Third Misconceptions Seminar Proceedings (1993)

Paper Title: Going Beyond the Written Word - what performance assessment can tell us about concept understanding
Author: Harmon, Maryellen

Abstract: This paper will discuss the science portion of some of the early findings of a study of multiple choice and alternative forms of assessment presently being conducted in urban schools by Boston College, Center for the Study of Testing, Evaluation, and Educational Policy. The study, under the title Urban Development Assessment Consortium (UDAC) is targeted for 11 major urban school districts and aims to develop assessments that will meet two requirements: to provide diagnostic information to classroom teachers and local school administrators for the improvement of instruction, and at the same time to provide a monitoring system to district evaluators and policy boards. Teachers need information on what students are thinking and how they are personally making sense of what is taught, information that is not provided merely by the student's ability to select a correct answer on a multiple-choice test. School districts need information on the strengths and weaknesses of schools in effecting student learning so as to know where to allocate resources, what kinds of teacher training to provide, where innovations are more effective than prior programs and should be supported, and where there may be waste of resources in ineffectual programs. Districts and institutions of higher education presently make high stakes decisions based on the results of commercially available multiple-choice tests which in fact do not sample higher order thinking or in-depth concept understanding, and tell little or nothing about students' abilities to solve the kinds of poorly structured non-routine problems they will meet in future life. Although the initial findings discussed here are very early and partial, and therefore any conclusions must be at best tentative, I believe some description of both process and results is relevant to the intent of this seminar for a number of reasons.

Keywords:
General School Subject:
Specific School Subject:
Students:

Macintosh File Name: Harmon - Performance Assessment
Release Date: 12-15-1993 C, 11-5-1994 I

Publisher: Misconceptions Trust
Publisher Location: Ithaca, NY

Volume Name: The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics
Publication Year: 1993
Conference Date: August 1-4, 1993
Contact Information (correct as of 12-23-2010):
Web: www.mlrg.org
Email: info@mlrg.org

A Correct Reference Format: Author, Paper Title in The Proceedings of the Third International Seminar on Misconceptions and Educational Strategies in Science and Mathematics, Misconceptions Trust: Ithaca, NY (1993).

 The original publication of this proceedings was supported by a grant from the National Science Foundation, and the transformation of this collection into a modern format was supported by the Novak-Golton Fund, which is administered by the Department of Education at Cornell University.
----- -----

GOING BEYOND THE WRITTEN WORD

WHAT PERFORMANCE ASSESSMENT CAN TELL US ABOUT CONCEPT UNDERSTANDING

MARYELLEN HARMON

Cornell University, Ithaca, New York
August 1-4, 1993

Few people need to be convinced that student prior conceptions o ability to hear and understand the science and mathematics concepts teachers are trying to instill. However, teachers are often unaware of the misconceptions students are harboring. This is particularly true if they are students who score well on tests and to all appearances are knowledgeable, articulate, and able to apply what they know.[1]

This paper will discuss the science portion of some of the early findings of a study of multiple choice and alternative forms of assessment presently being conducted in urban schools by Boston College, Center for the Study of Testing, Evaluation, and Educational Policy. The study, under the title Urban Development Assessment Consortium (UDAC) is targeted for 11 major urban school districts and aims to develop assessments that will meet two requirements: to provide diagnostic information to classroom teachers and local school administrators for the improvement of instruction, and at the same time to provide a monitoring system to district evaluators and policy boards. Teachers need information on what students are thinking and how they

[1]The Harvard Private Universe Project provides illustrations and there are many other examples in the literature.

are personally making sense of what is taught, information that is not provided merely by the student's ability to select a correct answer on a multiple-choice test. School districts need information on the strengths and weaknesses of schools in effecting student learning so as to know where to allocate resources, what kinds of teacher training to provide, where innovations are more effective than prior programs and should be supported, and where there may be waste of resources in ineffectual programs. Districts and institutions of higher education presently make high stakes decisions based on the results of commercially available multiple-choice tests which in fact do not sample higher order thinking or in-depth concept understanding, and tell little or nothing about students' abilities to solve the kinds of poorly structured non-routine problems they will meet in future life. Although the initial findings discussed here are very early and partial, and therefore any conclusions must be at best tentative, I believe some description of both process and results is relevant to the intent of this seminar for a number of reasons.

