Exploring the Dynamics of Complex Problem-Solving With Artificial Neural Network-Based Assessment Systems

CSE Technical Report 444

Karen Hurst and Adrian Casillas
UCLA School of Medicine

Ronald H. Stevens
UCLA School of Medicine and
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
University of California, Los Angeles

September 1997

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90024-6511
(310) 206-1532

This work was supported by a Robert Wood Johnson Foundation grant (#23223) to A. M. Casillas and a National Science Foundation grant (NSF # ESI-9453918) to R. H. Stevens. The findings and opinions expressed in this report do not necessarily reflect the policies or opinions of the Robert Wood Johnson Foundation or the National Science Foundation.

Publication of this technical report was supported under the Educational Research and Development Center Program, cooperative agreement number R117G10027 and CFDA catalog number 84.117G, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

EXPLORING THE DYNAMICS OF COMPLEX PROBLEM-SOLVING WITH ARTIFICIAL NEURAL NETWORK-BASED ASSESSMENT SYSTEMS[1]

Karen C. Hurst[2,3], Adrian M. Casillas[3,4], and Ronald H. Stevens[3,5]

Abstract

Assessment of cognitive models developed by students in complex scientific disciplines ideally captures the progressive and dynamic nature of learning. We have developed a computer-based performance assessment system based on the production-system model of problem solving (Newell & Simon, 1972) in order to explore assessment of medical student problem-solving skills. We are able to track student data acquisition within groups of related concepts (concept domains) through a process called "search path mapping," which reveals differences between successfully executed performances and unsuccessful strategies. We were able to identify both changes in the overall strategic approach (i.e., selecting the correct concept domain) and refinements in problem-solving strategies (i.e., more efficient use of concepts within a domain).

While search path mapping allows for the assessment of student strategies, it is also time-consuming and potentially subject to reviewer bias. We therefore automated the search path map analysis by using supervised back-propagation artificial neural networks (ANNs), trained with data from previous medical student performances. These artificial neural networks could discriminate strategy improvement as well as hypothesis utilization in problem solving. To provide an indicator of learning, we compared the number of hypotheses utilized by students on practice and examination problems in the same concept domain. We found that medical students used information more efficiently during the examination, resulting in fewer hypotheses generated. Although the overall utility of ANN-based analysis was seen, this supervised network did not identify certain specific cases determined by search path mapping to be more "expert" than the training data set. Our ANN also misinterpreted rare cases of irrelevant concept usage. These studies demonstrate the utility of ANN-based assessments as an adjunct to traditional forms of assessment and suggest additional studies needed for this approach to become part of a comprehensive evaluation system.

---

1 The authors would like to thank Dr. Robert Allen for his critical review and Stanley Chen, John Stallings, and Peter Wang for their expert technical assistance.

2 Molecular Biology Institute, UCLA School of Medicine

3 Department of Microbiology and Immunology, UCLA School of Medicine

4 Department of Medicine, Division of Clinical Immunology and Allergy, UCLA School of Medicine

5 Graduate School of Education and Information Studies/National Center for Research on Evaluation, Standards, and Student Testing CRESST, University of California, Los Angeles

Introduction

In the development of science education technology, authenticity and validity of content must be assessed in conjunction with cognitive models developed by students. Ideally, such models are dynamic, continuously combining information and understanding into a framework where a hypothesis can be formulated, tested, and modified (Alexander & Judy, 1988; Groen & Patel, 1988; Peverly, 1991). Traditional studies of students' mental models, such as protocol analysis, have provided fundamental information about the nature and structure of mental models in novices and experts (Simon, 1995). However, this information has had relatively little impact on everyday instruction and evaluation in the classroom. This may be due to the lack of timely, informative feedback provided to teachers based on these in-depth cognitive research studies.

We have been exploring the use of computer-based assessment technologies in order to bridge this gap, and to provide a dynamic and generalizable tool for large numbers of students across various domains. These technologies provide the opportunity for assessment to be consistent with learning as a continuous process. The understanding of isolated concepts does not comprise comprehensive domain knowledge, yet it is the basis for successful problem solving. Concepts and principles are linked to form structures called "concept domains" or "knowledge structures" (Glaser 1984, 1990, 1992). Concept domains can be linked into larger and often overlapping groups in disciplines such as immunology. Proficiency in a specific discipline results from a mental representation or model of the field containing not only information, but also the skills necessary to organize and utilize this information. Relevant usage of concept domains reflects a cognitive ability that allows adaptation and generalization of knowledge in solving unfamiliar problems (Brown, Bransford, Ferrara, & Campione, 1983; Chi & Glaser, 1984), leading to what is known as "expertise." Reasoning and appropriate usage of information within concept domains allows for the formulation and modification of hypotheses during thoughtful problem solving, identification of analogous problems, and self-monitoring during testing of the hypothesis (Campione, 1990; Sugrue, 1994).

Diagnostically useful assessments should accurately identify the spectrum of student performances. Several levels of student performance in problem solving can demonstrate the varying degrees in understanding of experimental data and

underlying principles (reviewed in Sugrue, 1994). The lowest level of student performance lacks concept-unifying principles, which is seen as disorganized searching of data. The next level shows an understanding of basic concepts, but without an organized reasoning process, such as hypothesis generation. Other students form hypotheses that may be inappropriate to the presented problem, where hypotheses are formed in irrelevant domains, or where experimental data are incorrectly interpreted despite a correct hypothesis. Finally, students with an effective problem-solving strategy show both an understanding of concept-unifying principles and the ability to focus this information by making an appropriate hypothesis. Ultimately, a good problem solver equipped with information from a specific concept domain should be able to formulate and verify a hypothesis within that concept domain.

Prior to pursuing an analysis of student problem-solving skills, the problem itself should be carefully assessed. The content must be relevant to the concept domain (Millman & Greene, 1989) and be designed at an appropriate level for the intended audience (Sugrue, 1994). The test items, at a minimum, should contain sufficient information to solve the problem. The problem structure must be flexible enough to allow application of forward reasoning (i.e., planning and executing a strategy) and domain-specific activity (i.e., hypothesis formation) to distinguish stronger, "more expert" performances from weaker, "less expert" performances (Glaser, Raghavan, & Baxter, 1992).

Once the learning task or problem is established, how is student performance measured? A sequence of individual test item selections, often based on a mental model, can develop into a coherent strategy (Simon, 1995). Furthermore, the use of certain key concepts may be used to ascertain pivotal areas of understanding. Although the selection of informative concepts cannot be used as the sole parameter in performance assessment, it may offer insight into specific search strategies. Whether the student solves or does not solve the problem is useful as an outcome measure, but this also fails to give insight into the problem-solving process. Likewise, a student's score becomes a useful assessment tool only when used in the context of a strategic approach. Examination of hypothesis generation through strategic choice of domain-specific test items is a reliable way to identify student performance on a specific problem, or development of general problem-solving skills (Sugrue, 1994). A truly dynamic performance standard accommodates the process of transition across levels of student understanding.

This becomes the challenge for developing computer-based learning assessments. Standards that truly allow the assessment of transition from novice to expert and that incorporate other assessment parameters would enrich the performance standard setting. For this reason, we have explored using artificial neural networks (ANNs) to identify patterns in student problem-solving performance.

Artificial neural networks allow for rapid and continual assessment in areas that are categorically ill-defined, or where performance patterns are hidden within the data (Reggie, 1993; Weinstein et al., 1992). Previous studies aimed at recognizing and understanding expert strategies in medical student problem-solving performances have revealed the utility of ANNs (Stevens & Lopo, 1994). Supervised back-propagation ANNs trained with medical student performances on immunology or infectious disease problems correctly identified the problem-solving outcomes of other students in over 85% of cases (Stevens & Najafi, 1993). These studies did not, however, explore the use of ANNs (a) for documenting the dynamics of students' hypothesis formation and rejection within a problem, or (b) for measuring student progress. This report specifically addresses these two issues.

