LOOK ACADEMIC PUBLISHERS
OPEN ACCESS

INTERNATIONAL JOURNAL OF ENVIRONMENTAL & SCIENCE EDUCATION
2017, VOL. 12, NO. 2, 177-193

Assessing the Validity of Multiple-choice Questions in Measuring Fourth Graders' Ability to Interpret Graphs about Motion and Temperature

Mehmet Dulger and Hasan Deniz

Teaching & Learning Department, University of Nevada Las Vegas, Nevada, USA.

ABSTRACT

The purpose of this paper is to assess the validity of multiple-choice questions in measuring fourth grade students' ability to interpret graphs related to physical science topics such as motion and temperature. We administered a test including 6 multiple-choice questions to 28 fourth grade students. Students were asked to explain their thinking in writing for each question. In addition, we interviewed all 28 students and asked them to justify their answer for each question by thinking out loud. We found that a significant number of students were not able to provide appropriate explanations for their correct answers. Interestingly, however, a significant number of students were able to provide appropriate explanations even though they initially selected an incorrect response. As a result of this study, we suggest caution in using multiple-choice questions as a single data source to assign grades or to make other important decisions about student achievement.

KEYWORDS
Assessment, multiple-choice questions, elementary students, validity

ARTICLE HISTORY
Received 27 October 2016
Revised 10 December 2016
Accepted 28 December 2016

Introduction

A wide range of adoption of the Next Generation Science Standards (NGSS Lead States, 2013) presents a challenge for science educators to develop curricula and assessments that are aligned with the NGSS. The new standards conceptualized learning around three dimensions: the science and engineering practices, the crosscutting concepts, and the disciplinary core ideas of life sciences, physical sciences, earth and space sciences, and engineering and technology. The NGSS focuses on a limited number of core ideas in science by adopting the notion of learning as developmental progression. In NGSS, same concepts are revisited with increasing levels of sophistication at K-2, 3-5, 6-8, and 9-12 grades. The NGSS provides an overarching framework but it does not specify curriculum and assessment methods. Therefore, the new standards must be translated to curriculum, instruction, and relevant assessments. The Committee on Developing Assessments of Science Proficiency in K-12 was charged with developing assessments aligned with the NGSS:

CORRESPONDENCE Mehmet Dulger dulgerm@unlv.nevada.edu

© 2016 M. Dulger & H. Deniz
Open Access terms of the Creative Commons Attribution 4.0 International License apply. The license permits unrestricted use, distribution, and reproduction in any medium, on the condition that users give exact credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if they made any changes. (http://creativecommons.org/licenses/by/4.0/)

> The committee will prepare a report that includes a conceptual framework for science assessment in K-12 and will make recommendations to state and national policy makers, research organizations, assessment developers, and study sponsors about the steps needed to develop valid, reliable, and fair assessments for the framework's vision of science education. The committee's report will discuss the feasibility and cost of its recommendations. (p. 2)

The history of standardized written examinations in the United States goes back to the 1850s and through the development of mental test in the 1900s, it gained a pivotal role for "classifying and placing students by ability" (United States Congress Office of Technology Assessment, 1992, p. 110). Standardized testing became an integral part of the education system as public education became more accessible to the masses in the second half of the nineteenth century (Kaestle, 2013). DeBoer (1991) highlights that the effect of the development of standardized tests in the science areas was to focus attention on the more easily measurable outcomes of education in the 1950s. Since content mastery was the easiest thing to measure, this dominated science education in the progressive era (DeBoer, 1991).

Testing has been used widely until today, and the impact of accountability was recognized starting from the late 1970s (Longo, 2010) to date, especially since the No Child Left Behind (NCLB) Act passing into law in 2002. NCLB mandates each school system to write a set of Adequate Yearly Progress (AYP) objectives in mathematics, language arts, and science. With the No Child Left Behind (NCLB) act – and the [Barack] Obama administration's "blueprint" which places similar weight on test scores – testing became the main source for the decision of restructuring of schools, rehiring teachers and decisions related to student graduation (Brickhouse, 2006). The emphasis on testing is tremendous from elementary schools to high schools. Based on students' performance of meeting these objectives, states may sanction schools by lessening funds and taking over the schools (Aydeniz & Southerland, 2012).

Secretary of Education, Arne Duncan, at the 2013 American Educational Research Association Meeting stated that state assessments in mathematics and English often fail to capture the full spectrum of what students know and can do and that students, parents, and educators should know there is much more to a sound education than picking the right answer on a multiple-choice question. Supporters of standardized testing claim that testing is a valid and reliable indicator of student learning and that testing is an effective system in ensuring that minimum academic competencies are attained by all students (Greene & Winters, 2003). Despite accepting the possible benefits testing may bring to schools, other researchers claim that basing high-stake decisions such as rehiring teachers, restructuring schools, and decisions related to student graduation based on a single test does not ensure the quality of science education in classrooms (Brickhouse, 2006).

