for clues to a strategy or tries various strategies, the search goal and prior experience impose some order on this performance. It is this inherent order in student performances and the resulting coherent patterns that emerge from performance data aggregation that can help us make meaning of such performances. By studying such patterns and their dynamics over time, educators can assess student understanding and not just their recall of memorized facts (Fischer and Bidell, 1997). They can also use such patterns to identify the ways learning "breaks down" or goes awry" in order to develop interventions to improve that learning (Messick, 1989).

For 12 years the IMMEX™ project at UCLA has been using computer technologies to accelerate the adoption of alternative learning and assessment modes in education, first in medical education and then in K-12 schools (Stevens, et Al., 1996). IMMEX™ software is a set of problem-solving authoring and performance data collection and analysis tools, which inherits much of its theory from case-based and problem-based learning (Stevens, 1991). Each case opens with a problem-framing prologue and then presents a problem space that students can explore for items that support or refute a changing series of hypotheses. The data variables in IMMEX™ are loosely structured as menu items or button links and can be selected in any sequence. These items, and the problems posed, defines the problem space for the students and researchers. The user interface contains few "seductive details" as to prevent interference (Harp and Mayer, 1998) yet are engaging to students.

To align these tasks with student abilities, the problem spaces contain many potential solution paths yet have a finite structure emphasizing the development of linkages among problem space variables. Thus, while the approach is constructivist, it is not so open ended and complex as to overwhelm users (Mayer, 1997). Most problems also have a single solution. While Hart (1994) and others have expressed concerns regarding problems with single solutions, our five-year experience shows no evidence from student performances, notes associated with problem solving, or survey data to suggest that students are viewing the problems sets through the small lens of "a single right answer". This is in part due to the broad perspective of a problem space created by IMMEX™, and by the large number of parallel case forms in each problem set. These instances, sometimes as many as sixty, enable students to study the domain from different viewpoints and gain extensive problem solving experience.

The IMMEX™ case development process is strongly teacher centered and is iterative in nature with multiple levels of feedback and refinement guiding development. The cases are first structured by addressing pedagogical and curricular needs, and learning goals. Classroom implementation and curricular integration schemes also shape the development of each problem so as to maximize student learning and to scaffold other teachers in their use of technology and case-based reasoning. The design process is similar to the construction of discipline models, learning models, and student models (Mislevy et Al., 1999), through which teachers and educators structure the problem space. The case topics emphasize relevant, real-world tasks (Los Angeles Times, 1997) that are appropriate for, and engage the majority of students. In parallel, this process also addresses infrastructure issues by providing training and materials to enhance the capacity for educational systems to use these tools.

A core IMMEX™ feature is the use of performance data for discovering the learning paths that students develop and use in a case-based reasoning environment. Teachers, students and researchers have access to continual and varied student performance information at different

levels of detail and abstraction. These range from simple efficiency and proficiency measures, such as percent solved, to sophisticated artificial neural network clustering. As expected from the format of IMMEX™, students and teachers perceive this system more as a tool for reasoning and integrating information than as a system for learning new facts (Hurst, et Al., 1998). Nevertheless, performance on IMMEX™ cases requires both content knowledge and reasoning ability. In fact, course grades correlate at an intermediate (r = .3 to .5) level with the percentage of IMMEX™ cases solved over the course of a year where between 50-100 cases were performed (Stevens, 1991). Predictive validity studies also suggest that the impact of IMMEX™ problem solving and course grades on Advanced Placement (AP) Chemistry scores is high. These two variables predict over 60% of the variance in AP Chemistry scores for girls and 42% for boys. Additionally, performance on IMMEX™ cases seems to contribute to the improvement of students' general biology problem-solving skills (Palacio-Cayetano et al., 1999; Palacio-Cayetano, 1997). Moreover, in contrast to many pencil and paper assessments, IMMEX™ performance seems less gender biased and recent results suggest that the IMMEX™ format may also be less tightly linked to student reading skills as measured by SAT 9 reading scores (Vendlinski, 2001).

While measures of student proficiency and efficiency are useful, there is a need to go further and document patterns of student scientific reasoning if the eventual aim is developing valid, reliable and efficient techniques for assessing the reasoning of individual students. IMMEX™ begins to accomplish this by generating maps of students' searches of the problem space (Stevens, 1991). In these maps, each student action is represented by a rectangle that is colored to visually relate items closely linked by content, concept or type (library resources, expert help, etc.). These icons are organized in different configurations and lines connect the sequences of items selected by the students while performing the case. In Figure 1 two search path maps are shown for an IMMEX™ problem in high school genetics called *True Roots*. The map on the left is an example of a very extensive search strategy and the one on the right, of a more refined strategy. Without knowing the details of the problem set, it is clear from the quantity and ordering of the test selections that these maps represent quite different strategies.

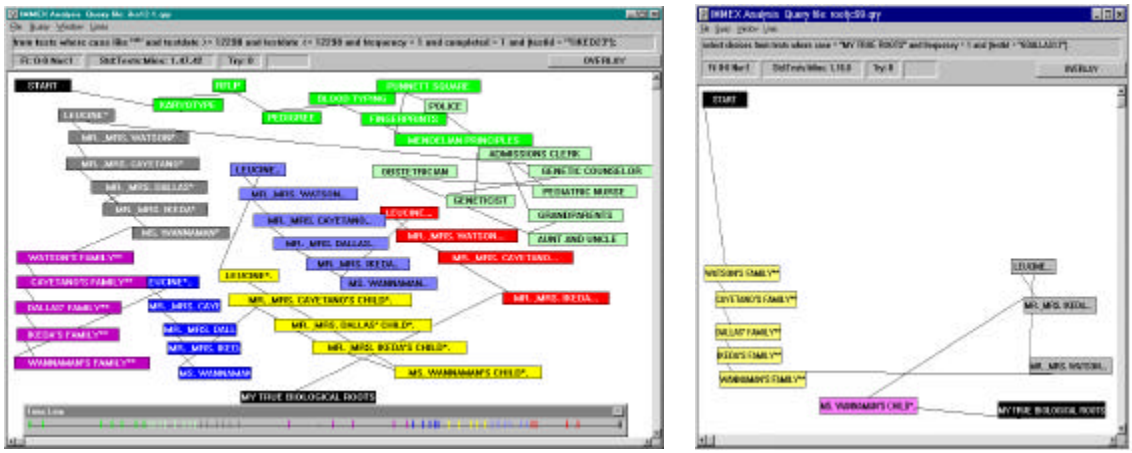


Figure 1. Sample IMMEX™ Search Path Maps.

