Experimental Validation of Learning Accomplishment

Allen Klinger
School of Engineering and Applied Science
University of California, Los Angeles
Los Angeles, California 90095-1596

Abstract

This paper reports on educational assessment: measures to validate that a subject has been learned. Outcomes described are from actual UCLA Computer Science courses, but the approach is independent of subject matter. There is a bibliography describing application of the methods presented here to other subjects and school levels. That bibliography summarizes an extensive literature including assessment in distance learning and elementary school situations. The text here outlines ideas and derivations, and the references enable deeper understanding, but a reader can use these procedures without either.

The methods are not the author's original invention. The following describes a way to apply them and to display students' learning. The paper contributes new ways to indicate student achievement and distinguish individuals with subject mastery from others tested. This is by figures shown here that enable teachers and students to understand and apply this form of testing.

The measures use probability and are based on concepts of information. Although both are discussed in the outline of derivations, it is unnecessary to understand either concept to test this way. A summary figure, the discussion of a practical implementation, and the references presented give the reader ability to vary weights shown here. That summary figure displays scoring based on logarithmically weighting the subjective or belief probability when individuals make one of thirteen possible responses. These are to large numbers of questions that can be completed with any of three candidate statements, one obviously false to anyone who knows the material. Three possible answers are intermediate between the remaining possibly true statement completions. They enable expression of partial knowledge, relative belief that either of the other two are more likely true or an inability to distinguish between them.

Each of the thirteen answers can readily be understood as subjective probability statements. The summary figure relates their weights to statistical concepts such as risk and loss. This applies to guiding students to answer with their true beliefs. Otherwise such mathematical ideas are not needed by testing participants. The practical effect is that thirteen responses enable converting three alternative statement completion questions into a means to evaluate whether complex material has been absorbed.

The method is generally applicable. It is related to earlier work on a computerized learning system called Plato that handled a wide variety of subjects at many educational levels. This paper describes ways to use an unconventional assessment approach to rapidly determine concepts not yet absorbed. New methods presented here are those I developed in classes I taught. Ideas and tools in this paper could empower others to expand and enrich their teaching and the learning processes it is to assist.

Acknowledgement. *Stephen Seidman commented on earlier drafts. Many thanks to him and others who supported this work: James Bruno, Martin Milden, Thomas A. Brown, E. Richard Hilton, Charles W. Turner, and David Patterson.*

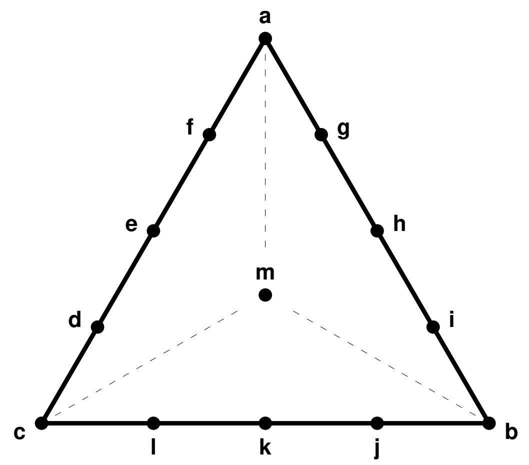
Introduction

Assessment methodology is usually thought of as divided between thorough methods that involve time consuming evaluation of student essays and rapidly scored multiple choice tests. There is a third way. It involves multiple choice where response corresponds to beliefs. That approach turns questions into a subtle tool to evaluate how much is known.

Knowledge and information are similar so one could start with logarithmic weighting [1-3]. Information theory uses that weighting to measure received transmissions. There probability is associated to different messages. The most is learned when one knows which specific value was sent from a group that were equally probable. Decision theory [4, pp. 14-15] adds the ideas of loss and risk. Risk involves average loss: it enables rational choice between actions when one expresses relative certainty about knowledge.

While a background in probability, statistics, information theory and decision theory is useful, that is not necessary either to design or take tests that use procedures explained below. A summary of the key weights I've made part of class use appears here as Figure 1. The following describes how those values relate to statements about belief in answers. Explanations relate them to logarithmic weighting and minimizing risk, something that translates into the traditional effort to gain the maximum score on a test.

1.a Three-Answers and Thirteen-Responses



If totally sure use a definite, **a**, **b**, **c** response.
If totally unsure use an **m** response.