Test Pressure on Teachers, Students and Instruction

Increased test pressure affects teachers and classroom practices in various ways. When test results are 'high stakes' for teachers, they exert significant pressure on teachers, which is then transferred to students, even if the tests are not high stakes for students (Harlen, 2013). Teachers from different states

reported that teachers give more attention and time to tested content area, thus causing to de-emphasize on or neglect untested subject areas (Jones, Jones, Hardin, Chapman, Yarbrough, & Davis, 1999; McMillan, Myran, & Workman, 1999).

Exerting high pressure for student performances on a test can reduce the instruction to test preparation and therefore minimize the skill that teachers bring to classrooms (Hillocks, 2002). Teachers from both high-stake states and low-stake states report that their state testing program prompt them to teach in ways that contradict their own notions of sound educational practice (Abrams, Pedulla, & Madaus, 2003). When test results are at "high stakes" for teachers, teachers focus their teaching on the test content, train students in how to pass tests and feel impelled to adopt teaching styles that do not match what is needed to develop a real understanding (Harlen, 2013). In addition, teachers reported more than 77% decrease in morale and 76% stated that teaching became more stressful after the implementation of the North Carolina state-testing program (Jones et al., 1999). Science teachers are not immune to unintended consequences of testing. Osborne, Simon and Collins (2003) point out that teachers have been observed to adopt a transmission style of teaching even though this is not what they believe to be the best for helping students' understanding and development of skills.

Alexander (2010) found that national tests place pressure on children and teachers, narrows the curriculum, and diminishes the goals of learning and children's self-esteem. Harlen (2013) performed a systematic literature review based on published research within the years of 1980-2008. In this extensive review supported by the Danish Clearinghouse for Educational Research, Harlen (2013) summarized the main themes as follows:

- A narrowed down or distorted curriculum experienced by the students: teachers simplifying demands on students' thinking; facts and mechanical skills are emphasized at the expense of creative and aesthetic activities
- More teaching time being allocated to matters included in tests at the expense of those not included
- Teaching becoming devoted to teaching to the test and rote learning. (p. 28)

According to Harlen's (2013) summary, the influences of tests on students are dramatic:

- The mere announcement of a test starts emotional reactions such as nervousness and fear, especially among girls
- Students prepare for the test by learning by heart and memorizing sentences
- For high achievers motivation increases while low achievers lose their motivation
- A student's test result can influence future motivation and self-efficacy. (p. 28)

Testing: What is it measuring?

Competencies refer to students' mental abilities and skills—their thought processes—and cannot be observed directly. Thus, it necessitates educators to build certain assessment tasks that are used to "measure" students' thought processes by performing a task to infer whether a student possesses a particular competency related to a content (Gilmer, Sherdan, Oosterhof, Rohani & Rouby,

2011). The problem of measuring student progress is a critical issue. This depends on what we believe the progress is and how we measure it. If we believe "progress" is learning substantial amount of scientific knowledge, then our method of assessment might heavily rely on quantitative measures such as standardized testing. On the other hand, if we value students' acquisition of scientific inquiry and critical thinking skills, we might oppose to the administration of standardized tests as a single source of evaluation (Aydeniz & Southerland, 2012).

Articles from the literature either showed or supported that the level of knowledge measured through high stakes testing is mostly lower level of knowledge skills (Aydeniz & Southerland, 2012; Huber & Moore, 2002) and interestingly students' ability to transfer learning from one test to another is questionable—which means we cannot know whether they learned or not (Amrein & Berliner, 2002). According to Bailey (1998), traditional assessments do not measure students' understanding directly and they are one-shot, speed-based, norm-referenced and designed to measure what learners can do at a particular time. They do not inform us about the student progress and any type of difficulties they face during the test.

Mere emphasis on learning through objective facts with some practices results in "superficial" level of student engagement instead of richly structured knowledge and upper-level thinking skills (Huber & Moore, 2002; Kohn, Thompson, Ohanian, & Eisner, 2001; Morgenstern & Renner, 1984). Pellegrino, Chudowsky, and Glaer (2001) underscored that most state assessments rely on multiple-choice formats, test vocabulary and factual knowledge rather than application of concepts or problem-solving skills. In addition, the type of knowledge and skills tests measured and what teachers believe was important for students to learn—such as scientific inquiry skills—were not measured with standardized testing (Shepard & Dougherty, 1991; Aydeniz & Southerland, 2012).

Chudowsky and Pellegrino (2003) pointed out that standardized tests were constructed to rank individuals and measure general proficiencies. Due to its "outdated conceptions of learning," testing experts need to work on improving the nature of achievement testing. In addition, advancements in measurement and technology "have expanded the capability to collect and interpret more complex forms of evidence about student performance." Web-based Inquiry Science Environment (WISE) can be given as an example for this kind of measurement of student understanding as it collects student inputs from a variety of activities to deduct student level of content understanding.

Alternative Methods for Measuring Student Learning

Researchers have suggested some solutions for "teaching to test". Since it is inevitable that students will be tested and since their performances have substantial effects on not only their academic future, but also teachers' and schools' future, one researcher aimed to negotiate "student learning" and "teaching to test." Hammerman (2005) claimed that by using methods and concepts in National Science Education Standards advocating inquiry-based science, we can prepare students not only for the test, but also give them critical thinking skills. Longo (2010) supported the same claim with 5E inquiry methodology while teaching biology content for high school students. In addition, Nowak (2007) showed that the inquiry-oriented Problem-based Learning (PBL)

approach can be used to support content knowledge acquisition if teacher-directed instruction is embedded within PBL approach.

