Improving Educational Assessment By Incorporating Confidence Measurement, Analysis of Self-Awareness, and Performance Evaluation: The Computer-Based Alternative Assessment™ (CBAA™) Project

Dr. Jody Paul

jody@acm.org

> What do you know?
> How sure are you?
> How sure should you be?
> Can you use what you know?

The Computer-Based Alternative Assessment (CBAA) project is concerned with finding ways to answer these questions for all levels of students from elementary through university. Project activities involve developing and testing alternative means to enhance educational assessment with *confidence measurement, analysis of self-awareness,* and *performance evaluation.* The research demonstrates how computer technology can help to achieve more revealing, more accurate, and more useful assessments of students' knowledge while maintaining the benefits of standardization. This paper describes the philosophy, architecture, design and implementation of the confidence-measurement, analysis of self-awareness, and performance-assessment components of the CBAA system.

Introduction

Although assessment is central to the educational process, only a small fraction of the potential benefit is typically obtained. The primary impediment to realizing greater benefit has been the infeasibility of implementing more effective alternatives in the resource-limited settings typical of modern educational environments. The Computer-Based Alternative Assessment (CBAA) project demonstrates the exploitation of computer and hypermedia technologies to overcome serious limitations of traditional assessment methods. The CBAA project vehicles and methods represent the synthesis of research and practice in Experimental Computer Science and Engineering, Cognitive Science, Education, and Human-Computer Interaction.

The fundamental goal of the CBAA project is to improve the value of assessment by addressing each of the following: provide more useful experiences for students, achieve more valid assessments of students' knowledge, produce comparable measures, and assess students' abilities to apply knowledge in solving practical problems. While the project seeks to address these in all settings, the primary target is the institutional education environment,

including elementary, secondary and post-secondary institutions. Key problem characteristics of such contexts are large student populations and limited labor resources. The main potentially mitigating factor in these settings is the increased presence of computers with graphical user interfaces. Caution is needed, however, because computerization sometimes diminishes impact and thus is not always the best implementation choice. (Preliminary investigation of the use of CBAA concepts and tools in an early childhood education environment shows promise for use in discovery learning and assessment. Further study is necessary to determine the appropriateness and applicability of CBAA methods to this age group.)

The combination of goals and environmental characteristics suggests the following desiderata for assessment processes and vehicles. The vehicles should be multi-modal, addressing visual, aural, and kinesthetic dimensions. The experience should be engaging and should address both left- and right-brain activation. Personalized feedback should be provided, with rapid turn-around where pedagogically appropriate. Concerns of producing comparable measures and achieving cost-effectiveness require assessment to be standardized. Compared to the prevailing testing procedures and methods, assessment should be more revealing, with lower administrative overhead.

The most prevalent vehicle for standardized testing is the traditional multiple-choice test. This vehicle exhibits many desirable characteristics, such as standardization, automated scoring, and low administrative cost. Unfortunately, traditional scoring, which treats students' responses as absolute (effectively a 0 and 1 based probability distribution), begs the question: "Is a student's knowledge black and white?" How can a student express belief in the *likelihood* that an alternative may be correct? Further, how can a student's *ability to carry out a process* be traced and evaluated? Addressing these questions requires going beyond traditional multiple-choice testing techniques.

The CBAA project demonstrates that improved assessment is achievable with the introduction of an architecture that addresses the design of cost-effective confidence-measuring and performance-testing vehicles using computers found in typical educational settings. Confidence measurement is implemented using the CBAA Triangle; performance testing by using interactive simulation. Enhancing assessment with these components makes it possible to develop a more comprehensive picture of students' knowledge and meta-knowledge.

The additional information made accessible by these components permits discrimination between finer-grained states of knowledge, disclosure of students' ability to apply their knowledge, and effective determination of how aware students are of their own knowledge states. The base information provides a better composite indication of students' content knowledge. Calculation of a realism function determines the degree to which students appear to overvalue or undervalue their knowledge. The determined lack of realism can also be used to provide a better composite indicator by computationally adjusting for reporting bias. Detected response patterns provide additional information useful in diagnosing content misconceptions and learning difficulties.

The following section provides an overview of educational assessment and the use of hypermedia technology to enhance assessment. This is followed by a discussion of confidence measurement, a presentation of the CBAA confidence-reporting vehicle, an in-depth treatment of relevant scoring systems, and an exposition of the CBAA scoring system. Next, the CBAA confidence measurement architecture is presented, along with guidelines for preparing assessment materials and a discussion of how to use the collected data to determine students' knowledge and self-awareness. This is followed by a brief introduction to the performance assessment aspects of the CBAA project. Next, a discussion of the use of cognitive models for analysis of both confidence-measurement and performance-assessment data is presented. This paper concludes with a statement of work-in-progress.

Assessment

The educational experience can be enhanced by using assessment methods as techniques for evaluation and as guides for instructors and administrators in curriculum design and teaching methods [1]. Unfortunately, traditional standardized assessment methods do not discriminate between finer-grained states of knowledge nor do they adequately reflect the ability of students to *apply* what they've learned. In addition, since the assessment instrument significantly influences instruction, alternative assessment methods are needed to better address fundamental educational goals. Past attempts to address these problems and goals on a large scale have proven infeasible primarily due to the high costs of providing adequate, standardized materials and controlled, responsive environments. The CBAA project demonstrates alternatives that exploit the characteristics of modern hypermedia-capable computer systems to achieve the desired goals in a cost-effective way.

Value of Assessment

Direct contributions of assessment include: establishing and revealing status ("knowing what you know"), diagnosis of weaknesses ("knowing what you don't know"), comparative assessment with larger populations (answering "Where do I stand?"), assimilation of knowledge into internal cognitive frameworks ("pulling it all together"), and the exercise of higher-order cognitive abilities (such as application of knowledge in alternate contexts and synthesis of discrete concepts). Accurate self-assessment is also essential to support metacognition, which empowers students to regulate and control their own learning and for which laboratory studies show positive correlations with studying and achievement [2, 3].

Indirect contributions, typically viewed in the context of instructors and administrators, include establishing and revealing students' knowledge states for problem diagnosis (of individual students, instructors, institutions, curricula, or instructional methods), grading, certification, and value-added measurement (such as required for accreditation and funding).

Standardization vs. Validity

Standardization, necessary to meet these objectives, has traditionally been at odds with the goals of accurately revealing finer granularity of knowledge state or the ability to utilize knowledge appropriately. For example, multiple-choice tests, the most widely-used method, "can mask a large variety of states of knowledge and can introduce guessing on the part of the student" [4]. Further, such tests are of limited value in determining students' higher-order cognitive skills, such as the ability to apply what they have learned in different problem-solving contexts [5, 6]. Resnick and Resnick [7] report that the nature of standardized tests now in use is "fundamentally incompatible" with the goal of improving students' higher-order abilities and that alternative assessment methods are needed.

Resource Limitations

Although traditional assessment methods fail to adequately achieve the desired objectives [8], alternatives that use traditional assessment techniques have prohibitive resource requirements. For example, achieving timely, customized feedback requires an impractical student-teacher ratio. Likewise, providing all students with complete, controlled laboratory environments is often costly (e.g., expensive instruments or materials), sometimes dangerous (e.g., hazardous materials or dangerous procedures), or otherwise infeasible (e.g., hypothetical or extra-terrestrial environments).

Teaching to the Test

Another significant aspect of assessment is that it drives instruction ("teachers teach to the test"). This has a subtle but profound impact: assessments that do not involve higher-order cognitive abilities, such as problem solving, do not encourage teachers to emphasize those abilities. Evidence suggests that because typical uses of standard assessment methods do not address these levels of cognition, such abilities are neglected in general [7, 9] and that there exists the need for a "reformulation of assessment to help not hinder the effort to teach thinking" [9].

Hypermedia Technology and Assessment

The use of hypermedia-capable computer systems enables more effective ways to achieve the desired assessment goals, even in resource-limited educational settings. A well-integrated hypermedia support-base provides interactive multimedia capability coupled with the ability to link and navigate through domain and pedagogical information. Appropriate use of multimedia extends the *involvement* of the student (bringing additional senses into play and employing both "left-" and "right-brain" faculties), facilitates *confidence-measuring* protocols, and provides practicable *performance assessment* (evaluating students' ability to solve problems that include the performance of particular tasks or procedures [6]). This also makes it easier for students to use the system, enhances the presentation of problem context, and supports the simulated environments used for performance assessment.

The CBAA project investigation of the feasibility of developing such alternative assessment tools blends techniques from education, artificial intelligence and human-computer interac-

tion [10-13]. The developed assessment products combine essentials of educational assessment and performance testing [1, 5, 6, 8, 14-17] with student modeling [18-20] and principles of interface design [21-24]. The project has adopted the goal of developing an architecture that applies to a large variety of subject areas, thereby providing additional economic advantage (since the same physical systems can be used for multiple application areas).

Confidence Measurement

The goal of *confidence measuring* assessment is to more accurately measure students' true knowledge states. Typical multiple-choice examinations require students to respond with what amounts to a probability distribution restricted to 0 and 1 values such as in the following example.

```
Which form of testing should be performed after making
  a minor modification to a module of a working system?

            A. Top-down testing
            B. Regression testing
            C. Ad-hoc testing
```

A choice of "B" is interpreted as P_A=0 (probability that A is correct is zero), P_B=1 (probability that B is correct is one), P_C=0 (probability that C is correct is zero). By restricting responses to this 0 or 1 distribution, we lose the ability to discriminate between states of knowledge such as "I strongly believe B to be correct," "I believe C to be incorrect but can't distinguish between A and B," and "From what I know, each alternative seems equally likely to be correct."

