How can paper and pencil tests and performance assessments be used to effectively measure different knowledge types and skills in science?

María Figueroa
Stanford University, California, USA.
Universidad de los Andes, Bogotá, Colombia
mafiguer@stanford.edu

Rafael Montenegro
Universidad de los Andes, Bogotá, Colombia
ra.montenegro38@uniandes.edu.co

Abstract

Science teaching has changed from a text-based to an activity based (hands-on) approximation. This has generated an increase in research and development of different types of assessments. Assessment of science learning requires instruments and techniques that are aligned with the methodology used, and the depth and complexity of what students understand and can do in this discipline. Therefore, assessments need to include a large range of types of tests, formats and instruments. This study compares the results of 5th grade students using three different types of instruments after studying a unit on electric circuits using a hands-on approach. Student learning was assessed with a multiple-choice test, a hands-on performance assessment, and a computer-based performance assessment. Given the changes in the way science is taught, the comparison of these three instruments provides useful information about the viability, feasibility and practicality of using different assessments to measure students' knowledge and practical skills in science. An additional analysis is done comparing the interchangeability of the two types of performance assessments, with implications in classroom and large-scale implementation of these assessments.

Key words: Science assessment, performance assessment, computer simulation, multiple-choice tests, electric circuits.

Introduction and Background

Science teaching has changed from a text-based to an activity based (hands-on) approximation. Since the methodologies used to teach science have changed, the way we assess what students are learning also needs to be modified. The different approaches in methodology and evaluation have generated an increase in research and development of different types of assessments in science teaching and learning (Ayala, Shavelson, Lin and Shultz, 2002; Haertl, 1999; Shavelson, Baxter and Pine, 1992). Some of these assessments include: multiple-choice tests, open ended questions, and performance assessments to name a few. Understanding if these tests can be correlated can provide additional information in science assessment research.

This study measures students' knowledge using two different types of assessments: a traditional multiple choice that includes open-ended questions and a performance assessment. Performance assessment tests can be presented in many forms including conducting experiments; performing mathematical calculations, extensive essay writing, and performing computer simulations (Elliott, 1995).

The benefits of performance testing are well documented: Ayala (2002), argues that science performance assessments can measure different types of knowledge including declarative, procedural and schematic knowledge. Ruiz-Primo and Shavelson (1996) say that performance tests produce high level of reasoning processes, since these tests are closely related to what students and scientists do in the lab. Haertel (1999) argues that these tests not only show how students learn, but also, students show higher engagement in learning. Elliot (1995) argues that performance assessments provide evidence of what students know and are able to do. Quellmatz (1999), says that the evidence gathered during the performance provides insights to students' thinking, and at the same time introduces students to authentic real-world problems, which allows them to show how they can apply academic knowledge to practical situations. For the reasons mentioned above performance assessment tests are alternatives to measure students' knowledge in electrical circuits.

The comparison of the results in two types of performance assessments and the multiple choice test, can provide information as to the uses of these types of instruments and the feasibility of applying and interchanging them in an effort to assess students' knowledge in science. Additionally, this study explores how students who perform high and low in the multiple-choice test do in the performance assessments.

Methods

This section describes the participants of the study, the types of assessment, and the procedures.

Participants and School Context

The participants of the study were fifth-graders of four Colombian public schools who finished the school year in November 2009. These schools are located in very low SES neighborhoods in Bogotá, Colombia. Each school has approximately 1200 students from K – 11, with an average of 40 students in each classroom. The communities composing the student body these schools have several problems including undernourishment, interfamily violence, sexual abuse, use of drugs, and low motivation for studies. The ages of students who participated in the study ranged from 10 to 12 years of age. Two classes were selected in each school. Each of the classes has the same teacher, since there is only one science teacher in fifth grade in each school.

Description of Instruments Used

Instrument 1-Multiple Choice and Constructed Response Items

The development of the paper and pencil assessment was organized in several steps: 1) item development, 2) piloting the items and doing think-alouds, and 3) item revision.