Researchers proposed to measure students' science process skills as an indicator of their science achievement. Since students' multiple-choice test score depends not only on students' understanding of science content, but also their reading ability (Scott-Jones & Clark, 1986; Tolman, Sudweeks, Baird & Tolman, 1991), this indirect measurement of students' science achievement lead researchers to propose hands-on tests to measure students' science process skills. Hands-on tests are less abstract and more concrete when compared to multiple-choice tests. Thus, it appears to be more developmentally appropriate to use hands-on tests to measure science process skills of upper-elementary students (Saturnelli & Repa, 1995). Fifth and sixth grade students explained that they enjoyed doing hands-on tests and also added that they felt more like a "project" or an "activity" rather than taking a test (Hamilton, 1994). Students expressed, "you don't have to think as hard, or study, or memorize lots of stuff like on most tests" (Hamilton, 1994, p.13).

Standards for Educational and Psychological Testing (SEPT) defines validity as "the degree to which all the accumulated evidence supports the intended interpretation of the test scores for the proposed purpose" (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999, p. 11). SEPT proposes five types of evidences related to the validity of a test score. Four of them are based on test content, the test's internal structure, the relations of test scores to other variables, and the consequences of testing. The fifth one is validity evidence based on response process to collect evidence related to the validity of a test (AERA et al., 1999). This type of validity evidence has been overlooked due to its cost and time-consuming nature (Schafer, Wang, & Wang, 2009; Sireci, Han, & Wells, 2008).

Despite researchers' and governmental and professional organizations' attempts to advocate inquiry-based instruction informed by constructivist teaching and learning principles, the continuous implementation of standardized testing has boiled the education system down to a test score (Kersaint, Borman, Lee, & Boydston, 2001; Lee & Luykx, 2005). Attempts to create an accountability system for teachers and schools by using test scores as a major component of the system are greatly likely to shape the education system in the near future. This study, for this reason, investigates to what extent students' answers in a multiple-choice test reflect their understanding of science content.

Methodology

Participants

Twenty eight fourth grade students from a charter school in Las Vegas participated in the study. The charter school started in the 2007–2008 academic year. The school was designated as a high achieving school in the 2007–2008 and 2008–2009 academic years and the school met adequate yearly progress (AYP) requirement in the 2009–2010 academic year. Total number of upper-elementary students were 334 and 126 of them were fourth graders. There were 32 Asian, 8 Black, 62 White, 20 Hispanic, 2 Hawaiian/Pacific Islander, and 1 Indian/Alaskan at the time of data collection. Data were collected during Spring 2011.

Data Collection

Students were asked to take a test that included 6 multiple-choice questions. Students were asked to answer three questions about motion (Keeley & Harrington, 2010) and three questions about temperature prepared by the researchers (see Appendix 1). The three temperature questions were reviewed by two science educators before the actual data collection. The first three questions measured students' ability to interpret representations of distance and time, and the other three questions measured students' ability to interpret graphical representations of temperature and time. Students were given an ample amount of time to answer each question and were asked to provide an explanation for their answers. All 28 students were also interviewed and asked to elaborate on their written answers on the day and the day after the test was administered.

Data Analysis

All 28 students' responses to these 6 multiple-choice questions were marked as correct or incorrect. All of these 6 questions had only one correct answer. Students' verbal explanations for each question during the interviews were transcribed in verbatim. Students' written and verbal explanations for each question were qualitatively analyzed by the two researchers. Researchers agreed that four categories captured the variety of student explanations for each question. These categories are "correct explanation," "incorrect explanation," "partially correct explanation," and "insufficient explanation." This allowed us to create 2 (correct; incorrect) X 4 (correct exp; incorrect exp; partially correct exp; insufficient exp) data table (See Table 1). This provided us with 8 possible categories. For example, a student could mark the correct choice for one question but his/her explanations could fall into any of the four categories mentioned above. Similarly, a student could mark the incorrect choice for a question but his/her explanations could also fall into any of the four categories. Student interview excerpts representing 8 categories are presented in Appendix 2. The two researchers separately classified students' explanations for each question by using four categories. Then the two researchers compared their classifications with each other. Researchers reached 95% agreement in classifying students' explanations. Disagreements were fully resolved by referring back to data and discussion.

Results

The multiple-choice test assumes that students, who make the correct choice in the test, did so as a result of correct student understanding backing that selection and vice versa. However, the results of this study showed that from a total of 94 correct choices of student responses, 13 of them fell into the categories other than correct understanding. In addition, from a total of 73 incorrect choices of student responses, 28 of them fell into the categories other than incorrect understanding. Therefore, the results indicated that student answers in a multiple-choice test do not fully reflect students' true competency that the test is claiming to measure.