Search path maps have both instructional and assessment potential and teachers use them in multiple ways. For teachers who are authors, these maps provide a validity check ensuring that the cases perform as intended. Also, by comparing earlier maps to later ones, a teacher can judge student progress through refinements of problem solving approaches over time as shown below. Other teachers provide these maps to students to encourage reflection and have them write essays

analyzing their own progress (Lawton, 1998). Search path maps are particularly important for examining and promoting the metacognitive aspects of problem solving such as persistence, elimination of alternative hypotheses, efficiency, confidence and certainty.

Similar maps can be constructed by the aggregation of multiple student performances and these maps begin to reveal performance patterns that repeat across different students, classrooms and teachers. Here the thickness of the lines is proportional to the number of students making a selection. An example of such a group map is shown in Figure 2 for one classroom.

When analyzing student IMMEX™ performances, the complexity of the task is immediately obvious. As the number of menu items in an IMMEX™ problem increases, so to does the way a student may move through the problem space. If one considers the order in which a student viewed different pieces of information, the number of paths through the problem space increases exponentially. Even if one were only to consider what information a student used when solving a problem, the number of strategies a student could use to solve the problem is a multiple of the number of information items available. Such a large number of possibilities make it almost impossible for humans to find the patterns that make student performances similar with any consistency. While the complexity of these maps is high, it is also apparent that 1) all sequence connections are not used equally, and 2) there are distinct patterns to the way that students approach and solve these cases.

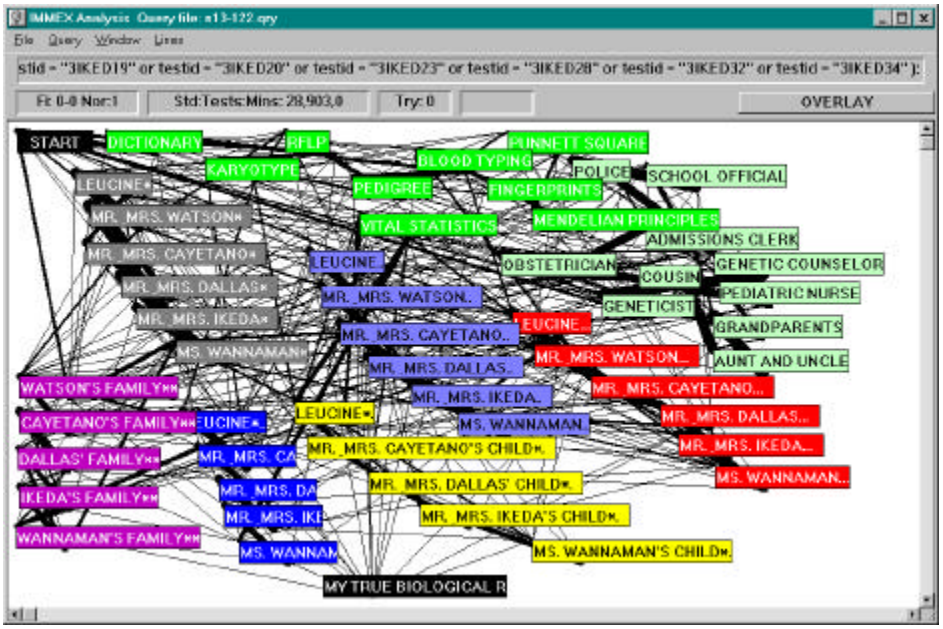


Figure 2. A Sample Group Search Path Map.

Currently, hundreds of IMMEX™ problem-solving simulations are being created in many disciplines (science, math, history, etc.) by teacher/expert teams and are being used in local schools. Consequently, a major challenge is to develop valid and reliable performance classifications of student problem-solving strategies to facilitate their greater use for instruction and assessment by teachers in the classroom.

We have used two approaches to generate these classifications: teacher and educator defined rubrics; and unsupervised neural network classifications. Each method provides a different view of contextual information and granularity and can be productively combined. When teams of teachers, educators, content experts and case authors develop scoring rubrics, performance

guidelines are first established and agreed upon for the case and representative student work. These rubrics are then used to score large numbers of search path maps and the rubrics are refined for performances where the inter-rater reliability is low. While more time intensive than artificial neural network analysis, these rubrics contain important person-specific and context-specific elements. Sample rubrics can be found at http://www.immex.ucla.edu/K-12RubricMainFrame.htm.

We also use unsupervised artificial neural networks to cluster strategies based on the sequence of student actions during problem solving. This is a powerful classification approach that we have broadly used (K-12 to medicine) to derive valid and interesting performance models inductively from the sequence of actions chosen on the task, even when no a priori model exists (Stevens and Najafi, 1993; Casillas, et Al., 2000). In this manuscript and presentation we review our methodology, address validity and reliability issues and review significant findings from two different tasks: high school science; and medical licensure.

Tasks And Problem Spaces

To examine how broadly neural network analysis can be applied to assessment situations we have utilized both low stakes, relatively simple problem-solving datasets (IMMEX™), as well as highly complex high stakes datasets like the National Board of Medical Examiners' (NBME) Clinical Case Scenarios (CCS). (Stevens, et Al., 1996; Casillas, et Al., 2000)

Typical IMMEX™ cases contain between 40 and 120 different information items that students can access when solving IMMEX™ problems. For example, a high school genetics problem called *True Roots* contains 52 such menu items. On the other hand, current implementations of the CCS simulations encompass a much larger and more detailed problem space of over 2500 resources or action items. The number of items actually viewed by students prior to completing the case can be as few as 1 (guessing) to virtually every item in IMMEX™; whereas for the CCS simulations, the number of selections has been as few as 60 and as large as 275. With both simulations however, the number of unique selections made are often a restricted subset of the entire problem space. For instance, although the two current CCS cases under evaluation contain a set of 6683 student performances, only 836 unique tests of the 2500 possible tests/actions were selected.