If see one definitely wrong **a**, **b**, **c** value, choose among the remaining five on the opposite line.

Indicate preference between two **a**, **b**, **c** values by **d**, **f**, **g**, **i**, **j**, **l** responses.

If two of **a**, **b**, **c** seem equal, respond **e**, **h** or **k**.

1.b Values Between a and m Indicate Belief

Choice	*Letter*	*Points*	*Fraction*
Correct	**a b c**	30	1.000
Wrong	—	–100	0.000
Uninformed	**m**	0	0.769
Near Right	**d f g i j l**	20	0.923
Between Two	**e h k**	10	0.846
Near Wrong	**d f g i j l**	–10	0.692

The above point values make it advantageous to give **a**, **b**, **c** answers *only when more than 90% certain.*

The fractions can be used for scoring. They represent the amount of the range of points from [-100, 30] to award to any of the thirteen possible answers.

Figure 1. Knowledge Assessment and Response Scoring

1.c Loss Matrix

	a	**b**	**c**
a	30	-100	-100
b	-100	30	-100
c	-100	-100	30
d	-10	-100	20
e	10	-100	10
f	20	-100	-10
g	20	-10	-100
h	10	10	-100
i	-10	20	-100
j	-100	20	-10
k	-100	10	10
l	-100	-10	20
m	0	0	0

The definition of risk is expected loss.

Rows show points for answers when the respective column headings are true.

The loss-matrix values cause minimum risk only if **a, b, c** is chosen *when more than* 90% *certain.* Otherwise an **m** response has score advantages. People who recognize their lack of knowledge can get a reasonably high examination score.

Figure 1. Knowledge Assessment and Response Scoring

Although the basics have been in use for many years [5] instructional delivery innovations, such as video and internet or web use to facilitate remote learning, make it especially relevant today. I first saw it applied to a wide variety of subjects at Rand under the Plato system more than twenty years ago. There are commercial systems continuing that work today. [Plato refers to a project on computer-assisted learning initiated at the University of Illinois and involving Control Data Corporation, currently being applied to distance-learning.] Digital computers offer ways to extend procedures described in this paper to large numbers of people. Nevertheless the paper and pencil implementation of assessment within classes is in my opinion a more useful process. To display that, the following describes tests to measure university student accomplishment at undergraduate and graduate levels, summarizing the results, and comparison with more conventional evaluation procedures.
In [3] a cogent operations research analysis makes an argument in favor of logarithmic weighting of test results versus the usual all or nothing scores, say one for a correct answer, zero for incorrect. Let's call one or zero weights, *right or wrong* to keep the educational emphasis in view. Figure 1 presents another approach, weights that range

between zero and one because there are other possibilities than purely right or wrong. [References [1-3] start with logarithmic weighting. Several other reference and bibliographic items contain the words *admissible* and *reproducing*. They describe the mathematical quality scoring must possess, namely to ensure that on the average any response pattern that differs from an expression of one's true knowledge yields lower scores.] The weights strongly penalize guessing and reward accurate statement of exactly what one knows.

Both what is being weighted and statements about one's knowledge are actually subjective probabilities. When certain about an answer, one assigns unity to it and zero to all others. When totally uncertain, one is divided equally among the alternatives. The system rewards honest statements of total uncertainty and penalizes belief in the truth of something that is false. The next section describes this in practice.

Responding

In Plato at Rand responses to three alternatives were input to a computer by using a light pen. The response was initiated by pointing to a place in the middle of a triangle displayed on a computer video monitor. Any one of the points on the triangle or within it was immediately converted into a decision about one's belief, expressed as a triple of probabilities [6]. Figure 1 shows a triangle with **a**, **b**, and **c** vertices; each label corresponds to one of three completions to a statement. The diagram indicates other labels to signify less certainty. Group them into sets to highlight their similarities. We've five cases: α = {**a**, **b**, **c**}, β = {**d**, **e**, **f**, **g**, **h**, **i**, **j**, **k**, **l**}, γ = {**e**, **h**, **k**}, a difference set $\beta - \gamma$ = {**d**, **f**, **g**, **i**, **j**, **l**}, and δ = {**m**}. If someone believed **a** false they could give a **k** response. That would signify three subjective probabilities, conveniently written as a vector $(p_a, p_b, p_c) = (0, 0.5, 0.5)$. An **m** indicates subjective probability (1/3, 1/3, 1/3). When tests have an obviously wrong statement completion a person who knows something should quickly move to five alternatives. E.g., if **c** is nonsense choice should be among **a**, **b**, **g**, **h**, **i** , i.e., the three elements of β on the line joining the two valid letters. An **m** signifies inability to distinguish between the three completions. The α responses show belief that some **a**, **b**, or **c** is definitely correct. They express subjective certainty: this is assigning probability 1.0 to one completion and zero to the other two.

