A Value-Based Approach for Quantifying Student's Scientific Problem Solving Efficiency and Effectiveness Within and Across Educational Systems

Ron Stevens, Ph.D.
IMMEX Project, UCLA
5601 W. Slauson Ave #255
Culver City, CA 90230
immex_ron@hotmail.com
310-649-6589

Abstract

The challenge addressed in this study is: 'What is a suitable description of problem solving that can capture important cognitive and performance information about an individual's problem solving, yet provide rapid and meaningful comparisons within and across science domains and educational systems?' While such a measure would have practical benefits at many levels of education, there are also theoretical reasons to support these developments. In this manuscript we first discuss the need for developing assessments of problem solving and focus on creating metrics to track the development of these skills over prolonged periods of time. Next, we describe the IMMEX problem solving environment that provides a wide range of online problem solving experiences for students from middle school through medical school. Then, by using a combination of machine learning tools, we describe a value-based metric of problem solving that allows assessment of problem solving across scientific domains, levels of education, and educational systems. Lastly, we show how this measure can be used to identify classrooms where students' progress at developing these skills is not progressing as predicted by other achievement scores.

Introduction

Supporting students' ability to effectively solve problems is viewed as a national educational priority. However, on the most recent Program for International Student Assessment [PISA] test, American 15-year olds ranked 24th out of 29 developed nations in Mathematics literacy and problem-solving (Augustine, 2005, OECD, 2004). While all stakeholders in science education recognize the need for developing effective problem solvers, classroom teachers find it difficult to quantify students' strategic thinking in ways that can rapidly inform instruction (Pellegrino, et al. 2001). The current challenges for such assessments relate to the cognitive and non-cognitive differences among students, the difficulty of generalizing across problem solving content domains, the design and development of appropriate tasks, and finally the speed, scale, and conceptual accessibility of assessment data.

Strategic problem solving is influenced by many variables, such as: students' prior knowledge and skill, cognitive and metacognitive abilities, task characteristics, gender, ethnicity, classroom environment (Fennema, et al. 1998), as well as affective variables such as motivation and self

efficacy (Mayer, 1998). An additional complication is that the acquisition of problem solving skills is a dynamic process characterized by transitional changes over time as experience is gained and learning occurs (Lajoie, 2003).

While the challenges for developing assessments of problem solving are substantial, the real-time generation and reporting of metrics of problem solving efficiency and effectiveness could fulfill many of the purposes for which educational assessments are used such as grading and feedback for improving learning. (Bennett, 1998, Atkin, et al. 2001). In addition, such metrics could help target interventions for students, focus professional development activities for teachers, and influence training, implementation and support decisions throughout school districts (Spillane, et al. 2002).

An important starting framework for any assessment is construct validity (Messick, 1989). The tasks from which behaviors are extracted as evidence of skill acquisition must be accurate, appropriate for the audience and appropriate for the cognitive / behavioral constructs purported to being measured. The assessments concurrently developed should have concurrent, divergent, discriminant and predictive validity and be reliable, scalable, adaptable and extensible. They should also be understandable by teachers and students (Reckase, 1998).

The groundings for this study are based on the theories of strategy development (Anderson, VanLehn, 1996, Schuun & Reder, 2001, Schuun et al., 2001, Haider & Frensch, 1996) and skill acquisition (Ericcson) and can be organized around the following principles:

- Each individual selects the best strategy for them on a particular problem and individuals might vary because of learning in the domain and/or process parameter differences;
- People adapt strategies to changing rates of success. Note that the base rate of success is not the same for all people on a task or for an individual on different tasks.
- Paths of strategy development emerge as students gain experience; and,
- Improvement in performance is accompanied by an increase in speed and reduction in the data processed.

We believe that the paths that students employ while navigating an IMMEX task provide evidence of a strategy, which we define as a sequence of steps needed to identify, interpret and use appropriate and necessary facts to reach a logical conclusion or to eliminate or discount other reasonable conclusions. From these strategies a student demonstrates understanding by consistently, and efficiently deriving logical problem solutions.

Our framework for assessing problem solving skills uses and addresses the following metrics:

- How well / rapidly were the problems solved? (*easy to assess, but contains little strategic information*)
- Can hard / easy problems be solved? (*more difficult to assess; IRT estimates can be useful*)
- What problem solving strategy was used? (*more difficult to assess*)
- Are the problem solving strategies improving with practice? (*more difficult to assess*)
- What strategy will the student next use? (*hard to assess*)

And then of course is the challenge of how to generalize across domains and educational systems.

Methods

What is IMMEX?

The IMMEX™ Project upon which this proposal is based, hosts an online problem solving environment and develops and delivers scientific simulations and probabilistic models of learning trajectories that help position students' scientific problem-solving skills upon a continuum of experience (Stevens et al, 2004, 2005, 2006).

To illustrate the system, a sample chemistry task is called *Hazmat* which provides evidence of a student's ability to conduct qualitative chemical analyses. The problem begins with a multimedia presentation, explaining that an earthquake caused a chemical spill in the stockroom and the student's challenge is to identify the chemical. The problem space contains 22 menu items for accessing a Library of terms, the Stockroom Inventory, or for performing Physical or Chemical Testing. When the student selects a menu item, she verifies the test requested and is then shown a presentation of the test results (e.g. a precipitate forms in the liquid) When students feel they have gathered the information to identify the unknown they can attempt to solve the problem.

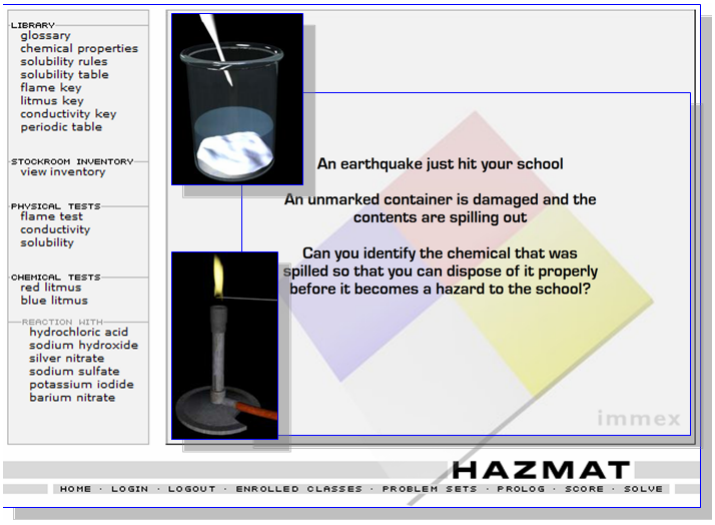


Figure 1. This screen shot of *Hazmat* shows the menu items down the left side of the main "Hazmat" window on the screen and a sample test result (the result of a precipitation reaction).