Data Collection and Processing

As with all assessment techniques, the data collected should adequately reflect the complexity of the task, be informative, be practical (another production curve), and produce results that are easily interpreted and predictive of future performance. The analysis of search path maps from hundreds of different simulations has revealed patterns of sequential test usage that have produced such results. The important variables in our analysis include 1) the items selected, 2) the order of selection and 3) the latencies between selections. Given the large number of tests and the relatively limited selection from the entire problem space, a matrix of sequential steps becomes sparse and large as illustrated in the Figure 3. In order to compress spatial information into a more compact dimension we encode the sequence of steps taken into a hexadecimal value represented by a color. The resulting pattern then depicts test items encoded by a color that is then related to the order in which the item was taken. This compresses the 2-D performance structure into a linear, color-encoded structure that captures both test item and sequence (Figure 3).

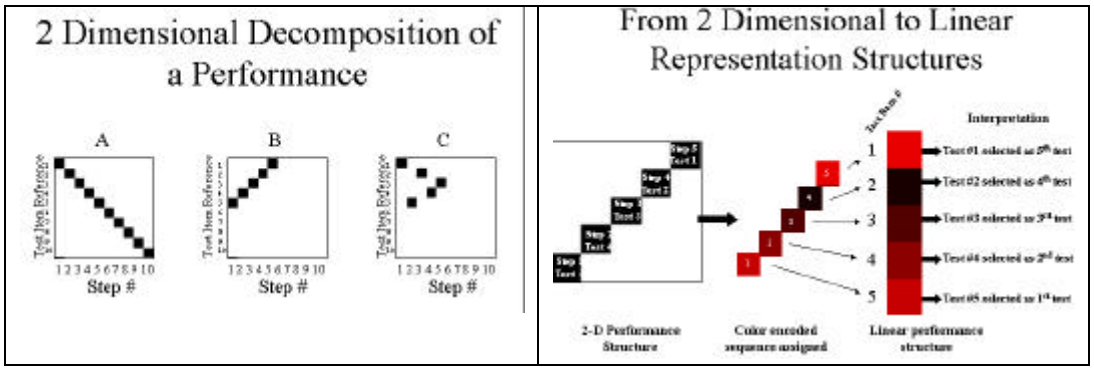


Figure 3. Generating a Linear Dataset From A Sparse Matrix. A performance can be represented visually in two dimensions by decomposing its structure into the specific items that compose it and the step at which each one was taken. For test items 1 through 10 taken in sequence (i.e. Test Item 1 taken first, Test Item 2 second, and so on) the graphical representation shown in A would result. In performance B, Test Item 5 is taken first, followed by Test Item 4, etc. results in a distinct representation. In performance C, the sequence of Test Items is 1, 5, 2, 4, 3. In each case, the two dimensional space needed to represent the performance in 10 by 10, and the majority of the area (90 to 95%) is "white space." In Figure 3b, this data has been color encoded into a linear structure.

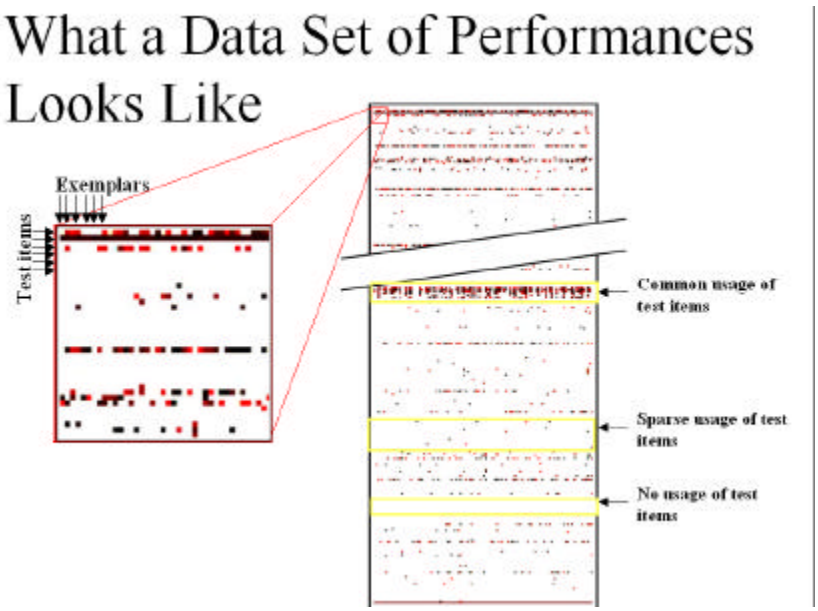


Figure 4. Visualizing A Preprocessed Data Set. An input set of data consists of a series of performances or exemplars that are displayed in columns. The test items that were selected in each performance are represented as a pixel or colored square in a position on the y-axis that corresponds to a unique value for that particular test item. The color of the pixel is determined by the order in the performance that that item was taken. This data set serves as the input for the artificial neural network and is read as 1xn where n is the total number of unique test items in all performances of the database from which the exemplars are taken. Thus, novel performances can be used in new input data sets since test items not represented in one set can be represented in subsequent data sets.

The resulting dataset of performances consists of the colored coded sequence information aligned with the different test items available and can be visualized, much like a compressed version of a search path map. Test items that are used by large numbers of students on problem sets appear as

a series of highlighted bands, and the colors within the bands provide an indication of sequence heterogeneity. Again, for all the problem sets examined, we observe specific patterns of test item usage (Figure 4).

Artificial Neural Network Analysis

The consistent patterns of test item usage across a wide range of problem solving scenarios suggest that unsupervised artificial neural networks would be a productive approach for constructing categories of strategy types for each problem set that could be used for research and reporting. Our studies have repeatedly confirmed this hypothesis on multiple tasks. Although neural network modeling makes it difficult to understand how the classification model works, we are more interested in inductively determining strategy types (i.e. discovering structure) and changes in strategy types, than we are in deriving and understanding a particular classification model at this time.

