HOW TO EVALUATE SCIENCE PROBLEM SOLVING IN A COMPUTERIZED LEARNING ENVIRONMENT? CONSTRUCTION OF AN ANALYZING SCHEME

Zvia Fund

ABSTRACT
This paper describes the construction of a '*co*mputerized *s*cience *pro*blem *s*olving' scheme, which enables analysis and evaluation of the effectiveness of science problem-solving by junior high-school students working in a computerized learning environment. The scheme was based on observations of 187 students as they solved qualitative science problems taken from a specific computerized learning environment. Students were also interviewed before and after the problem solving. The scheme is presented on two levels. The large-scale comprises 11 main categories, each sub-divided into sub-categories to yield the detailed-level. The sub-categories were based on a repertoire of activities found in the observation protocols, and were approved by external judgement and a validation process. The detailed-level scheme enables evaluation and statistical analysis of the participants' problem-solving effectiveness, providing substantial evidence for the construct validity of the scheme, and demonstrating its potential as a valid analyzing and evaluative tool for computerized science problem solving.

KEYWORDS
Science problem-solving, computerized learning environment. analyzing scheme, effectiveness score, problem solving protocol

INTRODUCTION

Recent years have seen the development of many computerized learning environments, which enable students to engage in interactive computerized learning activities such as solving problems. For research purposes, there is a need for a tool that will enable the researchers to analyze, evaluate, and assess such activities.

This paper describes the construction of a two-level scheme: a large-scale scheme, which served as an analyzing tool, and a detailed scheme, which served as a coding instrument to evaluate the effectiveness of students' problem solving. The construction of the scheme was part of a wider research project (see Fund, 1999, 2002), and was based on observation of 187 students solving science problems in a computerized learning environment called 'Inquire and Solve' (Educational Technology Center, Israel), while their problem-solving activities were transcribed. The computerized environment is a micro-world, which combines a problem-solving environment with a simulation of laboratory experiments. The construction of this '*co*mputerized *s*cience *pro*blem *s*olving scheme' (COSPROS) is demonstrated by a description, analysis and evaluation of the solving protocols of two students. The use of the detailed scheme with the normative sample of 187 students, which revealed significant differences among the treatment groups, provides a substantial and external evidence for its construct validity (Messik, 1994).

RESEARCH ORIENTATION

A theoretical issue addressed is the debate between cognitive and situation-oriented (situative) perspectives. The cognitive perspective on knowledge emphasizes general cognitive abilities (such as reasoning, planning, and solving problems) along with general problem-solving heuristics (Greeno, Collins & Resnick, 1996). The situative perspective looks at knowledge as determined by unique people and their unique environments. The latter perspective focuses on processes of interaction among individuals and their physical and technological systems (Anderson et al., 2000; Greeno, Collins & Resnick, 1996).

In the construction of our tool we try to navigate between common features of different learning environments or contexts, and the requirements of a specific environment and a certain science domain. At each level of the instrument we incorporated characteristics of both generalization and specificity, thus integrating the cognitive approach and the situative approach.

Complex activities, such as problem solving, should be studied in the context in which they occur, which addresses the methodological issue of collecting and analyzing verbal data (i.e., explanations, interviews) as well as non-verbal data (observations and problem-solving protocols) (Chi, 1997). In the current research, observing the students' actions and activities while manipulating the computerized tools, which yields some sort of transparency of the solving process (Fund, 2002), supplied the non-verbal data, while the verbal data included spontaneous verbal comments, as well as questions and answers with the interviewer. In applying the constructed COSPROS scheme to analyze such multi-faceted data, both methods of verbal analysis and protocol analysis (Chi, 1997) are actually employed.

METHOD

The Computerized Learning Environment

The Inquire and Solve computerized micro-world, combines a problem-solving environment with a simulation of laboratory experiments. It consists of 60 qualitative science problems, of which 42 were found to be adequate for the science curriculum of the present (seventh grade) research population. Each problem presents a question represented by textual and graphical components (e.g., 'Which vessel contains the greatest amount of air: 1, 2 or 3?', 'One of the coils in this system is made of copper, another of iron, and a third of aluminum. What is coil no. 2 made of?', 'In which gas compound, 1 or 2, do the particles move more quickly?') .