Item development.

The items used in this assessment come from different sources. Some of the items were developed during a six-day workshop with teachers, scientists and science educators. Participants of the workshop were guided to develop two types of items: some that were close to the electric circuits module (proximal) and others that matched national education standards of Colombia (distal).

Participants were also directed to develop items that tapped into three types of knowledge: (1) declarative knowledge (factual, conceptual knowledge) or "knowing that" (e.g. what materials conduct electricity?); (2) procedural knowledge (step by step investigations) or "knowing how" (e.g. how to interpret a graph); and (3) schematic knowledge (knowledge used to reason about) or "knowing why" (e.g. explain why one bulb in a circuit may turn off all other bulbs in the same circuit?)

Other distal items were provided by ICFES, the Colombian institute that carries out all standardized testing in the country.

Pilot testing and think alouds.

The items were tested in different pilot studies including one in Panama that provided information about the difficulty of the items; another in a city close to Bogotá, a third in the city of Cali, and the fourth one in Bogotá. Items from each pilot were analyzed using the statistical software Iteman (Item Analysis) in order to identify items that were not working well as well as the reliability of the scores. Iteman also provided information regarding the difficulty level of the items as well as their discrimination index.

Item revision.

All the information gathered in the pilot studies and the think-alouds allowed for a final selection and improvement of items and its classification. The final instrument was a paper and pencil test composed of 33 questions, 31 of which were multiple choice and 2 constructed response questions. In order to counterbalance the effect of the order of the questions, three booklets were produced. The reliability of the results of this assessment was 0.72.

Test administration and data collection.

Four Master of Education students were trained in the implementation of this test. All test givers participated in a four-hour training session where they received and read a test implementation manual and were provided with logistical and technical information about the data collection. The training and manual aimed at standardizing test administration and unifying instructions and protocols.

All tests were given under standardized testing conditions and classroom setup. In each classroom, two test givers were present to administer the test. They followed the instructions found in the training manual and reported any irregularities that occurred during the implementation in a provided form. Data collection of the four schools occurred in November, 2009.

Based on student performance in the paper and pencil test, the top ten and the bottom ten students were selected from each school according to their scores. The selection of top and bottom students evaluated how students with extreme differences in level of achievement in the paper and pencil test, would perform in the performance assessments.

Instrument 2-Performance Assessment: Hands-on

Description of instrument

The hands-on electric circuits performance assessment was taken from the Stanford Education Assessment Laboratory website http://www.stanford.edu/dept/SUSE/SEAL/Assessments) and translated into Spanish. A performance assessment in general, includes a challenge, a response, and a scoring system. In this case, the challenge required students to identify the components hidden inside of six black mystery boxes when given materials such as a wire, a bulb, and a battery. There was no specification of the steps to be taken. Responses were registered in a notebook and collected after the assessment ended. The time each student spent responding to the assessment was recorded in the notebook.

Pilot testing and think alouds.

In order to revise the clarity of the language, which could have been affected during the translation, the hands-on performance assessment was piloted with sixth grade students. We asked the students to think-aloud while they were completing the assessment.

Test administration and data collection.

Students were selected after they responded the paper and pencil test. As mentioned above, the ten top and ten bottom performers in that test were selected to participate in this part of the study. All students, except two, responded to both the hands-on and the computer-simulation performance assessments. Students were randomly assigned to start with either the hands-on or the computer-simulation assessment.

Data collection occurred in two separate days for each school. During the first day, ten randomly assigned students answered the hands-on assessment in the library, while the other ten answered the computer-simulation in the computer lab. The second data collection occurred 12 to 15 days after the first one. After each data collection, approximately three students were asked to provide retrospective information regarding the assessment and their performance (e.g. What was the most difficult part of this assessment and why?). These interviews were tape recorded.

Instrument 3-Performance Assessment: Computer Simulation

Description of instrument.