Methods

Problem Design and Implementation

We have developed a computer-based performance assessment system based on the production-system model of problem solving and have used it to evaluate medical student diagnostic skills in multiple clinical disciplines (Stevens, 1991; Stevens, Kwak, & McCoy, 1989). These problems are clinically relevant and require an understanding of basic science concepts in the medical microbiology and immunology course at the UCLA School of Medicine. Each problem consists of a patient history (starting condition) and approximately 75 items of patient data, which the students can access in an uncued manner. A student has the option to solve the problem at any point by selecting from a list of over 40 possible solutions (goal condition). While this study focuses on a medical discipline, it is important to note that search, and the cognitive paradigm of a starting condition, a goal condition, and resources to transit these two states are general components of problem solving (Newell, 1990) and can be applied to many disciplines. In this regard, we are conducting similar studies in a variety of different disciplines and across broad levels of education (Stevens, 1995).

In this study, two forms of a problem were created that were identical except for the opening scenario (patient case history). These problems were designed to test the students' knowledge of T-cell receptor signaling (specifically CD3 zeta chain), an important regulatory step of most successful immune responses. One of the problems was contained in a set of 8 practice problems that were performed by students in the six weeks prior to the midterm examination. The second problem was contained in a set of 6 problems, 2 of which had to be solved by each student for a portion of his or her grade. The subjects in this study were the 21 students who both completed the practice problem and received the same problem on the examination.

Seven concept domains relevant to the identification of a specific immunological defect (Figure 1A) can be spatially oriented in related groups: a. integrins, b. immunoglobulin recombination, c. antibody production, d. histocompatibility, e. interleukin production, f. signaling, and g. T-cell receptor complex (Figure 1A). These domains collectively define the problem space and contain all of the necessary information required to solve a problem. For example, a problem with a defect in interleukin 2 (IL-2) production requires specific information from test items in the interleukin production domain (Figure 1A, domain *e*). Students who correctly identify the type of problem should focus their choice selections within that domain. It should be noted however that the information in other concept domains will often provide some indirect evidence of the problem solution and (limited) broad search not only is justified but may be desirable.

Search Path Map Analysis

The problem space represents the author's model of a test platform, and conscious efforts are made during the problem design to make the problem space as broad as possible to encompass a multitude of student strategies. The student selects data from a menu structure where test items are presented by type of test; for example, flow cytometry studies, antibody tests, western blots, etc. Each test item "costs" the student 50 points from a starting score of 3000. Every test item selected, the order of selection and the time interval between selections are automatically recorded into a database during problem solving (Stevens et al., 1989). Using IMMEX::Analysis (Stevens, 1991) the pattern of selections chosen by the student can be recreated as a search path map (Figure 1B) and visually

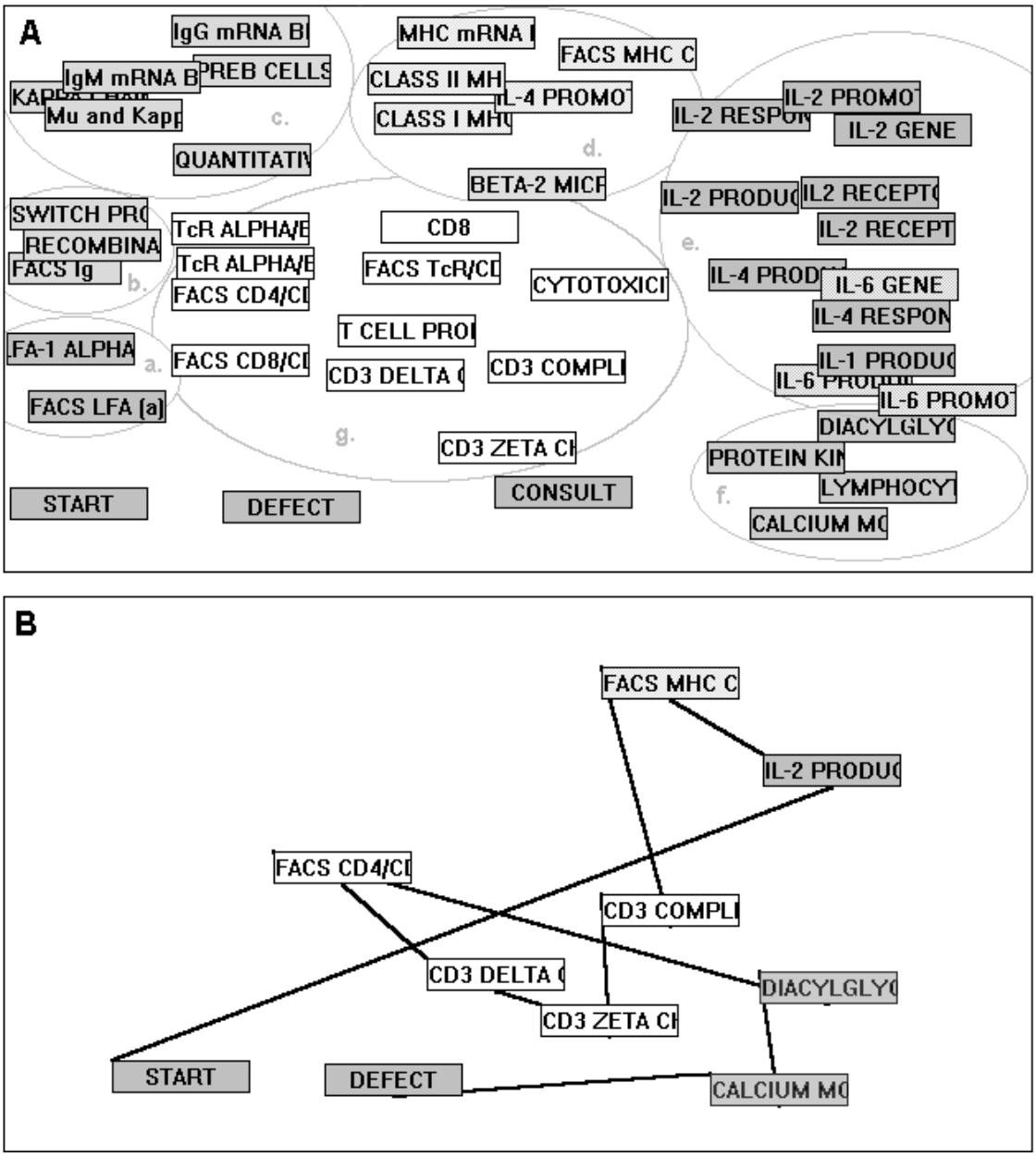

Figure 1. Search path mapping. Chart of concept domains (A), with each box representing a test item that can be chosen by students. The concept domains are labeled: a. integrins, b. recombination, c. immunoglobulin production, d. histocompatibility, e. cytokines, f. signaling, and g. T-cell immunity. Choices in the relevant domain for these problems are shown in white. Sample search path map (B) shows the order of tests in a single student performance. A line connects the top left corner of the first test ("START") and the center of the second test ("IL-2 PRODUCTION"), etc.

compared to the entire problem space (Figure 1A). Each test item selected is linked by a line connecting the upper left-hand corner of the first test item box to the center of the subsequent selection item box. For example, in Figure 1B, the student's first choice is "IL-2 PRODUCTION" followed by "FACS MHC" then "CD3 COMPLEX," etc. The final selection is the solution, labeled "DEFECT" in our immunology problem set. An additional feature of the software is the identification of students who utilized any particular data item. The search path map can also be used to analyze students grouped by score, problem number, solution (solved/not solved) or a combination of these parameters.

Artificial Neural Network Analysis

The neural network for this study was trained to recognize six major concept areas in immunology with more than 200 high-scoring student performances from one medical school second-year class. Outputs were generated for each step in student performance in this study. High output weights (approaching 1.0) indicate a close match of the student performance to those in the training set.

The neural network architecture used was a back-propagation supervised learning network. In supervised learning, pairs of inputs (student data) and outputs (solutions) are presented to the network. The network takes each input and produces its own output, which it compares to the correct output. The network learns by making corrections to the connections between input and output based on the error between the two outputs. As training progresses, the amount of error is minimized (Lawrence, 1993).