Various tools, represented by icons, are provided. Using these tools, the learner is able to find the answer by 'performing' the experiment, observing its results, collecting missing data from available sources, and deciding which data are relevant to the current problem. The most important tools are listed below:

> *Camera* – allows the learner to move from one episode of the simulation to another (2–5 episodes per problem). *Magnifying glass* – provides information about specific graphical parts of the experimental system, and enables the learner to obtain more information or more refined information. *Data pages* – A simulation of a data book gives as much as six kinds of information (e.g., boiling point, density, tendency to be notched, and scales of proportional values for physical properties). *Watch* – Various measuring tools are provided (e.g., thermometer, voltmeter, current meter, manometer). *Answer flag* – Upon presenting the suggested solution, the learner receives the appropriate feedback (correct/not correct). *Guide* – textual guidance for the suggested steps.

Participants

The 187 participants were seventh-grade students studying from the same science textbook, and had worked with the Inquire and Solve computerized environment approximately once every 2 weeks for 6-

months. The wider study included several additional treatments (for details of the broad study, see Fund, 2002).

Procedure

This study includes data from interviews with and observations of all 187 participants. Students were individually interviewed for about 25 minutes, and their problem-solving activities (including spontaneous remarks, questions, and explanations) were observed and transcribed. Before the problem-solving session began, each student was given a sheet of paper and was asked to use it (for note taking), whenever needed, during the solving process. A "mathematics and reading comprehension" questionnaire was administered before the beginning of the study to sort the participants into three academic ability levels.

THE INSTRUMENT - "COMPUTERIZED SCIENCE PROBLEM SOLVING SCHEME"

The Basic Primitive Model

Previous research states that effective science problem solving comprises three major steps: Initial problem analysis (also problem description or problem representation); construction of a solution; and checking the solution (or self-monitoring) so that it can be appropriately revised (if necessary).

These three stages are too general to describe different solving processes or to distinguish among them, and further sub-division into categories is required. This led to the design of the overall 'large-scale' COSPROS scheme, described below.

The Large-Scale COSPROS Scheme

The scheme presented in Table 1 below consists of eleven main categories, as well as two additional categories ('intentional learning' statements and non-relevant noises). In 'intentional learning' statement we mean explicit verbalization about intended steps or the current solving process. Such statements are very important in any learning task, and serve as evidence for mental effort, motivation and willingness to learn and understand (Bereiter & Scardamalia, 1989). Non-relevant noises are technical errors in manipulating the computerized environment, or waiting for instructions from the interviewer (e.g., "shall I begin?").

Eight of the eleven main categories involve cognitive skills while three incorporate meta-cognitive skills, referring to self-assessment and finding the error.

Table 1. The Large-Scale COSPROS Scheme

Stages	Main Categories of cognitive and meta-cognitive skills	Comments *
Initial problem analysis	*Initial analysis:*	
	Finding the goals of the problem	*Cognitive, resource-ind.*
	Collecting data for problem description	*Cognitive, resource-dep.*
	Translation into scientific language:	
	Global: identifying the subject of the problem	*Cognitive, resource-ind.*
	Specific: mapping the problem subject into natural language	*Cognitive, resource-ind.*
Construction of a solution	Collecting missing data	*Cognitive, resource-dep.*
	Using the collected data in the problem (reasoning is required)	*Cognitive, resource-dep.*
	Reaching a solution	*Cognitive, resource-ind.*
Checking the solution	Self-assessing the problem-solving process	Meta-cog., *resource-ind.*
	Assessing the final answer	Meta-cog., *resource-ind.*
	Explaining the method of solution	*Cognitive, resource-ind.*
	For incorrect solution: finding the error and its causes	Meta-cog., *resource-ind.*
Others	Intentional learning statements	
	Non relevant noises	

* Each category is either cognitive or meta-cognitive (*meta-cog*.), resource-dependent (*resource-dep*.) or resource-independent (*resource-ind*.) (see explanation below).