Nevertheless, there are important architecture and training considerations when classifying with unsupervised neural networks. These include training times, numbers of exemplars, implementation of conscience, momentum, etc. One of the most critical parameters is the number of neural nodes, as this limits the maximum number of different strategy clusters that one can obtain. While we often determine this value empirically for each problem set, it generally scales with the complexity of the simulations. For instance, IMMEX™ simulations often map to 25 strategy clusters (nodes), while CCS simulations appear better mapped to 100. Keeping the network unconstrained to avoid forcing different performance strategies together in a single node is an important guideline. This can be determined in pilot studies by verifying the existence of nodes containing few, if any, exemplars. A second important observation is that while there are a relatively few (comparatively speaking) ways to solve a problem, there are many more ways to miss one. Therefore, if the research goal is to study successful problem solving, fewer nodes may be needed to account for failure.

The reliability of the clustering is established by training parallel (between three and five) neural networks with the preprocessed datasets and generating estimates of the degree of co-clustering of cluster members across the different networks. For relatively simple IMMEX™ problems such as *True Roots*, the co-clustering of members across three independent neural networks approaches 99%. For more complex problem sets like the chemistry problems *Hazmat* and *Desperately Seeking Solution* where there are cases of different complexity and difficulty, the co-clustering efficiency was lower (75-90%), but could be improved to the 95% + level through a combination of unsupervised and supervised neural network training (Vendlinski and Stevens, 2000). Lastly, with highly complex cases like the CCS dataset there is a higher frequency of performances that fail to consistently co-cluster (Figure 5).

Comparison of Performance Class Across 3 Networks

Classification	N = 100	Network Co-clustering
Class 1	28 %	1 = 2 = 3
Class 2a	42 %	1 = 2
Class 2b	6 %	2 – 3
Class 2c	6 %	1 = 3
Class 3	18 %	No co-clustering

Figure 5. Inter-network Reliability of Clustering for the CCS Performance Data. Class designation for the 100 exemplars used in this series is shown in tabular form. Numbers in the co-clustering column refer to each of the three networks used to analyze these performances.

The overall methodology of our approach is summarized in Figure 6.

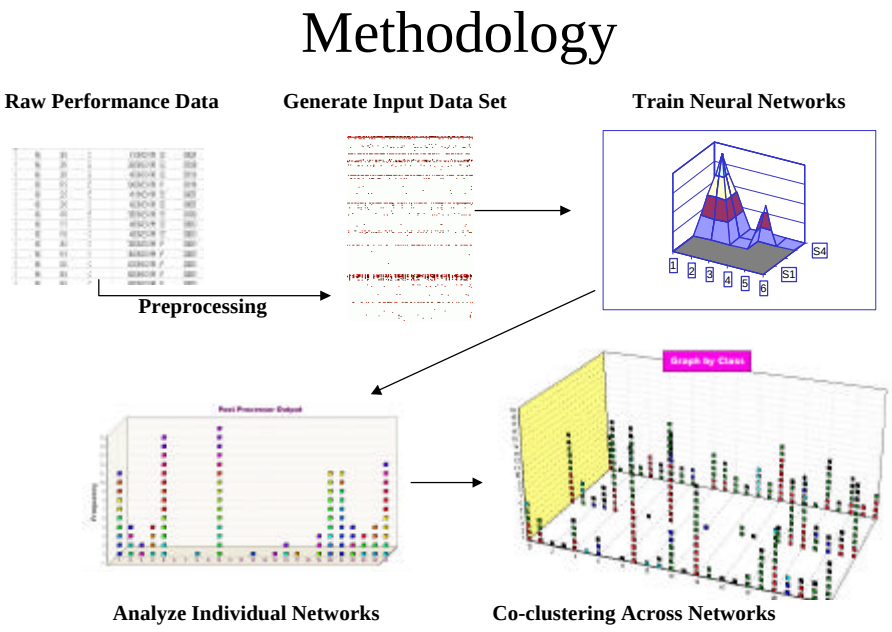


Figure 6. Neural Network Methodologies. Raw performance data provided by IMMEX™ or the NBME is parsed to obtain each performance's test items in sequence. The data is preprocessed to produce a digital input used to train unsupervised artificial neural networks. In the current analysis, 100 exemplars from one case were used to train three networks. Each network was trained with 300 epochs and output projected onto a 5x5 space. The clusters assigned by the ANN were then reported on a 2-D graph of frequency at each node. The performances at each of the three networks were then analyzed in order to assign a class of performance based on co-clustering across the networks. Performances that always co-clustered were assigned Class 1, performances that never co-clustered were assigned Class 3, and performances that co-clustered on at least two networks were assigned Class 2.

Validation, Reliability and the Information Value of The Neural Network Performance Model

One of the strengths of automatic cluster detection with neural networks is that it is an undirected knowledge discovery technique that can be used to inductively derive performance models. Validation of the resulting model however, becomes more important than for tasks where a careful task analysis during the task construction has limited the number of variables to be tracked (Mislevy, et Al, 1999). We accomplish this by performing search path map analysis on each cluster. In the following analysis, for example, 1625 high school student performances of the five cases of the *True Roots* problem set were clustered by unsupervised learning on a 25 node neural network. The student population included regular, honors and AP students from 25 different teachers. *True Roots* is a relatively easy case series for the students with an average solution frequency of 77%. The clustering of student performances at each neural net node, along with the solution frequency is shown in Figure 7. While the performances are widely dispersed across the network, there are several nodes with few, if any, performances indicating that the network was not constrained.

NODES	# Solved	#Unsolved	
1	2	0	100%
3	2	0	100%
4	13	0	100%
11	1	0	100%
12	2	0	100%
17	33	1	97%
14	38	2	95%
19	29	2	94%
23	43	3	93%
7	87	11	89%
2	22	3	88%
10	35	5	88%
25	49	7	88%
9	62	9	87%
24	96	17	85%
16	26	6	81%
15	84	20	81%
22	49	12	80%
8	63	16	80%
6	42	11	79%
21	27	9	75%
5	44	15	75%
13	91	37	71%
20	18	9	67%
18	289	183	61%

Figure 7. Distribution of Performances and Solution Frequency at Each Neurode

We next performed search path map analysis of each of the 25 nodes to document the basis of the clustering. Representative examples are shown in Figure 8, and the entire 25 node output can be found at http://www.immex.ucla.edu. The six search path maps illustrate three different general strategy types identified following the clustering. Some of the clusters, such as Nodes 7 and 15 (top) represent strategies where students for the most part chose a single data group to sample, Node 7 emphasizing DNA Typing (blue) and Node 15 (red) focusing on Blood Typing. Other

nodes of this strategy type focused on Fingerprinting or Pedigree analysis. This strategy was termed *limited* in that, for the most, part the students did not have sufficient information to solve the case (i.e. premature closure) and generally had lower solution frequencies. Nodes 9 and 23 in the middle group of Figure 8 represent a second strategy group. This group shows evidence of integrated search, and data reduction and elimination across different test groups. Multiple clusters of this strategy type existed with different compositions of test selections, and this integrative approach was often associated with higher solution frequencies. The next clusters, Nodes 18 and 8 illustrate overt guessing and extensive undirected search, respectively. These strategies, and nodes with similar lean or prolific strategies, always contained the lowest solution frequencies.