Alternative scoring schemes have been proposed in attempts to account for and reflect partial information and confidence. The typical "Number Right" scoring scheme assigns the same value to those who have complete information and those who select the correct response by guessing. It likewise groups together "those with complete misinformation, those with partial misinformation, and those who guess and select an incorrect response" [25]. A common replacement is the use of a "Correction for Guessing" formula, which claims to account for random guessing. Unfortunately, the correction-for-guessing approaches do not take into consideration the effects of partial knowledge [26], do not give credit for partial knowledge, and do little to encourage students to report their true levels of knowledge [27]. Methods intended to provide students the opportunity to report levels of information and misinformation were introduced over 45 years ago, notably "Elimination Scoring" [28], where students are asked to eliminate the incorrect responses, and "Inclusion Scoring" [25, 29], in which students are asked to choose the smallest subset of answers that includes the right answer. While these two schemes are found to be more reliable than both "Number Right" and "Correction for Guessing" [30-32], they tend to add confusion for test takers and produce inconsistent results [25]. Most attempts at using confidence weighting techniques have generally failed to improve reliability and validity coefficients while demanding relatively larger amounts of testing time [33-36].

To overcome the limitations of these traditional and modified testing structures, a *confidence-reporting* vehicle was developed for the CBAA project. The basic assessment model centers on the use of three-alternative questions and the sixteen-region response template shown in Figure 1, the CBAA Triangle. Proximity to a vertex corresponds to the degree of belief that the answer indicated by that vertex's letter is correct. In practice, the regions are colored for easy identification and, as the student moves the pointer over a region, graphical and textual feedback provide the interpretation of that region.
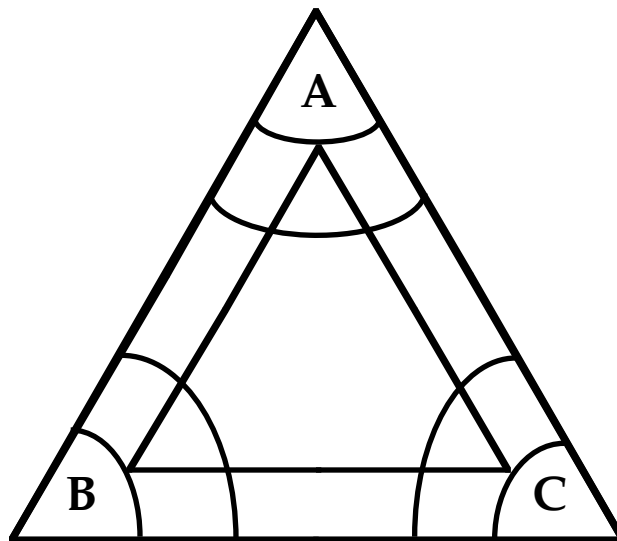

Figure 1. The CBAA Triangle confidence-measuring response template

The decision to use sixteen regions is based on experiences reported with the infinite precision probability space used in the foundational work at RAND [4, 37], and the results of field testing with alternatives ranging from four regions to 4500 regions. The sixteen-region template provides sufficient information for useful discrimination among students' knowledge states, avoids artifacts from minutia obsession and mechanical coordination (such as students' motor-skills in trying to fine-tune positioning of a pointer), and exhibits intuitive correspondence between the visual regions and their interpretations.

The decision to use three-option multiple choice items follows the evidence reported in educational and psychological literature indicating greater efficacy of using three options rather than a higher or lower number of alternatives [38-41]. Although surprising to many people, the research consistently shows that the three-alternative test is more effective than two-, four-, or five-alternative tests. (Because this is somewhat counter-intuitive, this research is repeated every few years, so essentially the same result appears in the literature every four or five years.) The basic finding is that if total test time is considered as a constant, related to the total number of options across items in the test, then three options per item produce the most reliable test scores. That is, the use of three-options is optimal with respect to time and validity. Note that there are conditions under which the founding assumption is invalid, such as stems that require students to read extended paragraphs.

The interpretation of each selectable region is shown in Figure 2. The values shown indicate the strength of belief that the correct answer is "A", labeled P_A. Values are analogous with respect to each vertex for answers of "B" or "C". Distance from a vertex is directly proportional to the degree of belief that the answer corresponding to that vertex is incorrect. Selection of a given region coincides with a three-element vector, $<P_A, P_B, P_C>$.
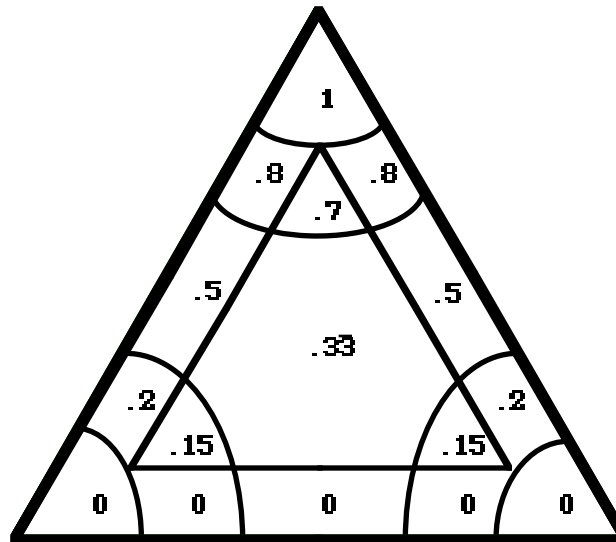

Figure 2. The CBAA Triangle, showing strength of belief P_A
(that the correct answer is "A") associated with each region

Admissible Scoring

> If students represent their degree of belief concerning information, such as in alternative responses on a multiple-choice test, how do we assign credit?

In any assessment endeavor, the scoring method used significantly impacts validity and reliability. As such, there has been substantial investigation concerning alternative scoring methods, especially with respect to multiple-choice test structures [25-32]. As discussed earlier, none of the accepted scoring schemes ("Number Right", "Correction for Guessing", "Elimination Scoring, "Inclusion Scoring") adequately addresses partial knowledge or reporting of confidence.

A student who indicates 40% certainty for an item that is actually incorrect cannot be said to be wrong, but not exactly right either. This student deserves more credit than one who gave the same selection an 80% chance of being correct, but less credit than one who reported only a 10% belief in its correctness.

Any scheme allowing students a wider-range of responses must also encourage students to report their true knowledge states. That is, we would like the scoring scheme to encourage students to accurately reflect their states of knowledge, in this case by responding with their perceptions of the probability distribution.

Students who report their true beliefs in the likelihood of correctness should reap higher rewards than those who "shade" their reporting one way or the other. For example, a student who believes that given selection has a 40% chance of being correct, should receive more credit, on average, by indicating 40% certainty and less for either reporting 30% or 50% certainty. Scoring systems that exhibit these desired properties are called "admissible," "reproducing," or "proper." More simply, they are "scoring systems which encourage honesty." The development of an admissible scoring scheme for the CBAA project is based on the work performed at RAND in the early 1970s concerning the evaluation of intelligence reports from multiple sources with varying degrees of confidence [4, 37, 42].

To illustrate how an admissible scoring system is developed, consider assigning credit according to a scheme analogous to wagering. Students select from among various wagers at various odds corresponding to the likelihood of an item being correct. The more-knowledgeable students would, over a period of time, gain more credit than the less-knowledgeable ones.

If there are $f(u)du$ wagers available at the correct odds of $\frac{1-u}{u}$, a student who believed the likelihood of an item being correct to be p would accept all wagers at odds better than $\frac{1-p}{p}$ that the item is correct. Similarly, that student would accept all wagers on the item not being correct at odds better than those appropriate for probability $1-p$.

Consider the following payoff formula that covers a student choosing among n alternatives where p_i is the probability of the ith alternative.

$$\text{payoff if } i^{\text{th}} \text{ event occurs} \quad = \int_0^{p_i} f(u)\frac{1-u}{u}\,du - \sum_{j \neq i}^n \int_0^{p_j} f(u)\,du$$

$$= \int_0^{p_i} \frac{f(u)du}{u} - \sum_{j=1}^n \int_0^{p_j} f(u)\,du$$

The flaw with this scheme is that the student will be able to secure a positive payoff by simply assuming equal probability for all alternatives. That is, it is possible to "game" the system to guarantee making a profit even when absolutely ignorant about the information involved in the test.

The formula may be adjusted so that total ignorance corresponds to zero payoff by simply requiring the student to take the odds on wagers placed at probabilities greater than $\frac{1}{n}$, and to offer the odds on wagers placed at probabilities less than $\frac{1}{n}$, as follows:

$$\text{payoff if } i^{\text{th}} \text{ event occurs} = \int_{\frac{1}{n}}^{p_i} \frac{f(u)du}{u} - \sum_{j=1}^n \int_{\frac{1}{n}}^{p_j} f(u)\,du$$

Putting this formula into practice requires deciding on the explicit definition of the function $f(u)$. Two choices of merit dealt with in the research literature include quadratic scoring or "Brier score", $f(u) = u$ [43], and logarithmic scoring, $f(u) = 1$ [44]. These yield the following payoff formulae:

Quadratic: payoff if i^{th} event occurs

$$= \int_{\frac{1}{n}}^{p_i} du - \sum_{j=1}^{n} \int_{\frac{1}{n}}^{p_j} u \, du$$

$$= p_i - \frac{1}{n} - \sum_{j=1}^{n} \frac{p_j^2 - \left(\frac{1}{n}\right)^2}{2}$$

$$= p_i - \frac{1}{2} \sum_{j=1}^{n} p_j^2 - \frac{1}{2n}$$

Logarithmic: payoff if i^{th} event occurs

$$= \int_{\frac{1}{n}}^{p_i} \frac{du}{u} - \sum_{j=1}^{n} \int_{\frac{1}{n}}^{p_j} du$$

$$= \log(p_i) - \log\left(\frac{1}{n}\right) - \sum_{j=1}^{n} \left(p_j - \frac{1}{n}\right)$$

$$= \log(np_i) \quad \text{where } \sum_{j=1}^{n} p_j = 1$$

The quadratic scoring system corresponds to the traditional concept of least-squares optimization: the best score is achieved by minimizing the squared difference between the selection and the actual outcome. The logarithmic scoring system corresponds to the maximum-likelihood method of statistical estimation, an efficient method for statistical selection of accurate forecasters.