As can be seen in Table 1, the eleven *main* categories might be sorted by resource-dependency into: (a) resource-independent categories, appropriate for almost any computerized learning environment; and (b) resource-dependent categories (2^{nd}, 5^{th}, and 6^{th}), with unique characteristics related to the specific computerized learning environment. This scheme might serve as an analyzing tool, to find out which categories are performed in the solving process and which are missing, as illustrated below. Yet, sub-division of each category is necessary so the scheme might differentially evaluate students who utilize the same categories but in different ways and effectiveness.

TWO CASE STUDIES

Eli And Shani are both medium-academic level students (according to the mathematics and comprehension questionnaire), but their approaches to the problem (described below) are different. Their problem-solving protocols are analyzed using the large-scale scheme. The appropriate categories and some interpretation are added (*in italic*).

The Problem

Electrical conductivity at the most basic level is presented on-screen as follows: "One of the coils is made of copper, the other is made of iron. What is coil 2 made of?"

The problem consists of four separate on-screen episodes, depicting an electrical circuit with its components. The coils are connected to the circuit one at a time, and the switch is off and on for each coil intermittently. Comparing the current, the student has to conclude that coil 1 has higher electrical

conductivity than coil 2. The 'data pages' must be used to obtain values for the electrical conductivity of copper (6 on the scale) and iron (2 on the scale), leading to the answer that coil 2 is made of iron.

The solution requires that the student understand the meaning of closed and open electric circuits, and infer the relationship between electrical conductivity and intensity of the current in the circuit. (i.e., that a higher conductivity enables a higher current to flow, leaving all the other variables unchanged.)

Protocol: Case 1 - Eli

Eli begins by checking the menu, and identifying the main subject (*identifies the subject using an external hint*--3rd category). Then he reads the on-screen question and copies it to his note-taking paper (1st category). Now he begins to collect data for problem description (2nd category), by taking the magnifying glass (*a computerized tool)*, and putting it carefully and systematically on every circuit's component, in each of the four episodes. When he notices a new measure in the current meter he rechecks it, as well as rechecking any information while constructing the solution (*self-assessment during the solving process*--8th category). He systematically formalizes--by writing a detailed story of any important data he collects while solving the problem (2nd and 5th categories). While solving, he verbalizes many spontaneous comments, (e.g., "I always begin with it to find out what I'm given in the problem", or "coil 2, 0.2 ampere. I've just finished, it's the last episode"--*both implying 'intentional learning' statements*). To the interviewer's question of how he knows it, he answers: "there are two coils, and I almost know what the current is for each coil (he *maps the experiment to the main subject*--4th category) and I'm coming to a conclusion". Seeing an unknown word he asks the interviewer about its meaning (e.g., "what is ampere?"). Before going to the next stage he reads silently his own notes (2nd category), and explains spontaneously: "First I collect all the data; when I have it, I turn to the question and answer it" (*Demonstrates a mental distinction between describing the question and constructing the solution*).

The construction of the solution stage includes performing the 5th category by taking the data pages tool, knowing the relevant property to look for (electrical conductivity) and the relevant materials (iron and copper). He compares their values but comes to the inverse conclusion ("iron conducts better"). By re-scanning all the episodes and rechecking the data pages (*self-assessment during the solving process*--8th category) he recognizes his wrong conclusion and amends it, saying "Higher value means greater conductivity, so copper conducts better" (*uses the collected data in the* problem--6th category). He says he knows the answer (7th category) and spontaneously writes a detailed correct answer (10th category). At the end he says he wants to check the answer, and does it by using the answer flag tool (*self-assessment at the end of the solving process*--9th category), and writes the feedback "correct answer" (*documented confirmation*).