These analyses helped validate the unsupervised neural network clustering in several ways. First it provided a library of strategy types that can be expected when large numbers of student perform the *True Roots* problem set. Scoring sessions to construct/refine scoring rubrics being developed, in turn, can incorporate this real-world information. Second, these maps help relate problem solving success to different strategy styles suggesting opportunities for student intervention. Third, they provide a framework for following student progress by tracking node transitions as students perform multiple clones of a problem set.

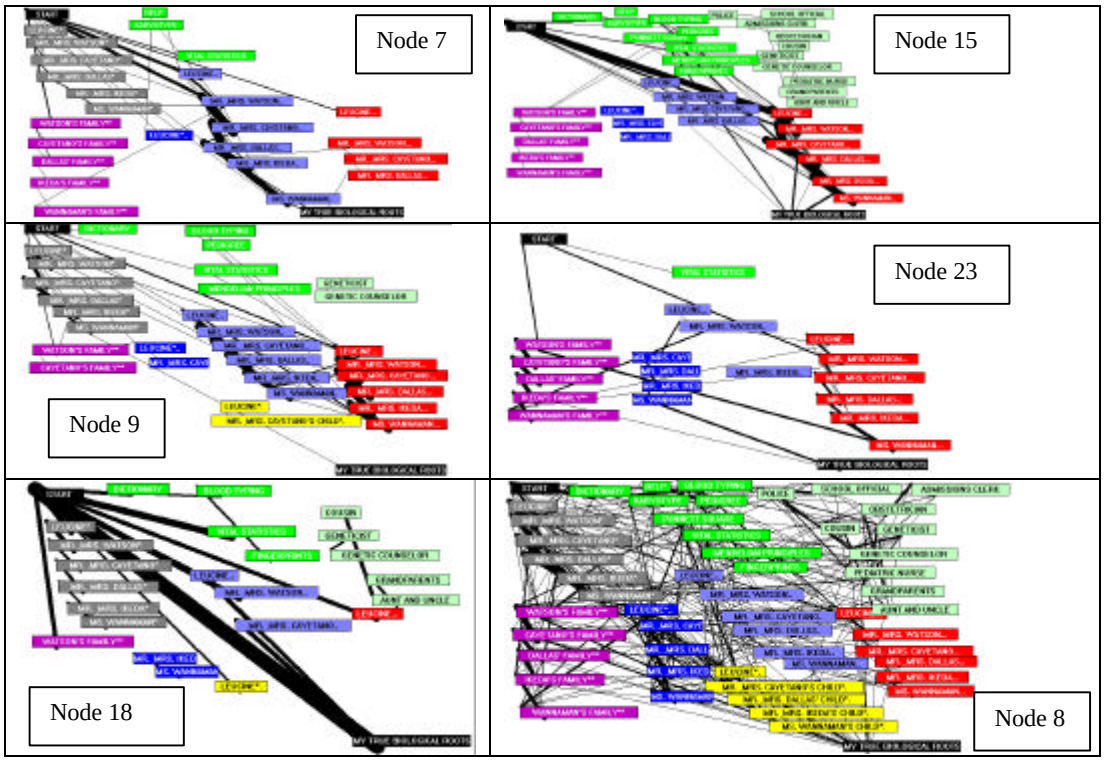


Figure 8. Search Path Map Analysis of Student Performances At Different Neural Network Clusters. Following the clustering of student performances by unsupervised learning, search path maps were generated for each node. In each of the figures the lines have been normalized by 2X to highlight the dominant patterns.

Without Intervention, Students' Use Of Strategy Types Changes Slowly

Examination of thousands of student performances in many domains and across multiple grade-levels of education (elementary through medical school) has indicated that many search

strategies/approaches are applied while solving IMMEX™ problems. Students differentially use these strategies, often unconsciously (Reder & Schunn 1996), and through this use, they can develop higher order strategies. In following these strategies over time however, it has become apparent that, for each problem-set, there are transitional states that students may or may not pass through as they develop expertise. Neural network data analysis can be used to follow how students arrive at, and depart from these states enabling researchers to use this information in a predictive fashion. These results are derived from the work of Vendlinski (2001) where the analysis was conducted with a high school, analytical chemistry IMMEX™ problem set called *Hazmat*.

In these studies a combination of unsupervised and supervised artificial neural networks was used to identify 25 discrete solution strategies used by the students to solve this problem. As was true in *True Roots*, the presence of a node with no performances suggested that the network was not constrained and, therefore, was able to cluster all unique strategies. The clusters produced provided important insights into student ability to apply qualitative chemistry concepts to a real world situation.

Unlike the strategies students used in *True Roots*, the various *Hazmat* strategies yielded diverse percentage solve rates. In spite of this diversity however, the strategies and the solution rate yielded by a particular strategy was similar regardless of student ability (college undergraduate, AP chemistry or first year high school), teacher, or the time of year the problem was seen. More importantly, as in *True Roots*, inspection of the strategies suggested three overarching categories that described the type of strategy used. As seen in the figure below, students who chose very few items of information before attempting to solve the problem (i.e. "limited" strategies) were unlikely to solve the problem. Similarly, students using strategies that investigated large amounts of information (i.e. "prolific" strategies) were also unlikely to solve the problem correctly.

