Tech., Inst., Cognition and Learning, Vol. 5, pp. 325-337
Reprints available directly from the publisher
Photocopying permitted by license only

Quantifying Student's Scientific Problem Solving Efficiency and Effectiveness

RONALD H. STEVENS[1*] AND VANDANA THADANI[2]

[1] *UCLA IMMEX Project, Culver City, CA, USA*
[2] *Loyola Marymount University, Los Angeles, CA, USA*

Using online problem-solving tasks and machine learning tools, a measure has been developed to quantify the effectiveness and efficiency of students' problem solving strategies. This measure can be normalized across problem solving tasks allowing the efficiency of problem solving to be measured across individuals, classes, schools and science domains. This extensible approach has relevance for helping teachers to teach, students to learn, and administrators to make intelligent, data-driven decisions via documentation of students' problem solving progress.

Keywords: Problem Solving, Assessment, Artificial Neural Networks.

INTRODUCTION

Promoting students' ability to effectively solve problems is viewed as a national educational priority (Augustine, 2005). However, teaching problem solving through school-based instruction is no small task and many teachers may find it difficult to quantify and assess students' strategic thinking in ways that can rapidly inform instruction.

Part of the assessment challenge is cognitive. Strategic problem solving is a complex process with skill level being influenced by the task, the experience and

*Corresponding author: immex_ron@hotmail.com

knowledge of the student, the balance of cognitive and metacognitive skills possessed by the student and required by the task, gender (Fennema et al, 1998), ethnicity, classroom environment and overall ability constructs such as motivation and self efficacy (Conati, 2002). It is further complicated as the acquisition of problem solving skills is a dynamic process characterized by transitional changes over time as experience is gained and learning occurs (Alexander, 2003).

Other challenges are observational in that assessment of problem solving requires real-world tasks that are not immediately resolvable and that require individuals to move among different representations. Assessment also requires that performance observations be made that are revealing of the underlying cognition and can also be effectively reported (Pellegrino et al, 2001). Tasks meeting these criteria are becoming more common in science classrooms, and with the increasing technology capabilities, the cognitive granularity of the assessments can become detailed (Heffernan & Koedinger, 2002). However, granularity can come at the cost of generalization, ease of implementation, and clarity of understanding. Finally, there are the technical challenges of speed and scale; speed relating to how rapidly valid inferences can be made and reported from the performance data, and scale in how multiple content domains and grade levels can be effectively compared (Bennett, 1998).

While the challenges for developing problem solving assessments 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, including evaluation, policy development, grading, and feedback for improving teaching and learning (Bennett, 1998, Atkin, et al. 2001, Spillane et al., 2002).

There are a number of challenges in building assessments that can provide useful feedback for any kind of learning, much less problem solving. First, the findings from these assessments should be very quickly available. For instance, performance assessments—while useful for assessing higher levels of thinking—might take middle and high school teachers a week or more to score. Second, results from the assessment should be clearly linked to interventions that teachers might use with individual students or the class as whole. And third, data should be comparable across learning events so that students and teachers can track growth (or lack of growth) in learning.

In this paper we focus on trying to describe diverse problem solving results in terms of outcomes that are comparable across learning events and different problem solving tasks (Mislevy et al, 1999). In doing so, we take an approach that combines the efficiency of the problem solving solution, as well as its

correctness. These are components of most problem solving situations and have close ties to the pattern oriented modeling (Grimm et al, 2005) cost/benefits (Zeidner, 1991) and neuroeconomics (Stigler, 1961) literatures where they have been applied across diverse domains and disciplines in business and healthcare (O'Conner et al, 2006). In essence we are seeking to measure the maximizing of outcomes with the minimal consumption of resources.

DATASET

The IMMEX™ Project 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. Students access resource data such as experimental results, reference materials, advice from friends and / or experts, etc. to solve the problem. Their exploration of these resources is unconstrained in that they choose how many (or few) resources they use and in what order. Every IMMEX problem set includes a number of cases—parallel versions of the problem that have the same interface and resources, but present different unknowns, require different supporting data and have different solutions. The IMMEX database serializes and mines timestamps of which resources students use. We then use machine-learning tools to build layers of student performance models that are used to assess student problem solving skills (Stevens et al., 2004, Stevens et al., 2005, Stevens et al., 2006). The dataset for this study included 154 classes from 64 teachers (mostly middle school) across 27 schools, with 79,146 problem performances.