Eli's solution consists of the first ten categories and the intentional learning (12th category). The 11th category is missing since the answer was correct. Intuitively his solving process seems to be quite effective, but it is far from an accurate evaluation of his solution, implying the scheme construction is not yet finished. In the next protocol more categories are missing, and quite a poor solving process is exhibited.

Protocol: Case 2 - Shani

Shani's initial problem analysis begins with reading the on-screen question (1st category). Collecting data for problem description (2nd category) is then performed, by taking the magnifying glass, putting it on some of the circuit's components in three (out of four) episodes, (*skipping some components and missing the last important episode --interpreted as a random search*). She makes no comment while collecting the data and writes nothing. Before she begins the construction of the solution she rereads the on-screen question (1st category), then uses the data pages tool (5th category) knowing the relevant property to look for (electrical conductivity), but she scrolls the tabular information up and down, not focused on specific materials. S*he probably does not know the relevant materials since she has not written any note*). She rereads once again the on-screen question (1st category) (*probably as a way to see the names of the materials again*). She takes data pages again (5th category), but scrolls it as before,

unfocused (she *has just looked at the names of the materials, but does not connect the two kinds of information so as to use the data pages effectively*). At that point she says "I don't know the answer", but still makes another trial and looks again at the graphical information of episodes 1 and 2 (*but not of episodes 3 and 4, implying a random search for information*--2nd category). She cannot solve the problem yet, since she still misses the current meter in episode 4 and the electrical conductivity scale of iron and copper. Although, she takes the answer flag tool, gives iron as a guessed answer (*ineffective utility of the* 9th category), gets the feedback "you can't answer yet", but does nothing else. To the interviewer's question about her answer she gives the following explanation, while showing it on the screen: "When we put coil 1 the bulb lights up, and in coil 2 it doesn't light up. In the data pages it's written iron - 2, and here camera2 (*episode 2*) it's written 2, so it gives hints. Iron conducts better, and here (*shows episode 2*) the switch is on and the bulb lights up."

Shani's explanation was entirely wrong; she merely interpreted what she saw so as to fit her answer. She used surface and external characteristics of the problem (e.g. number of the coil and number on the conductivity scale in the data pages to bolster her explanation), which characterizes poor solvers and novices, as compared to the substantial characteristics of a problem that expert and good solvers refer to (Chi, Feltovich & Glaser, 1981). The overall solution consisted of utilizing the 1st, 2nd, 5th, 9th and 10th categories quite ineffectively, using the computerized tools in a technical manner, while the other required categories (3rd, 4th, 6th, 7th and 8th categories)--which induce incorporating cognitive resources--were missing. Additionally, external representation (written notes) of the problem or of the collected data and intentional learning statements were not found in the solution, which makes such a solving process, intuitively, not effective as compared to Eli's.

In order to conduct an accurate evaluation of the two protocols and enable their comparison, more sub-division is needed to differentiate between effective and non-effective activities which are accommodated in the same category. In the 5th category (collecting missing data), for example, taking the data pages tool when the learner does not know the relevant property and/or the materials' names (see Shani's solution), is obviously less effective than doing so when the desired property and the materials' names are already known (see Eli's solution). Thus, subdivision of each main category to construct a detailed scheme should be the next step in constructing the scheme, as is further elaborated in the detailed COSPROS scheme.

THE DETAILED COSPROS SCHEME

Each one of the first eleven main categories was subdivided into specific sub-categories or codes, which cover all possible instances of the category.

Such sub-division should, by definition, be more context-dependent. In the current research accounts they should be specific to the observed 'Inquire and Solve' environment. We tried to find general principles for the sub-division of each category, leaving the examples of the sub-categories to refer to the current environment and research. Doing so, some generalization is inserted into the context-dependent detailed scheme, which enable other researchers to adapt the detailed scheme as is, or perform required changes to adjust it for other computerized learning environments. As an illustration the 1st main category 'finding the goals of the problem' is sub-divided and presented in Table 2 below.