Every element of $\beta - \gamma$ is either close to or far from the correct completion. In terms of probability, either an element in $\beta - \gamma$ is close, say preferring **a** to **b** and believing **c** to be false [probability (0.75, 0.25, 0)] when **a** is correct, or far. In the far case the numbers are the same but **b** is the correct completion. Assignments of letters to points around the triangle is clockwise. Following apex **a** is **b** at the lower right triangle vertex, then **c** at lower left. The line between **c** and **a** has successively **d**, **e**, **f** as we move from the former to the latter, hence **d** is 1/4 of the distance toward the apex. The scoring scale ranging from negative 100 to positive 30 comes from a linear scaling of logarithms of subjective probabilities. The logarithms of probabilities 0.75, 0.5, and 0.25 scale linearly onto point values of +20, +10, and -10. The other point values, 0, 30 and -100 have consistent probabilities under the inverse linear scaling.

The fractional values are determined from the relative position of each score within the interval [-100, 30]. For example, since 0 occupies a position the 100/130 = 0.769 part of the distance from -100 to 30, an **m** response is associated with the fraction 0.769. Each of the other letters has a similar fraction associated to the corresponding point score for alternative correct **a**, **b**, **c** completions. These fractions appear in Figure 1; they have not been used in previous applications of this approach.

Figure 1 also presents the loss matrix. It yields the risk or expected loss from different decisions. *I.e.*, a value found by multiplying components of vector (p_a, p_b, p_c) by the loss entries in one matrix row. By choosing two neighboring letters and equating their risks it is possible to calculate what subjective probabilities lead to worthwhile definite answers, **a**, **b**, or **c** responses. E.g., consider a and f . Here one is assigning $p_b = 0$ and hence $p_c = 1 - p_a$. The result of equating the sums for rows a and f is quickly solved for p_a and it is found to be 0.9. This is the reason why Figure 1 contains "give **a**, **b**, or **c** responses only when more than 90% sure."

Experience

My previous teaching of courses at UCLA confirmed that the multiple response method discriminates between those with firm, some, and meager understanding of a subject. These experiments were created to show the idea to others. Since there are many sources on information and rational decision making, e.g., [7-10], the goal was to create a single page free of theory to supply needed tools for individuals to develop their own experience with this assessment approach. Figure 1 is that summary.

In past UCLA teaching I used this assessment method in undergraduate data structures; introductory Pascal programming; general students' computer seminar; mathematics for non-majors; continuing or adult education; and graduate courses. The following experiments come from two courses taught Fall '96 in the engineering school computer science curriculum, one graduate, the other for seniors. Graduate students had some exposure to loss and risk; undergraduates had minimal probability, most of them none. Both groups were able to work with multiple responses after less than fifteen minutes class discussion.

Initiating Multiple Response Assessments

The introduction emphasizes the essential features of this approach: partial credit, the need for answers to reflect knowledge, and the possibility of low total score unless one avoids guessing. Instead of probability, the discussion uses such terms as *elimination of a statement completion seen to obviously be wrong, preferring one response to another, etc.* To reduce guessing I stress the *greater than 90 % sure* statement. I reinforce that by a single question presented to the class: something no one knows. I compile the overall class responses and score them as in the Figure 1 fractions. At a later meeting there is a trial quiz, *e.g.*, five different statements, each with three completions. Students learn the high cost of mistakes on overall scores, and to express partial knowledge or lack of understanding. They gain points from m responses.

Two of the five graduate course initial quiz items follow; after them is a question from the undergraduate computer project design course:

1. A *pattern* is:
   a. A textured image.
   b. An element from a set of objects that can in some useful way be treated alike.
   c. An immediately recognizable waveform or alphanumeric symbol.
2. A *feature* is:
   a. An aspect of a pattern that can be used to aid a recognition decision process.
   b. A real number that characterizes an aspect of a pattern.
   c. The primary thing one observes in inspecting a pattern.