RESULTS

The central question driving this research is, "What is a suitable description of problem solving efficiency and correctness that can capture important cognitive and performance information about individual problem solving, yet provide rapid and meaningful comparisons within and across educational systems and science domains?" *Correctness* can be determined by assessing whether or not an outcome was successful, and this may be extended by Item Response Theory Analysis (IRT) estimates of difficulty (theta), to yield more refined performance estimates when cases of varying difficulties exist. *Efficiency* is another important component of problem solving which has been

somewhat more difficult to assess as constraints are involved, such as time, risks, costs, benefits, and available resources.

We postulate 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 solving the problem, 1 point for solving the problem on a second attempt, and 0 pts for missing the solution).

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

Not all resources available in a problem space are equally applicable to the particular problem at hand, and different combinations of resources will have different strategic value within the contexts of different problems. Thus, estimates of the quality of resources used are also required. This qualitative dimension is derived from strategic classifications derived from unsupervised artificial neural network (ANN) clustering of performances.

Artificial Neural Networks Supply Estimates of the Strategic Quality

The most common student approaches (i.e. strategies) for solving IMMEX problems 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, Stevens et al., 2005). The result is a topological ordering of the neural network nodes (generally 36) according to the structure of the data where geometric distance between nodes becomes a metaphor for strategic similarity. The strategic complexity of each node is visualized by a histogram showing the frequency of items selected for student performances classified at that node (Figure 1). 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. In Figure 1A there is also a composite ANN topology map of performances generated during the self-organizing training process. 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. 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.

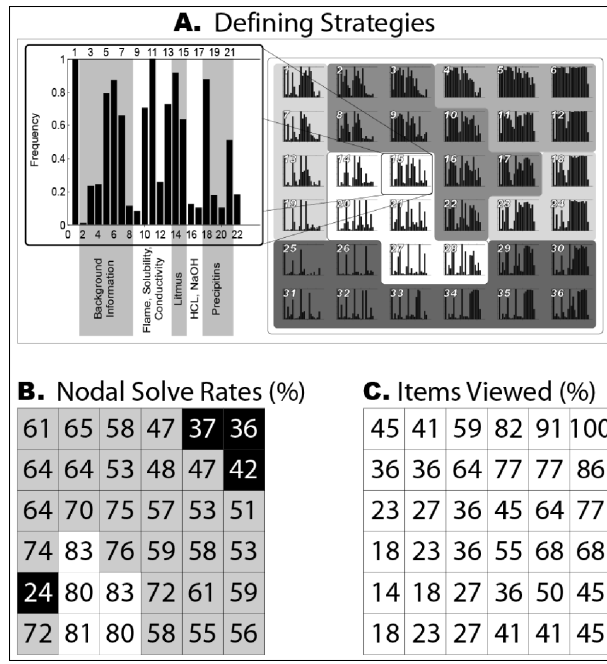

FIGURE 1
Sample Neural Network Nodal Analysis for Identifying Strategies. A) The selection frequency of each action (identified by the labels) is plotted for the performances at node 15, characterizing the performances for this node and relating them to performances at neighboring nodes. This figure also shows the item selection frequencies for all 36 nodes where the nodes are numbered in rows, 1-6, 7-12, etc. B) The solution rate for each node is listed with the lowest solved rates in black and the highest in white. C) The values indicate the proportion of tests selected during performances at each node.

As shown in Figure 1B, not all strategies result in the same outcomes. Some of the strategies such as those represented by nodes 5, 6 and 12 are neither efficient (many items selected), nor effective (low solve rate) and 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. The proportion of tests selected at each node is then calculated using 50% as a cutoff value. Thus from the ANN we can derive the efficiency components needed for the EI measure. 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.

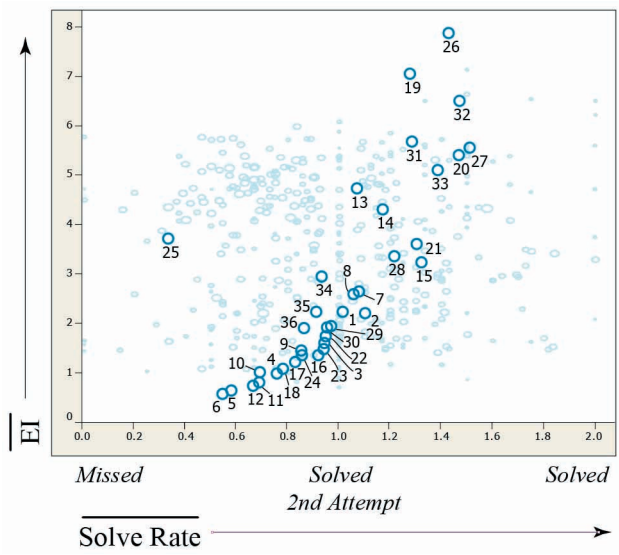

FIGURE 2
Relationships Between EI and Solve Rates The numbered symbols plot the average EI and solve rate for each of the 36 nodes of the ANN from Figure 1. The lighter symbols in the background show the average EI and solve rate values for ~30,000 students who performed between 4 and 30 IMMEX cases.