One reason is the role assessment plays in educational reform. Although assessment-driven instruction is theoretically unacceptable, backwards, "the tail wagging the dog," that tests do have this power is an established fact. We would all agree that change in curriculum and pedagogy should come first, to be followed by appropriate changes in assessment to show us how effective the instruction has been.

However, another study recently completed by Boston College has documented what everybody also already knows. What is on the test is what is taught; teachers, particularly teachers of minority and at-risk students, are profoundly influenced in what they teach and how they teach it by the needs created because students must face high stakes testing, and if the way it is tested rewards only or primarily recall, then teachers focus on drill of facts, believing this will help students achieve on that kind of test. So my first reason for introducing the issue of testing into a conference on misconceptions is my concern for reform in education, a reform inhibited in the USA by continued use of present-day commercially available multiple-choice tests. If we want change in education to occur, the kinds of tests that measure instructional effectiveness must be designed to provide sound information on the depth of concept understanding, the latent misconceptions currently held, and students' ability to use higher order thinking skills.

The second reason flows from this. To create, administer, score, and analyze alternative assessments requires costs in time and money that are orders of magnitude beyond those required by machine-scannable nationally-normed tests in the United States. In addition, issues of standardization of administrative and rating procedures, and of validity, reliability, and inter-rater agreement are still unresolved. And so even those school districts that are trying to improve education by using hands-on and other innovative

approaches in science find it easy to continue using the commercially available machine-scannable tests. The security of numbers, percentile scores that can be used to compare schools (and even sometimes teachers) is comforting, never mind that 95% of the math and 80% of the science items on such tests sample only recall of factual information and application of mechanical procedures in routine types of problems and provide almost no information on students' thinking.[2]

The challenge facing us in the UDAC project therefore has been to provide school districts with tests that in fact do yield information about student's thinking, their misconceptions, and their ability to solve problems, while at the same time meeting the equity requirements of freedom from gender and ethnic bias, reliability, inter-rater agreement, and reasonable cost in time and other resources. And, in addition, to do this without losing focus on the primary purpose of assessment: to inform the classroom teacher about student thinking so that he/she can shape instruction to effect concept development.

I shall not go into details of the entire study - analyses are not all completed yet but a series of papers will be presented at AERA next spring after all the data have

[2]For complete analysis of the six most commonly used standardized test batteries see Harmon, M. and Mungal, C. (1992). *Standardized and text-embedded science and mathematics tests,* part of a recent study of standardized and publisher supplied testing released by Boston College, October 15, 1992 and available from the Center for the Study of Testing, Evaluation, and Educational Policy at Boston College.

been analyzed and supplemented by another round of testing in a different city. Nor will I report here on the other three subject areas tested, but I shall describe the types of tests administered in science, the results obtained through alternative assessments, and focus especially on what performance assessments have to say about the present conceptual understanding of a particular set of students and their ability to generate new understandings from the assessment process itself.

The design of the UDAC project calls for assessing students in 7-10 schools in each of the districts that have chosen to participate in the study. The tests will be given over to each district as well as technical assistance to inservice teachers and administrators in corresponding curriculum and pedagogy changes since the whole project is within the framework of school restructuring. To date, a first round of tests in Reading, Writing, Science, and Mathematics have been created, piloted in Boston Public Schools, revised, and re-administered in the spring of 1993. The entire test battery consists of a cluster of multiple choice items from the most recently released NAEP items (National Assessment of Educational Progress) (10 to 25 items depending on which intact cluster is chosen each round from NAEP, and including any open-ended items that appear in that cluster), short and long open-ended essay items created by UDAC, and performance items also created by UDAC. Testing covered one 45 minute period per day over three days. The

testing days were not necessarily consecutive and in some cases were a fortnight apart but the order in which the different modes were administered was maintained in all schools. Administration and scoring were done by project staff in the fall administration. In the spring, teachers administered NAEP and open-ended items, and teams consisting of a teacher, a Boston College UDAC staff person, and a community person administered the performance assessments, completed observation records independently, then arrived at consensus ratings. They scored all alternative assessments, both open-ended and the written questions of the performance assessments, again using consensus to resolve discrepant ratings. The computer analyses (not yet complete for all grades) will show us what level of interrater agreement was present before consensus.