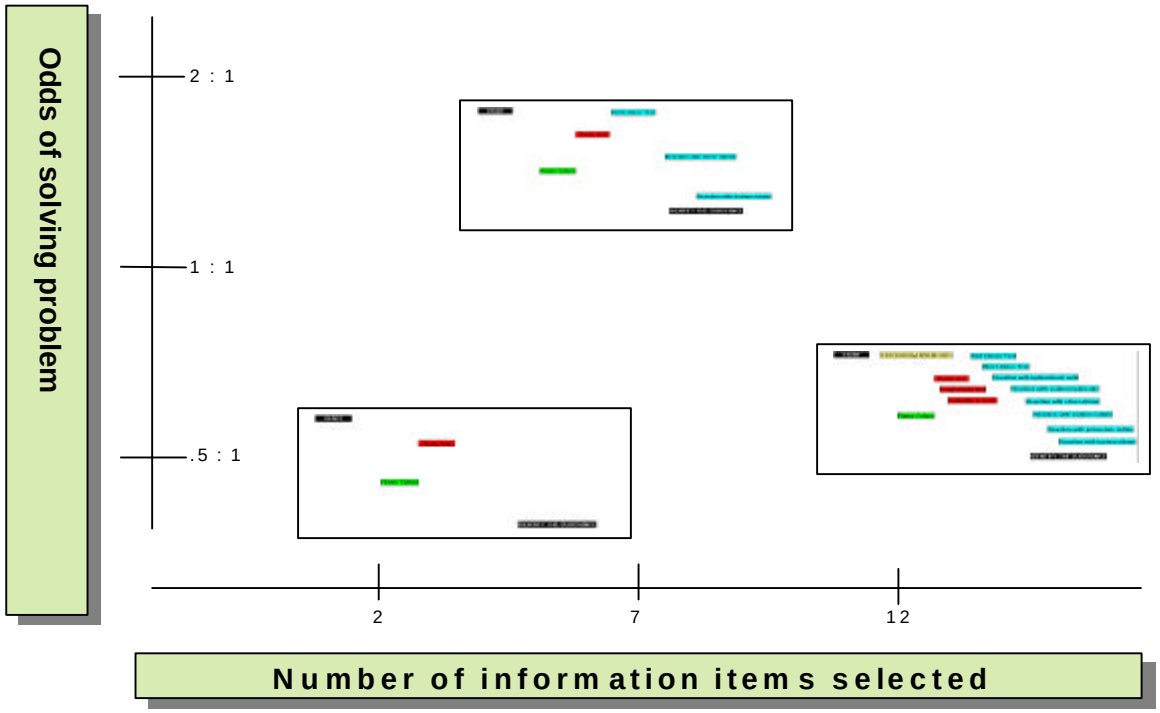


Figure 9. This figure illustrates one method to evaluate various *Hazmat* solution strategies.

Finally, student strategies that were characterized by the use of information that was germane to the problem solution (i.e. "efficient" strategies) had better than even odds of correctly solving the problem. Hurst, et al. identify similar strategy types in the problem solving strategies of medical students (Hurst, et Al. 1998).

The student performances were then ordered and plotted so transitions between categories could be seen. As shown in Figure 10, most students who begin with a limited strategy, remain there despite little success at solving problems. These students appear unlikely to improve their problem-solving abilities without intervention. While a number of students who begin with prolific strategies do transit to a more effective strategy on their own, most seem to fall back into using more profuse strategies in subsequent attempts to solve IMMEX™ cases. These students might also profit from problem-solving interventions. Lastly, most students who begin solving problems efficiently continue to use similar strategies in subsequent cases.

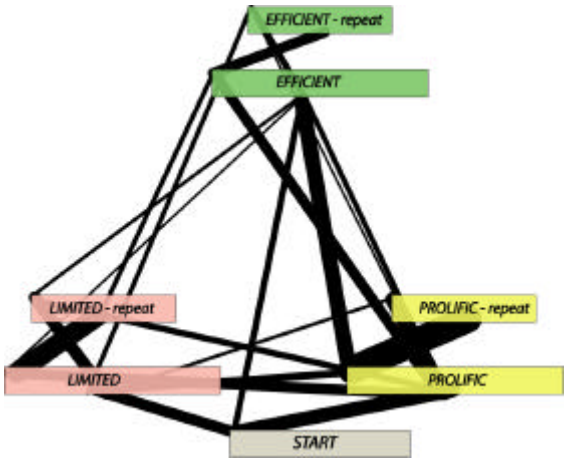


Figure 10. This figure shows the probabilistic graph of transitions students make between various strategy types over time when they solve *Hazmat* problems. The type of strategy a student used to solve his or her first problem is represented by the line from start to the appropriate rectangle. Subsequent student performance transitions are represented by lines from the upper left corner of the strategy type a student move from to the center of the strategy type rectangle the student moved to. The figure represents students who use the same type of strategy repeatedly, as a line between a strategy type and its adjacent "- repeat" rectangle.

These trends are even more apparent when the first strategy type chosen by a student is

	"Limited" strategy on most problems	"Efficient" strategy on most problems	"Prolific strategy on most problems
"Limited" strategy on first problem	30	8	6
"Efficient" strategy on first problem	5	35	8
"Prolific" strategy on first problem	5	9	23

Figure 11. This table compares the category of the strategy a student used to solve his or her first qualitative Chemistry IMMEX™ problem to the category of the strategy a student used most often to solve different versions of the same problem. There is significant statistical evidence ($\chi^2 = 70.5$; d.f. = 4; $p < .001$) to conclude that, without teacher intervention, the number of students who repeatedly use the same strategy to solve problems is not random.

juxtaposed against the most frequent strategy type the student used to solve problems. As shown in Figure 11, the first strategy type is significantly correlated with the most frequent strategy type chosen by the student.

Using the approach described here, we can not only identify students who successfully make the transition from less successful to more successful strategies (and perhaps why), but also the probabilities that students will make such transitions without teacher intervention. When interventions are attempted, we can use this information to test the effectiveness of those interventions and individual student improvement. Such tools also provide important feedback to educators and allow them to document the overall effects of their teaching approaches and methods.

The power of this approach is that probabilities can now be derived for students at each strategy level and interventions can be tested for their effect on changing these probabilities where appropriate. These diagrams also provide a means for documenting effects of overall classroom teaching approaches and methods.