When the EI for the 36 nodes of the ANN are plotted against the average solve rate, the distribution in the foreground of Figure 2 is obtained. Strategies represented by nodes 5, 6 and 12 (lower left hand corner) represent poor

outcomes with extensive resource utilization whereas those represented by nodes 26 and 32 are very effective and efficient. Node 25 at the left center of this figure most likely represents guessing as the solve rate is very low and only a few tests are ordered (from Figure 1). When students perform multiple cases in a problem set, an average placement on this map can be generated by determining the ANN node represented by each strategy, and averaging the associated EI values, and then plotting this value vs. the average solve rate. The lighter symbols in Figure 2 show such averages for ~30,000 students, and illustrate the diversity of strategic efficiency and outcomes.

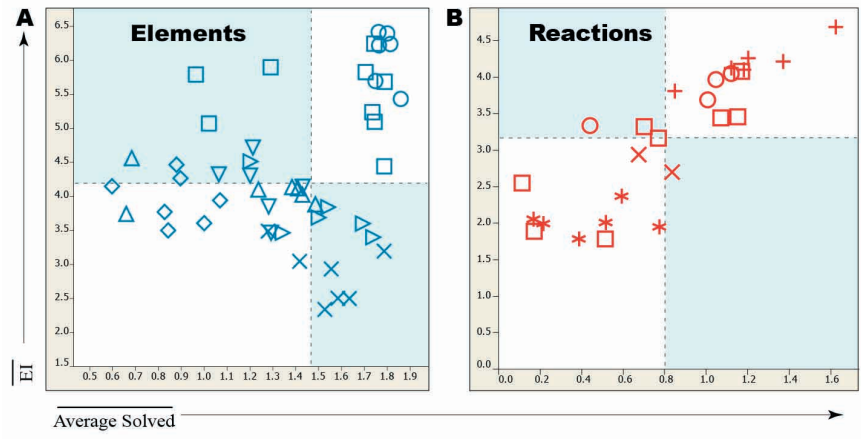

FIGURE 3
Middle School Classroom Distributions of EI and Solve Rate. The student EI and Solved values on the middle school chemistry problem sets (Elements and Reactions) were aggregated for 52 classes of seven teachers. The symbol types denote the classrooms of each teacher. The dotted lines indicate the overall EI and Solve Rate averages, and partition the strategy space into four quadrants.

Using the above approach, it is also possible to aggregate the data for different classrooms and different teachers. Figure 3 shows such an aggregation for two middle school problem sets. Here, each symbol represents the classrooms of one teacher. As shown by the similar shapes in the figures, different classrooms of the same teacher often clustered together on the quadrant maps.

Given the across-classroom performance differences, a teacher-by-class comparison of student progression was then performed using four teachers.

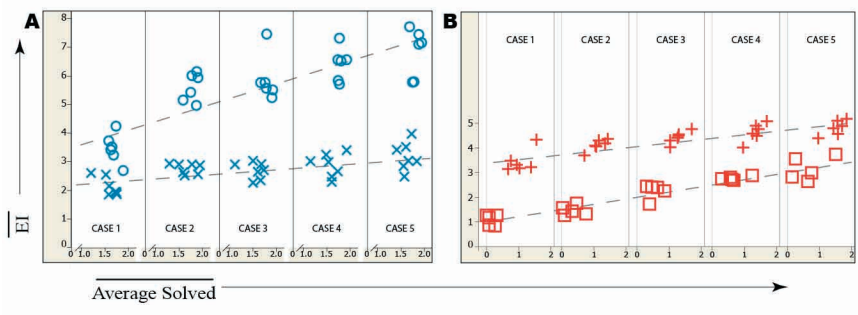

FIGURE 4
Student Improvements in Classroom EI and Solve Rate with Practice. The EI and solved rates of the classes of two teachers for Elements (X, O) and for Reactions (+, □) are plotted for the first 5 case performances for each problem set. The dotted lines plot the class means for the different teachers.