Results

Characteristics of Successful Strategies

Students who solve these problems correctly access information within a relevant concept domain and minimize searches in unrelated concept domains. Search path maps from successful strategies of one exemplary student during the practice (Figure 2, thin lines) and the examination (Figure 2, thick lines) revealed a focused set of test item choices within one domain. The student begins the practice performance by choosing T CELL PROLIFERATION followed by other tests in the correct domain, leading to the DEFECT (solution), where the correct answer (defect in the CD3 zeta chain) was chosen. A search strategy focused on

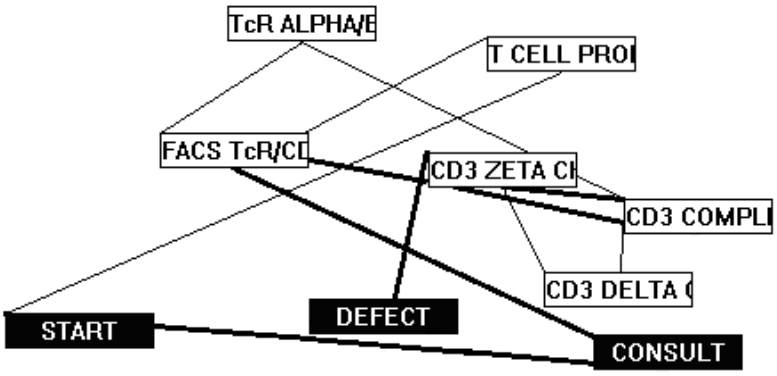

Figure 2. Two performances by the same student. The test items chosen on the practice test are connected by thin lines, and order of tests on the exam are connected by thick lines.

the correct domain indicates that the student had an understanding of the problem. In addition, the variation in the order of test items selected between performances reflects the use of a slightly different strategy in each case and suggests that this student's domain and problem understanding was quite rich. Of the six test selections chosen on the practice problem and the four chosen on the examination, only three items were common to both performances. This observation that different but conceptually related strategies can result in the same successful outcome is not unusual. In fact, of 1153 successful problem performances obtained from students at three medical schools, only 10 were duplicates, indicating that a wide range of successful strategies exists (Stevens et al., unpublished data).

There Are Identifiable Unsuccessful Strategies

Search path mapping revealed two general approaches pursued in unsuccessful problem-solving performances. The first, which accounts for approximately 50% of the missed problems, shows the inability of a student to develop a focused search strategy within the problem space, resulting in a great deal of test selections from nonrelevant concept domains (Figure 3A). This unfocused problem-solving strategy suggests that the student did not have a general understanding of the case and was conducting an exhaustive search in unrelated concept areas. During this search, the student did select important information from the relevant T-cell domain (Figure 3A, white boxes) but was unable to realize that this information was crucial.

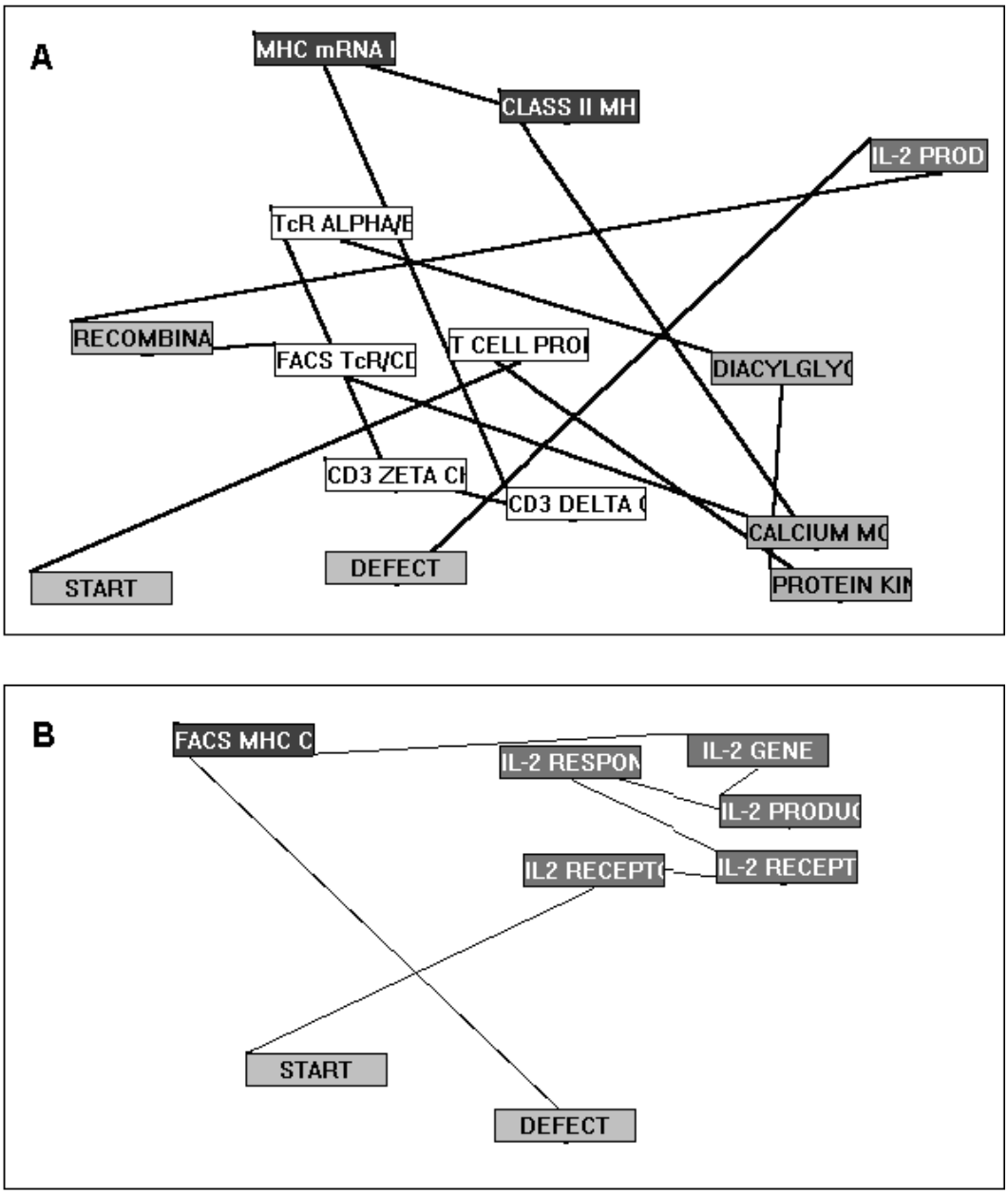

Figure 3. Two strategies leading to incorrect solutions. (A) Unfocused problem-solving strategy. This student chooses many items, including those in the correct domain (shown in white), but does not group choices by concept domain. (B) Focused problem-solving strategy. Student J9 chooses six tests in one (irrelevant) domain before choosing an incorrect solution.

In a second approach, students who missed the problems seem to have coherent yet inappropriate strategies. Such students conducted a thorough exploration of an inappropriate, although often closely related, domain before

leaving it for the next domain or choosing an incorrect solution. Although they misinterpreted the nature of the problem, they were able to execute a focused, domain-specific search (Figure 3B). Here, a thorough search is initiated in the interleukin 2 (IL-2) concept domain. Although the solution to this problem is not an IL-2 defect, it is clear that the student tested this possibility extensively by choosing five of six IL-2-specific test items within the interleukin domain (compare Figure 3B with domain *e* in Figure 1A). After completing this search, the student proceeded to choose an incorrect solution. Both student performances presented (Figures 3A and B) resulted in an incorrect solution, but whereas the second student showed a lack of thorough understanding of the concept domains, the first student demonstrated essentially no understanding by misinterpreting the appropriate test data.

Documenting Student Progress

Search path mapping can be most revealing of student progress when students demonstrate a focused strategy in the relevant domain on one problem after previously attempting to solve a similar problem by diffuse searching or incorrect concept domains. By comparing practice and examination performances, we noted strategic improvement for a number of students similar to that of student A1 (Figure 4). Here, the practice performance (Figure 4, thin lines) was an incomplete exploration of at least three concept domains. Exploration of the pertinent domain is also incomplete, and student A1 chose an incorrect solution. By contrast, the examination performance of student A1 (Figure 4, thick lines) shows a strategy transition with early recognition of the problem. This recognition leads the student to search within the relevant domain and choose the correct solution. In this example, search path mapping of practice and exam performances documents a significant change in this student's approach to this problem.