The logarithmic scoring system also exhibits a significant relationship with information theory. This relationship is apparent when we look at the calculation of expected profit:

$$\text{Expected profit} = \sum_{i=1}^{n} p_i \log(np_i)$$

$$= \log n - \left(-\sum_{i=1}^{n} p_i \log p_i\right)$$

<p align="center">Entropy</p>

The quantity labeled *entropy* represents the expected amount of information to be conveyed by revealing which of the alternatives is actually correct. This correspondence between scoring scheme and information-theoretic measure demonstrates that a student's reward, on average, equals the amount of knowledge the student possesses about the material in question.

Finally, the logarithmic scoring system has the strong advantage that it depends solely on the probability assigned to the alternative that is actually correct. All other admissible scor-

ing systems (on more than two alternatives) depend on both the probability ascribed to the correct alternative and on the way the probability is divided among the incorrect alternatives [45].

Brown and Shuford [37] demonstrated that people who know they are to be rewarded according to admissible schemes are encouraged report the probabilities they believe in "rather than shading them one way or the other to exploit the scoring system." In essence, any response which varies from a student's true belief requires placing bets that the student considers unrewarding or failing to place bets the student considers rewarding. That is, this scheme is admissible and is one of the "scoring systems which encourage honesty" [37].

The CBAA Scoring System

<p align="center">A student's reward, on average, equals the amount of knowledge the student possess about the material in question</p>

The CBAA scoring system is an adaptation of the logarithmic admissible scoring system for exactly three alternatives. The calculated score is given by the following formula, where p_X is the probability ascribed to alternative X, n is a normalization constant, and k is a range constant.

$$\text{Score if A is correct} = n + k \log(3p_A)$$
$$\text{Score if B is correct} = n + k \log(3p_B)$$
$$\text{Score if C is correct} = n + k \log(3p_C)$$

For example, Figure 3 depicts scores associated with regions in the case where the correct alternative is "A". In this case, constants were chosen to provide scores in the range 0 to 100 (n=62, k=23.7). The value inside each region corresponds to the score awarded to a student who selects that region in the case where the correct alternative is "A". The values corresponding to the situations where alternatives "B" or "C" are correct are symmetric with respect to the appropriate vertex. Another natural choice of constants are those that yield scores in the range –150 to +100 with a center region value of 0, shown in Figure 4. Note that as p_X approaches zero, scores approach negative infinity. We must therefore choose a meaningful lower bound on p_X, as demonstrated by the minimum scores in these examples.

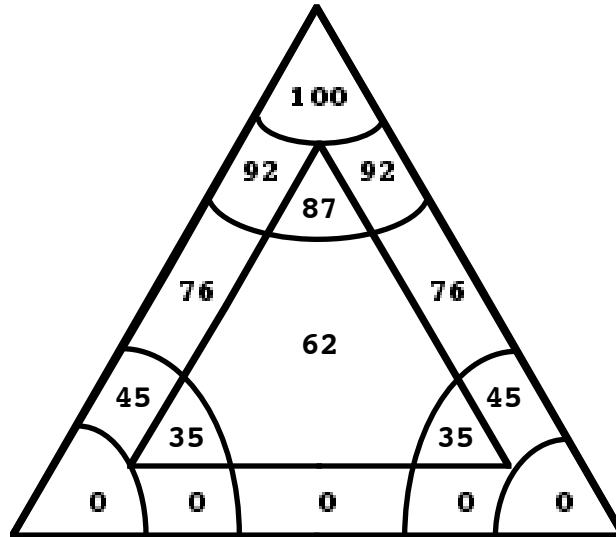Figure 3. The CBAA Triangle, showing values awarded for each selectable region in the case where the single correct answer is "A" where n=62, k=23.7
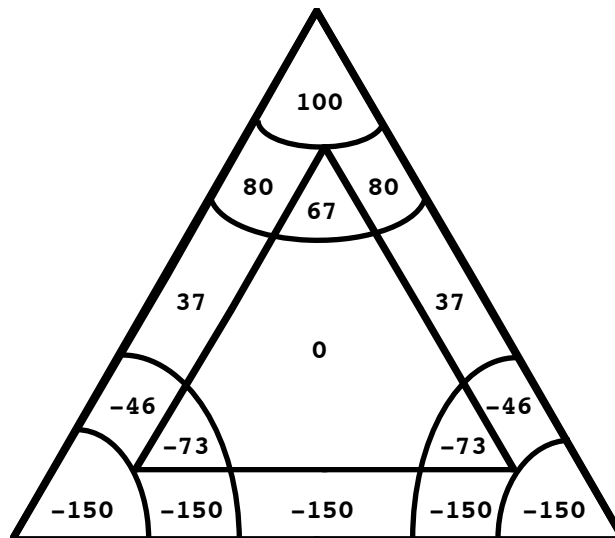

Figure 4. The CBAA Triangle, showing values awarded for each selectable region in the case where the single correct answer is "A" where n=0, k=63

To see the effect of admissibility, let's look at two examples using the scoring values from the CBAA Triangle in Figure 3. First, consider the case where the student's belief corresponds to the probability vector $<P_A, P_B, P_C>$ with values <.8, .2, 0>. The appropriate region of the triangle is that just below and to the left of the top vertex, shown in Figure 5 as region iv. This region yields a score of 92 if the correct answer is A, 45 if the correct answer is B, and 0 if the correct answer is C. Combining the student's assessment of the likelihood of correctness of each alternative with these values (92×.8 + 45×.2 + 0×0) yields the expected reward of 83. As shown in Figure 6, selecting this region yields a higher expected score than selecting any of the adjoining regions (labeled ii, iii, iv, and v in Figure 5).
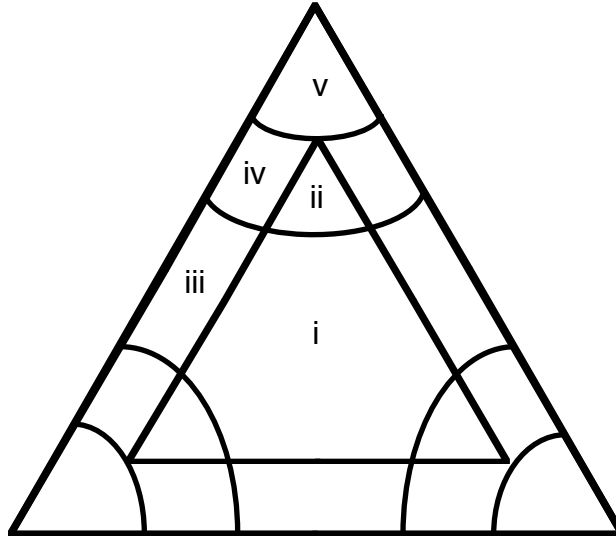
Figure 5. Regions of the CBAA Triangle considered
in context of belief vector <.8, .2, 0>

$$\begin{array}{lll}
\text{Region i} & 63 = & \big(63 \times .8 + 63 \times .2 + 63 \times 0\big) \\
\text{Region ii} & 77 = & \big(87 \times .8 + 35 \times .2 + 35 \times 0\big) \\
\text{Region iii} & 76 = & \big(76 \times .8 + 76 \times .2 + 0 \times 0\big) \\
\text{Region iv} & 83 = & \big(92 \times .8 + 45 \times .2 + 0 \times 0\big) \\
\text{Region v} & 80 = & \big(100 \times .8 + 0 \times .2 + 0 \times 0\big)
\end{array}$$

Figure 6. Expected payoffs for regions i–v from Figure 5
with belief vector <.8, .2, 0>

Figures 7 and 8 show the calculation of expected payoff information where the student's belief vector is <.33, .33, .33>, indicating complete uncertainty. Once again we see that choosing the region that corresponds most closely to the actual state of belief, region i, results in the highest expected score, 63.
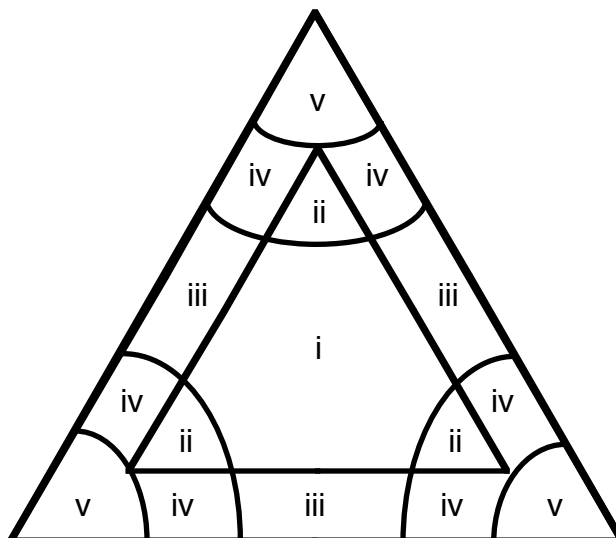
Figure 7. Regions of the CBAA Triangle considered
in context of belief vector <.33, .33, .33>

$$
\begin{array}{lll}
\text{Region i} & 63 = & \left(63 \times .33 + 63 \times .33 + 63 \times .33\right) \\
\text{Region ii} & 52 = & \left(87 \times .33 + 35 \times .33 + 35 \times .33\right) \\
\text{Region iii} & 50 = & \left(76 \times .33 + 76 \times .33 + 0 \times .33\right) \\
\text{Region iv} & 46 = & \left(92 \times .33 + 45 \times .33 + 0 \times .33\right) \\
\text{Region v} & 33 = & \left(100 \times .33 + 0 \times .33 + 0 \times .33\right)
\end{array}
$$

Figure 8. Expected payoffs for all regions from Figure 7
with belief vector <.33, .33, .33>

The essential character of the CBAA admissible scoring scheme is that it rewards honest reporting of belief and recognizes that acknowledged uncertainty is preferred to belief in the truth of something that is actually false. Because it is critical that students understand that they are to be rewarded according to this type of scheme, the CBAA confidence measuring system includes an animated interactive tutorial and practice test, described in the following section.