The undergraduate course included this:

3. A *project* is:
   a. A clear description of a product that one knows will be a success.
   b. A business term describing a useful way to involve several people in a large task.
   c. An effort that generally requires more than one person and always involves accomplishing a series of substantial parts.

Questions 1 and 2 involve definitions from class lectures. The method of composing questions includes using such obvious statements, and plausibly completing them with a near right but wrong item, and another clearly wrong statement as well as the correct words. Here 3 has c correct while b is nearly right but.wrong because of the word *useful*. Many such questions *aggregated* are an effective way to measure learning when coupled to the multiple response system. Composing questions this way is easy for the instructor: all that is needed is to begin by recording key points taught each week. Quizzes can be weekly. Many quizzes can be expanded into longer examinations. The examinations with more than thirty questions can separate students based on their actual learning.

Overview of Results

The primary issue one seeks to measure is: *did the students actually acquire the technical knowledge we presume was transmitted.* An index that represents a numerical quantification of learning results from adding up the Figure 1 second column fraction scores obtained from all questions. This measure validates accomplishment two ways. First, it gives a methodology to differentiate between obtaining *general understanding* and acquiring *exact knowledge*. This is the point of the analysis in [3]. However what is new here is overwhelming classroom evidence in experiments that *the measure places individuals into two distinct groups*. In the better performing group one individual had a near perfect response pattern. This near-perfect score validates the instruction and shows that others could learn. The outcome is reasonable since all questions are simple for one who knows the material. The inferior performance group contains those lacking some understanding of one or more fundamental points. Low scores indicate misunderstandings that produce questions, interaction, and discussion that supports active learning, a further benefit of the method. The experiments also compare multiple-response measures with traditional means: examinations including problems to solve. This was done twice, in the middle and end of the term. Homework and term papers reinforced the comparison: student work was rated similarly by the different systems. The graduate course also included an open-ended question done over a five-day period. Grades on this item were almost identical with the scores from thirty multiple-response questions assigned to be answered within three hours. Once again, a traditional means of evaluating students was consistent with the use of multiple-response questions.

A simple view of the index compares the aggregate value to that from reporting no knowledge. Any consistent declaration of *no knowledge* about posed questions receives a numerical score of 0.769 or 76.9%. Examining whether a score falls above or below that value determines whether learning actually occurred. Dividing students relative to the 76.9% line is a simple use of the information about accomplishment. Far more can be seen in diagrams such as the Figure 3 scatter plot, a visual showing actual class sample statistics. A tabular display can also enable useful comparisons of individuals to their peers.

Numerical Assessment Data

The experiments involved student scores for a trial quiz and a midterm in two Fall 1996 classes. Table 1 shows the graduate class trial quiz having four main groups. This is evident from a histogram of these scores given here as Figure 2, where the advanced and marginal groups stand out. The trial quiz differentiates four in the lowest two groups, labeled X and Y, and five superior performances, labeled E, in a population of twelve. The two four in the lowest two groups had scores *below* what they could have achieved by uniformly answering **m**, *know nothing* about this topic. The category labeled F gained only a slight amount above that level. But all five in category E saw the statements as obvious.

Table 1. Graduate Course Quiz Data - Five Definition Questions

	Score (%)	Wrong (#)	No Knowledge (#)	Group
	56.9	2	0	X
	60	2	0	X
	67.7	1	0	Y
	72.3	1	1	Y
	76.9	1	0	F
	80	1	0	F
	80	1	0	F
	93.8	0	0	E
	93.8	0	1	E
	93.8	0	0	E
	95.4	0	1	E
	96.9	0	0	E
μ	80.6	0.8	0.3	
σ	13.7	0.7	0.4	

To see how stable the results were, limited time midterm examinations using the thirteen choices had take home sections. The scores from the essays and problem solutions in those sections correlated almost perfectly with multiple choice measures. The experiment validated my past experience, where in over a decade I found consistency of fifty item final, and thirty question midterm scores *with grades on projects and papers*, in many different courses, including data structures, and pattern analysis.