The recognition of the nature of the problem is one aspect of learning achievement, but the recognition of more subtle improvements within generally good strategies is equally important. What, if any, improvement can be detected where the relevant domain has already been identified by the student or where many choices within the correct domain are used by the same student with different outcomes? In the search path maps for student G3 (Figure 5A), there are areas of overlap between the practice and exam performances. In both cases, an

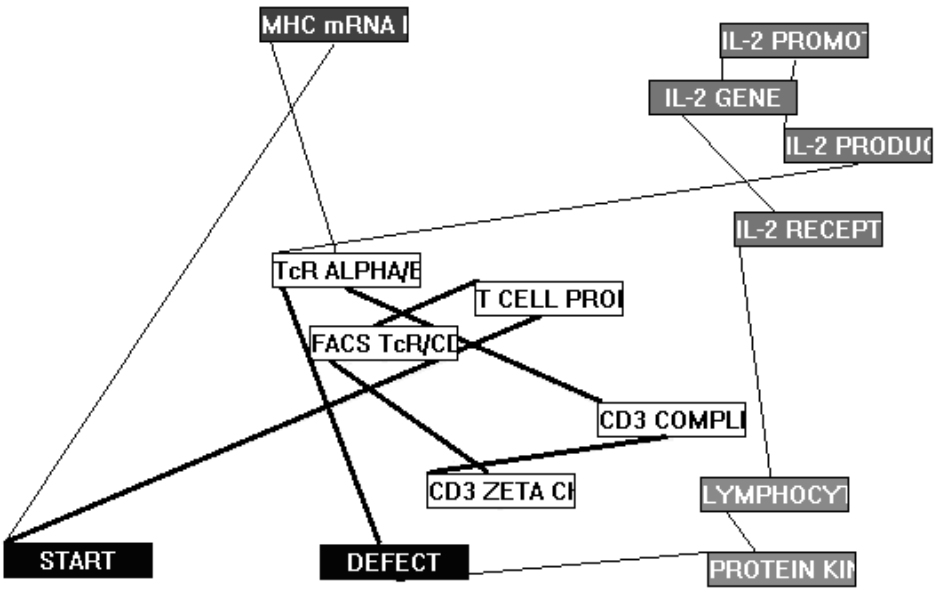

Figure 4. Transition from incorrect hypothesis to correct hypothesis. On the practice test (thin lines), student A1 develops a strategy, choosing tests from two irrelevant domains, but is not able to solve the problem. On the exam (thick lines), A1 uses a more focused strategy, choosing tests from the correct domain only and solving the problem.

irrelevant domain was sought out first, and other non-informative test items were selected. Upon analysis of the relevant test items selected (T-cell receptor/CD3 complex), two groups of test items can be identified for the practice and exam performances (Figure 5B). The practice performance included items from the relevant domain that are not as closely related conceptually as the items selected in the examination performance. This search path map data shows that even within the relevant domain, there are still notable improvements, which are evidently the result of an improved focus. In this example, the transition from a good strategy to a better strategy is seen as the student recognizes a distinctive group of relevant tests (Figure 5B).

While it is important to note this type of student progress, there are certain drawbacks. The major problem to this approach for analysis and assessment is that it is an enormously time-consuming process. Subtle changes in domain-specific activity may be difficult to discern, especially where the test selections are closely related.

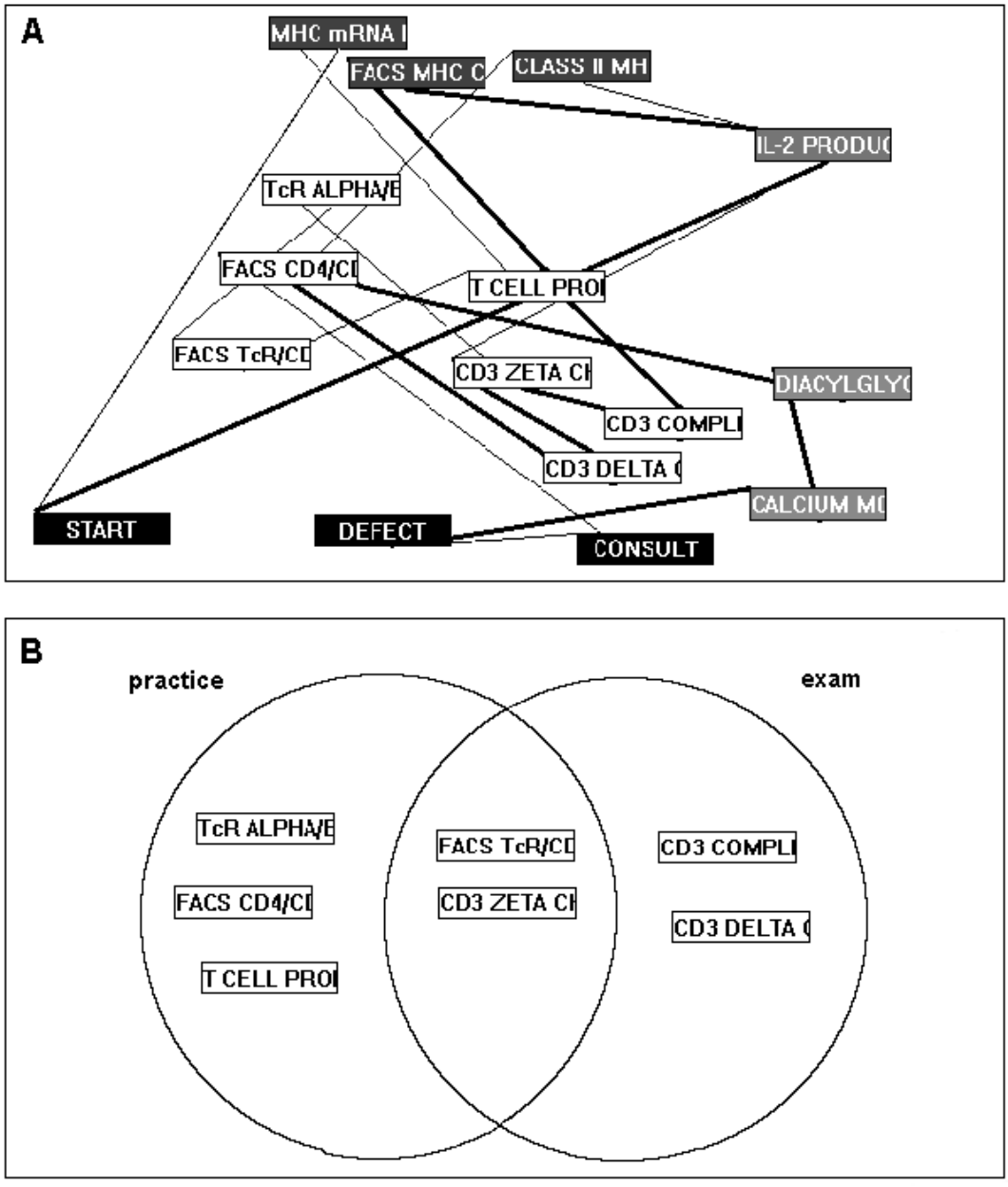

Figure 5. Transition from unfocused strategy to more focused strategy. This student chose many of the same test items on the practice (thin lines) and exam (thick lines), but follows a more organized and domain-specific search for relevant information on the exam.

Automation of Analysis Using Artificial Neural Networks

The above limitations of search path map analysis, namely time constraints, low objectivity, and difficulty in quantification of results, prompted our exploration of applying ANNs for assessing patterns of student performance. We have

previously used neural network analysis to identify patterns in student hypothesis generation and problem solving (Stevens & Najafi, 1993) and have shown the sensitivity (86%) and specificity (100%) of this approach to be high.

We first considered performances where changes in strategic approach were apparent by search path map analysis and where a student was unable to solve the practice problem, but was successful on the exam. Differences between practice and exam performances showing a transition from incorrect concept domain searches to relevant concept domain search activity were analyzed. Revisiting student A1's performance, an unsuccessful practice performance (Figure 4, thin lines) is followed by a focused, successful strategy (Figure 4, thick lines). Our ANN then assessed these search paths, and the outputs were charted (Figure 6A and B). In the practice performance (Figure 6A), the ANN detected the highest outputs at OUTPUT 1 (cell signaling domain) and OUTPUT 5 (interleukin domain). This indicates that the student solved the problem with a pattern of test selections, which the trained ANN recognized. There was an initial hypothesis of a cell signaling defect at OUTPUT 1 (ANN output ~0.58), and next, a hypothesis of an interleukin production defect at OUTPUT 5 (ANN output ~0.56). As the correct domain is represented by OUTPUT 3, neither of the hypotheses was correct. On the examination performance (Figure 6B), this student utilized one correct hypothesis. The ANN recognized this as a single strategy highly resembling the successful strategies of the ANN training set for this problem where the final weight at OUTPUT 3 was 0.95.