Table 1. Number of students in each category across 6 questions
(The following Abbreviations are used in the table: C: Correct, I: Incorrect, NR: No reply, CE: Correct Explanation, IE: Incorrect Explanation, PCE: Partially Correct Explanation, InsE: Insufficient Explanation)

	Question 1		Question 2			Question 3		Question 4		Question 5		Question 6	
	C	I	C	I	NR	C	I	C	I	C	I	C	I
CE	14	3	6	1	1	-	1	22	-	17	2	22	1
IE	-	4	3	6	-	-	24	-	3	1	4	-	4
PCE	-	6	2	5	-	-	3	1	1	-	2	-	1
InsE	1	-	2	2	-	-	-	1	-	2	-	-	-

In question 1, 14 students marked the correct choice and they had correct explanation. However, 3 students had correct explanation even though they selected the incorrect choice. 6 students had partially correct answers even though they selected the incorrect choice for question 1. 4 students had both incorrect choice and incorrect explanation for question 1.

Table 2. Summary for number of students in each category

	Correct Answer	Incorrect Answer
Correct Explanation	81	8
Incorrect Explanation	4	45
Partially Correct Explanation	3	18
Insufficient Explanation	6	2

The nature of multiple-choice questions does not reveal student thought processes during the test. Each multiple-choice question is either scored correct or incorrect. Multiple-choice questions do not account for partially correct understandings. Our results indicated that almost 36% of student responses did not get any credit although they demonstrated correct or partially correct explanations. On the other hand, the multiple-choice test gave almost 14% of student responses a full credit despite the fact that they could not provide any evidence of understanding for the underlying competencies measured in the test.

Using multiple-choice questions may be an effective way to assign students scores or grades, but this study questions the validity of test scores in interpreting students' understanding. Our results indicate that it is possible for students to make correct explanations for a multiple-choice question even though they have marked the incorrect choice. It is also possible that students may mark the correct choice for a question and at the same they can fail to make a correct explanation for their answer. More interestingly, some students can make partially correct explanations even if they select the incorrect choice. Our results show that the complexity of student learning process may not be fully captured through using only multiple-choice questions.

All 28 students did not answer question 3 correctly. Students misunderstood the graph in question 3 as a picture rather than a graph. This phenomenon is well documented in the literature and it is called as "picture-of-the-event" or "iconic graph difficulty" (Berg & Boote, 2015). In our study, 5 out of 6 multiple-choice questions included a graph. We did not observe this phenomenon for the rest of the questions including graph interpretation.

Discussion

Our findings show that multiple-choice questions may not be fully appropriate to assess student learning especially at the elementary grades. Therefore, it would be beneficial to assess student learning using a variety of question formats. A relatively new report titled "Developing Assessments for the Next Generation Science Standards" (National Research Council, 2014) supports our position by suggesting that a variety of question formats including questions requiring students to provide explanations for their answers should be used to assess student understanding.

Our findings indicated that the multiple-choice question format allowed us to efficiently measure fourth grade students' ability to interpret graphs with regard to motion and temperature. However, this obvious scoring efficiency was compromised when students provided incorrect explanations for their correct answers and vice versa. The traditional multiple-choice question format, without any room for written student explanations, is not conducive to tap into student reasoning while assessing knowledge. In short, the traditional multiple-choice test administrator does not have any idea about how the students have arrived at their answers. Our findings showed that some students demonstrated bits of knowledge—which were not enough to correctly answer the question—but these bits of knowledge were not appreciated in a test comprised of traditional multiple-choice questions. Therefore, the test scores do not fully inform us about student progress and what particular difficulties students encounter during the test (Bailey, 1998).

Researchers pointed out that the multiple-choice question format might not require deep understanding of the tested content (Biggs, 1973; Beard & Senior, 1980; Entwistle & Entwistle, 1992). The multiple-choice question format may also be an impediment to assess students' conceptual understanding. We have evidence that some students incorrectly answered certain questions even though they had correct or partially correct explanations for the questions. We also have evidence that some students picked the correct choices but were not able to provide correct explanations for their choices. The multiple-choice question format introduces the element of luck into the assessment process. This particular format randomly rewards some students and punishes others. Multiple-choice questions are commonly used because multiple-choice questions' scoring mechanism offers certain affordances such as efficiency and objectivity. One can efficiently score a test including only multiple-choice questions in a timely manner, but our findings indicate the multiple-choice question format does not afford the desired full objectivity. Our findings make us question the validity of scores obtained from a traditional multiple-choice test. Therefore, people using multiple-choice questions should keep in mind that students' test scores may be a reflection of their luck combined with their conceptual understanding.

Limitations

The study was conducted with 9-year old 28 fourth grade students. We should keep in mind that some students may not be developmentally ready to be tested on subjects requiring abstract thinking. At least some of our participants were likely to be in the concrete operational stage according to Piaget. Therefore, they may not be developmentally ready to engage in abstract graph interpretation questions. According to Piaget, students at the concrete operational stage (ages 7-

11) can use logical operations to solve problems involving concrete objects and events immediately present, but they may not be developmentally ready to engage in abstract graph interpretation questions (Saturneli & Repa, 1995). Therefore, our findings may not be generalizable to a different age group and grade level. Our test mostly included graph interpretation questions about motion and temperature. There should be more research assessing the validity of the multiple-choice question format for topics other than motion and temperature graph interpretation ability. Graph reading and interpretation questions require higher order thinking compared to memorization and recall questions. Therefore, the multiple-choice question format may work better when it is used to measure memorization and recall that do not require higher order thinking.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Mehmet Dulger Teaching & Learning Department, University of Nevada Las Vegas, Nevada, USA.