To determine how simulated clinical performance classifications related to ratings that were derived independently by the NBME, a training set of 100 randomly selected performances on an emergency infectious disease case was used to train an unsupervised 10x10 output node network. The training data was then analyzed by the same network to identify where the performances clustered on the neural network (Fig. 12a). We observed major clusters of student performances at nodes 10 (n=11), 25 (n=8), 29 (n=9), 43 (n=5), and 44 (n=7). Since the NBME had previously rated these performances from 1 (worst) to 8 (best), it was possible to determine if any nodes represented clusters with specific ratings. Nodes 43 and 44 were composed exclusively of highly rated performances (rating ≥ 7) (Fig. 12b), while low ratings (i.e. <4) showed limited clustering (Fig. 12c), of a few performances per node. In fact, lack of clustering by the ANN was associated with poorly rated performances.

To better explore the complexity of clustering we observed and to further understand the student strategies assigned to these clusters, representative performances were reconstructed as search path maps. Analysis of test usage revealed that the degree to which a student searched within a limited set of clinical studies closely related to the type of infection associated with the overall rating (Fig. 13). Performance ratings of 7 and 8 (i.e. at nodes 43 and 44) were uniformly associated with the selection of a series of many diagnostic

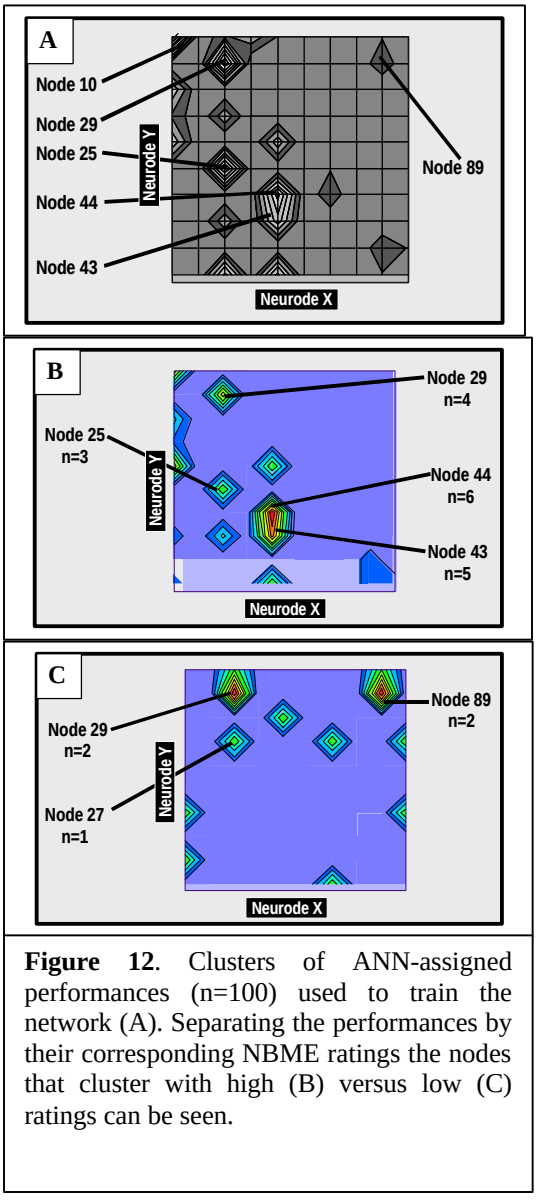


Figure 12. Clusters of ANN-assigned performances (n=100) used to train the network (A). Separating the performances by their corresponding NBME ratings the nodes that cluster with high (B) versus low (C) ratings can be seen.

studies associated with the site of infection (Fig. 13a) while performance scores below 3 (i.e. node 89) were associated with minimal or no test ordering in that domain (Fig. 13b). As expected, the lowest ratings showed no usage in the critical test domain suggesting a failure to recognize the

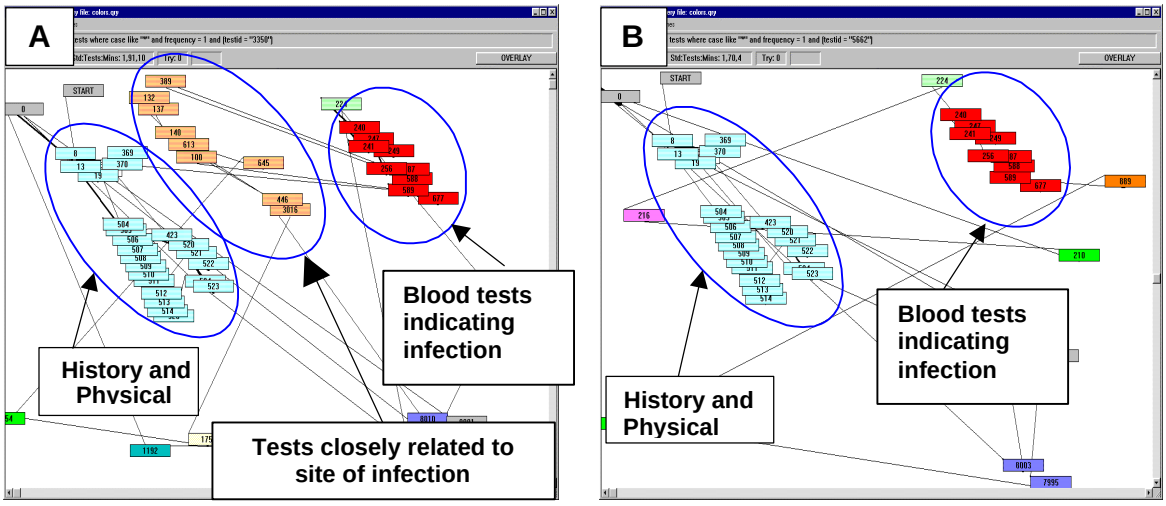


Figure 13. The limited Search Path Maps for performances at Node 43/44 and Node 89. Search path maps showing a specific group of tests crucial to solving the case are indicated as "tests closely related to the site of infection." Performances were selected based on where they were clustered by the ANN: those at node 43/44 (A) also had high NBME-assigned ratings (7 or 8) and those at node 89 were rated as failures (B). Note the lack of specific usage of "tests closely related to the site of infection" in (B).

exact source and cause of the infection. While the neural network consistently clustered some performances in agreement with the NBME-assigned ratings, other nodes contained a range of performance ratings. This finding was characteristic of the cluster at node 29. The search path map analysis for this node (not shown) indicated the use of excessive test item selections throughout several domains.