On the problem set *Elements* (Figure 4A), all classes of both teachers improved their average solve rates with practice, but the classes of one teacher (O) showed greater strategic improvement that did the classes of the other (X). A different progress pattern is shown in Figure 4B for the *Reactions* problem set where the classes of the two teachers being analyzed differed in both the starting EI and solved rates, but improved across both dimensions at similar rates on subsequent cases. These results suggest a consistent teacher component to student's strategic development.

Combining Strategic Efficiency and Correctness into a Single Quantitative Value

Figure 3 also suggests a way of generalizing student performance further to provide a single value for the student position on the efficiency plots. The shading in the plot indicates quadrants defined for the average solve rate and EI of the problem sets, and such plots can be generated for any of the dozens of IMMEX problem sets being used. A Quantitative Value (QV) that combines strategic efficiency and correctness can be assigned to these quadrants that help describe and generalize the problem solving efficiency and outcomes across problem sets. For instance, the upper left corner contains approaches with low outcomes, but high efficiency, i.e. guessing. We assign these a value of '1'. The lower left corner nor effective approaches but suggests that students are at least trying. We assign these a value of '2'. Students in the lower right corner are not

efficient, but they are effectively solving the problems; we assign these performances a value of '3'. Finally, students in the upper right corner are both efficient and effective and get assigned a value of '4'.

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 generated across problem sets over the course of the school year, and across the middle school grades. By normalizing the vertex of the quadrant to the average EI and average solve rate for each problem set, it also becomes possible to compare QV's across problem sets (Stevens, 2007).

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

We next determined 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 five problem sets spanning the domains of chemistry, math, and biology, allowing correlations to be made for IRT, EI and QV (Stevens, 2007). For 119 of these students the California Achievement Test scores in Reading, Language and Math were also available. Using these aggregated values, a multiple regression analysis was conducted to evaluate how well the IRT, EI and QV predicted CAT Math scores. The linear combination of the the measures was significantly related to the standardized scores ($F(3,118) = 24.5,\ p < .001$). The sample multiple correlation was .57 indicating that approximately 32% of the variance in the CAT scores could be accounted for by these measures. The QV ($r=.17$) and IRT ($r=.32$) scores both contributed significantly ($p<.001$) to the prediction of CAT Math scores while EI was not correlated.

We examined these findings further by hypothesizing that if teachers were preparing their students well for problem solving a moderate positive 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 attempted to solve 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 5). 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 M-SS performing classes, most students were at QV = 2. These were students who appear to be looking extensively at the data but repeatedly failing to solve the problems during the school year, suggesting that their teachers were not preparing them to carefully select and synthesize data.

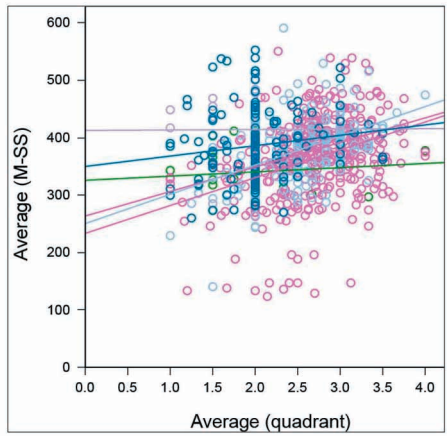

FIGURE 5
Student Quadrant Values from Six Different Middle School Teachers. The students (n=775) from the classes of six teachers performed between 25 and 60 IMMEX problem cases during a year. The average QV is plotted vs. the student's Math CAT scores.

DISCUSSION

The goal of this study was to develop and begin validating an assessment of scientific problem solving that could provide teachers and students with feedback about the latter's problem solving skills. Our aim was to develop measures that could capture problem solving process as well as outcomes, could provide meaningful comparisons within and across science domains, and could simultaneously acknowledge the strong contributions of content knowledge to problem solving. Moreover, we wanted the findings from these assessments to be rapidly available and clearly linked to interventions that teachers might use with individual students or the class as whole.

The EI and QV measures—which have only become possible because of the expanding library of IMMEX simulations and the vast number of performances collected longitudinally from students in different domains—have these properties.

These relatively simple constructs—when paired with individual student outcomes and examined across domains and school organization—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.