CBAA Confidence Measurement Architecture

The CBAA confidence measurement architecture incorporates multiple communication modes and media, the sixteen-region triangle for expressing confidence, and analysis of student responses based on realism functions and cognitive models. It permits the development of a more comprehensive picture of students' knowledge and meta-knowledge that allows discrimination between finer-grained states of knowledge and determination of how aware students are of their own knowledge states. These features are combined with real-time interpretation and individualized feedback to make a significant step toward more effective educational assessment.

Several prototypes were developed that use multimedia vehicles (text, graphics, sounds and audio-visual sequences) to present problems, alternatives, and feedback. The prototypes were implemented on the Macintosh™ platform using HyperCard™ as the integration substrate and QuickTime™ for real-time audio-visual presentation. Appropriate use of multiple modalities extends the involvement of the student, makes it easier for students to use the system, and enhances the presentation of problem context. In some assessment situations this is mandatory, such as when a student must follow a score while listening to a musical composition or make observations of timed chemical reactions.

In the CBAA confidence-measuring prototypes, support for the use of the triangle and feel for the scoring system are reinforced by giving students visual and dynamic feedback about the regions, their selections and corresponding scores. Figure 9 shows the color scheme used

for regions of the triangle which helps to distinguish the regions and to indicate region simi-
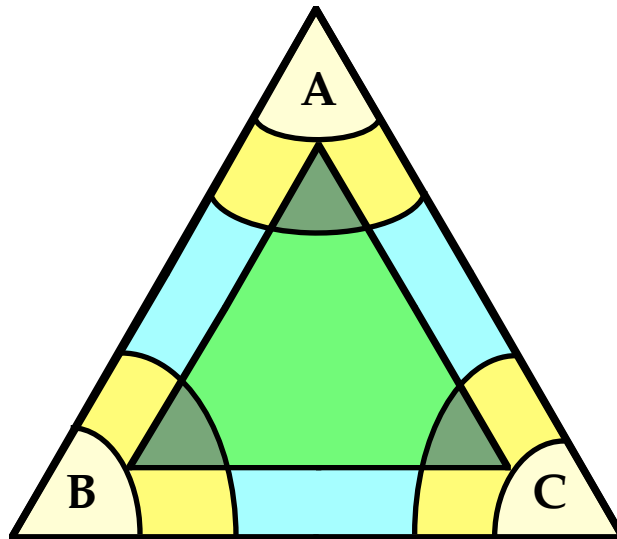larities.

Figure 9. The CBAA Triangle color scheme as displayed in prototypes

When the student moves the cursor over any region of the triangle, pop-up text is displayed
that reinforces the region's interpretation. In addition to the pop-up text, the actual scores
that would be awarded for choosing the region are shown in a score-box. Figure 10 shows
pop-up text associated with eight of the sixteen regions of the triangle. Figure 11 shows the
score-box when the cursor is over the region associated with the pop-up text "Probably A,
Possibly B or C". These dynamic feedback elements are delivered as a natural consequence
of cursor positioning and require no additional action on the part of the student.
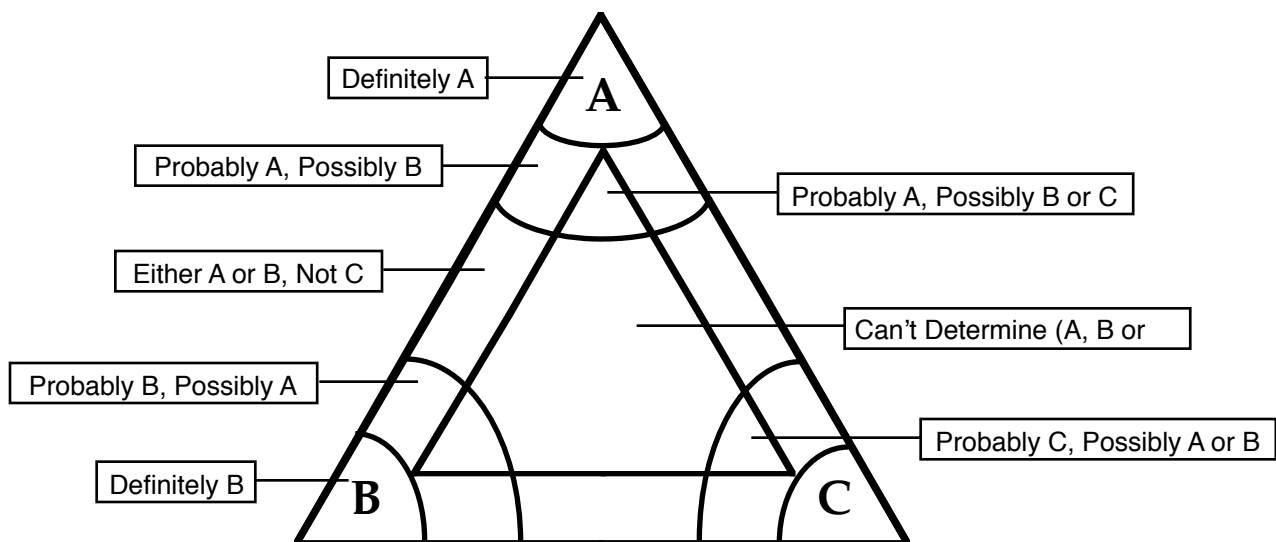

Figure 10. Sample text pop-ups showing the associated
regions of the CBAA Triangle

Value if correct answer is:		
A	B	C
87	35	35

Figure 11. CBAA Triangle score-box shown for
"Probably A, Possibly B or C" (Region iii of Figure 5)

In the initial field trials of the CBAA confidence-measuring system, students were given minimal instruction prior to the actual testing exercise. As a test of the intuitiveness of the interface, they were simply given a practice exercise consisting of six questions of general trivia knowledge. These early experiments suggested the potential value of initial instruction in the use of the CBAA triangle. After investigation and testing of several alternatives, there now exists a ten-minute interactive animated tutorial in which students learn about the principles of confidence reporting and the CBAA system. The tutorial introduces confidence reporting initially in the form of a bipolar (True-False) example, then extends the concept to the three-way multiple-choice CBAA triangle. The tutorial also includes an informal presentation of the scoring scheme that demonstrates and reinforces the fact that the highest scores result from accurate reporting of confidence or belief.

Preparing CBAA Triangle Materials

Guidelines for creating materials for the CBAA Confidence Measuring system are largely equivalent to those for designing any effective assessment, and those for multiple-choice assessments in particular [46-49].

Questions should be developed that assess the broad range of cognitive skills and learning objectives, including higher-order thinking. Theoretical, prescriptive, and descriptive methods and models may be used to provide guidance and to check the range of cognitive behaviors covered by an assessment. Bloom's [50] Taxonomy of Educational Objectives provides a useful hierarchy of six learning levels (knowledge, comprehension, application, analysis, synthesis, evaluation). Ory and Ryan [48] present example multiple-choice questions corresponding to each of these levels. Haladyna [46] offers a descriptive typology for writing and classifying multiple-choice items based on a content dimension (fact, concept, principle, procedure) and cognitive operation dimension (recalling, defining, predicting, evaluating, problem solving), and also provides examples of multiple-choice questions for each combination of content category and cognitive behavior in the typology.

Measuring a student's ability to apply what was learned in different problem-solving contexts may conflict with the desire to reuse items from one testing situation to the next, since it generally requires the generation of problem situations that are novel to students. The CBAA architecture provides the basis for parameter-driven, automated, dynamic problem generation that is capable of addressing this difficulty. Care must be exercised in its use, however, to maintain consistency and comparability required for standardization.

One of the greatest difficulties in the construction of assessment vehicles is reducing ambiguity. This is common to all assessment forms and is not unique to the multiple-choice structure or the CBAA architecture. Gause and Weinberg [51] present useful ambiguity revealing and reducing heuristics that are quite effective when applied to the design and construction of assessment items.

Each item in the assessment should be constructed to measure a specific skill. Thus, each question would focus on a single problem, principle, fact, concept or procedure. Question sequences can then be used to assess complex multistage thinking processes or skills, more clearly and with greater discrimination, by designing each item within the set to address a different aspect.

Care must be taken to ensure that both of the incorrect choices seem plausible, each representing the result of an identified, common but erroneous, path of reasoning. These should not be designed to "trick" students, but should appear reasonable to those who have not mastered the material. Similarly, if domain content is being tested, basic common sense or logic should be insufficient to distinguish between correct and incorrect choices.

The limit of only two incorrect choices in the CBAA multiple choice structure is a key factor in enabling the greater number of questions that may be administered in the same amount of time, which has been demonstrated to improve the validity of results and distinguish between finer-grained states of knowledge.

The availability of multimedia technology extends the domains to which such testing may be applied, such as those requiring audio clips, dynamic graphics, full-motion audio-visual sequences, or the integration of these. For example, CBAA prototypes address music appreciation, software engineering, and cognitive science. Multimedia may also be used to enhance student engagement in the assessment process, such as through more dynamic presentations, or to overcome other limitations, such as by providing spoken audio presentations of written text.

Using the Collected Data

The information collected using the CBAA triangle technique is rich with significance. A simple interpretation of the base information provides a rough, composite indication of students' content knowledge. For example, using the scoring scheme shown in Figure 3, average scores near 62 are representative of a deficiency in domain knowledge. Scores significantly below that level indicate misconceptions and false beliefs rather than simple lack of knowledge.

A realism function [37] or external validity graph [4] can be used to detect how aware a student is of his or her own knowledge state. Such calculations allow the determination of whether a student appears to overvalue or undervalue his or her knowledge. The realism function further enables the computation and disclosure of both the loss in score attributable to this bias in assignment of probabilities (also called "lack of realism" or "labeling er-

ror") and the loss attributable to lack of information about the subject matter. The determined lack of realism can also be used to provide a better composite indicator by computationally adjusting for reporting bias.