NAEP items in all four subjects comprised one period. UDAC written items, two to four short essay answers in each of the four subjects plus two or three longer items in one of the four subjects, were answered by each student in the second testing period. Matrix sampling was used so that one-fourth of the population received long items in each of the subject areas but no student had to answer long items in more than one area. This same group then participated in the performance assessment in that subject on a third day. Performance assessments in science used a circus model, with students functioning in groups of three or four to complete at least two tasks of the three provided, and respond as a

group to related written questions. Examiners were free to ask questions such as "Can you explain to me what you are doing?" or "Why are you doing that?" if clarification was needed during the actual performance or if they noted that students who had worked industriously had not, in fact, answered any of the reflection questions. These interventions were few, and all interventions, even paraphrasing a word, were noted on the examiners' sheets.

Examiners were provided with checklists which identified the concepts, various higher order thinking skills, communication skills, and cooperative group skills demonstrated by students as they worked. Tests were provided in both Spanish and English. Where bilingual or monolingual Spanish students were present at least one observer understood that language. We felt this necessary, as the most useful information came from listening to students dialogue as they worked. A large number of bilingual students discussed the task and arrived at consensus in Spanish but chose to write their answers in English.

WHAT DO THE RESULTS SHOW US

Results for the UDAC open-ended items at the primary level were disappointing but not surprising. One question required students to read a straight-line graph which related the length of the shadow of a stick to the time of day, then in a following question to explain why the shadow got shorter and shorter as the time moved from 8:00 AM to noon. All

items were scored on a scale from 0-4, where 0 meant no attempt, 1 a naive response or misconception, and 4 a complete, well-written, and excellent response. Primary students were quite able to read the graph and select the correct length for the shadow at 11 AM (mean score 3.43 on a scale of 0-4, median and mode both 4; frequency of correct responses 31; N=67) but most were totally unable to explain why the shadow got shorter as the sun rose (mean score for explanation: 0.96, median and mode both 1; 2 students provided adequate explanations.) A number of students could not distinguish the symbolic from the real-world, and said the shadow would be shorter because the "line was going down" or "someone must have been pushing the stick further and further down into the ground" or "the line was going down bcause the sun was going down." Others simply said "the earth was turning," or "the sun was moving," but could make no connection between that statement and the length of the shadow. At the middle school level 23 students out of 25 answered the multiple-choice question correctly but only 2 of the 25 could explain the phenomenon.

A second question described a cook with two pots of potatoes boiling. He turned the heat lower on one, but still kept it boiling. He thought he should leave the other on high because he needed those potatoes first. About one-third of the primary students said his thinking was wrong but only 10 out of 67 explained that boiling water does not get hotter

when one applies more heat: "boiling is boiling and the temperature will be the same in both pots."

A multiple-choice problem-solving question asked students to select the correct design to determine the effect of fertilizer on plant growth. Although 10 primary students selected the correct experimental design (~15%) only one could explain why. However,8 were able to explain why they had rejected one or more of the other alternatives. Half of these eight had selected an incorrect design, i.e. one that manipulated the wrong variable, a fact which seems to indicate that while they were able to reject some obvious errors, they still had difficulty identifying just what was to be tested.