Students access resource data in an open-ended manner from experimental results, reference materials, advice from friends and / or experts, etc. to solve the problem. The IMMEX™ database serializes timestamps of how students use these resources.

For assessments and measurements to be useful, the tasks must be accurate, appropriate for the audience and appropriate for the constructs purported to being measured (Mislevy et al, 2001). The construction of IMMEX problems follows design specifications that emphasize valid and current content, and ensure that there are many ways to succeed or fail in problem solving (Stevens & Palacio-Cayetano, 2001).

Layers of IMMEX Assessments

Behind the scenes, IMMEX uses rich machine learning tools to build models of student's performance that address the frameworks above. The solve rates, time spent on each case of a

problem set and the number of problems performed are all reported in real-time to both students and teachers.

The next layer takes advantage of the multiple problems in each problem set that vary in difficulty. *Hazmat*, for instance has 38 'clones' that contain acids, bases, and different compounds that may or may not show a positive color by flame testing. From the different item difficulties, and thousands of student performances, we have developed item response theory models (IRT) that provide ability estimates that take into account not only if a problem was solved or not, but also the difficulty of the problem.

Item Response Theory Estimates of Student Ability.

The first layer of our data analytic system uses estimates of student ability (theta) as determined by IRT. Item response theory relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of a positive response (such as solving a case). Unlike classical test theory item statistics, which depend fundamentally on the subset of items and persons examined, IRT item and person parameters are invariant. This makes it possible to examine the contribution of items individually as they are added and removed from a test. It also allows researchers to conduct rigorous tests of measurement equivalence across experimental groups.

Using IRT, pooled data for students are used to calibrate all of the items and to obtain a proficiency estimate for each student (Figure 2).

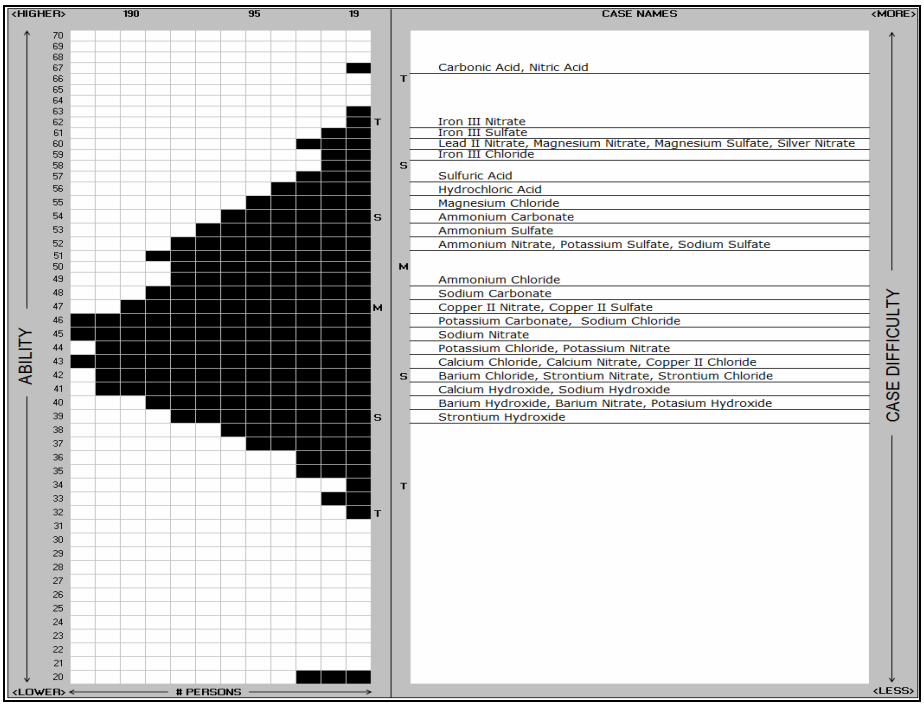


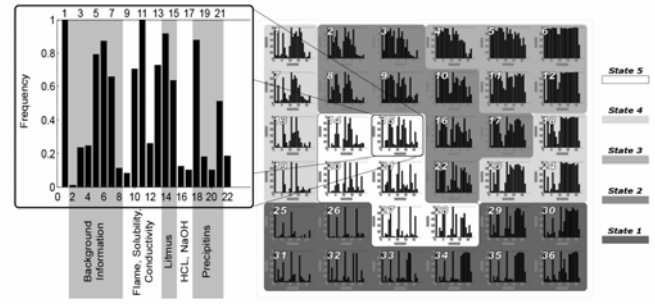
Figure 2. Levels of Problem Difficulty. The case item difficulties were determined by IRT analysis of 28,878 student performances. The problem difficulty begins with the easiest at the bottom and increases towards the top. The distribution of student abilities is shown on the left with the highest ability students at the top, decreasing downwards. For each graph, M indicates the mean, S, the standard deviation, and T two standard deviations.

As expected, the flame test negative compounds are more difficult for students because both the anion and cation have to be identified through a more extensive chemical analysis. The distribution of student abilities provides an important validity check of the appropriateness of the content for the intended audience, and suggests that the cases in the problem set present an appropriate range of difficulties to provide reliable estimates of student ability by IRT. However, while IRT is useful for ranking the students by the effectiveness of their problem solving, it does not provide a strategic measure of a student's problem solving performance.

The next layer of IMMEX analytic tools, artificial neural networks, extends the IRT measures by providing evidence as to how the students solved the problem and where the problem solving process succeeded or failed. In this regard we begin to answer whether the tasks are appropriate for detecting strategic differences between students.