A particular pattern of responses may provide additional information about the knowledge state of the student, especially useful in diagnosing content misconceptions, reasoning errors, and learning difficulties. This information can be used to generate appropriate feedback, including differential or custom navigation through problems where standardization is not required [52]. The CBAA architecture supports the integration of cognitive models to help interpret this data. Applying cognition-based methods helps to diagnose students' difficulties and provide customized corrective and directive advice. Note that tools based on cognitive models apply to both the confidence-measurement and performance-assessment components and are discussed in the following section: *Analysis Using Cognitive Models.*

The interactive computer-based architecture also provides the optional availability of immediate total or partial feedback. Care is necessary to appropriately exploit this technological enabling, as the pedagogical value of the immediacy and degree of disclosure depends greatly on the educational objectives and situational characteristics [35, 53-55].

Assessment of Bias

A valuable property of the CBAA confidence-measuring system is the ability to determine reporting bias. Automated analysis of collected data can be used to furnish results to students that enable them to gain more accurate awareness of their own knowledge states, that is, a means to relate their confidence reports to reality. The more realistic perspective provided is a fundamental aid to improving students' use of knowledge as well as to improving their ability to acquire knowledge that they lack.

The mechanism adopted for CBAA is the calculation of a *realism function* based on the proposals of Brown and Shuford [37] and Sibley [4]. The characteristic feature is a comparison between reported confidence and the relative frequency with which each particular confidence level is associated with a correct alternative. Each time a student selects a region of the CBAA triangle, three confidence reports are recorded corresponding to the probability vector $<P_A, P_B, P_C>$. The relative frequency computed for each confidence value is the ratio of the number of times that value was assigned when the associated choice was correct to the total number of times that value was assigned:

$$\text{Relative Frequency} \left(P_i \right) \;=\; \frac{\#\ \text{assigned } P_i\ \text{when choice was correct}}{\text{total } \#\ \text{assigned } P_i}$$

As an example, consider just those times during a test when a given student assigned a confidence of 70% to any alternative. Suppose those alternatives turned out to be correct 70% of the time. In this case, we have identified a student who is an unbiased and realistic reporter of his or her own knowledge within that sample set and the bounds of sampling variability. If those alternatives turned out to be correct more than 70% of the time, we

would say the student "undervalues" that part of his or her knowledge. Likewise, if those alternatives turned out to be correct less than 70% of the time, we would say the student "overvalues" that part of his or her knowledge.

An *external validity graph,* such as the sample shown in Figure 12, is a plot of this relationship for all assigned confidence values and their relative frequencies. According to Brown and Shuford [37], several hundred to several thousand observations are required to calibrate an individual's performance using the external validity graph directly. However, the nature of the CBAA confidence-measuring system allows the relationship between assigned confidence values and relative frequencies to be approximated satisfactorily by a simple straight-line linear regression, a *realism function*. Brown and Shuford [37] report that "with the least-squares estimation procedure, stable estimates may be obtained with as few as twenty or more probability estimates."
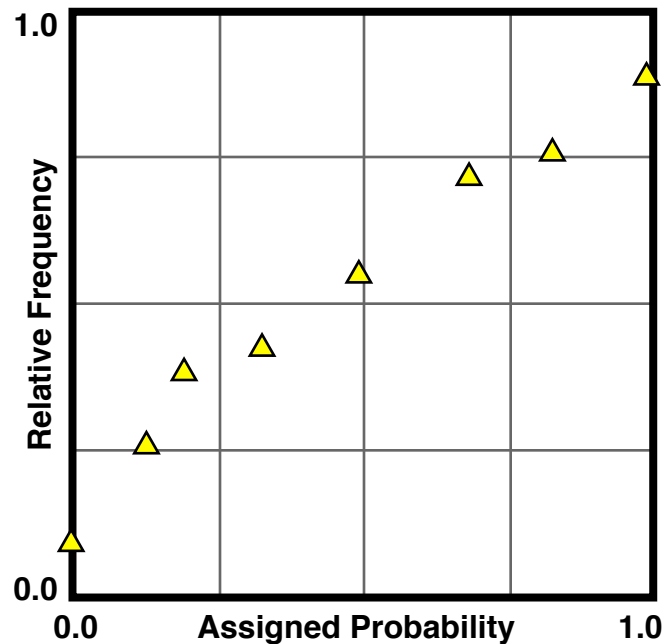

Figure 12. External Validity Graph

Simple least-squares regression, however, does not take into account the varying amounts of information that contribute to the points in the external validity graph. For example, if the student assigns the probability value 0.5 fifteen times and the 0.8 value only twice, then the point corresponding to the relative frequency of data assigned probability value of 0.8 should contribute significantly less to the realism function than the point corresponding to the 0.5 value. A more accurate linear estimation procedure developed for CBAA that adjusts for this contributory weighting is shown in Figure 13.

Let p_i be an assigned probability

r_i be the computed relative frequency for p_i

z_i be the number of contributors to r_i

To determine $m \ \& \ b$ (slope & intercept) in $y = mx + b$:

$$\texttt{If} \ \left(\sum p_i^2 z_i == \left(\sum p_i z_i \times \sum p_i z_i \Big/ \sum z_i\right)\right) \ \texttt{then} \ m = 1$$

$$\texttt{else} \ m = \frac{\sum p_i r_i z_i - \left(\sum p_i z_i \times \sum r_i z_i \Big/ \sum z_i\right)}{\sum p_i^2 z_i - \left(\sum p_i z_i \times \sum p_i z_i \Big/ \sum z_i\right)}$$

$$b = \frac{\sum r_i z_i - \left(m \times \sum p_i z_i\right)}{\sum z_i}$$

Figure 13. CBAA weighted least-squares linear estimation procedure

The primary interpretation of the realism function, $y = mx + b$, concerns the computed slope value, m. As shown in Figure 14, fully unbiased and realistic reporting of confidence corresponds to the ideal line with intercept of zero and slope of one, $y=x$. If the slope is less than one, the student appears to overvalue his or her information. This is readily seen when absolute confidence (a probability of one) is assigned to an alternative that is in fact not correct and thus the relative frequency of occurrence is some value less than one. Likewise, this is evidenced by an assigned probability of zero to alternatives that are correct, which yields some positive relative frequency of occurrence. Similarly, if the slope is greater than one, the student appears to undervalue his or her information. The amount by which slopes deviate from one corresponds to the degree to which students tend to overvalue or undervalue their own knowledge.
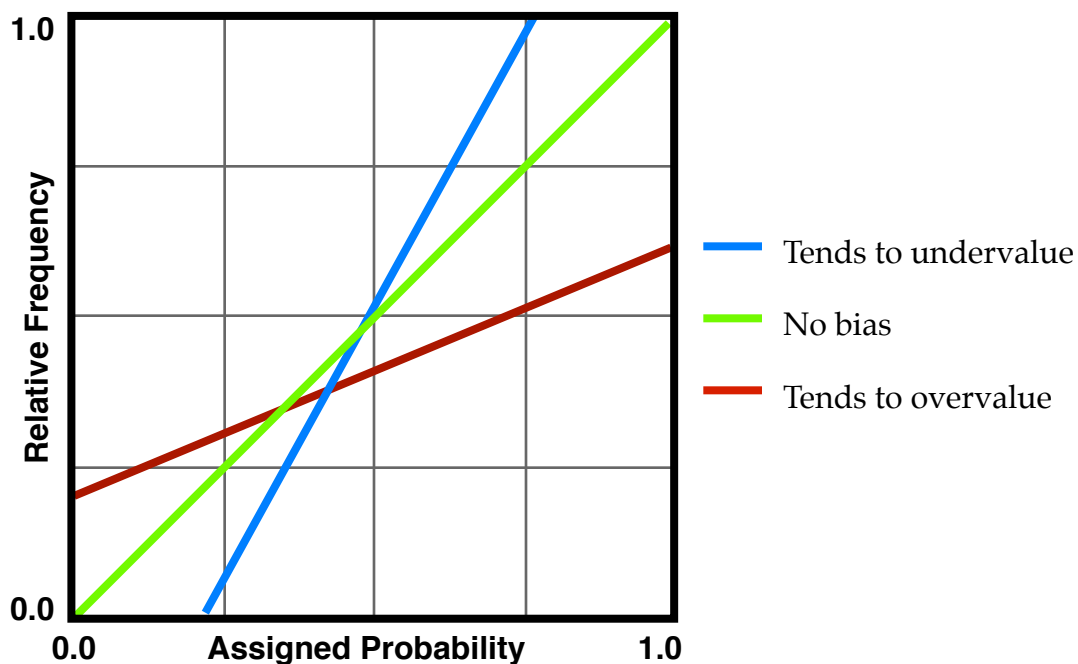

Figure 14. Sample realism functions and their interpretations

Correction for Bias

Once the CBAA system detects a lack of realism (reporting bias) on the part of a student, the computed realism function can be used to determine the score that he or she could have achieved if there had been no bias. This is accomplished by transforming the student's assigned probabilities using the realism function and recalculating his or her score. That is, we adjust each of the student's reported probability values, x, as though he or she had responded with $mx+b$, and then recompute the score.