Hasan Deniz, PhD, is an Associate Professor of Science Education and currently a faculty member of UNLV College of Education.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice, 42*(1), 18-29.

Alexander, R. (2010). Children, their world, their education. *Final report and recommendations of the Cambridge Primary Review* (p.316). London: Routledge.

AERA, APA, & NCME. (1999). *Standards for education and psychological testing*. Washington, D.C.: American Educational Research Association.

Amrein, A., & Berliner, D. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives, 10,* Retrieved from http://epaa.asu.edu/ojs/article/view/297.

Aydeniz, M., & Southerland, S. A. (2012). A national survey of middle and high school science teachers' responses to standardized testing: Is science being devalued in schools. *Journal of Science Teacher Education, 23*, 233–257.

Bailey, K. (1998). *Learning About Language Assessment: Dilemmas, Decisions, and Directions*. Heinle & Heinle, Pacific Grove, CA.

Beard, R. M., & Senior, I. J. (1980). *Motivating students.* London, UK: Routledge & Kegan Paul.

Biggs, J. B. (1973). Study behavior and performance in objective and essay formats. *Australian Journal of Education, 17*, 157–167.

Brickhouse, N. W. (2006). Celebrating 90 years of science education: Reflections on the gold standard and ways of promoting good research. *Science Education, 90*(1), 1–7.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42*(1), 75-83.

DeBoer, G. E. (1991). *A history of ideas in science education: Implications for practice*. New York: Teachers College Press.

Entwistle, A., & Entwistle, N. (1992). Experiences of understanding in revising for degree examination. *Learning and Instruction, 2*, 1–22.

Gilmer, P. J., Sherdan, D. M., Oosterhof, A., Rohani, F. & Rouby, A. (2011). Science competencies that go unassessed. *Online Submission*. Retrieved from http://www.cala.fsu.edu.

Greene, J. P., & Winters, M. A. (2003). *Testing high stakes tests: Can we believe the results of accountability tests?* (Report 33). New York: Manhattan Institute Center for Civic Innovation.

Hamilton, L. S. (1994). *An investigation of students' affective responses to alternative assessment formats*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Hammerman, E. (2005). Linking classroom instruction and assessment to standardized testing. *Science Scope, 28*(4), 26-32.

Harlen, W. (2013). *Assessment & inquiry-based science education: issues in policy and practice*. Global Network of Science Academies.

Hillocks, G. (2002). *Testing trap: How states writing assessments control learning*. New York: Teachers College Press.

Huber, R. A., & Moore, C. J. (2002). High stakes testing and science learning assessment. *Science Educator, 11*(1), 18-23.

Jones, G., Jones, B., Hardin, B., Chapman, L., Yarbrough, T., & Davis, M. (1999). The impacts of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan, 81*(3), 199-203.

Kaestle, C. (2013). *Testing policy in the United States: A historical perspective*. Published by the Gordon Commission on the Future of Assessment in Education. Retrieved from http://www.gordoncommission.org/publications_reports/assessment_education.html.

Keeley, P., & Harrington, R. (2010). *Forty-Five New Force and Motion Assessment Probes*. NSTA Press.

Kersaint, G., Borman, K. M., Lee, R., & Boydston, T. L. (2001). Balancing the contradictions between accountability and systemic reform. *Journal of School Leadership, 11*(3), 217 – 240.

Kohn, A., Thompson, S., Ohanian, S., & Eisner, E. (2001). Fighting the tests: A practical guide to rescuing our schools. *Phi Delta Kappan 82*(5). 248-357.

Lee, O., & Luykx, A. (2005). Dilemmas in scaling up innovations in science instruction with nonmainstream elementary students. *American Educational Research Journal, 42*(3), 411 – 438.

Longo, C. (2010). Fostering creativity or teaching to the test? Implications of state testing on the delivery of science instruction. *Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *83*(2), 54–57.

McMillan, J.H., Myran, S., & Workman, D. (1999, April). *The impact of mandated statewide testing on teachers' classroom assessment and instructional practices*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Morgenstern, C. F., & Renner, J. W. (1984). Measuring thinking with standardized science tests. *Journal of Research in Science Teaching, 21*(6), 639-648.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. Washington, DC: National Academies Press.

Nowak, J. (2007). The Problem with using problem-based learning to teach middle school Earth/Space science in a high-stake testing society. *Journal of Geoscience Education, 55*(1), 62-66.

Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education, 25*, 1049-1079.

Pellegrino, J. W., Chudowsky, N., & Glaer, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Saturnelli, A. M., & Repa, J. T. (1995, April). *Alternative forms of assessment in elementary science: The interactive effects of reading, race, economic level and the elementary science specialist on hands on and multiple-choice assessment of science process skills*. Paper presented at the Annual Conference of the American Educational Research Association, San Francisco, CA.

Schafer, W., Wang, J., & Wang, V. (2009). Validity in action: State assessment validity evidence for compliance with NCLB. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 173–194). Charlotte, NC: Information Age Publishing.