Artificial Neural Networks

As students navigate the problem spaces, the IMMEX database collects timestamps of each student selection. The most common student approaches (i.e. strategies) for solving Hazmat are identified with competitive, self-organizing artificial neural networks using the students' selections of menu items as they solve the problem as the input data (Kohonen, 2001, Stevens et al, 2004, 2005). The result is a topological ordering of the neural network nodes according to the structure of the data where geometric distance becomes a metaphor for strategic similarity. Often we use a 36-node neural network, in which each node is visualized by a histogram (Figure 2 A). The histograms show the frequency of items selected for student performances classified at that node. Strategies so defined consist of items that are always selected for performances at that node (i.e. with a frequency of 1) as well as items ordered more variably.



Nodal Solve Rates (%)					
61	65	58	47	37	36
64	64	53	48	47	42
64	70	75	57	53	51
74	83	76	59	58	53
24	80	83	72	61	59
72	81	80	58	55	56

Figure. 3. Sample Neural Network Nodal Analysis for Identifying Strategies. a.) The selection frequency of each action (identified by the labels) is plotted for the performances at node 15, thus characterizing the performances for this node and relating them to performances at neighboring nodes. The nodes are numbered in rows, 1-6, 7-12, etc. This figure also shows the item selection frequencies for all 36 nodes (Stevens et al., 2004, 2005, 2006). b) This figure shows the solution rate for each node with the lowest solved rates in black and the highest in gray.

In Figure 3A there is also a composite ANN topology map of performances generated during the self-organizing training process. Each of the 36 matrix nodes represents where similar student problem solving performances were automatically clustered by the ANN procedure. As the neural network was trained with vectors representing the items students selected, it is not surprising that a topology developed based on the quantity of items. For instance, the upper right hand of the map

(nodes 6, 12) represents strategies where a large number of tests have been ordered, whereas the lower left corner contains strategies where few tests have been ordered. As shown in Figure 3b, not all strategies result in the same solve outcomes.

Once ANN's are trained and the strategies represented by each node defined, new performances can be tested on the trained neural network, and the node (strategy) that best matches the new performance can be identified and reported. The strategies can be aggregated by class, grade level, school, or gender, and related to other achievement and demographic measures.

Hidden Markov Models

On their own, ANN-defined categories provide a strategic snapshot of any one particular student's performance, but they really don't provide much information on whether or not the students are improving their problem solving skills. To obtain this information, IMMEX borrows the ideas of Hidden Markov Modeling from digital signal processing. First we postulate that there are a number of states that students will go through as they begin to refine their problem solving skills; this often emerges from the cognitive task analysis conducted in parallel with the construction of the problem set. For instance, most students who are engaged in the task will initially conduct a rather thorough exploration of the problem space. At the other pole of competence, highly effective and efficient students will exhibit refined and parsimonious strategic approaches. Between these two poles will be other more transitional, or dead-end states that reflect individual progress (or lack thereof) towards improving skills.

For IMMEX problems we often choose 5 hidden states. Then, similar to the training of the ANN's, many sequences of strategies (ANN nodal classifications) are presented to the HMM modeling software (Murphy, 2001) which then constructs probabilistic progress models (Stevens et al., 2004, 2005, 2006). This process is shown in Figure 4.

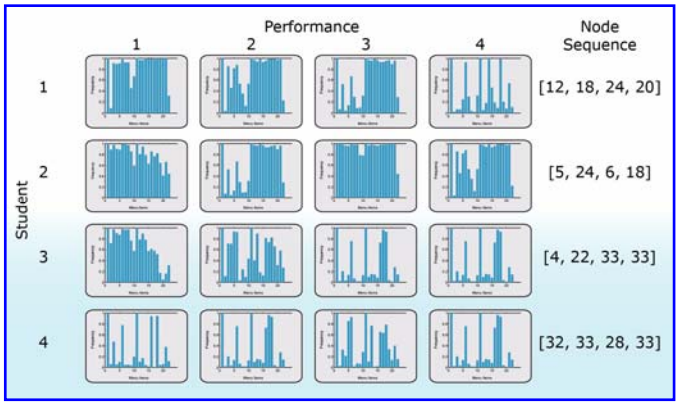


Figure 4. Developing Learning Trajectories from Sequences of Strategies.

Student #1 began by an extensive exploration of the problem space which was refined on subsequent performances. Referring back to Figure 3, the first performance was classified at node 12, the second at node 18 and the third at node 24. The data reduction occurring with each subsequent performance primarily reflected less reliance in the background resources in the Library section of the problem. Student #2 began similarly, but made slower progress in reducing the data accessed. Students #3 and 4 rapidly transited to efficient strategies and stabilized there.

Hidden Markov Modeling provides two important perspectives on problem solving progress. First, an emission matrix maps the five states back to the most likely symbols they contain, i.e.

the ANN nodes. Second, the processing generates a transition matrix which provides the probability of transiting from one state to another. This is shown in Figure 5.

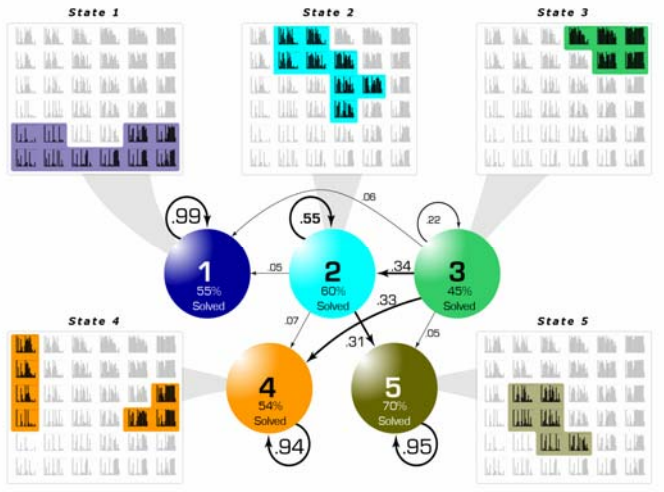


Figure 5. The Emission and Transition Matrices from the Hazmat HMM Model.

The emission and transition matrices resulting from training a HMM with 1790 sequences of students performances show the strategies most likely to represent the different states as well as the probability of transiting from one state to another. The overall solution frequency of each state is shown inside each ball.

State 3 represented the most frequent starting point for students and the strategies most often associated with it showed extensive testing. Approximately a fifth of the students at this state will remain there on the subsequent performance, with a third transiting to either State 2 or State 4. State 2 is also a transitional state with approximately one third of the students next transiting to State 5 which is a stable state characterized by the highest solved rate.