Another question used at 4th, 8th, and 10th grade levels provided a picture of a U-tube open to the atmosphere in both arms, with water being poured in on one side, and asked the students to draw on the picture just where the water level would be if there was not enough water to fill the tube to overflowing. The most common response was that the water would stack up on one side: "if you want water on both sides you have to pour it in on the other side too." In many cases both primary and middle grade students focussed on the shape of the tube, drew uneven levels in the two arms, and said "it will look that way because that is the way the tube is shaped." At the primary level 8 students drew the water level correctly on the sketch provided; of these 3 were able to

explain why the water levels would be the same in both arms. At the middle school level, while 9 students out of 25 (~36%)[3] produced an acceptable drawing, no student was able to explain it adequately. Most of the students in their explanations focussed on the shape of the U-tube, and said the water level would look that way because "that's the way the tube is bent."

The particular questions cited above were selected from NAEP released multiple-choice items,to which we then added one or more questions requesting explanation of the choice. Other short questions were created by staff or adapted from other sources and included questions on plant growth and nutrition, erosion, food chains, protective coloration and adaptation, sound, electricity, and the evaporation-condensation cycle. Full reports will be available by Fall, 1993, but the results even at this early stage of analysis raise questions about the validity of basing decisions about students' conceptual understanding on their responses to individual multiple-choice items, even those as well-constructed as NAEP.

Longer questions were built around scenarios requiring the students to problem solve, design experiments, explain which variables "should be kept the same" and why, identify

[3] The small sample is due to the fact that other data are still being processed. Summaries will be reported later.

possible sources of error, draw conclusions and apply them to new instances.

One such question involved comparing two brands of batteries to verify claims made on TV that one brand "outlasts all others." Although primary students had difficulty identifying what exactly their experiments were supposed to test,(mean 1.22, mode 0, N=50) about 20% achieved at least an adequate rating. They fared less well in describing a "plan for an experiment" to find out the answer, (mean 0.90, mode 0) or what conclusions they could draw. About 17% of the middle school students were able to identify the question, and produce an acceptable design, but only 1 student understood the concept of controlling certain variables and none could identify possible sources of error.

How well students performed on problem-solving questions seemed to be context dependent. For example, middle school students were able to perform much better on a mechanical question than they had on the battery question. For the mechanics questions they needed to compare the distances cars would travel beyond the ramp when they were started simultaneously down ramps with different angles of incline. In this case, 52% could state the problem and a hypothesis in scientific terms, 32% identified possible sources of error in design or measurements, and 40% drew valid, well-expressed conclusions.

We recognized that there were several possible curricular, pedagogical or even developmental explanations, for some of the results. At the primary level it might be tempting to conclude that science is not being taught in most of these schools to any degree,or is not taught in ways that develop concept understanding. The latter seems closer to the correct hypothesis. However, we also recognized that we were asking students to operate at an abstract level, and use their own visual imagination to transform the prose scenario into something they could handle, and we were also asking them to provide written prose explanations for the concepts underlying certain multiple choice questions. Although there was enough "creative scoring" to be reasonably sure we understood what the student was trying to say, the very fact of writing out prose answers can be daunting for students who are accustomed to worksheets, do not customarily articulate whole answers in class, and are not immersed in writing across the curriculum.

And so we turned to performance assessments. We used the same tasks at both 4th and 8th grade levels but the requirements for design and recording of data were more elaborate for the 8th grade, there was less scaffolding for the responses, and there were more, and deeper reflection questions to be answered in writing. The tasks included a leaf sorting task, and two problem solving tasks: one in which students were given equipment and asked to determine the effect of temperature on the rate of solution of a

tablet, the other in which they were given glasses, a pitcher of water, a geoboard, strings and rubber bands, and a wooden stick and asked to construct a musical instrument, and play "Three Blind Mice" or another melody of their choice for one of the examiners. After they had accomplished that part of the task, they were also asked to explain orally how they obtained tones of higher and lower pitch, what problems they had encountered tuning their instrument, and how they had solved them. The written questions asked them to describe in a paragraph all they had learned about sound during this task, and what new questions they would like to explore.