We also wished to determine whether the ANN could discriminate among more subtle changes in strategy within a relevant concept domain. In the performance of student G3, we used search path mapping to see a change in focus within the relevant concept domain (Figure 5). The ANN then analyzed these search paths and outputs were charted (Figure 6C and D). In the practice and exam performances, the ANN outputs in the correct domain (OUTPUT 3) were 0.38 and 1.0, respectively. This change in final output weight suggested the possibility that the ANN could detect the subtle changes in strategy occurring within a concept domain. We also noted that exam performances by both student A1 and student G3 refined their hypothesis utilization, since each student used two working hypotheses in practice performances and only one during the exam. In addition, in both cases, the practice performances were unsuccessful.

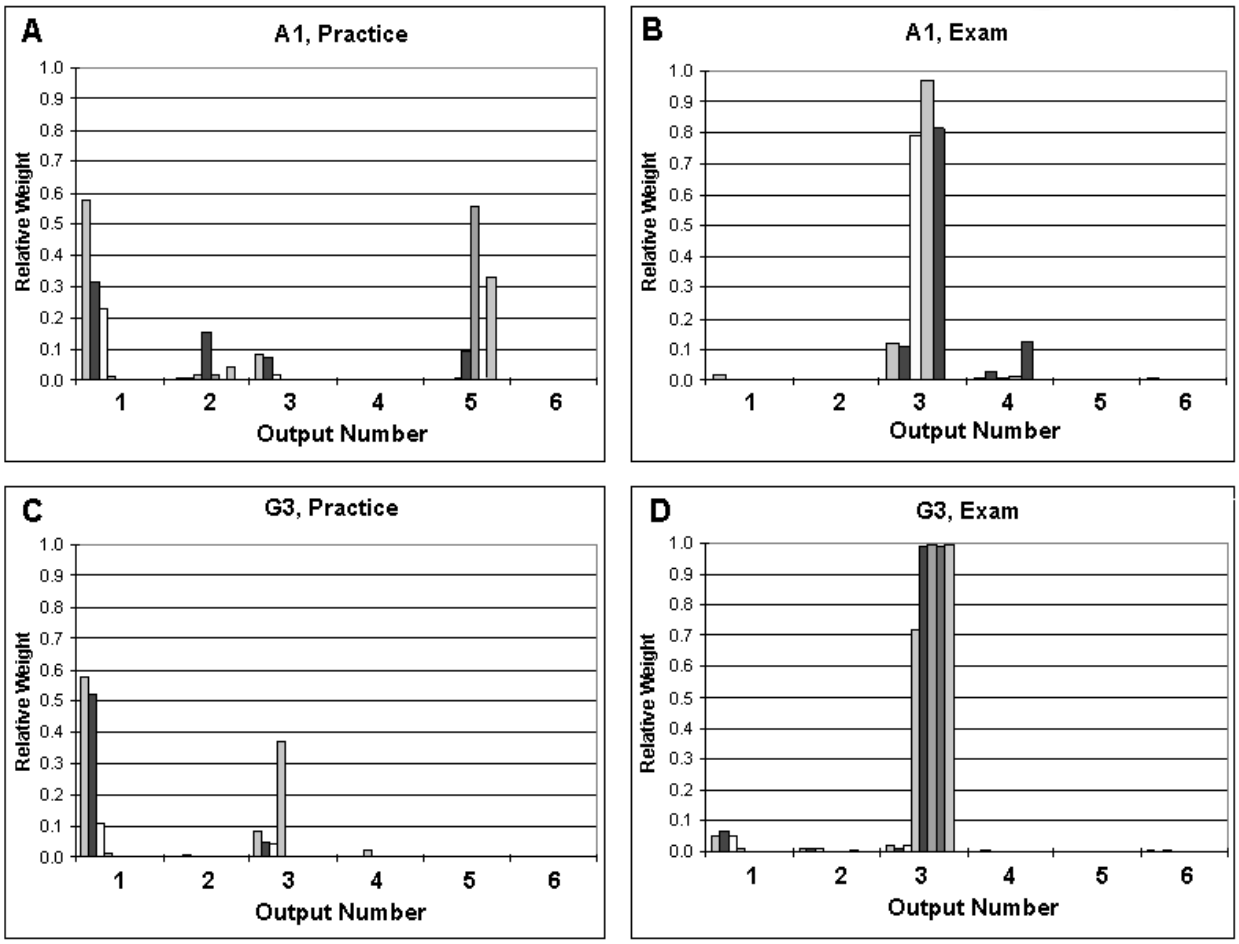

Figure 6. Artificial neural network output showing strategy formation. Each output number represents a concept domain within the immunology problem set. The relative weight indicates the match between this student performance and performances in the training set. Each bar represents one test item. In these problems, a high relative weight in output 3 indicates a hypothesis or strategy using the relevant test items in an order recognized by the neural network as "ideal." During the practice session (A), student A1 chooses tests recognized by the neural network as outputs 2 and 5 (not correct). On the exam (B), the student chooses tests recognized as output 3 (correct). Similarly, student G3 has a practice performance (C) recognized by the neural network as output 1 (incorrect) shifting to output 3 (correct), but an exam performance (D) with only one hypothesis at output 3 (correct).

Evidence for Hypothesis Utilization and Refinement in Successful Strategies

In order to gain a more complete sense of refinement in hypothesis utilization, we looked for examples of student performances where the practice and exam problems were both solved, regardless of score. The ANN profiles were compared to their respective search path maps. For student J2, the practice performance (Figure 7B) revealed ANN results initially high at OUTPUT 4