By plotting the proportion of students at each state for each performance of a problem set, trajectories can be developed that visualize progress. This is shown in Figure 6 for students working individually as well as in groups.

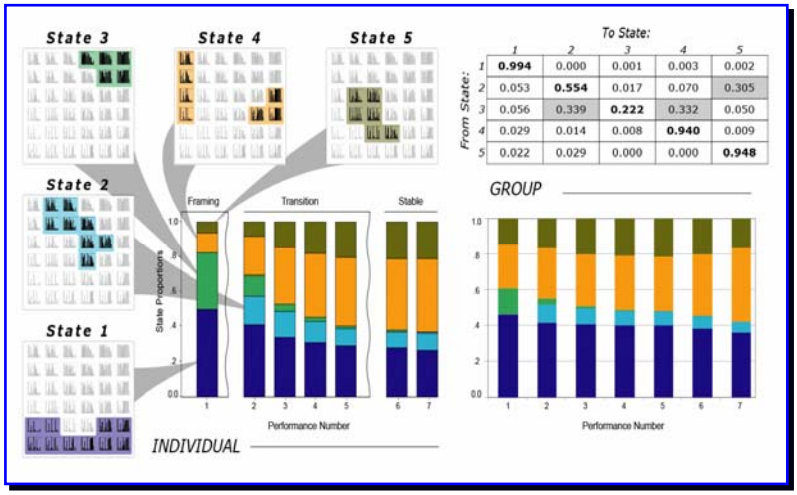


Figure 6. Modeling Individual and Group Learning Trajectories. This figure illustrates the strategic changes as individual students or groups of students gain

experience in Hazmat problem solving. Each stacked bar shows the distribution of HMM states for the students (N=1790) after a series (1-7) of performances. These states are also mapped back to the 6 x 6 matrices which represent 36 different strategy groups identified by self organizing ANN. The highlighted boxes in each neural network map indicate which strategies are most frequently associated with each State. From the values showing high cyclic probabilities along the diagonal of the HMM transition matrix (upper right), States 1, 4, and 5 appear stable, suggesting once adopted, they are continually used. In contrast, students adopting State 2 and 3 strategies are more likely to adopt other strategies (gray boxes) (Stevens et al. 2004).

One finding that has been observed from middle school through the university is that students initially conduct extensive explorations of the problem space, and then begin to refine their strategies as they gain experience. Consistent with models of skill acquisition (Ericsson, 2004), after relatively few problem performances (generally 5-7), most students stabilize with preferred strategies. Students often continue to use these stabilized strategies for prolonged periods of time (3-4 months) when serially re-tested (Stevens, 2006).

Using this modeling approach we have shown that higher ability students stabilize strategies more slowly than low ability students, perhaps suggesting a richer repertoire of approaches to draw from (Stevens et al., 2004). Although males and females solve the same number and proportion of attempted problems, there are significant strategic differences across gender at both the ANN and HMM modeling levels (Stevens & Soller, 2005). This apparent disconnect between the problem-solving outcomes and the strategies used has also been observed in Bayesian Network models of factors influencing IMMEX™ strategies and outcomes (Stevens & Thadani, 2006).

One 'Big Problem' with Probabilistic Models

Another component of our assessment framework asks whether the assessment measures are easily understood by students, teachers and parents. While the ANN and HMM provide detailed performance and progress information at both the student and classroom level, there is still a serious challenge for their widespread adoption by teachers: Each problem set has its own ANN topology and HMM state transitions. Therefore, a teacher who implemented six different IMMEX problem sets with her classes would have to understand six different performance and six different progress models, each composed of 36 nodes and 5 states respectively! This is unreasonable to expect from teachers.

We therefore began to explore alternative ways of representing the problem solving process that could draw on ANN and HMM models, and could replicate many of the findings of these models, but yet still could be easily understood, and compared across problem sets. The approach we have followed is to express problem solving as a value that combines the efficiency and effectiveness of the process. We first postulated that the students demonstrating high strategic efficiency should make the most effective problem solving decisions using the least number of resources available, whereas students with lower efficiency levels would require more resources to achieve similar outcomes and / or will fail to reach acceptable outcomes. As problem solving skills are refined with experience, this should be reflected as a process of resource reduction (Haider et al., 1996).

The core components of strategic efficiency for resource utilization are therefore 1) the quantity of resources used vs. the quantity available, 2) the value of the resulting outcomes expressed as a

proportion of the maximum outcomes, and 3) the quality of the data obtained. The first two components can be represented by Equation 1 where we define a resource-utilization Efficiency Index, termed EI. For IMMEX™ problems the maximum outcome is 2 (e.g. 2 points for the correct answer, 1 point for the correct answer on a second attempt, and 0 pts for missing the solution). This equation yields a simple exponential curve with a minimum approaching 0 where there are no / poor outcomes with extensive resource utilization and a varying maximum depending on the value of the absolute quantity of resources available.

$$EI_R = \left(\frac{\text{obtained outcome}}{\text{max outcome}} \Big/ \frac{\text{resources used}}{\text{resources available}} \right) \quad (1)$$

Not all of the available resources in a problem space are equally applicable to the particular problem at hand, and different combinations of resources will have different values within the contexts of different problems. Thus, estimates of the quality of resources used are also required. This qualitative dimension is derived from classifications resulting from unsupervised artificial neural network (ANN) analysis described above. When the EI for the 36 nodes of the ANN are plotted against the average solve rate, the distribution in Figure 8 is obtained.