The aggregation and generality of these measures are also their greatest shortcoming. First, this approach tries to express a very complex process in an uncomplicated understandable way and in doing so it relies on the aggregation of performances into strategies (ANN classifications). While the co-clustering of performances into the same group occurs with a high frequency (90-95%) when multiple neural networks are trained with the same dataset, the co-clustering is not absolute (Stevens & Casillas, 2006). Depending on the nodal locations on the ANN topology map this may or may not significantly affect the QV. Next, the quadrant boundaries are set solely from the mean values for solved and EI, which while empirical, we feel is justified from the large datasets. Finally, the numbering of the quadrants is unconventional as traditionally they are numbered from the upper right quadrant. From analysis of a large number of problem sets, in combination with classroom observations we feel that our 1-4 numbering better reflects the degree of problem-solving sophistication of students; guessing is to be valued less than extensive navigation of the problem space (quadrants 1 and 2), and efficient as well as effective navigation is valued more than just effective. Nevertheless, the measures replicate many of the student learning dynamics, and effects of contextual influences such as gender and collaborative learning effects, of prior nominal modeling approaches (Stevens et al, 2004; Stevens et al, 2005).

If made available to teachers and students in real time, these measures would enable both to monitor problem-solving progress across problem sets, semesters, and even years. Such an analysis and reporting of problem solving efficiency could support learning at several levels. First, since the quadrant designations are based on both strategy and outcome, they also embed implications for strategic change: Students low in outcomes and high in efficiency (guessers) must explore the resources more extensively; those low in outcomes and low in efficiency are failing to use the results of their search effectively; and those high in outcomes but low in efficiency should start thinking about the relative utility of resources. Although students' designations at each of these levels might be determined by more than one cause (motivation, understanding), their performance data narrows the range of teacher interventions (or students' strategies for self-correction) that have to be entertained.

Teachers could also use the QV measure to track class progress as a means of monitoring their own teaching. Class averages, particularly when stable across problem sets or cases, can indicate that teaching needs to change, perhaps to provide increased motivational or pedagogical support. Also, between-teacher differences in quadrant distributions suggests applications for targeting teacher professional development in ways that address trends in class-level problem solving. Finally, the automated nature of this assessment technique provides an opportunity to assess students rapidly, at a scale that has traditionally only been possible through standardized testing.

The model used to derive EI is also extensible to other problem solving situations where there are constraints like costs, time, risks, etc. This could be easily accomplished by substituting the denominator of Equation 1 with the time used / time available, or the costs / funds available. Because all problem solving, as opposed to problem posing, involves constraints the analysis could be applied to situations other than hypothetical-deductive 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.

Supported in part by grants from the National Science Foundation (NSF-ROLE 0528840, DUE Award 0126050, ESE 9453918) and the U.S. Department of Education (R305H050052).

REFERENCES

Alexander, P. A. (2003). The Development of Expertise: The Journey from Acclimation to Proficiency. Educational Researcher, 32 (8), 10-14.

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

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. (2002). Probabilistic Assessment of User's Emotions in Educational Games. Journal of Applied Artificial Intelligence. 16(7-8), 555-575.

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

Grimm, V., Revilla, E., Jeltsch, F., Mooij, W., Railsback, S., Thulke, H-H., Weiner, J., Wiegand, T., DeAngeles, D. (2005). Pattern-oriented Modeling of Agent-based Complex Systems: Lessons from Ecology. Science, 310, 987-991.

Haider, H., and Frensch, P.A. (1996). The Role of Information Reduction in Skill Acquisition. Cognitive Psychology 30: 304-337.

Heffernan, N. T., & Koedinger, K. R. (2002). An Intelligent Tutoring System Incorporating a Model of an Experienced Human Tutor. Proceedings of the Sixth International Conference on Intelligent Tutoring Systems, Biarritz, France.

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

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999). A Cognitive Task Analysis, with Implications for Designing a Simulation-based Assessment System. Computer and Human Behavior, 15, 335-374.

O'Connor, A., Mulley, A., Wennberg, J., (2006). Standard Consultations are Not Enough to Ensure Decision Quality Regarding Preference-sensitive Options. JNCI, 95: 570-571.

Pellegrino, J., Chudowsky, N, Glaser, R. (2001). *Knowing What Students Know*. National Academy Press

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., & 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). 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. (2007). A Value-Based Approach for Quantifying Student's Scientific Problem Solving Efficiency and Effectiveness Within and Across Educational Systems (in press).

Stevens, R.H., and Soller, A. (2005). Machine Learning Models of Problem Space Navigation, The Influence of Gender. ComSIS 2 (2): pp. 83-98.

Stigler, G. J. (1961). The Journal of Political Economy, volume 69, page 213.

Zeidner, J., & Johnson, C.C. (1989). The Economic Benefits of Predicting Job Performance (IDA Paper P-2241). Alexandria, VA: Institute for Defense Analysis.