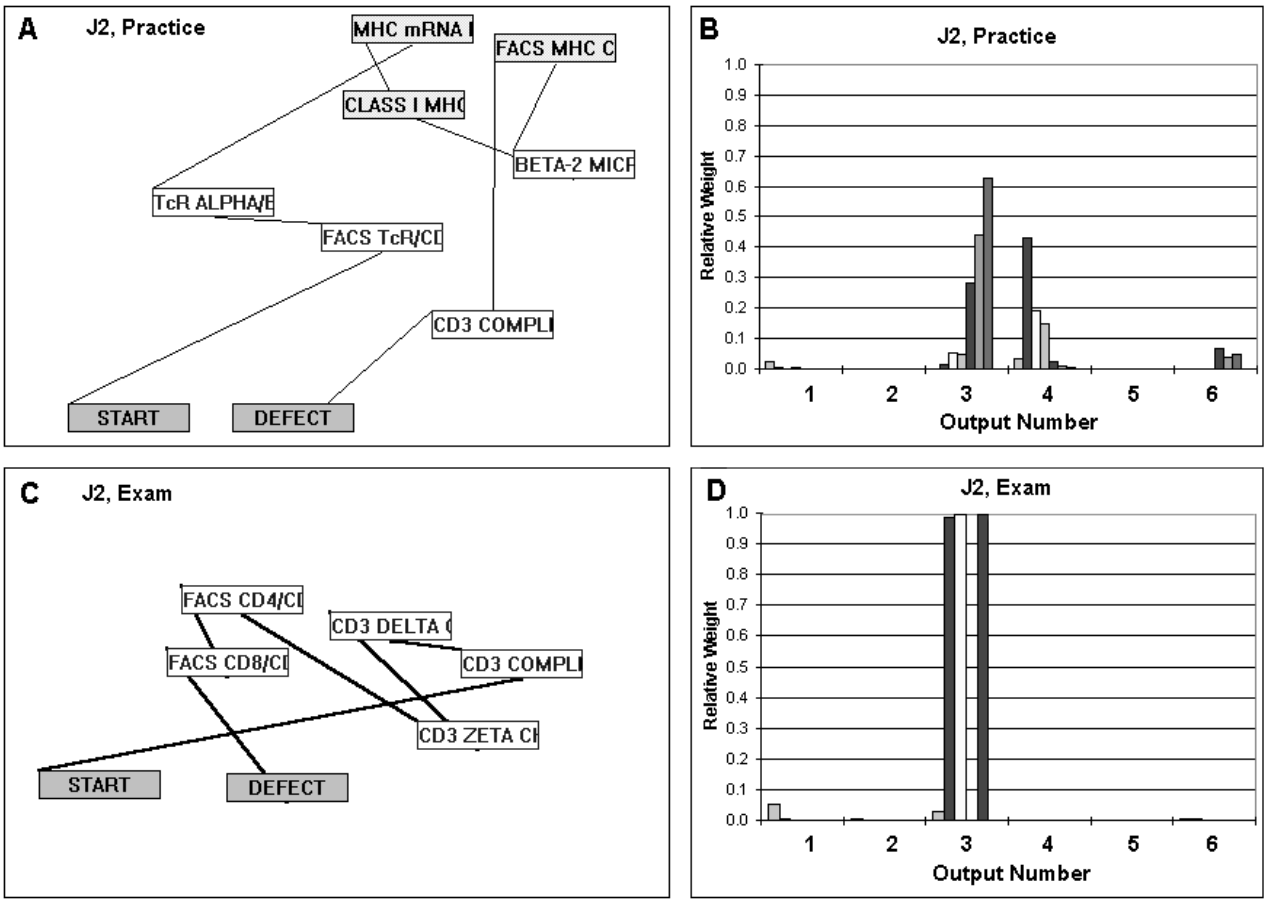

Figure 7. Shift from two-hypothesis strategy to one-hypothesis strategy. (A) Student J2 chooses test items from a domain initially recognized as output 4, then recognizes the correct domain, which is indicated by a higher relative weight in output 3. When J2 performs the similar problem on the exam (B), the correct domain is recognized earlier, and the neural network recognizes the performance as output 3 only.

followed by a subsequent decline in this output with a corresponding increase in OUTPUT 3 (correct domain). This indicated that the ANN recognized the first hypothesis, which corresponds to OUTPUT 4 (histocompatibility domain), the student's rejection of this hypothesis, and the subsequent development of a new hypothesis characterized by increasing values at OUTPUT 3. The search path map associated with this performance (Figure 7A) illustrates the fact that the student sampled the correct domain, but quickly went on to another domain (histocompatibility) before developing a strategy within the correct domain (T-cell receptor/CD3 complex). On analysis of the student's examination performance (Figure 7D), the ANN output occurs only in the area of OUTPUT 3 without the

development of other hypotheses by the student. The search path map for this performance (Figure 7C) reveals that the student focused on the relevant domain and chose items that specifically dealt with the CD3 complex. This differs from the relevant searching in the practice performance where the student used T-cell receptor data, which is closely related but not the focus of the problem. The improvements in test item selection are also reflected in the ANN output, which assigns a nearly perfect output weight (~1.0) to the exam performance compared to the lower weight (<0.65) assigned in the practice performance. These observations indicate that refinement of hypothesis utilization as well as the improvement in the execution of these hypotheses can be detected and quantified by our ANN.

A Caution: The Need to Correlate Some ANN Outputs With Search Path Maps

The use of ANNs to automate the analysis of search path maps can yield rich information about the utility and refinement of hypothesis formation in student performances. It must be kept in mind that supervised ANNs assign outputs as a function of the training set of performances that created the network, and if sufficient diversity is not included in the training set, then the ability of the neural network to "generalize" will be reduced. We were able to identify certain performances where an ANN did not predict what is evident from the search path map. This mismatched ANN output was seen in two settings where either inappropriately high or low output was generated for irrelevant domain searches or where an unusually low number of tests were required by the student to solve the problem. The first case is illustrated by student J11, who solved the problem, yet conducted a search strategy within an irrelevant domain (Figure 8A). Here, the student searched only within the (incorrect) interleukin domain, but chose the correct solution. The ANN output suggests the possibility that two distinct hypotheses were formulated by the student since OUTPUT 2 is high initially followed by an eventual decline in output and an increase in OUTPUT 3 (Figure 8B). We believe that in this case the ANN recognized the jump to the correct solution as a separate strategy. This type of misinterpretation is not representative of a large number of student performances.

At the other extreme is the case where a low ANN output for the relevant problem results despite a search path with specific focus within the pertinent

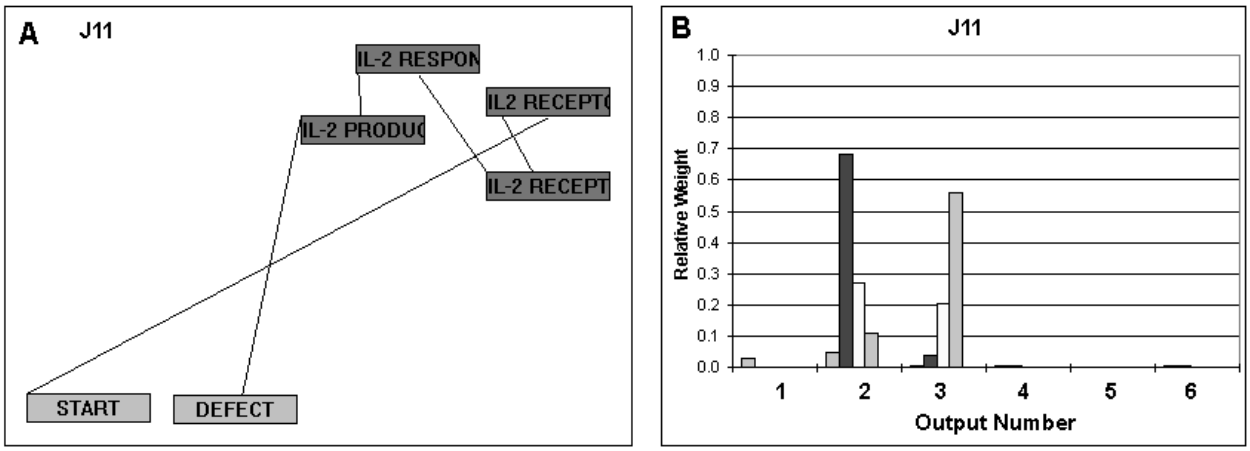

Figure 8. Weak correlation between search path mapping and neural network output. This search path map (A) shows the student exploring an irrelevant domain, correlating to high relative weight in output 2 (B). The jump from output 2 to output 3 is not detected on the search path map, as it resulted from a student choosing the correct solution ("DEFECT") only on the second attempt.

domain. In the practice and exam performance by student W2 (Figure 9A and C), it is obvious that the student needed very few test items within the pertinent domain to solve the problem correctly. The solution was correct in both cases, suggesting an extraordinary understanding of the problem compared to the majority of high-scoring medical students whose performances actually trained the network. The ANN output of student W2's performances is low in both the practice and exam setting with values of <0.3 and <0.1, respectively (Figure 9B and D). Including performances such as those of student W2 in the training set for the ANN could increase the predictive value of the ANN in such cases, as would the inclusion of expert performances. These examples illustrate the possible discrepancy that may exist when comparing ANN output to the actual search path maps.

A Proposed Scoring Rubric Based on ANN-Analyzed Search Path Maps

Our rubric for scoring these student performances is based on a combination of search path map data and the corresponding ANN outputs (Figure 10). We are able to demonstrate at least four types of performance based on the following criteria: (a) solved/not solved, (b) domain-specific activity (choice of test items within concept domains), and (c) relevant ANN output related to hypothetical construct within a given concept domain (quality of domain-specific activity). The