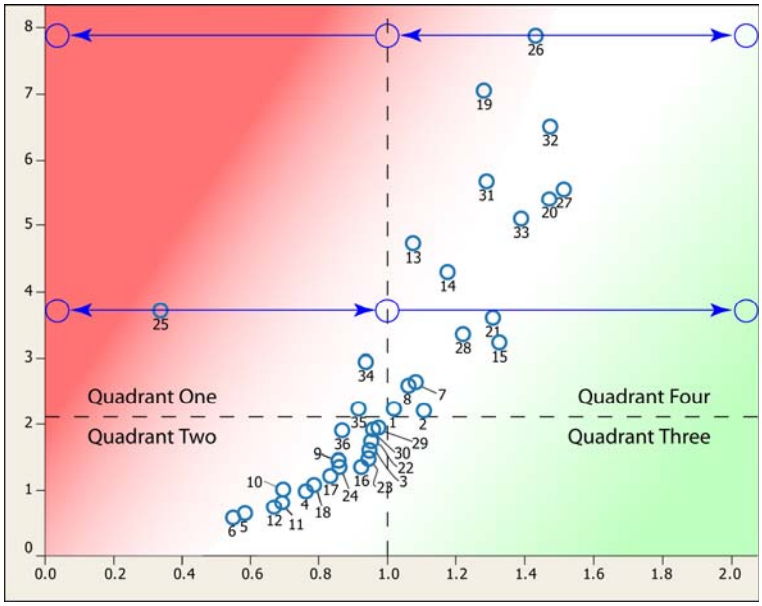


Figure 7. Distribution of ANN Obtained by Plotting the Average EI vs. the Average Solve Rate. In this figure, we have plotted the EI for each of the neural network nodes (numbered according to Figure 3) vs. the solve rate for each node. These values are calculated from the data shown in Figure 3 and quadrants are established based on the mean solve rate (0.98) and EI (2.15) for the entire dataset.

Some of the strategies such as those represented by nodes 5, 6 and 12 are neither efficient (low IE), nor effective (low solve rate) for solving the problem. Relating back to Figure 3, these strategies are characterized by a detailed examination of the problem space, often without solving the problem. Other strategies, represented by nodes 26 or 19, have high solve rates, with limited use of the laboratory tests.

By dividing the data set into quadrants using the average EI and the average solve rate, a more general metric of problem solving can be obtained which we term a Quadrant Value (QV). When new performances are obtained, the QV measure can be constructed through a four step process which can be automated to provide real-time measures as new performances are generated.

- Step 1 – develop strategy categories with ANN,
- Step 2 – calculate an efficiency index for each ANN node,
- Step 3 – develop a quantitative quadrant value (QV) grid from average solve rate, and average EI,
- Step 4 – calculate QV for new student performances.

In applying the steps above for calculating a QV for any particular student performance, suppose a student's performance is classified at node 26 (top right). The effectiveness of this student's performance can be either greater or less than the average value for this node (which is around 1.5). If the student solved the case on the first try then it would still be classified in Quadrant 4. If, however it was missed it would be classified in Quadrant 1. Similarly for a performance at Node 25, if the student solved the case on the first attempt with this strategy it would also be rated as QV4. For an individual student, the QV metric therefore represents his or her proficiency in using resources to solve scientific problems effectively, abstracted across the specific problem sets administered to the student. As described shortly, this metric can be carried forward across problem sets over the course of the school year, and across the middle school grades. When this procedure is applied to a dataset of student performances of *Hazmat*, the distribution in Figure 8 is obtained.

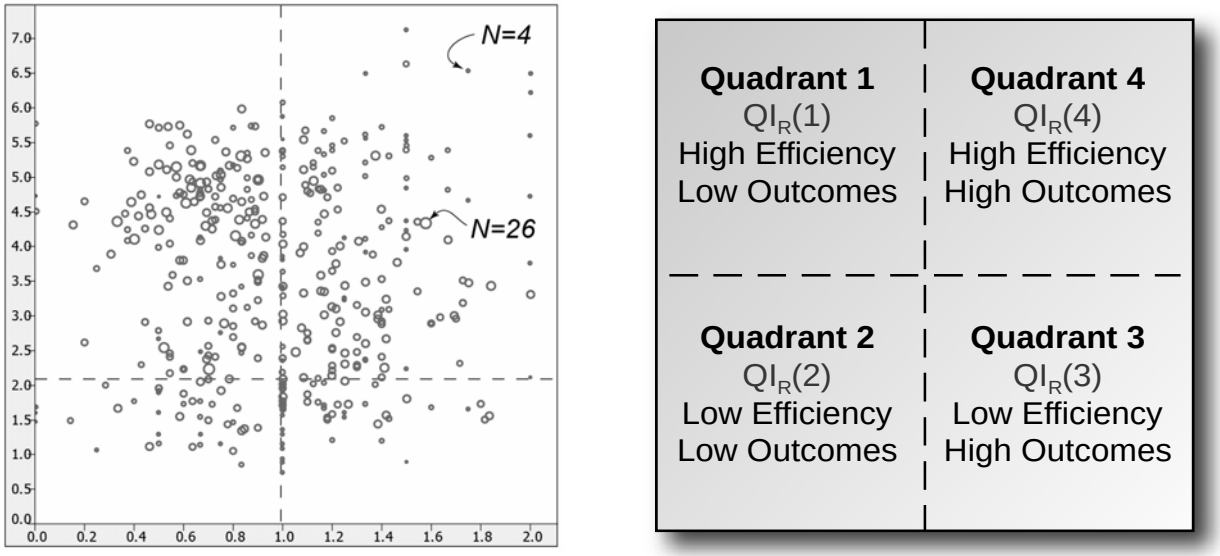


Figure 7. Defining Quantitative Values (QV) of Problem Solving. a) Plot of the average EI of student performances vs. the solve rates where students performed 4-30 cases of middle school IMMEX problems (n~30,000 performances). b) Definition of QV based on the vertex created by the average EI and solved rates for the data.

As shown in Figure 2, *Hazmat* contains cases of different difficulty levels. The most difficult compounds (e.g., the flame test negative compounds), had the lowest solved rates and EI values;

that is, the more difficult the problem, the more data students required to derive a successful outcome. As expected, the correlation of EI and the IRT problem difficulty was strong but negative (r = -.887, p = .000).

To determine if QV values could replicate student's learning progress similar to HMM, we first performed chi-square analysis of QV with the HMM States shown in Figure 6. This analysis indicated that HMM State 1 mapped closest to QV1, the low efficiency HMM State 3 mapped to QV2 and QV3, while the HMM state with the highest solution frequency, State 5, most closely mapped to QV4 (Table 1). (χ^2 = 7568, df= 12, p = .000).

Table 1. Crosstabulation comparisons of HMM states and performance quadrant distributions. Values in the cells represent the percentages of different states in each quadrant.