For the performance assessments both process and product were evaluated by three observers (or sometimes more, if there were visitor observers) who circled the room identifying first how groups approached a problem and what their conversations revealed of the prior knowledge they brought to it. Using a prepared checklist, observers then noted evidence of relevant concepts and misconceptions, appropriate scientific thinking skills such as problem solving, hypothesizing and predicting, planning or designing and conducting an experiment, measuring, recording and analyzing data, developing hypotheses, interpretations, and generalizations, and evaluating at some level either the effectiveness of their design, or its results and possible sources of error. In addition, their oral communication and cooperative group skills were observed. Observation records consisted of simple checks, with blanks if an observer did

not see evidence of a concept or skill, and notes if there was clear evidence of misconceptions or evidence that an expected skill was not present. Groups recorded their data as they worked, then discussed and answered the reflection questions as a group after they had cleaned up their station to be used by another group. Immediately after the testing period, observers discussed the evidence by which each had determined his/her ratings and agreed on a consensus rating for each concept or process category for each group. The students' written responses were scored later, using the same 0-4 scale used for essay questions and the same consensus process. Then results of the process and product evaluations were combined to give an overall profile of the level of the group's functioning.

WHAT WE LEARNED FROM THE RESULTS

We learned that in these particular schools the students brought with them almost no prior knowledge of the concepts in the domains sampled: sound and music, matter and energy, heat and temperature, sorting, laboratory procedure and experimentation (one middle school and one bilingual elementary were the exception in demonstrating their experience in planning and conducting experiments and controlling variables). But we also found that while few planned their experiments in advance, and all seemed to be at the "messing about" stage of discovery, many groups moved rapidly by experimenting to discovery of the essential

concepts embedded in the problem posed.(42% of the groups on the chemistry task, 53% sound, 47% on characteristics of leaves, 66% on classification). We also found that their performance and use of higher order thinking skills were context dependent.as well as dependent on previous experience working with equipment. For example, more than half the groups worked out their musical instrument, and succeeded in playing a tune of their choice, recognized the correct relationship between length of the air column or string and pitch and were able to manipulate pitch at will. But none of the other concepts in the area of sound were "discovered" in spite of time and opportunity, and misconceptions remained about pitch, volume, echoes, and effect of method of production on pitch. Students were having too much fun playing with their instruments, and were not at the stage even to be interested in other questions or generalizations.

In chemistry, on the other hand, every discovery raised new questions, such as "what difference would it make if we stirred one glass and not the other?" or "does the size of the tablet make a difference?" or "why is there a scum on top when we dissolve it in cold water, but not in hot?" and a few of the groups repeated the experiments trying to find out. In chemistry there was much more evidence of experimental design,(9 of 19 groups) very careful observing and recording (7 of 19 groups) and good measuring skills (10 of 19 groups). We attributed this care to their fascination with real equipment and the feeling that they were "acting like

scientists," feelings not generated by sorting leaves or even playing music with glasses. Also in the chemistry experiment students were willing to attempt the generalizations and applications to new tasks, although at this stage 4th grade students were usually in error and drew unjustifiable conclusions. (Success rate on interpreting data, hypothesizing causes, and applying to new cases: 4 of 19 groups.) Were this an embedded assessment within instructional time, these erroneous generalizations could have been pushed by questions like "how could you check that out?"

In general, process observations were supported by the written records, though each mode was suited to evaluate some areas not accessible to the other. The leaf sort was the simplest task. Fourteen of the 17 groups showed good observing and sorting skills although only half the groups could identify explicitly the categories by which they were sorting. Differences were easier for students to identify than likenesses among the leaves but the written record showed 15 of the 17 groups had a good conceptual grasp of classification,critical characteristics, likenesses and differences. About half the groups demonstrated good group skills.

Higher order thinking skills were required by the two problem solving tasks, by far the most exciting for the

students.[4] However, it was clear that students did not have the habit of reading directions through, began messing about before doing any planning or experimental design, even though prompted by the direction sheet to write their plan first, and that the exercise was clearly a discovery task developing concepts rather than an assessment of concepts already developed. In spite of that, 17 of the 19 groups demonstrated excellent skills in observing and recording and more than half the groups developed an acceptable written plan (sometimes only after the first attempt at the experiment proved a failure) 10 of the 19 groups measured the time, volume, and temperature with accuracy and precision. The written record showed that about half the groups could articulate the relationship between rising temperature of the solvent and rate of solution (at the middle grade level some groups could express it in mathematical form but only one or two used ratios and none chose to graph it) but none of the primary students could develop a feasible hypothesis for the phenomenon observed. There were weaknesses in all groups across all schools in the areas of analyzing data and expressing concepts and hypotheses in writing. Students did make the connections, however, between what they had just discovered about the effect of temperature on rate of solution and real-world applications like taking aspirin or alka seltzer. "Does that mean my daddy should take alka