Course Completion

This section concerns how the multiple response measures indicate summary or overall level of accomplishment. The framework was a ten week term, lecture/recitation class. and an eleventh final examination meeting. The primary data is a table with components derived from a spreadsheet of the data. Data from the midterm is shown here in Table 2 and Figure 3. More information helps one compose the final grade. I created tables that start with a row of average values of eight quantities: overall score, followed by the numbers of answers that were correct, incorrect, no choice **m**, preferred a correct answer, equally comfortable with correct and wrong answers (**e**, **h**, or **k**), preferred an incorrect answer, and then the sum of the correct and prefer-correct values. The left-most column simply lists the students' scores. Then I varied the tables along lines in Table 2 below where italics show "outside one standard deviation interval on either side of the sample mean value" to highlight different kinds of student accomplishment.

Table 2. Graduate Course Multiple Response Midterm Detail -Thirty Questions

	Score	Correct Correct	Near Near	Correct & Know	Didn't	Wrong
	66.9	16.	1.	*17.*	*9.*	3.
	70.	21.	0.	21.	*9.*	*0.*
	72.6	16.	2.	*18.*	*7.*	2.
	79.5	19.	0.	19.	5.	3.
	86.9	22.	1.	23.	3.	3.
	90.8	17.	6.	23.	1.	3.
	92.	17.	1.	*18.*	*0.*	*6.*
	92.3	19.	5.	24.	1.	2.
	92.6	*15.*	8.	23.	*0.*	3.
	93.3	*28.*	0.	*28.*	2.	*0.*
	95.4	*27.*	0.	*27.*	1.	1.
μ	84.8	19.7	2.2	21.9	3.5	2.4
σ	10.	4.2	2.7	3.5	3.3	1.6

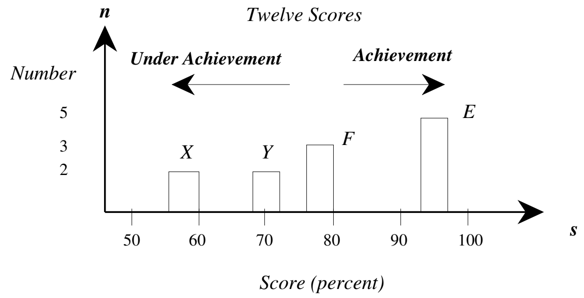


Figure 2. Graduate Course Quiz Histogram

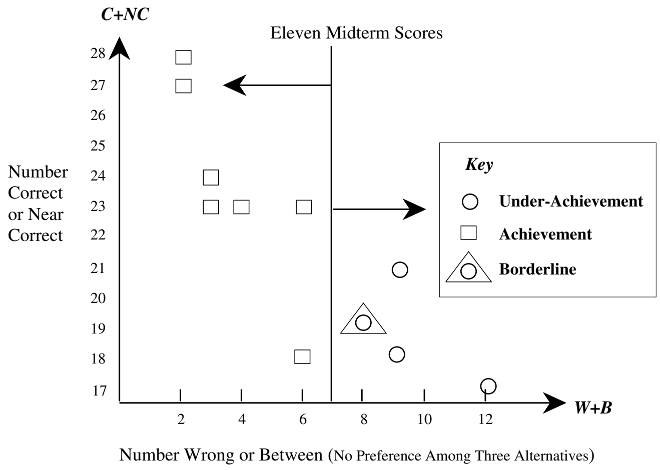


Figure 3. ***Scatter Plot of Graduate Course Midterm Detail***

Other material not included here confirms the general value of testing, although some students with poor performance during the term did well on twenty multiple response questions. Overall higher course grades were earned by students who tested well on the final and the quizzes, when the most important measure was project work. This leads to a positive conclusion about the multiple response measures: it seems clear from these experiments that this method yields more than right or wrong scoring, and that it is an efficient and engaging way to organize student and instructor interaction.