	HMM State					
QV	**1**	**2**	**3**	**4**	**5**	**Total**
1	44	9	1	36	10	100
2	23	13	32	25	7	100
3	25	15	27	25	8	100
4	35	11	0	33	21	100

Learning trajectories developed from the QV distributions over multiple *Hazmat* performances (Fig. 9) showed that initially most students were rated as QV2 or QV3, both representing low efficiency with variable outcomes, but with experience transited to QV1 or QV4 representing higher efficiency. Similar to the stabilization of the HMM states, the quadrant classifications stabilized after 4-5 cases. Also similar to the HMM state findings, there were significant differences across quadrants for gender (χ^2 = 34.33, df= 3, p = .000) or whether students were working individually or in collaborative groups (χ^2 = 227.45, df= 3, p = .000). These validation studies suggest that the single QV measure converges on many aspects of the ANN performance and HMM progress models.

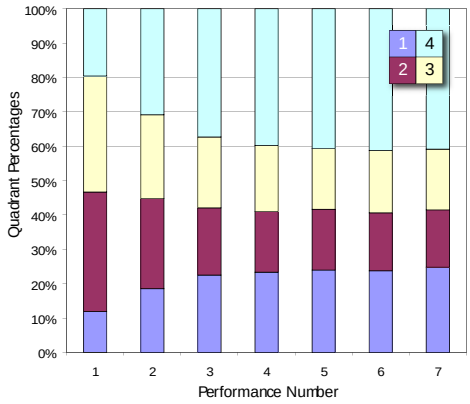


Figure. 9. QV Models of problem solving progress. The quadrant distributions were calculated for ~25,000 *Hazmat* performances. Each stacked bar represents the percentages of students in the different quadrants after different numbers of case performances.

The Final Wrinkle: IMMEX Problem Sets Differ in Difficulty and the EI

All IMMEX problem sets are not of equal difficulty when measured broadly across middle school. Also, from the nature of the denominator in Equation 1, the maximum EI for different problem sets can also vary over a wide range as a greater number of resources available, the

higher the potential EI. While QV is beginning to represent a value-related metric, to satisfy our challenging question we still need to be able to compare across problem sets. We satisfy this need by normalizing the EI and solve rates for each problem set to the average values for the dataset of 5000+ performances. When the classroom data from the different problem sets are normalized this way, the quadrant structures shown in Figure 10 resulted.

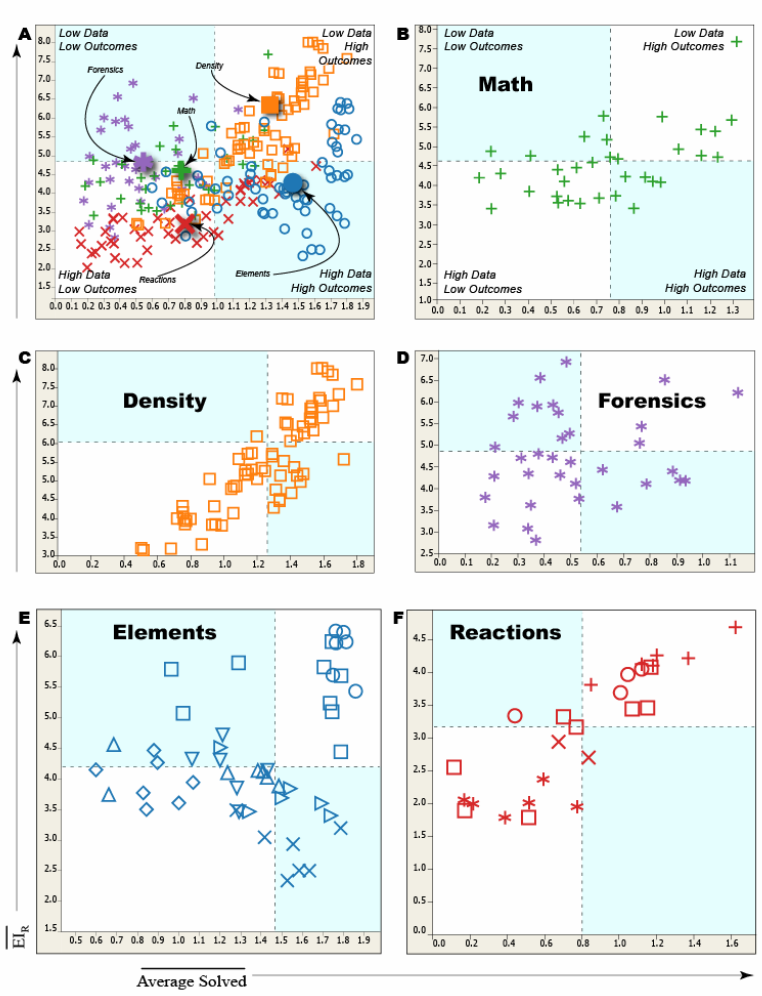


Figure 10. A) Average EI and Solve Rates for Five Middle School Problem Sets.

The mean EI and solve rates for five different middle school problem sets are plotted for over 100 different classrooms. The larger symbols indicate the mean EI and solve rates for each of the problem sets. B-F) The individual quadrant profiles are shown for each of the problem sets.

As shown by the similar shapes in the figure, different classrooms of the same teacher often clustered together on the quadrant maps. We explored this further with two sets of teachers using two different chemistry problem sets, Element Identification (Elements) and Chemical Reactions (RXN). To gauge progress, we plot the EI and solve rates after the first through fifth case performances. For the problem set Elements there were significant differences in the trajectory slopes indicating differential progress, whereas for RXN the classrooms of one teacher started lower than those of the other, but both improved similarly.

Correlations of EI, IRT and QV with California Achievement Test scores

We next sought to determine how a student's problem solving performance over a year correlated with another measure of their ability, the California Achievement Test scores.

A sample of students (N=137 representing ~3500 problem solving performances) performed cases from all five problem sets allowing correlations to be made for IRT, EI and QV. For 119 of these students the California Achievement Test scores in Reading, Language and Math were also available.

A multiple regression analysis was first conducted to evaluate how well the IRT, EI and QV predicted CAT Math scores. The linear combination of the thee measures was significantly related to the standardized scores (*F(3,118) = 24.5, p = .000*). The sample multiple correlation was .571 indicating that approximately 32% of the variance in the CAT scores could be accounted for by these measures. The QV (*r=.173*) and IRT (*r=.326*) scores both contributed significantly (*p=.000*) to the prediction of CAT Math scores while EI was not correlated.