The computer-simulation performance assessment was developed based on the hands-on performance assessment. The development took approximately four months, including revisions and pilot tests. The computer programmers worked under our direct supervision to replicate as closely as possible the hands-on electric circuits performance assessment. However, the labels of the boxes in the computer simulation were different than the labels of the boxes in the hands-on performance assessment (e.g. Box D in the computer simulation was equivalent to Box A in the hands-on assessment). The assessment is in Spanish and has audio to help students follow the instructions. A screenshot of the computer-simulation is presented in Figure 1.

Pilot testing.

The assessment was piloted with sixth graders at the ICFES office. After the pilots, students were asked to provide feedback in an interview, which was recorded. The information was then used to fine-tune the assessment.

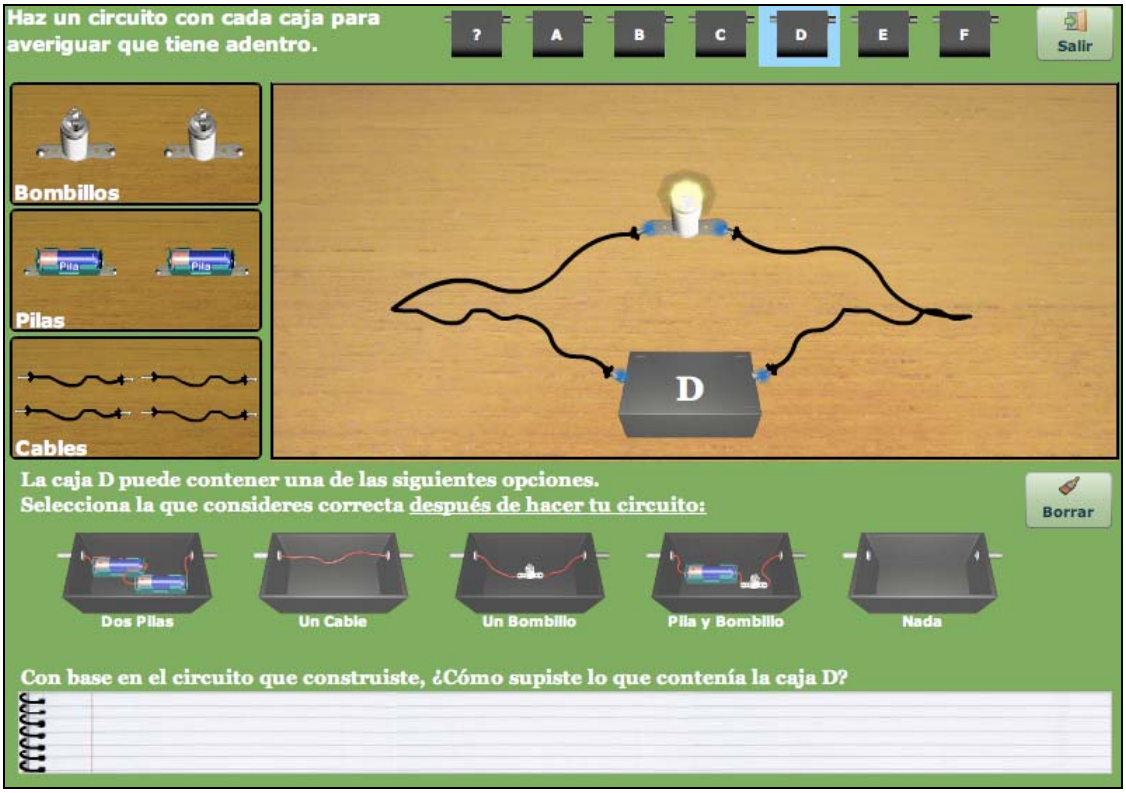

Figure 1. *Screenshot of the Computer-Simulation Performance Assessment*

Data analysis

We used all instruments mentioned above and compared the performance of students of different abilities (high and low) in the two types of tests. In order to compare the scores of these two studies we will see the relationships between the assessments. Descriptive statistics provided information about correlations between paper and pencil and performance assessments.