The difference between a student's unmodified score and the one corrected for bias represents the loss due to inability to correctly assess the value of one's own knowledge. Also known as labeling error, this loss represents the portion of the difference between a perfect score and the earned score due to inappropriate use of the knowledge the student already possesses. The remaining difference is accounted for by knowledge that the student has not yet acquired. Sharing this analysis with the student helps increase individual awareness of any tendency to inappropriately value knowledge, and is reinforced by showing the effect on earned score. A sample report of this information is shown in Figure 15. The set of textual responses, corresponding to degrees of bias, include those shown in Figure 16. Another visualization method for the same information is shown in Figure 17. This depicts the correspondence between actual and perceived knowledge based on the analysis of bias [4].
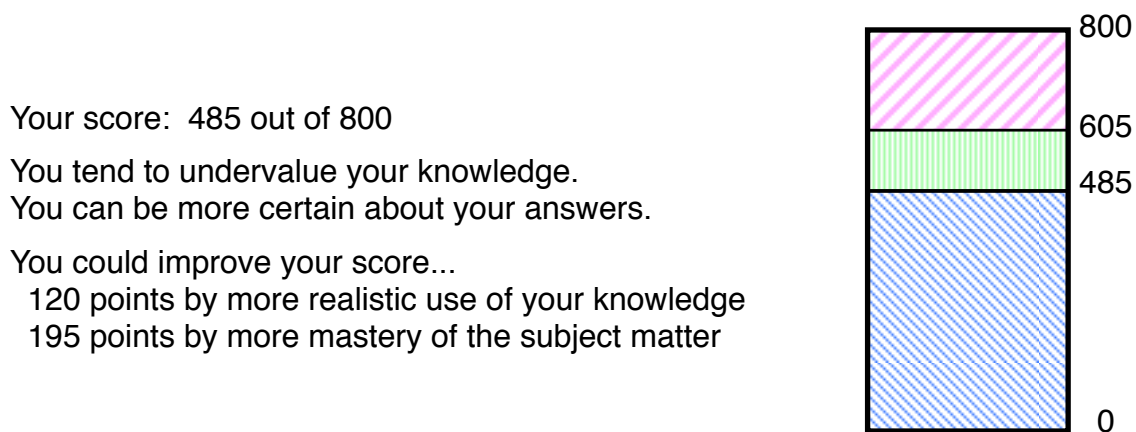
Your score: 485 out of 800

You tend to undervalue your knowledge.
You can be more certain about your answers.

You could improve your score...
  120 points by more realistic use of your knowledge
  195 points by more mastery of the subject matter

Figure 15. Sample feedback to student concerning score and interpretation

You tend to considerably overvalue your knowledge. You should be much less certain about your answers.

You tend to overvalue your knowledge. You should be less certain about your answers.

You tend to slightly overvalue your knowledge. You should be a bit less certain about your answers.

You tend to accurately value your knowledge. (You have a good idea of what you actually know.)

You tend to slightly undervalue your knowledge. You can be a bit more certain about your answers.

You tend to undervalue your knowledge. You can be more certain about your answers.

You tend to considerably undervalue your knowledge. You can be much more certain about your answers.

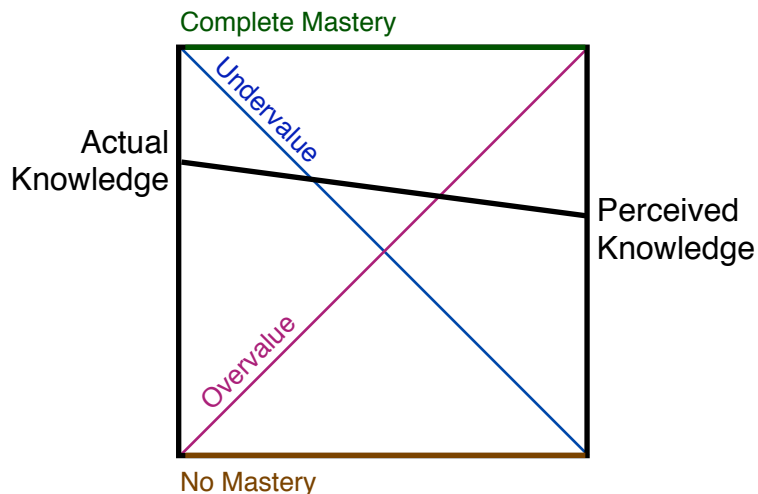Figure 16. Textual feedback selections for different degrees of bias

Figure 17. Visual comparison of relative information as determined and as perceived

The calculation of score adjusted for bias provides a more accurate indicator of student knowledge than the unmodified score. By using the realism function to interpret responses, we are in effect saying that when the student reported value x, he or she really meant to have reported $mx+b$. This correction for reporting bias typically yields a more authentic measure of the student's knowledge, independent of his or her timidity or boldness in reporting.

Performance Assessment

Another dimension of educational evaluation addressed by the CBAA project concerns the ability to do performance assessment [5, 6, 47, 56], in which the student is given a task to perform in a "virtual world" provided by the system. For example, where physical observables must be simulated (e.g., an interview or a chemical reaction) a video sequence may be displayed. The use of such simulation is indicated when actual performance tests are impractical due to cost, danger, the serious consequences of mistakes, or the impossibility of arranging actual performance situations. By simulating performance conditions the system controls most of the variables in the testing situation and we can standardize the assessment across students and administrations [6].

Exploratory performance-assessment prototypes in the CBAA project address chemistry and software engineering domains. Figures 18 and 19 show snapshots of task-performance displays and interaction from the chemistry version [57]. In the chemistry prototype, students carry out qualitative analysis to determine which ions are present in an unknown solution. This problem-solving task is suitable for high-school or first-year college chemistry laboratory assessment. Figure 20 shows a snapshot from the software-engineering version [58]. In the software engineering prototype, the student observes a customer explaining his concerns, then develops a top-level data-flow-diagram (DFD) using a palette of diagraming tools.
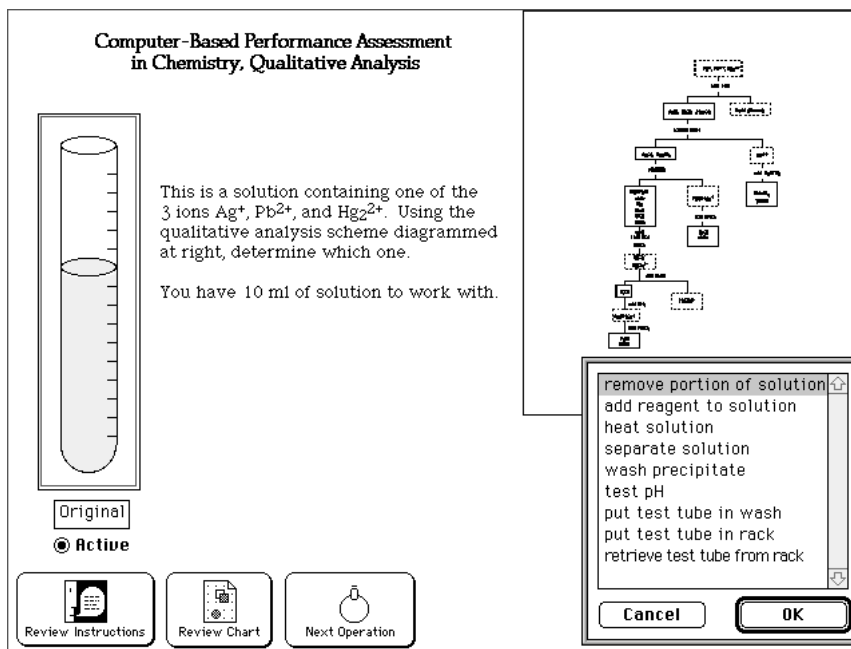
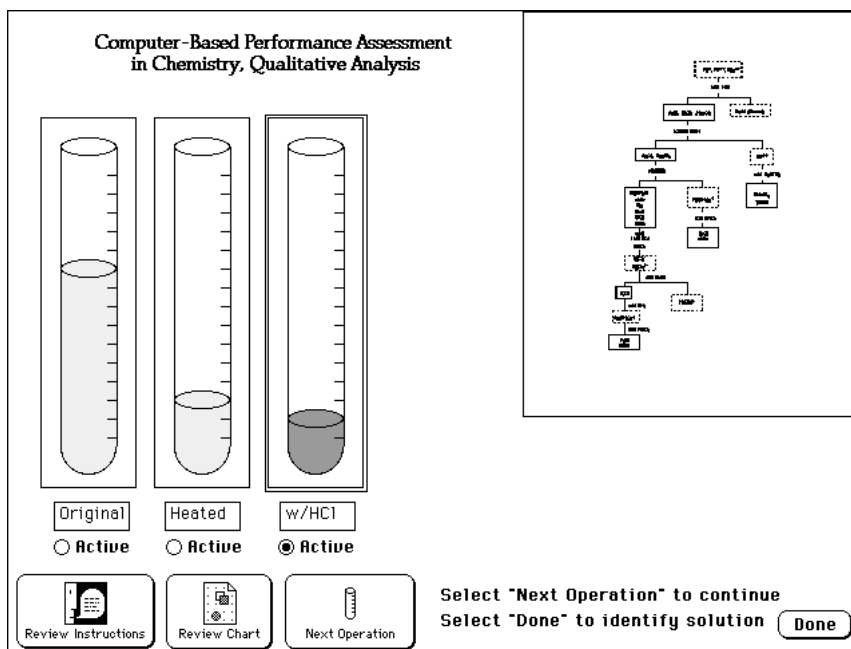Figure 18. Sample task-performance environment — Chemistry (1)

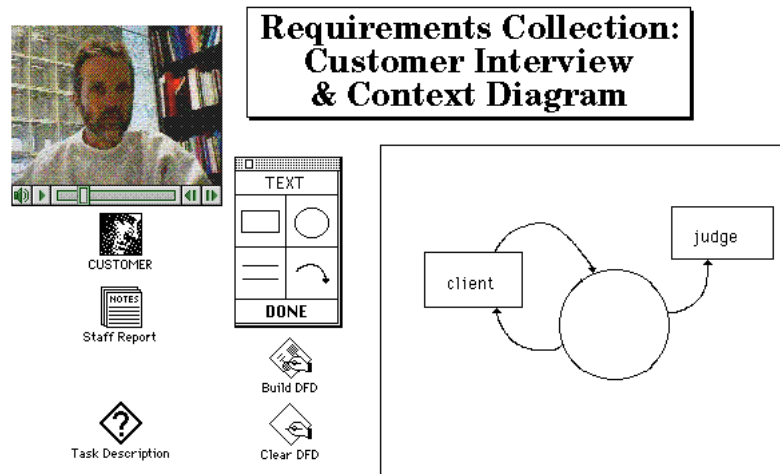Figure 19. Sample task-performance environment — Chemistry (2)

Figure 20. Sample task-performance environment — Software Engineering

A key issue that arises when attempting to assess the task performance of a student concerns *what to measure.* We could choose to look only at the result achieved, but that information may not be sufficient. For example, it may be possible to achieve the result by inappropriate means, such as using an insertion sort in a programming assessment when a quicksort is requested. Alternatively, we could look at the specific command sequence issued by the student. In this case, there may be an infinite variety of functionally equivalent sequences. The approach taken by the CBAA project, instead, is to look at intermediate-goal satisfaction, which presumes a set of necessary sub-goals that must be satisfied to achieve a higher-level goal.