Table 3. Across problem set performance metric comparisons. The dark cells are significant at the .01 level and the lighter cells are significant at the .05 level.

IRT

	Elements	Density	Forensics	Math	Reactions	CAT Reading	CAT Language	CAT Math
Elements	1.000							
Density	0.159	1.000						
Forensics	0.223	0.148	1.000					
Math	0.228	0.052	0.156	1.000				
Reactions	0.171	0.268	0.239	0.180	1.000			
CAT Reading	0.335	0.100	0.479	0.331	0.308	1.000		
CAT Language	0.341	0.132	0.398	0.328	0.346	0.751	1.000	
CAT Math	0.430	0.234	0.389	0.332	0.332	0.675	0.665	1.000

EI

	Elements	Density	Forensics	Math	Reactions	CAT Reading	CAT Language	CAT Math
Elements	1.000							
Density	0.240	1.000						
Forensics	0.173	0.176	1.000					
Math	0.169	0.122	0.241	1.000				
Reactions	0.066	0.128	0.052	0.417	1.000			
CAT Reading	0.137	0.159	-0.073	-0.010	0.127	1.000		
CAT Language	0.085	0.202	-0.141	0.004	0.090	0.751	1.000	
CAT Math	0.279	0.265	-0.047	0.069	0.136	0.675	0.665	1.000

QV

	Elements	Density	Forensics	Math	Reactions	CAT Reading	CAT Language	CAT Math
Elements	1.000							
Density	0.157	1.000						
Forensics	0.002	0.170	1.000					
Math	0.063	0.081	-0.075	1.000				

Reactions	0.168	0.042	0.294	0.146	1.000			
CAT Reading	0.293	0.109	0.447	0.079	0.195	1.000		
CAT Language	0.279	0.275	0.296	0.138	0.227	0.751	1.000	
CAT Math	0.266	0.327	0.301	0.061	0.190	0.675	0.665	1.000

Correlations were then performed for each metric across the five problem sets (Table 3). The strongest and most frequent correlations were outcome-based across the IRT scale and these ratings also correlated with scores on the three CAT tests. The EI correlations across problem sets were more variable suggesting that students were not necessarily using the same general strategic approaches across the different problem sets. The EI measures were also less well correlated with the standardized test scores. The QV for each problem set which (loosely) combines EI and IRT into a single measure was correlated with the standardized test scores, and was variably correlated across problem sets.

We examined these findings further by hypothesizing that if teachers were preparing their students well for problem solving a correlation should exist between problem solving metrics and the California Achievement Test (CAT) scores. For these studies the student population consisted of middle school students (N=775) from multiple classes of six teachers where the CAT mathematics scores (M-SS) were also available. The students performed 4-6 different IMMEX problem sets (between 25-60 different cases total) over a year's time. The QV measure was regressed for all performances against the M-SS test scores. A correlation between QV and the M-SS scores was seen for some teachers, but not for the others (Figure 11). This was not due to differences in the overall achievement levels of the students in the different classes; in fact, the two highest achieving classes (by the M-SS scores) were the most poorly correlated. In the lower performing classes, most students are at QV = 2. These are students who are looking extensively at the data but repeatedly failing to solve the problems during the school year, suggesting that their teachers are not preparing them to carefully select and synthesize data (Stevens & Thadani, submitted).

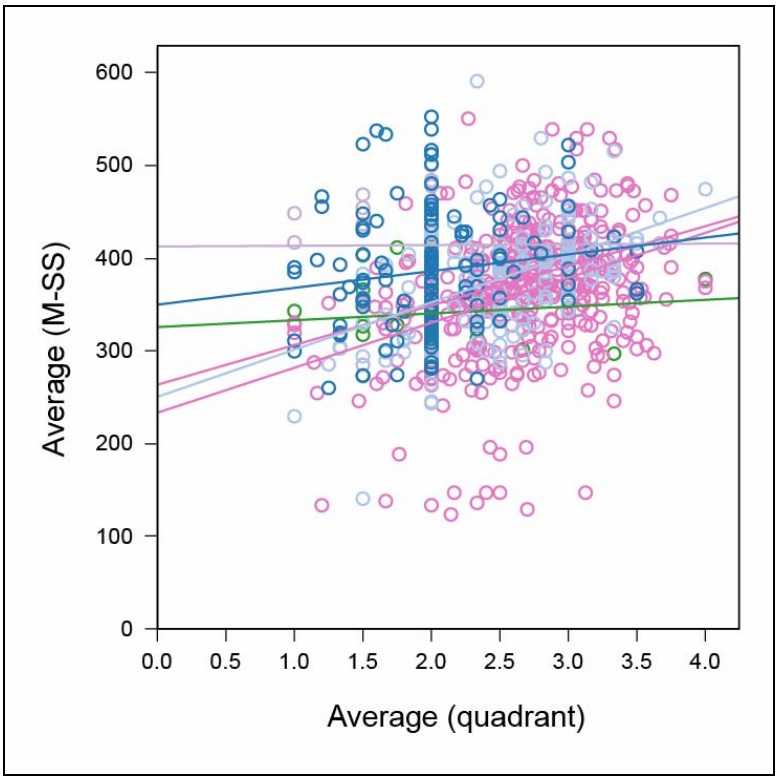


Figure 11. Student Quadrant Values from six different middle school teachers. The students (n=775) from the classes of six teachers performed between 25 and 60 IMMEX problem cases during a year. The average QV is plotted vs. the student's Math CAT scores.

Discussion

The goal of this study was to develop and begin to validate a value-derived measure of scientific problem solving that could be applied to, and compared across multiple problem solving situations. The study was motivated by the accumulating data on classroom problem solving that is being provided by the expanding library of IMMEX simulations. These performances are now being collected from hundreds of students in a longitudinal manner across different domains, and over semesters or years.

The relatively simple EI and QV constructs --when paired with individual student outcomes and examined across domains and school organizations--appear to have potential as rapid and meaningful indicators of problem solving on close-ended tasks. They are derived from real-world constructs, are easily reportable across educational systems, and can be normalized across tasks and domains. This generality is unusual as the performance data from most problem solving tasks is highly specific and difficult to aggregate across tasks or domains. Despite this generality, the measures replicate the student learning dynamics and effects of contextual influences of prior nominal modeling approaches.