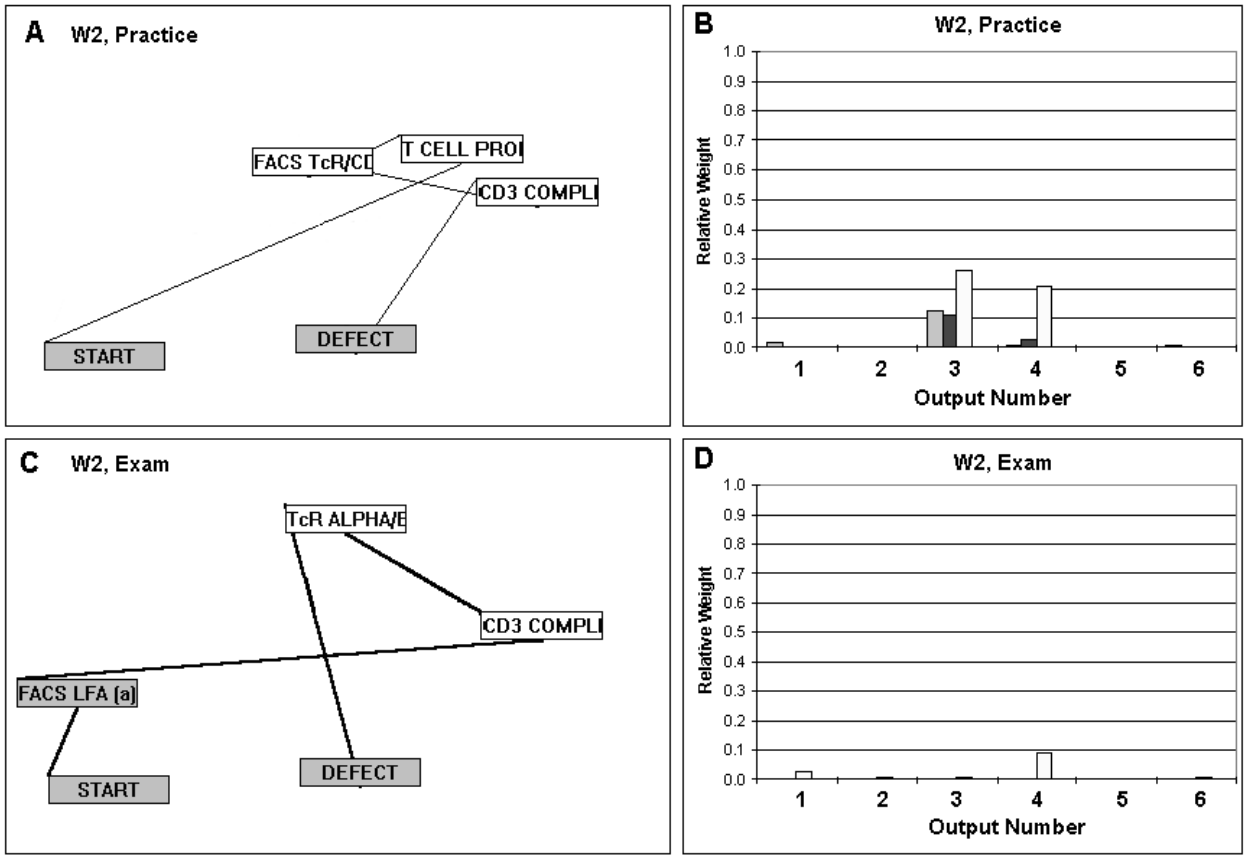

Figure 9. Expert performances are not represented in the training set for this neural network. Student W2 showed expert-like performance both on the practice (A, B) and exam (C, D), choosing a minimum of tests in the correct domain. As the neural network was trained on student performances only, it was unable to recognize this excellent problem-solving strategy as output 3 on either the practice (B) or exam (D).

highest quality performance on this scale results from a focused search within one relevant domain and is recognized by the ANN as highly related to the strategy used by the performances in the ANN training set (Figure 10A and B). This performance could arbitrarily be given a score of 4 (solved, with one hypothesis). Another successful strategy shown is where the student solves the problem after using one or more hypotheses, resulting in at least two distinct areas of significant ANN output (Figure 10C and D). Such a performance is assigned a score of 3 in our model. Another level could be predicted where students solve the problem without a discernible strategy, but there are no examples of this level in the student performances in this study. Unsuccessful strategies can similarly be categorized. A performance where the student does not solve the problem correctly, but formulates one or more irrelevant hypotheses can be scored as a 2

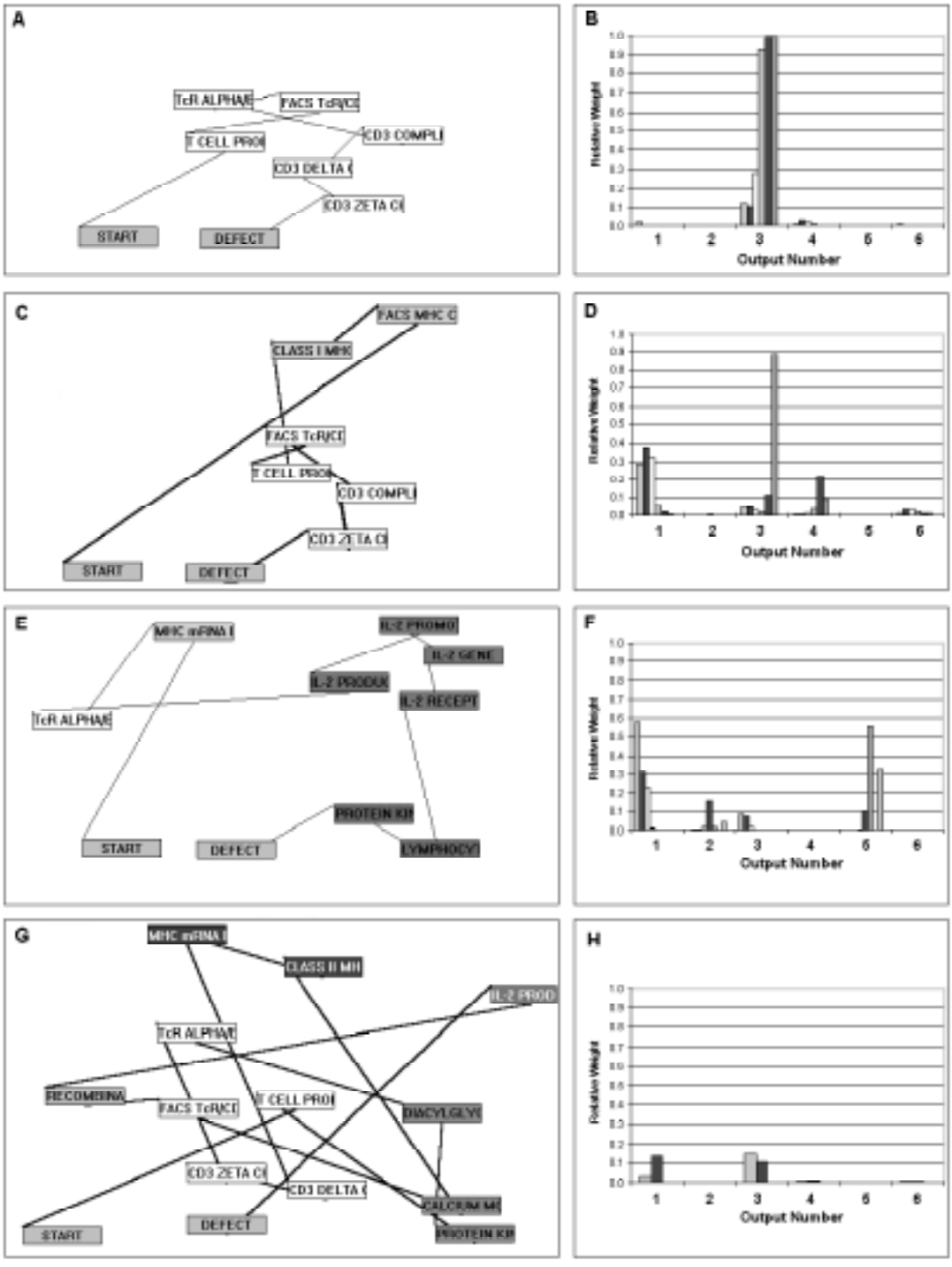

Figure 10. Hypothesis formation model. Search path maps (A, C, E, G) are paired with neural network outputs (B, D, F, H) to demonstrate levels of hypothesis formation. Level 1 (A, B): Solved, one hypothesis in the correct domain. Level 2 (C, D): Solved, more than one hypothesis, with the final hypothesis in the correct domain. Level 3 (not shown): Solved, no discernible hypothesis. Level 4 (E, F): Not solved, one or more hypotheses. Level 5 (G, H): Not solved, no discernible hypothesis.

(Figure 10E and F). Finally, where the student does not solve the problem and does not utilize hypotheses, a score of 1 is assigned (Figure 10G and H). By using this rating scale, we have shown an improvement in medical student performances from an indexed score of 2.48 ± 1.1 on the practice problem to a score of 3.38 ± .97 on the examination ($N = 21$, $p < 0.001$).

This proposed model favors solutions obtained through domain-specific activity and forward reasoning resulting in hypothesis utilization. It also de-emphasizes strict scoring based on the number of test items selected, although we have shown instances where score is clearly a useful adjunct to this analysis of strategy formation.