The ANN-based analysis of the CCS data set indicates that a high degree of complexity, adequately represented, can produce meaningful clustering of the input data. A degree of validity of the clustering is offered through the independent measure provided by the NBME raters and the associated strategy analysis in the search path maps showing the areas that the ANN was especially sensitive to in output designation.

References

Barrett, G.V., & Depinet, R.L. (1991). A Reconsideration of Testing for Competence Rather than for Intelligence. American Psychologist, 46: 1012-1024.

Berry, D.C., and Broadbent, D.E. (1988). Interactive Tasks and the Implicit-Explicit Distinction. British Journal of Psychology, 79: 251-272.

Casillas, A.M., Clyman, S.G., Fan, Y.V., and Stevens, R.H. (2000). Exploring Alternative Models of Complex Patient Management with Artificial Neural Networks. Advances in Health Sciences Education 5: 23-41.

de Neufville, R. (1990) Applied Systems Analysis: Engineering Planning and Technology Management. McGraw-Hill, NY.

Elstein, A.S. (1993). Beyond Multiple-Choice Questions and Essays: The Need For a New Way to Assess Clinical Competence. Academic Medicine 68: 244-249.

Fischer, K. W. and Bidell, T. R. (1997) Dynamic development of psychological structures in action and thought. The Handbook of child psychology: theoretical models of human development. Lerner. New York, Wiley. **1**:467-561.

Hart, D. (1994) Authentic Assessment. Addison Wesley, NY.

Harp,S.F., and Mayer, R.E. (1998). How seductive details do their damage: A theory of cognitive in science learning. Journal of Education Psychology, 90: 414-434.
Hudson, J., Fivush, R., and Kuebli, J. (1992). Scripts and Episodes: The Development of Event Memory. Applied Cognitive Psychology, 6: 625-636.

Hurst, K., Casillas, A., and Stevens, R. (1998) Exploring the dynamics of complex problem solving with artificial neural network-based assessment systems. CSE Technical Report No. 387. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Johnson-Laird, P.N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Kolodner, J. (1993). *Case-based Reasoning.* San Mateo, Calif.: Morgan Kaufman.

Kolodner, J.L., (1997). Educational Implications of Analogy. A View from Case-based Reasoning. American Psychologist, 52: 57-66.

Lawton, M., (1998) "Making the Most of Assessments," *Education Week, Technology Counts '98: 59*-62
Los Angeles Times (1997). Parents, Baby Reunited After Hospital Mix-up. October 8. P B4.

Maker, C. J. (1994) "Authentic Assessment of Problem Solving and Giftedness in Secondary School Students." *Journal of Secondary Gifted Education* (Fall 1994), 19-29.

Mann, R. and Jepson, A. (1993). Non-accidental features in learning. AAAI Fall Symposium on Machine Learning in Vision. Raleigh N.C.

Mayer, R.E. (1997). Multimedia learning: Are we asking the right questions? Educational Psychologist, 32: 1-19.

Messick, S. (1989). Validity, In R. Linn (ed.) *Educational Measurement* (pp. 13-103). Macmillian, NY.

Mislevy, R., Steinberg, L., Almond, R. (1999). On the Roles of Task Model Variables in Assessment Design. CSE Technical Report (500). Los Angeles, California. University of California, Center of Research on Evaluation, Standards and Student Testing.

Palacio-Cayetano, J. (1997) Problem Solving Skills in High School biology: the effectiveness of the IMMEX™ problem solving assessment software. Dissertation, University of Southern California.

Palacio-Cayetano, J., Allen, R., and Stevens R. (1999) Computer-assisted Evaluation – The Next Generation. *The American Biology Teacher* 61(7), 514 - 522.

Reder, L., and Schunn, C. (1996). Metacognition Does Not Imply Awareness: Strategy Choice Is Governed by Implicit Learning and Memory. In Reder, L.M. *Implicit Memory and Metacognition*., Lawrence Erlbaum, Mahwah, NJ.

Schank, R., (2000). Dynamic Memory Revisted. Cambridge University Press. NY.

Schwartz, J.L. (1989). Intellectual Mirrors: A Step in the Direction of Making Schools Knowledge-making Places. Harvard Educational Review 59: 1.

Stevens, R.H. (1991). "Search Path Mapping: A versatile approach for visualizing problem-solving behavior," *Acad. Med.* 66(9): S72-S75.

Stevens, R.H., and Najafi K. (1993). Artificial Neural Networks as Adjuncts for Assessing Medical Students' Problem-Solving Performances on Computer-Based Simulations. *Computers and Biomedical Research* 26(2), 172-187.

Stevens, R., Wang, P., Lopo, A. (1996). Artificial Neural Networks Can Distinguish Novice and Expert Strategies During Complex Problem-Solving. *JAMIA* vol. 3 Number 2 p 131-138.

Underdahl, J., Palacio-Cayetano, J., and Stevens, R. (2001) Practice Makes Perfect: Assessing and Enhancing Knowledge and Probelem-solving Skills with IMMEX Software. *Learning & Leading with Technology* 28(7), 26-30 & 98-100.

Vendlinski, T. & Stevens R.H. (2000). The Use of Artificial Neural Nets (ANN) to Help Evaluate Student Problem-Solving Strategies, In B. Fishman & S. O'Connor-Divelbiss (Eds.), *Proceedings of the Fourth International Conference of the Learning Sciences* (pp. 108-114). Mahwah, NJ: Erlbaum.

Vendlinski, T. (2001) Affecting U.S. education through assessment: new tools to discover student understanding. Dissertation, Massachusetts Institute of Technology.

Wiggins, G. (1993). *Assessing Student Performance: Exploring the Purpose and Limits of Testing. San Francisco*, Jossey-Bass.

Wolf, D. (1993). Assessments as an Episode of Learning, In R. Bennett & W. Ward (Eds.), *Construction vs. Choice in Cognitive Measurement* (pp. 213-240). Mahwah, NJ: Erlbaum.