Table 2. Sub-Categories Of The 1st Main Category ('Finding The Goals Of The Problem')

Sub-categories	**Examples**
Reading the problem:	
1.1* Reads the problem from the screen	
1.2 Reads the problem using own external representation of the problem	Reads the problem from own notes on the page
1.3 Reads the problem with initial processing of its meaning	Points to the screen while reading the problem, reads loudly, etc.
Making external representation of the problem:	
1.4 The external representation (on a page or in a computerized notebook) is a "primitive" copy of the problem	Copies the problem onto the page
1.5 Makes an elaborated external representation of the problem (the essence of the problem) on the page or in a computerized notebook	Records (onto the page) the names of materials and the question's key words (that have to be found)

* Codes of sub-categories

As can be seen in Table 2, the sequence of sub-categories (from 1.1 to 1.5) reflects the transition from a surface reading of the problem towards a deeper, more thoughtful reading of the problem (first principle of sub-division of the 1st category), with external representation of the problem (second principle of sub-division). Application of the sub-categories to the case studies cited above demonstrates the different qualities of these processes. For example, Shani read the question from the screen several times (sub-category 1.1), probably without extracting all the important information. Eli read and copied the question (sub-category 1.4), creating a simple external representation of the information. We observed other students, not described in the current paper, who read and pointed to the screen, or read the problem loudly, which reflected initial processing of the given information (sub-category 1.3). Even more elaborated instances included reading and creating an external representation of the *core* information such as the given materials names and the exactly required information (sub-category 1.5).

Effectiveness Scores For The Sub-Categories

The detailed scheme may be used to evaluate in a precise quantified manner the effectiveness of the solving process of any student, and thus allow comparison of any two protocols or between experimental groups. For that purpose, the sub-categories of each main category were arranged in ascending order of effectiveness, and each sub-category was assigned effectiveness score, both based on the judgment of three external judges (all teachers of the research students), after content validity process (98% agreement among judges). Such assignment was by no means context-dependent or particularly research-dependent. The judges (based on research literature) accounted for the research goals, and accordingly assigned higher effectiveness scores for specifically expected sub-categories and lower to others. Examples of expected sub-categories in the current research included constructing external representation of the problem or the collected data, predicting an answer, self-assessment during the solving process and at its end, etc. In assigning the effectiveness scores a variety of scoring scales, such as 2, 3, or 5-point scales, might apply depending on the evaluating purposes and the required differentiation. Accordingly, in the 1st category, for example, creating external representation is favored and hence accommodated within the more effective sub-categories (1.4 and 1.5), with effectiveness score of 4 using a 5-point scale (0-4).

Use of the detailed specific scheme as a research tool with our original group of 187 subjects is presented below.

MEASURING PROBLEM SOLVING EFFECTIVENESS

For analysis purposes, each solving protocols is divided into 'units of analysis' (any action or verbal statement), which are then individually analyzed, and ascribed to the corresponding sub-category (the coding process). For reliability purposes, 22 different protocols (305 units) and the resultant detailed scheme were given to three judges, who ascribed the units to the relevant sub-categories. At the first stage the inter-rater agreement was only 78% with disagreement of 8%, and 14% of the units were not coded by one of the judges. Accordingly, the disagreement and ambiguity were discussed and some sub-categories were redefined. A second coding process showed 98% of inter-rater agreement, resulting in a final detailed scheme, agreed upon by all the judges.

In any scoring scale, each student might be assigned a most effective (or least effective) category score, corresponding to the student's most (least) effective sub-category performed at least once, over all his/her observed problems. The total score is the sum of the scores for all the categories, as a percentage of the maximum available score. Applying a 5-point scale to the solving processes of Eli and Shani, (0= completely ineffective to 4=very effective; maximum 40 points, 4 points X ten categories) yields effectiveness scores of 97.5% for Eli (39 points) and 17.5% for Shani (7 points). More results are presented below.