Cognitive models may provide an essential element for effective use of this assessment technique. The following section, *Analysis Using Cognitive Models,* provides a discussion of the application of cognitive models to aid interpretation and generation of feedback for both the performance-assessment and confidence-measurement components of the CBAA architecture.

Analysis Using Cognitive Models

The identification and analysis of patterns of responses through the use of cognition-based methods provides additional information about the knowledge state of students. This information is especially useful in diagnosing content misconceptions, reasoning errors, and learning difficulties. It can also help improve the generation of appropriate feedback, including corrective and directive advice, and differential or custom navigation through the assessment space using principles from Computerized Adaptive Testing [52, 59, 60].

Models of reasoning developed by cognitive scientists, especially those in the area of Intelligent Tutoring Systems [18-20], can be used to improve diagnostic competency and to help further illuminate students' knowledge states. When appropriately applied, these aid in revealing the cognition that most likely underlies the observed patterns of behavior. Student

response patterns can be compared with those predicted by general student and learning models, as well as models of expert and buggy reasoning. Using such models may help students by identifying common misconceptions and reasoning flaws, and using the results to provide appropriate remediation advice.

The CBAA architecture focuses on domain-independent models in the confidence-measuring component to maintain greater generality of application. In contrast, domain-specific models of reasoning are likely to be necessary to provide the diagnostic power for effective interpretation of students' task-performance actions. Additional study is required to determine the most appropriate use of domain-specific and domain-independent models.

Data collected may also be used with case-based reasoning and case-based explanation methods [61-65]. These knowledge-based techniques detect meaningful response patterns, including those that might otherwise go unnoticed, and help generate interpretive, diagnostic, and advisory responses. They work by identifying similarities that exist in collected data and applying those discoveries to the interpretation of new data and to the improvement of future competence and performance.

In the CBAA architecture, data concerning a specific student assessment comprises an instance or episode stored in memory as a *case.* Integrated into the case structure are the recorded testing responses; interpretations and assessments; corrective and directive advice; and other case-specific recommendations. The collected set of all known cases reside in a knowledge-based memory structure called *episodic memory.* Cases are organized according to detected similarities, which are captured in memory organization structures called *generalizations.* Figure 21 shows a pictorial representation of a case and generalization. The upper region of each case represents the input data. The lower region represents the appropriate response, including case-specific analysis and feedback. Figure 22 depicts a snapshot of episodic memory, showing how generalizations work as memory organization structures for cases and more-specific generalizations.

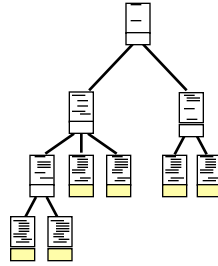Figure 21. Pictorial representation of cases and generalizations

Figure 22. Sample snapshot of episodic memory showing
six cases and four generalizations

The process of case-based reasoning with CBAA data works as follows. When a new case arrives to be processed, it is compared with those already in memory to find the closest matching case. The existing organization directs the matching process from the root, along the best-fit edges to more-specific generalizations, until the most appropriate leaf-node is found. This process is called *reminding.* The case thus located is the best match among known cases.

The matching case is retrieved and a detailed comparison with the input case is performed to determine similarities and differences. The results of the comparison are used to adapt the response portion of the retrieved case so that it applies to the input. This new response is output from the case-based reasoning component to be used in constructing the relevant feedback, such as advice to the student or diagnoses for the instructor.

The input case, newly completed by union with the computed response data, is now ready to be incorporated into episodic memory. The appropriate place to add this new case is already determined: as sibling to the just located, closest matching, previously known case. If a subset of the newly augmented set of siblings exhibit sufficient similarity to each other, a new generalization may be created thus reorganizing memory to facility future processing. In essence, a new common pattern has been detected of potential value to interpreting future assessment data.

Figure 23 depicts the progress of case-based reasoning processes in CBAA. A new case is received as input, triggering reminding of the closest matching case. A new response is created by adaptation from the retrieved case, produced as output, and integrated with the input case. Finally, the new case is incorporated into memory, in this case resulting in a new generalization begin added.
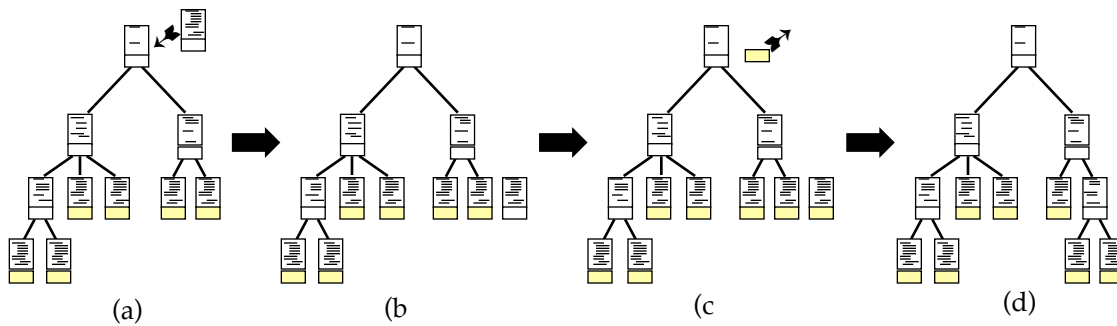
Figure 23. Sequential snapshots of episodic memory showing (a) input case and initial memory state; (b) reminding; (c) adaptation and output; and (d) dynamic memory adjustment including addition of a new generalization

The incorporation of intelligent interpretive, analytic and generative processing based on cognitive models provides the means to achieve better understanding of student, classroom, and institutional characteristics. A single implementation can address the continuous range of aggregation from individual student through entire population. The interpreted results provided by the system help students increase awareness of their own knowledge and better focus their efforts. These interpretations also provide educators and institutions valuable information concerning educational policy.

Cognitive model-based analysis may provide the key to gaining valuable understanding and insight that help students improve in specific competence and to become more effective life long learners.

Work in Progress

One of the current goals of the CBAA project is to migrate the concepts developed and demonstrated in stand-alone prototypes to implementation as network applications. Specifically, we face the challenge of maintaining the desired levels of engagement and interactivity while changing the delivery mechanism to web-based Internet and intranet vehicles which exhibit highly-varying and non-deterministic delay attributes. This has a significant impact on the design of the student interface and, to a lesser extent, on the interfaces for authoring and maintenance.

Another major goal is to improve diagnostic competency of the system through the integration of additional models of reasoning. Along with more sophisticated pattern-matching mechanisms, these will provide deeper levels of diagnosis and detection of student's thinking processes, especially as aids to the identification of misconceptions and the nature of mistakes made.

Formal evaluation is necessary to test the informal observations, representing a wide range of issues from the procedural (e.g., that fewer "clerical" errors are made in computer-based reporting compared with the use of pencil-and-paper answer sheets) to the cognitive (e.g., greater student engagement leads to less off-task "drift" during testing). Diverse field test-

ing is needed to determine the priority to assign to various implementation options, such as feedback alternatives or dynamic, parametric problem generation.

Finally, the CBAA project has merely scratched the surface of computer-assisted performance assessment. In particular, while the collection of task performance protocols has been addressed, there exists only the most limited interpretation mechanisms. The current focus is on the application of case-based reasoning and on reconciliation of collected data with various models of reasoning. The case-based reasoning and explanation frameworks are being used to provide the infrastructure necessary to support this effort.

References

[1]   P. W. Airasian, Classroom Assessment. New York: McGraw-Hill, 1991.

[2]   J. E. Barnett and J. E. Hixon, "Effects of Grade Level and Subject on Student Test Score Prediction," The Journal of Educational Research, vol. 90, pp. 170-174, 1997.

[3]   S. G. Paris and P. Winograd, "How Metacognition Can Promote Academic Learning and Instruction," in Dimensions of Thinking and Cognitive Instruction, B. F. Jones and L. Idol, Eds. Hillsdale, NJ: Erlbaum, 1990, pp. 15-52.

[4]   W. L. Sibley, "A Prototype Computer Program for Interactive Computer-Administered Admissible Probability Measurement," RAND, Santa Monica R-1258-ARPA, 1974.

[5]   R. W. Sweezey, Individual Performance Assessment: An Approach to Criterion-Referenced Test Development. Reston: Reston Publishing Company, Inc., 1981.

[6]   M. Priestley, Performance Assessment in Education and Training: Alternative Techniques. Englewood Cliffs: Educational Technology Publications, 1982.

[7]   L. B. Resnick and D. P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform," in Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction, B. R. Gifford and M. C. O'Connor, Eds. Boston: Kluwer Academic Publishers, 1992, pp. 37-75.

[8]   T. Kellaghan, G. F. Madaus, and P. W. Airasian, "The Effects of Standardized Testing," in Evaluation in Education and Human Services, G. F. Madaus and D. F. Stufflebeam, Eds. Boston: Kluwer-Nijhoff Publishing, 1982.

[9]   L. A. Shepard, "What policy makers who mandate tests should know about the new psychology of intellectual ability and learning," in Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction, B. R. Gifford and M. C. O'Connor, Eds. Boston: Kluwer Academic Publishers, 1992, pp. 301-328.