Scott-Jones, D., & Clark, M. L. (1986). The school experiences of black girls: The interaction of gender, race, and socioeconomic status. *Phi Delta Kanpan, 67*(7), 520-526.

Shepard, L. A., & Dougherty, K. C. (1991). *Effects of high-stakes testing on instruction*. Spencer Foundation.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*, 108–131.

Tolman, M. N., Sudweeks, R. Baird, H., & Tolman, R. (1991). What research says: Does reading ability affect science test scores? *Science and Children, 29*(1), 44-47.

United States Congress Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (Vol. 22). United States Government Printing Office.

Wadsworth, B. J. (1984). *Piaget's theory of cognitive and affective development*. NY: Longman.

Appendix 1: Multiple-choice Questions

1. Josey and her little brother Jack are walking side by side, eating ice cream cones. Josey stops to talk to a friend. While she is talking, Jack's ice cream cone starts to drip at a steady rate as Jack walks away. When Josey finishes talking to her friend and realizes that Jack is no longer next to her, she looks down and notices these drops of ice cream on the ground from Jack's ice cream cone:

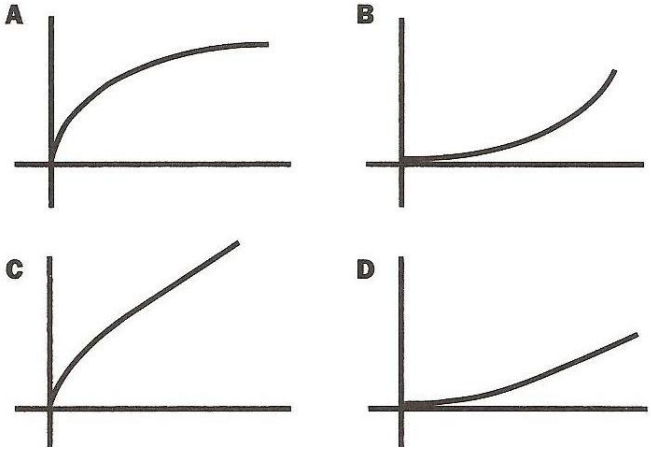
Josey needs help figuring out what Jack was doing. Circle the best answer that best shows how Jack moved (walked) while Josey stopped to talk to her friend.

A. The drips show that Jack started walking slowly and then went faster and faster.
B. The drips show Jack started out walking really fast and then slowed down and went slower and slower.
C. The drips show that Jack started out walking slowly, then walked faster and continued to walk at that same speed.
D. The drips show that Jack started out walking fast, slowed down, and then continued to walk at that same, steady speed.

Explain your thinking. Provide and explanation for your answer.

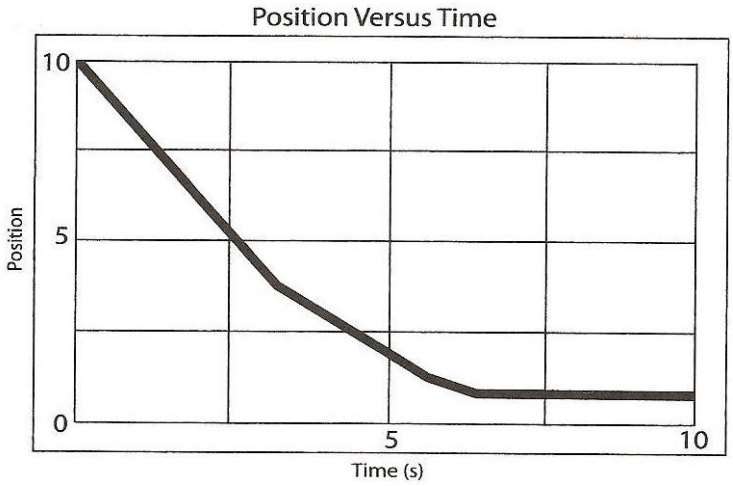
2. Josey and her little brother Jack are walking side by side, eating ice cream cones. Josey stops to talk to a friend. While she is talking, Jack's ice cream cone starts to drip at a steady rate as Jack walks away. When Josey finishes talking to her friend and realizes that Jack is no longer next to her, she looks down and notices these drops of ice cream on the ground from Jack's ice cream cone:

Josey needs help figuring out what Jack was doing. Which of the following position versus time graphs best show how Jack moved (was walking) while he was eating his ice cream cone? Circle the letter of the best graph.



Explain your thinking.
Describe how the graph you chose best matches Jack's motion.

3.



Jim and Karen have built a go-cart. They take their go-cart for a test run and graph its motion. Their graph is shown above. They show the graph to their friends. This is what their friends say:

Bill: "Wow that was a steep hill! You must have been going very fast at the bottom."

Patti: "I think you were going very fast at first, but then you slowed down at the end."

Kari: "I think you must have hit something along the way and come to a full stop."

Mort: "it looks like you were going downhill and then the road flattened out."

Circle the name you think the best describes the motion of the go-cart, based on the graph. Explain why you agree with that friend.

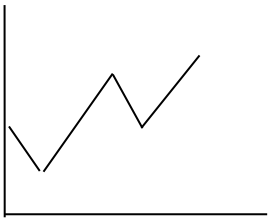
4. Jack is using a thermometer to measure the temperature of water in a can. He is heating and cooling the water in the can. He recorded his temperature measurements in the table below.