Discussion

We have described the development of an educational assessment tool that is generalizable, dynamic, and diagnostically functional. Within the problem space of applied immunology, we have analyzed student performance in response to a specific problem of immune deficiency due to defects in the T-cell CD3 complex. Student progress has been documented by a number of criteria including score, problem-solving strategy by search path mapping, and finally by supervised ANN analysis of the performance strategy to assess the utilization and refinement of hypotheses.

The number of test items chosen (equivalent to their raw score) probably has the least utility as a performance measure on both a theoretical and a practical level. A student might choose few tests, leading to an artificially inflated score as compared to a student with a more coherent strategy who chooses more test items. On other occasions it might be preferable—in medicine, for example—to perform a more expensive imaging procedure than to order multiple, less expensive, and less conclusive diagnostic procedures.

In our study, the students' raw scores on the examination were in fact lower than the scores of the practice performances (averaging 2340 vs. 2190, respectively) although strategic approaches documented by search path mapping improved. One (likely) explanation for this occurrence is the greater care or thoroughness that an examination evokes in students.

Through search path mapping, contrasting approaches to problem solving were apparent. In unsolved performances, two major patterns emerged, one

lacking focus, the other focused in an irrelevant area. Although both of these examples would be categorized as unsuccessful in terms of the ability to solve the problem, it is clear on visual inspection that the second category shows more rationale or strategy. This may reflect formation of a hypothesis, albeit the wrong one, in a specific concept domain.

Between the practice and the examination performances there was significant improvement in student performances, reflected by an improved focus on the pertinent concept domain. This improvement often coincided with the ability to solve the problem. This is exemplified by the performance of student A1 (Figure 4) where an unfocused search during practice improved on the exam.

As mentioned earlier, there are several limitations when assessing student strategies by search path mapping. First, this analysis requires expert review, making it difficult to scale from small groups of students. Second, the procedure is not quantitative. The previous examples demonstrate that, although a scale of performances can be identified, and a rubric for ordering these performances may be possible to develop, nevertheless assigning a numeric value to each performance would be difficult. Lastly, it is difficult to factor in the dynamics of the formation and refinement of hypotheses during the problem-solving process, which are demonstratively complex. To address these problems, we have applied the pattern recognition capabilities of ANNs, which are useful for the analysis of the patterns of performance that are not obvious based on search path mapping alone.

Artificial neural networks can begin to recognize when a strategy is being applied within a concept domain and can indicate by generated output the relative "strength" of the student's performance compared to those of peers or experts. Our studies indicate how ANNs can be used to monitor student improvement through the number and quality of the formulated hypotheses. It is also clear from our data that some students fail to progress, since their performance on the practice session was similarly poor on the examination. In this data set, 2 of 21 students did not improve their hypothesis formation from the practice to the exam. It is difficult to predict which students will do poorly on the examination because many students use improved strategies on the exam. This provides an argument for repeated testing to monitor student performance as a dynamic function of the learning process. By monitoring students over time, it is likely that variations in performance patterns will emerge. Students who are beginning to

understand a problem may shift from multiple hypotheses to a single hypothesis during problem solving, while others may be able to learn to recognize the correct domain earlier. Following this approach, instructors could easily identify student progress, as well as problematic concepts.

As a result of one of the limitations of this study—that is, the occasional inability to correctly identify "expert" performance—we are applying ANN technologies to better map the strategic nature of the transition from novice to competent to expert problem solving. These attempts have led to an increased ability to identify not only the proportion of a population performing at a high level, but also the differential strategic approaches leading to this higher level of performance (Stevens, Lopo, & Wang, 1996). Because both novice and expert performances may be deemed successful by simple solved/not solved rubrics, it is the subtle features within these performances that allow discrimination of the transition between the two states. These features may be embedded within student performance data and difficult to discern using many assessment tools, but can be approached through our ANN-based analysis.

We are also beginning to address the psychometrics of output weights. For simply optimizing the sensitivity and specificity of neural network decision thresholds, receiver operating characteristics (ROC) analysis has been very useful (Eberhart, Dobbins, & Hutton, 1990). However, the differences in strategy that result in a student performance receiving a final output weight of 0.7 (for example) instead of a final output weight of 1.0 are more subtle, and further studies are in progress to identify the nature of these differences over several types of immunology problems and several neural network raters (K. Hurst & R. Stevens, unpublished data).

The approach to educational assessment presented in this study serves to illustrate how the spectrum of student performance strategies can be practically assessed with supervised ANNs. These networks can accurately reflect relevant and sometimes subtle associations between concepts needed to solve problems. Furthermore, the method allows for dynamic assessment of a large number of performances, freeing valuable instructor time for modifying curricula to reach a maximum number of students in specific areas. This approach can be applied to many subject areas, provided that the problem sets are designed appropriately.

References

Alexander, P. A., & Judy, J. E. (1988). Interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research, 58*, 375-404.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In P. H. Mussen (Series Ed.) & J. H. Flavell & E. M. Markman (Vol. Eds.), *Handbook of child psychology: Vol. 3. Cognitive development* (4th ed., pp. 77-166). New York: Wiley.

Campione, J. C., & Brown, A. L. (1990). Guided learning and transfer: Implications for approaches to assessment. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 141-172). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H., & Glaser, R. (1985). Problem-solving ability. In R. J. Sternberg (Ed.), Human abilities: An information-processing approach (pp. 227-250). New York: W. H. Freeman.

Eberhart, R. C., Dobbins, R. W., & Hutton, L. V. (1990). Performance metrics. In R. C. Eberhart & R. W. Robbins (Eds.), *Neural network PC tools* (pp. 161-176). San Diego, CA: Academic Press.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39*, 93-104.

Glaser, R. (1990). Toward new models for assessment. *International Journal of Educational Research*, *14*, 475-483.

Glaser, R. (1992). Expert knowledge and processes of thinking. In D. F. Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics* (pp. 63-75). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, R., Raghavan, K., & Baxter, G. P. (1992). *Cognitive theory as the basis for design of innovative assessment: Design characteristics of science assessments* (CSE Tech. Rep. No. 349). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Groen, G. J., & Patel, V. L. (1988). The relationship between comprehension and reasoning in medical expertise. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 287-310). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lawrence, J. (1993). *Introduction to neural networks*. Nevada City, CA: California Scientific Software Press.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Peverly, S. T. (1991). Problems with the knowledge-based explanation of memory and development. *Review of Educational Research, 61*, 71-93.

Reggie, J. (1993). Neural computation in medicine. *Artificial Intelligence in Medicine, 5*, 143-158.

Simon, H. A. (1995). Near decomposability and complexity: How a mind resides in a brain. In H. J. Morowitz, & J. L. Singer (Eds.), *The mind, the brain, and complex adaptive systems* (pp. 25-43). Reading, MA: Addison-Wesley.

Stevens, R. H. (1991). Search path mapping: A versatile approach for visualizing problem-solving behavior. *Academic Medicine, 66*(9, Supplement), S72-S75.

Stevens, R. H. (1995). Problem solving in the sciences: An innovative software approach. In *Promising practices in math and science* (pp. 70-71). Washington, DC: U.S. Department of Education.

Stevens, R. H., Kwak, A. R., McCoy, J. M. (1989). Evaluating preclinical medical students by using computer-based problem-solving examinations. *Academic Medicine, 64,* 685-687.

Stevens, R., & Lopo, A. (1994). Artificial neural network comparison of expert and novice problem-solving strategies. *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care (*pp. 64-68).

Stevens, R. H., Lopo, A., & Wang, P. (1996). Artificial neural networks can distinguish novice and expert strategies during complex problem solving. *Journal of the American Medical Informatics Association, 3*, 131-138.

Stevens, R. H., McCoy, J. M., & Kwak, A. R. (1991). Solving the problem of how medical students solve problems. *MD Computing 8*, 13-20.

Stevens, R. H., & Najafi, K. (1993). Artificial neural networks as adjuncts for assessing medical students' problem-solving performances on computer-based simulations. *Computers and Biomedical Research, 26,* 172-187.

Sugrue, B. (1994). *Specifications for the design of problem-solving assessments in science* (CSE Tech. Rep. No. 387). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Weinstein, J. N., Kohn, K. W., Grever, M. R., Viswanadhan, V. N., Rubinstein, L. V., Monks, A. P., Scudiero, D. A., Welch, L., Koutsoukos, A. D., Chiausa, A. J., & Paull, K. D. (1992). Neural computing in cancer drug development: Predicting mechanism of action. *Science, 258,* 447-451.