Administering Tests

The basic material needed to administer a multiple response test is the diagram portion of Figure 1, which draws upon communications from J. Bruno, for its version of the triangle image, and thirteen letter labels. Once the choices are made by students the instructor presents the correct letter answers. Students exchange and score each others' papers. That quickly leads to the assessment index values. Students often total question points and compute exam score or percentage: hand calculators make this rapid. To see the scoring method, consider first five, then seven questions. Suppose **a** is the correct first question answer but the response was **g** because a belief in **a** of 3/4 and 1/4 for **b** seemed better than professing certainty in **a** and possibly losing all points. If **a** was correct this obtains a 0.923 fraction score. If questions two through five had **c**, **b**, **c**, and **b** correct completions but **c**, **h**, **h**, and **m** responses, the corresponding scores would be 1.000, 0.846, 0, and 0.769. That yields a total of 3.538 on the first five for an average score of 0.7076 or 70.76 %. With the same first five, a wrong guess on question six and a poor choice at the seventh would reduce the score significantly. If a seventh item got a 1/4 weight at correct completion: *e.g*., **c** correct but **f** response, the five question total of 3.538 becomes 4.230 for seven since 0 would be awarded for the error at sixth question and 0.692 for a poor choice at the seventh. The seven question result is thus 0.6043 or 60.43%which would correspond to a low D. This is down a full grade from the low C earned for the first five questions, because of lack of knowledge on the final two.

Conclusion

Increased production and availability of information and informal learning occurring from television and world wide web sources of varied quality and accuracy causes renewed interest in accurate assessment. This paper has described applying multiple response assessment in classrooms using just paper and pencil. The results include material on creating assessments using this method. They also show three new means to display score results that show low and superior achievement. Three technical discussions explain scoring. One is from the viewpoint of logarithmic weighting of subjective probability. Another describes bounding regions of such probability where one decisive answer should be chosen for on the average best performance in terms of point scores. Finally, an example displays how failure to understand a few key points can lower a grade a full letter value under this method of scoring. If this paper points others to an assessment process they can use it will be a successful effort.

References (RM or R signifies Rand Corp. Report)

[1] Hartley, R V, "Transmission of Information," *Bell System Technical Journal*, 7/28, p.535.

[2] Shannon, C E, Weaver, W., *The Mathematical Theory of Communication*, Urbana IL: '49.

[3] Brown, T A, "A Theory of How External Incentives Affect, and Are Affected by, Computer-aided Admissible Probability Testing," Rand Corp., Santa Monica CA, '74.

[4] Duda, R O, Hart, P. E., *Pattern Classification and Scene Analysis*, Wiley, NY, '74.

[5] Landa, S.,"CAAPM: Computer Aided Admissible Probability Measurement Plato IV," R-1721-ARPA, '76

[6] Sibley, W L, A Prototype Computer Program for Interactive Computer Administered Admissible Probability Measurement, R-1258-ARPA, 1973.

[7] Brown, T A, Shuford, E., "Quantifying Uncertainty Into Numerical Probabilities Reporting Intelligence," R-1185-ARPA, '73.

[8] Savage, L J, "Elicitation of Personal Probabilities & Expectations," *J. Amer. Stat. Assoc*. '71, 783-801.

[9] Good, I J, "Rational Decisions," *J. Royal Statistical Society, Ser. B, 14*, '52, pp. 107-114.

[10] McCarthy, J., "Measurement of Value of Information," *Proc. Nat. Acad. Sci*., '56, pp. 654-5.

Bibliography

Eisenberg, E., Gale, G. "Consensus of Subjective Probabilities," *Ann. Math. Stat., 30*, 3/59, pp. 165-168.

Epstein, E. S., "A Scoring System for Probability Forecasts of Ranked Categories," *J. Appl. Meteorology, 8* (6), 12/69, pp. 985-987.

Winkler, R. L., "Scoring Rules and Probability Assessors," *J. Amer. Stat. Assoc., 64*, '69, pp. 1073-1078.

Shuford, E, Albert, A., Massengill, H., "Admissible Probability," *Psychometrika, 31,* '66, pp. 125-145.

Shuford, E., Brown, T.A., "Elicitation of Personal Probabilities," *Instructional Sci. 4*, '75, pp. 137-188

Brown, T. A., Probabilistic Forecasts and Reproducing Scoring Systems, RM-6299-ARPA, 6/70.

Brown, T. A., An Experiment in Probabilistic Forecasting, R-944-ARPA, July '73.

Brown, T. A., Shuford, E., Quantifying Uncertainty Into Numerical Probabilities, R-1185-ARPA, '73.

Bruno, J. E., "Using Computers for Instructional Delivery and Diagnosis of Student Learning in Elementary School," *J. Computers in the Schools, 4* (2), '87, pp. 117-134.

Bruno, J. E., "Computer Assisted Formative Evaluation Procedures to Monitor Basic Skills Attainment Elementary Schools," *J. Comput. Childhood Educ., 2,* 91, pp. 79-103.