Results

We compared the total scores of the three instruments: multiple choice test, hands-on and computer simulation performance assessments. Analysis of the overall scores revealed a statistically significant association between the multiple choice test and the computer simulation, (r=.535, p=.000) as well as the multiple choice test and the hands-on performance assessment (r=.415, p=.000). We also found a statistically significant association between both performance assessments (r=.398, p=.001).

In this study, hands-on and computer-simulated versions of the Electric Mysteries assessment were administered to the same students in two different occasions (time 1, time 2). We found a positive correlation between multiple choice and both methods of performance assessments in time 1 (r=.468, p=.000), and in time 2 (r=.495, p=.000). In both times correlations are statistically significant.

As mentioned above, students were selected based on their performance in the multiple-choice test. There is an association between top students and their performance in the performance assessments ($r_{computer}$=.159, $p_{computer}$=.369) ($r_{hands\text{-}on}$=.066, $p_{hands\text{-}on}$=.713). However, when comparing the results between low performers in the multiple choice test, there is a negative association ($r_{computer}$=.-.111, $p_{computer}$=.517) ($r_{hands\text{-}on}$=-.373*, $p_{hands\text{-}on}$=.025). There is a statistically significant negative correlation between the performance in multiple choice test and the hands-on assessment.

Discussion and Conclusions

It is possible to design and administer instruments in science assessment that better reflect what happens in the classroom. During he implementation of the assessments, students mentioned that they were much more engaged with the performance assessments than with the multiple-choice test. Performance assessments were used in this occasion as a form of summative assessment, but they could be used successfully in a formative manner in science classrooms.

There was a significant correlation among the three types of assessments. This means that the results of the assessments are good indicators of what students know and what students are able to do, and that a variety of instruments provide more information about the students and the types of knowledge they have in science.

The use of performance assessments provided opportunities for students who did not have the skills to respond well in the multiple-choice test, to demonstrate their abilities, and knowledge. High scores in the performance assessments procedural and declarative knowledge of electric circuits. These different types of assessment allow teachers to observe different dimensions of their students and obtain a wider picture of their knowledge and skills. The results of this study provide evidence that the use of different instruments is essential for low performing students, but not necessarily for high performing students.

When comparing the multiple-choice test with the performance assessments and taking into account the time of implementation (multiple-choice test vs. hands-on first + computer second or multiple-choice test vs. hands-on second + computer first), we obtained high correlations. With this data, we have initial evidence that both the computer simulated and the hands-on performance assessments can be interchangeable.

The information gathered during this study provides preliminary data on interchangeability of science assessments. Further analysis should be done in order to compare students' performance according to knowledge types and to the proximity (close and distal) of the items.

References

Ayala, C. A., Shavelson, R. J., Yin, Y., & Schultz, S. (2002). Reasoning Dimensions Underlying Science Achievement: The case of Performance Assessment. Educational Assessment, 8(2), 101-122.

Brualdy, Amy (1988). Implementing Performance Assessment in the Classroom. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Elliott, Stephen (1995). Creating Meaningful Performance Assessments. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland

Haertel, Edward (1999).Performance Assessment and Education Reform. PHI DELTA KAPPAN, mayo 1999, 662-666.

Quellmalz, Edys, Patricia Schank & Thomas Hinojosa & Christine Padilla (1999). Performance assessment links in science. Practical Assessment, Research & Evaluation, 6(10).

Ruiz-Primo, Maria; Shavelson, Richard (1996). Rethoric and Reality in Science Performance Assessment: And Update. Journal of Research in Science Teaching, 33(10), 1045-1063

Shavelson, Richard; Baxter, Gail; and Jerry Pine. (1992). Performance Assessment, Political Rhetoric and Measurement Reality. Research News and Comments, mayo 1992,22-27.

U.S. Congress, Office of Technology Assessment. (1992, February). Testing in American schools: Asking the right questions. (OTA-SET-519). Washington, DC: U.S. Government Printing Office.