Temperature of Water Measured by Jack

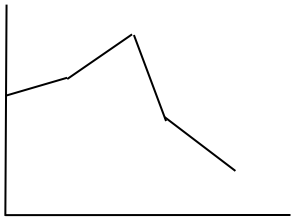
Measurement (Number)	Temperature (°C)
1	25
2	35
3	20
4	10
5	40

Can you help Jack to graph his recordings? Which of the following temperature versus time graphs best shows the recordings of Jack? Circle the letter of the best graph.

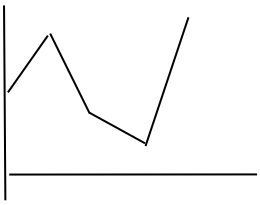
A



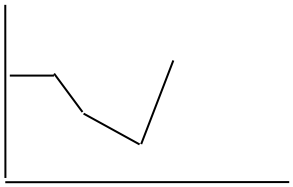
B



C



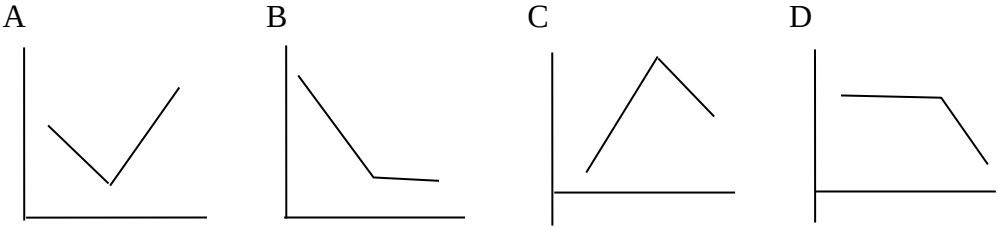
D



Explain your thinking:

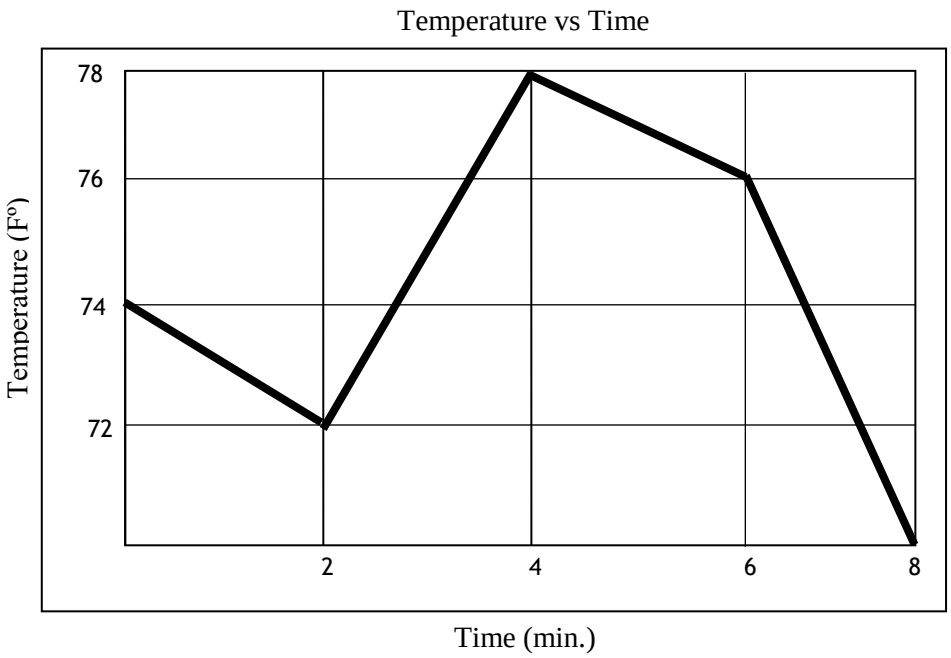
5. Jim is feeling thirsty after playing with his friends outside. He goes home and grabs a bottle of water from the refrigerator. But he realizes that the water is too cold for him to drink. He places the water bottle outside under the sun for a while. Then he realizes that the water is too warm and he puts some ice in the water bottle.

Which of the following temperature versus time graphs best shows how Jim`s drinking water temperature changes? Circle the letter of the best graph.



Explain your thinking:

6.



Jim and Karen are measuring the temperature of water in a can as they are heating and cooling the water. Their graph is shown above. Which of the following choices best shows how Jim and Karen heated and cooled the water in the can? Circle the letter of the best choice.