RESULTS

In the current research, a 5-point scale (0-4) and 2-point scale (0-1) were implemented. Two measures - maximal and minimal effectiveness - representing the highest and lowest scores, respectively, were derived from the 5-points scale, for each category, for each student. From the 2-points scale a third measure 'global effectiveness' was derived: each student got for a given *category* the score 1 ('effective'), if the student performed any of the effective sub-categories of the given category, at least once over all his/her observed solved problems, and 0 ('ineffective') otherwise.

These three measures--maximal, minimal and global effectiveness-- were evaluated for the 187 participants for each category and for the whole solution. The results were subjected to a 5x3 MANOVA analysis (5 treatments x 3 academic levels). The results for the *whole solution* indicated significant differences across all the three measures between the 5 treatments groups ($F(12, 450)=15.03$; $p<.001$), and between the three academic levels ($F(6,340)=7.00$; $p<.001$). A subsequent ANOVA (5 groups by 3 academic levels) analysis of the maximal effectiveness for the *whole solution* showed a significant differences between groups ($F(4,172)=52.11$; $p<.001$). ANOVA 5x3 analysis of maximal effectiveness and of global effectiveness for *each category* showed highly significant differences between the groups (all categories except 2nd category). Significant differences between academic levels were found in some categories (7th, 10th, and 11th categories). An interaction effect of treatment and academic level was found as well (5th and 9th categories). These results (not discussed in the current paper) are analyzed and discussed in Fund (1999, 2002). These provide additional substantial and external construct validity of the detailed scheme (according to Messik, 1994).

CONCLUSION

The two-level scheme, constructed and validated as described above, provides a reliable and valid analyzing tool. The scheme is useful for deriving various effectiveness measures (as well as other measures such as the 'length' of each category in percentages from the whole solution), and for measuring the effects of different treatments on cognitive and meta-cognitive skills in problem solving within a computerized science-learning environment.

The large-scale scheme is more general and thus less context-dependent than the specific detailed scheme. We believe the large-scale scheme is applicable in most computerized learning environment, for many researches, while the method to construct and to use the detailed scheme might be adapted and adjusted to fit the research specific goals and context. Other researchers might apply the general scheme

and additionally both the constructing methods of the detailed scheme and its application techniques. Yet, they probably should construct their own sub-categories, to fit their specific goals and researches. We still feel that a higher-level scheme, or meta-scheme, is required for science problem solving in any resource-rich learning environment. We call, therefore, for collaborative research with those who may wish to adapt any part of the scheme, so that we may together construct principles for such a meta-scheme.

REFERENCES

Anderson, J. R., Greeno, J. G., Reder, L. M. & Simon, H. A. (2000). Perspectives on learning, thinking and activity. Educational Researcher, 29(4), 11-13.

Bereiter, C. & Scardamalia, M. (1989). Intentional learning as a goal of instruction, in: Resnick, L.B. (Ed.). Knowing, Learning and Instruction. (pp. 361-392). NJ: Lawrence Erlbaum Assoc.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. The Journal of The Learning Sciences, 6(3), 271-315.

Chi, M. T. H., Feltovich, P. J., Glaser, R., (1981). Categorization and represantation of physics problems by experts and novices. Cognitive Science, 5, 121 - 125.

Fund, Z. (1999). Models of written communication as a cognitive support for computerized science problem solving. Paper presented at a conference on "Roles of Communicative Interaction in Learning to Model in Mathematics and Science", organized by the Leeds University Computer-Based Learning Unit, Corsica, France.

Fund, Z. (2002). Cognitive support in computerized science problem solving: Eliciting external representation and improving search strategies. In P. Brna, M. Baker, K. Stenning and A. Tiberghien (Eds.), The Role of Communication in Learning to Model (pp. 127-154). New Jersey: Lawrence Erlbaum.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R.

C. Calfee (Eds.), Handbook of Educational Psychology (pp. 15-46). New York: Macmillan.

Messik, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Reif, F. (1995). Millikan lecture 1994: Understanding and teaching important scientific thought processes, American Journal of Physics, 63(1), 17 - 32.

Fund Zvia
School of Education
Bar-Ilan University
52900, Ramat-Gan
Israel
Email: fundzv@mail.biu.ac.il