The resource utilization modeling is also extensible and by including other constraints such as time, risk, or cost, the plot of EI vs. solved rate could be expanded into other dimensions by substituting the denominator of Equation 1 with the time used / time available, or the costs / funds available calculations, etc. As all problem solving, as opposed to problem posing, involves

constraints the analysis could be applied to situations other than hypothetical-deductive problem solving.

Such an analysis and reporting of problem solving efficiency could support learning at several levels. If diagrams similar to Figure 7 were available to students in real-time (Stevens & Soller, 2005), with their performances highlighted, they could provide useful formative feedback through comparisons with other students in the class (or school, depending on the display aggregation). The quadrant designations could also help direct the targeting of specific suggestions to students before a subsequent case is attempted to enhance motivation (Conati & Zhao, 2004) or provide pedagogical support (Arroyo et al, 2004).

For teachers, the EI and QV measures can be used to track class progress and help them self-monitor their teaching. Also, the sensitivity of the quadrant distributions to different task representations by teachers suggests applications for targeting training in ways that address trends in class-level problem solving. Used in these ways, the EI and QV measures may help re-think the ways scientific problem solving is systemically assessed in the classroom, and how the impact of teaching these skills becomes quantified.

References

Atkin, M. J., Black, P., Coffey, J. (Eds.) (2001). Classroom Assessment and the National Science Education Standards. National Academy Press, Washington, DC.

Arroyo, I., Beal, C. R., Murray, T., Walles, R., Woolf, B. P. (2004). Web-Based Intelligent Multimedia Tutoring for High Stakes Achievement Tests. 468-477. James C. Lester, Rosa Maria Vicari, Fábio Paraguaçu (Eds.): Intelligent Tutoring Systems, 7th International Conference, ITS 2004, Maceiò, Alagoas, Brazil, Proceedings. Lecture Notes in Computer Science 3220. Springer 2004.

Augustine, N. R. (2005). National Academies Committee on Prospering in the Global Economy of the 21st Century, Rising Above The Gathering Storm: Energizing and Employing America for a Brighter Economic Future. National Academies Press, Washington, D.C.

Bennett, R. E. (1998). Reinventing assessment: speculations on the future of large-scale educational testing. Princeton, NJ: Educational Testing Service Report.

Conati, C., and X. Zhao. (2004). Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectiveness of an Educational Game. In *Proceedings of the 9th International Conference on Intelligent User Interface*, pages 6–13. ACM Press.

Ericsson, K. A. (2004). Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. Academic Medicine. 79 (10), S70-S81.

Fennema, E. Carpenter, T., Jacobs, V., Franke, M., and Levi, L. (1998). Gender differences in mathematical thinking. Educational Researcher, 27, 6-11.

Haider, H., and Frensch, P.A. (1996). The role of information reduction in skill acquisition. Cognitive Psychology 30: 304-337.

Kohonen, T. (2001). Self organizing maps. (3rd ed.) Springer Series in Information Sciences, Vol. 30, Springer Heidelberg, Germany.

Lajoie, S. P. (2003). Transitions and Trajectories for Studies of Expertise. Educational Researcher. 32 (8), pp. 21–25.

Mayer, R. E. (1998). Cognitive, Metacognitive, and Motivational Aspects of Problem Solving. Instructional Science 26 (1-2), pp. 49-63.

Messick, S. (1989) 'Validity' in Robert L. Linn (ed.). Educational Measurement 3rd Edition (American Council on Education & Macmillan, New York).

Murphy, K., (2004). Hidden Markov Model (HMM) Toolbox for Matlab. Available online at: http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

OECD (2004). Problem Solving for Tomorrow's World First Measures of Cross-Curricular Competencies from PISA 2003. Claire Shewbridge and Andreas Schleicher (eds) Programme for International Student Assessment, Organization for Economic Co-operation and Development (OECD). Paris, France.

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds.). (2001). Knowing what students know: The science and design of educational assessment. Washington, DC: National Academy Press.

Reckase, M.D., (1998). The Interaction of Values and Validity Assessment: Does a Test's Level of Validity Depend on a Researcher's Values? Social Indicators Research 45: 45-54.

Schunn, C.D., & Reder, L.M. (2001). Another source of individual differences: Strategy adaptivity to changing rates of success. Journal of Experimental Psychology: General, 130, 59–76.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy Implementation and Cognition: Reframing and Refocusing Implementation Research. Review of Educational Research, 72 (3), pp. 387-431.

Stevens, R. H., & Thadani, V. (2006). A Bayesian Network Approach for Modeling the Influence of Contextual Variables on Scientific Problem Solving. Intelligent Tutoring Systems, 8th International Conference, Jhongli, Taiwan, Proceedings. Lecture Notes in Computer Science 4053. Mitsuru Ikeda, Kevin D. Ashley, Tak-Wai Chan (eds.): Springer. pp. 71-84.

Stevens, R. H., & Thadani, V. (submitted). A value-based approach for quantifying scientific problem solving effectiveness within and across educational systems. Instructional Science.

Stevens, R., & Casillas, A. (2006). Artificial Neural Networks. In R. E. Mislevy, D. M. Williamson, & I. Bejar (eds), Automated Scoring of Complex Tasks in Computer Based Testing: An Introduction. Lawrence Erlbaum, Mahwah, NJ. (pp. 259-312).

Stevens, R., Johnson, D. F., & Soller, A. (2005 Spring). Probabilities and Predictions: Modeling the Development of Scientific Competence. Cell Biology Education 4 (1) pp 42–57.

Stevens, R., Soller, A., Cooper, M., and Sprang, M. (2004) Modeling the Development of Problem Solving Skills in Chemistry with a Web-Based Tutor. Intelligent Tutoring Systems.

J. C. Lester, R. M. Vicari, & F. Paraguaca (eds). Springer-Verlag Berlin Heidelberg, Germany. 7th International Conference Proceedings (pp. 580-591).

Stevens, R.H., and Soller, A (2005). Machine learning models of problem space navigation, The influence of Gender. ComSIS 2 (2): pp. 83-98.

VanLehn, K., (1996). Cognitive Skill Acquisition. Annual Review. Psychology. 47: 513-539