A. heat, cool, heat, heat
B. cool, heat, cool, cool
C. heat, heat, cool, cool
D. cool, cool, heat, heat

Explain your thinking:

Appendix 2: Excerpts from Student Interviews

Correct answer-correct explanation
Question 6

Student N: (reading question number 6 and her answer is B)
Author 1: why did you choose B?
Student N: well… this graph shows that it is cooling
Author 1: what is the initial temperature of this water?
Student N: 20
Author 1: of what? Is that Celsius or Fahrenheit?
Student N: Celsius
Author 1: and then what happened?
Student N: and then it heated up to 40° of Celsius
Author 1: it was 20 at the beginning then what happened here?
Student N: I mean cooled to 10°C and then it went up to 40°C.
Author 1: what about the time? When did this happen?
Student N: at four minutes
Author 1: and then
Student N: then it cooled to 30 degrees of Celsius in six minutes then cooled again to…I think 0°C in 8 minutes

Correct answer-incorrect explanation
Question 2

Student P: (reading question number 2 and her answer is D and she also read her explanations) because it shows that slows down. If he is going up more, so like he's getting tired.

Author 1: so, can you tell me the difference between A. and B.?

Student P: there's a shape difference like he is already almost up. Then it stops.

Author 1: what about B.?

Student P: it seems like then he just starts to go up.

Correct answer-partially correct explanation
Question 2

Student J: (reading question number 2 and his answer is D. he also read his explanations)
Author 1: Have you ever seen a graph before?
Student J: not really
Author 1: how did you come up with that answer? Do you know how to read a graph?
Student J: yeah
Author 1: so, what is the difference between A and D.
Student J: they are basically. this is position. So, he's position went dramatically up and just slowed down. And this one it was just slow the entire time.
Author 1: what about C.?
Student J: that means he dramatically way up in position, because as higher he goes…
Author 1: what would happen if it were a just straight line like here?
Student J: it means that he is just standing there. He would not have gone anywhere, because its position as he goes up that's how much far he is gone.

Correct answer-insufficient explanation
Question 4

Author 1: you chose C. in the question number four. How did you come up with that answer?
Student S: I used dots again this time
Author 1: what kind of dots?
Student S: tiny dots. I put five dots in each one.
Author 1: can you make dots for me in C.?
Student S: this is the six dots right here in A and I do not have any six on my measurement number. But on C. I put five dots because it could fit. And so I was like maybe C. is the answer because I could put 5 dots and it fits.
Author 1: when you look at the question, what does it tell you? Did you try to put these values into the graph?
Student S: no, I was just trying to put five dots.

Incorrect answer-correct explanation
Question 1

Student N: (reading question number 1) … and her answer is A.
Author 1: why did you choose A?
Student N: the drips showed that he kept going farther and farther away. It looks like he went faster and faster.
Author 1: do you think that is it always the case?
Student N: here it kind of went steady.

Incorrect answer-incorrect explanation
Question 3

Student N: (reading question number 3 and her answer is Mort and she is also reading her explanations)
Author 1: so, when I look at the graph I wonder what's happening at the beginning portion of the graph? What do you think that is happening here?
Student N: they are going downhill
Author 1: what is happening in the end?
Student N: it flats-out.
Author 1: do you think there is a go kart… when you put go kart here, it's going down fast fast fast and it slows down here. Do you think is it like that?
Student N: it shows that it's going fast at the beginning and then it kind of slows down a little bit.
Author 1: what would happen if the graph was starting from point 0 and am going straight up to here?
Student N: that would show… it kind of so-so
Author 1: is he running? is he going at the same speed? Is he stopping?
Student N: he's just going with the same speed all the way.
Author 1: what would happen if it were starting from and going like this to till here? It is just a horizontal line.
Student N: I think it would be like walking pace sort of.

Incorrect answer-partially correct explanation
Question 2

Student G: (reading question number 2 and her answer is C)
Author 1: why did you choose C.?
Student G: because it is the pretty straight one.
Author 1: can you tell me what the difference is between A. and C.?
Student G: that one is kind of bendy and that goes straight. It like slow...it is steep.
Author 1: what about this one?
Student G: that one starts fast and go slow.
Author 1: is this always the same speed?
Student G: yes
Author 1: can you tell me the difference C. and D.?
Student G: this one started creasing speed and then goes faster.
Author 1: the second part of C. and D… are they the same?
Student G: yeah
Author 1: what about the initial, first portion of these graphs? What is the difference as motion between C. and D.?
Student G: that one is kind of fast (C) and this one is slow (D).
Author 1: what do the drips tell us?
Student G: that slow and fast.
Author 1: so, what do you think is the answer?
Student G: D

Incorrect answer-insufficient explanation
Question 2

Student M: (reading question number 2 and his answer is B.
Author 1: what is he doing here? How is he moving?
Student M: same on the previous one
Author 1: what was that?
Student M: first he is tiptoeing, and walking a little bit faster and then started learning little bit…
Author 1: is he always increasing his speed?
Student M: no
Author 1: why?
Student M: he just continued at the same speed.
Author 1: you chose B. why did you choose this graph?
Student M: (no reply)
Author 1: what does it tell you? If I divided it by two, what does the first portion tell you?
Student M: it was straight and then curved a little bit
Author 1: okay. What kind of a movement is that?
Student M: (no reply)
Author 1: can you tell me the motion difference between A. and B.?
Student M: that is upside down to the other one.
Author 1: if you see a graph like this. This is position versus time graph. Let's say you saw a graph like this. This is zero, starting point and this is 10. Position is in meters and time is in seconds. So what do you think this graph means?
Student M: (no reply)
Author 1: it is a horizontal graph… does it tell you anything about the movement?
Student M: (no reply)