[4]In every school students wanted to keep on, repeat the experiment, or try to discover answers to new questions when we came to the end of the testing period.

seltzer with hot water?" They also made many side discoveries like the effect of particle size and of stirring on solution rates but were not ready to venture a guess as to reasons for what they observed. In general, the performance part of the assessment was successful in revealing students' problem solving ability, and good group skills, but in also revealing their inability to articulate in writing findings, hypotheses, relationships, and connections. It also revealed lack of experience with equipment, lack of instruction in planning and design, and of discussions of "what is going on here? why do you think this is happening?" as well as the fact of careful instruction in note-taking, observing, and recording. The means for most of the written summary and reflection questions were between 1.11 and 2.67, with means dropping to 0.071 (on a scale of 0-4) on questions involving hypotheses, evaluation and sources of error. Discussion, collaboration, and group self-organizing were quite good for 12 of the 17 groups and consensus decision making was practiced effectively by about half the groups and sporadically by a few others.

The questions on the chemistry task were carefully scaffolded to help the students stay on task. On the sound/music task all communicaton was oral with the exception of a very open paragraph in which the students were to describe what they had learned. For 4th grade students this was too open, and they were content with single sentences explaining the relation between pitch and length of

the column of air or of water. Nine of 17 groups discovered a relationship and could articulate it but many did not realize that it was the shortening column of air (as more water was poured in) that raised the pitch although they did recognize that where they struck the glass and its composition affected the timbre (which some called "echo"). Thus their understanding of pitch was still entangled with misconceptions and confusions between pitch and volume, between methods of production and pitch, and some private discoveries of their own about echoes and timbre that raised interesting new questions in discussion, questions they had not the skill to capture in writing nor the time to disentangle during the testing period. Once more the written record was disappointing, but with this task it was abundantly clear during the process evaluation that, while they could make careful observations and record them, at least half the groups could not put the pieces together and even the relationship between pitch and the length of the vibrating medium was on shaky ground, a finding not to be wondered at, as sound experiences had only just begun to be developed during this testing session.

Perhaps the most important category of findings was the data accrued by the researchers on feasibility and replicability of the performance testing process and the validity of its findings. Administration was done by the classroom teacher with a BC staff person assisting and a community person (business, parent, or other interested

educators) participating both in observing and scoring, and in consensus discussions on ratings. Using the consensus model for all ratings where there were any discrepancies between independent ratings proved to be quick and easy and was, in fact, both a means of teacher training in the nuances and subtleties of the concepts in question, and a political tool which provided a wealth of insights for members of the community. It was also costly in time. A team of 3 required about 6 hours to score 90 papers with open-ended answers in all subject areas, with at least 2 raters per paper. An experienced subject area specialist could score about 30 written records of the performance assessments in an hour, but teams of inexperienced raters took much longer. Community and Education Department interest was high, and sometimes there as many as 6 observers in the room along with media cameras and reporters. However, it was the designated "community member" of the administration team who participated in the consensus.

Our mandate was to develop alternative assessments which could be used to inform instruction and we have reason to believe that for those teachers who participated in observing and scoring there has already been a significant impact. We have succeeded in standardizing procedures, rubrics, and scoring processes, and have confidence in scores. An average of two or three ratings across all subjects for each child had to be resolved by consensus (independent ratings were also recorded for later statistical analysis). We still have

to find ways to make the team scoring process more efficient and to provide inservice for teachers in what is, for some, a whole new way of teaching but for all who participated in the spring round of testing, is now recognized as the way they must go if only the training and support can continue.