[10]   J. Paul, "Hypermedia-based Interactive Student-Assessment System (HISAS): Concept and Architecture," in Educational Multimedia and Hypermedia Annual, H. Maurer, Ed. Charlottesville: Association for the Advancement of Computing in Education, 1993, pp. 415-421.

[11]   J. Paul, "Multimedia Interfaces for Alternative Assessment Methods That Improve Engineering Education," in Proceedings of the IEEE First International Conference on

Multi-Media Engineering, Melbourne, Australia, July 6-8, 1994, M. Aldeen, Ed. Melbourne, Australia: IEEE Press, 1994.

[12]   J. Paul, "Improving Education Through Improved Assessment," in Proceedings of the 24th Frontiers in Education Conference, San Jose, CA, November 2-6, 1994: IEEE Education Society, ASEE Education and Method Division, IEEE Computer Society, 1994.

[13]   J. Paul, "Improving Education Through Computer-Based Alternative Assessment Methods," in People and Computers, vol. 9, G. Cockton, S. W. Draper, and G. R. S. Weir, Eds. Cambridge: Cambridge University Press, 1994.

[14]   A. Dworkin and N. Dworkin, Problem Solving Assessment. Novato: Academic Therapy Publications, 1988.

[15]   R. Freedle, Artificial Intelligence and the Future of Testing. Hillsdale: Lawrence Erlbaum Associates, 1990.

[16]   B. R. Gifford and M. C. O'Connor, Eds., Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction. Boston: Kluwer Academic Publishers, 1992.

[17]   R. J. Sternberg, "CAT: A program of Comprehensive Abilities Testing," in Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction, B. R. Gifford and M. C. O'Connor, Eds. Boston: Kluwer Academic Publishers, 1992, pp. 213-274.

[18]   P. H. Wood and P. D. Holt, "Intelligent Tutoring Systems: An annotated bibliography," SIGART Bulletin, vol. 1, pp. 21-42, 1990.

[19]   J. A. Self, "Bypassing the Intractable Problem of Student Modeling," in Intelligent Tutoring Systems: At the Crossroad of Artificial Intelligence and Education, C. Frasson and G. Gauthier, Eds. Norwood, NJ: Ablex, 1990, pp. 107-123.

[20]   H. S. Nwana, "Intelligent Tutoring Systems: An Overview," Artificial Intelligence Review, vol. 4, pp. 251-277, 1990.

[21]   M. Fleming and W. H. Levie, Eds., Instructional Message Design: Principles from the Behavioral and Cognitive Sciences, Second ed. Englewood Cliffs, NJ: Educational Technology Publications, 1993.

[22]   B. Laurel, Ed., The Art of Human-Computer Interface Design. Reading: Addison-Wesley, 1990.

[23]   D. A. Norman, The Psychology of Everyday Things. New York: Basic Books, 1988.

[24]   B. Shneiderman, Designing the User Interface: Strategies for Effective Human-Computer Interaction, Third ed. Reading, MA: Addison-Wesley, 1998.

[25]   D. Jaradat and N. Tollefson, "The impact of alternative scoring procedures for multiple-choice items on test reliability, validity, and grading," Educational and Psychological Measurement, vol. 48, pp. 627-635, 1988.

[26]   F. K. Abu-Sayf, "The scoring of multiple-choice tests: A closer look," Educational Technology, vol. 19, pp. 5-15, 1979.

[27]   R. B. Frary, "The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores," Applied Psychological Measurement, vol. 4, pp. 79-90, 1980.

[28]   C. H. Coombs, "On the use of objective examinations," Educational and Psychological Measurement, vol. 13, pp. 309-310, 1953.

[29]   P. L. Dressel and J. Schmid, "Some modifications of the multiple-choice item," Educational and Psychological Measurement, vol. 13, pp. 574-595, 1953.

[30]   L. S. Collet, "Elimination scoring: An empirical evaluation," Journal of Educational Measurement, vol. 8, pp. 209-214, 1971.

[31]   J. D. Gibbons, I. Olkin, and M. A. Sobel, "Subset selection technique for scoring items on a multiple-choice test," Psychometrika, vol. 44, pp. 259-270, 1979.

[32]   D. Jaradat and S. Swaged, "The subset selection technique for multiple-choice tests: An empirical inquiry," Educational and Psychological Measurement, vol. 23, pp. 369-376, 1986.

[33]   J. J. Diamond, "A preliminary study of the reliability and validity of a scoring procedure based upon confidence and partial information," Journal of Educational Measurement, vol. 12, pp. 129-133, 1975.

[34]   A. R. Hakstian and W. Kansup, "A comparison of several methods of assessing partial knowledge in multiple choice tests: II. Testing procedures," Journal of Educational Measurement, vol. 12, pp. 231-239, 1975.

[35]   G. S. Hanna, "A study of reliability and validity effects of total and partial immediate feedback in multiple-choice testing," Journal of Educational Measurement, vol. 14, pp. 1-7, 1977.

[36]   W. Kansup and A. R. Hakstian, "A comparison of several methods of assessing partial knowledge in multiple choice tests: I. Scoring procedures," Journal of Educational Measurement, vol. 12, pp. 219-230, 1975.

[37]   T. A. Brown and E. H. Shuford, "Quantifying Uncertainty Into Numerical Probabilities for the Reporting of Intelligence," RAND, Santa Monica R-1185-ARPA, 1973.

[38]   F. Costin, "Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests," Educational and Psychological Measurement, vol. 32, pp. 1035-1038, 1972.

[39]   F. M. Lord, "Optimal number of choices per item — A comparison of four approaches," Journal of Educational Measurement, vol. 14, pp. 33-38, 1977.

[40]   S. V. Owen and R. D. Froman, "What's wrong with three-option multiple choice items?," Educational and Psychological Measurement, vol. 47, pp. 513-522, 1987.

[41]   M. S. Trevisan, G. Sax, and W. B. Michael, "The effects of the number of options per item and student ability on test validity and reliability," Educational and Psychological Measurement, vol. 51, pp. 829-837, 1991.

[42]   T. A. Brown, "Probabilistic Forecasts and Reproducing Scoring Systems," RAND, Santa Monica RM-6299-ARPA, 1970.

[43]   G. W. Brier, "Verification of forecasts expressed in terms of probability," Monthly Weather Review, vol. 78, 1950.

[44]   R. L. Winkler, "Scoring rules and the evaluation of probability assessors," Journal of the American Statistical Association, vol. 64, 1969.

[45]   E. H. Shuford, Jr. et al., "Admissible Probability Measurement Procedures," Psychometrika, vol. 31, 1966.

[46]   T. M. Haladyna, Developing and Validating Multiple-Choice Test Items. Hillsdale: Lawrence Erlbaum Associates, 1994.

[47]   A. Oosterhof, Classroom Applications of Educational Measurement, Second ed. New York: Merrill, 1994.

[48]   J. C. Ory and K. E. Ryan, Tips For Improving Testing and Grading. Newbury Park, CA: Sage Publications, 1993.

[49]   L. A. Schoer, Test Construction: A Programmed Guide. Boston: Allyn and Bacon, 1970.

[50]   B. S. Bloom, M. D. Englehart, E. J. Furst, W. H. Hill, and D. R. Kratwohl, Taxonomy of Educational Objectives. New York: Longmans Green, 1956.

[51]   D. C. Gause and G. M. Weinberg, Exploring Requirements: Quality Before Design. New York: Dorset House Publishing, 1989.

[52]   H. Wainer, Computerized Adaptive Testing: A Primer. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.

[53]   G. W. Angell, "The effect of immediate knowledge of quiz results on final examination scores in freshman chemistry," Journal of Educational Research, vol. 42, pp. 391-394, 1949.

[54]   G. S. Hanna, "Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure," Journal of Educational Measurement, vol. 12, pp. 175-178, 1975.

[55]   T. L. Wentling, "Mastery versus nonmastery instruction with varying test item feedback treatments," Journal of Educational Psychology, vol. 65, pp. 50-58, 1973.

[56]   P. A. Supon, "Computer-Based Performance Assessment: An Annotated Bibliography," Department of Computer Science and Engineering, University of Colorado, Denver, Technical Report TR-34, September 1993.

[57]   P. A. Supon, "A Computer-Based Performance Assessment System for Chemistry, Qualitative Analysis," in Computer Science and Engineering. Denver: University of Colorado, 1994, pp. 94.

[58]   J. Paul, "Alternative Assessment for Software Engineering Education," in Software Engineering Education, J. L. Díaz-Herrera, Ed. New York: Springer-Verlag, 1994, pp. 463-472.

[59]   B. F. Green, Jr., "The promise of tailored tests," in Principals of Modern Psychological Measurement, H. Wainer and S. Messick, Eds. Hillsdale, NJ: Lawrence Erlabum Associates, 1983, pp. 69-80.

[60]   F. M. Lord, "Some test theory for tailored testing," in Computer-Assisted Instruction, Testing, and Guidance, W. H. Holtzman, Ed. New York: Harper and Row, 1970, pp. 139-183.

[61]   C. K. Riesbeck and R. C. Schank, Inside Case-Based Reasoning. Hillsdale: Lawrence Erlbaum Associates, 1989.

[62]   R. C. Schank, A. Kass, and C. K. Riesbeck, Eds., Inside Case-Based Explanation. Hillsdale: Lawrence Erlbaum Associates, 1994.

[63]   J. Kolodner, Case-Based Reasoning. San Mateo: Morgan Kaufman, 1993.

[64]   ICCBR-95, Proceedings of ICCBR–95, Case-Based Reasoning – Research and Development: First International Conference, Sesimbra, Portugal, October 23–26, 1995. New York: Springer-Verlag, 1995.

[65]   ICCBR–97, Proceedings of ICCBR–97, Case-Based Reasoning – Research and Development: Second International Conference, Providence, Rhode Island, July 25–27, 1997. New York: Springer-Verlag, 1997.