

Measuring Graph Comprehension, Critique, and Construction in Science

Kevin Lai¹ · Julio Cabrera² · Jonathan M. Vitale³ · Jacquie Madhok³ · Robert Tinker⁴ · Marcia C. Linn³

Published online: 21 April 2016
© Springer Science+Business Media New York 2016

Abstract Interpreting and creating graphs plays a critical role in scientific practice. The K-12 Next Generation Science Standards call for students to use graphs for scientific modeling, reasoning, and communication. To measure progress on this dimension, we need valid and reliable measures of graph understanding in science. In this research, we designed items to measure graph comprehension, critique, and construction and developed scoring rubrics based on the knowledge integration (KI) framework. We administered the items to over 460 middle school students. We found that the items formed a coherent scale and had good reliability using both item response

theory and classical test theory. The KI scoring rubric showed that most students had difficulty linking graphs features to science concepts, especially when asked to critique or construct graphs. In addition, students with limited access to computers as well as those who speak a language other than English at home have less integrated understanding than others. These findings point to the need to increase the integration of graphing into science instruction. The results suggest directions for further research leading to comprehensive assessments of graph understanding.

Keywords Assessment · Graphing · Graph understanding

Electronic supplementary material The online version of this article (doi:10.1007/s10956-016-9621-9) contains supplementary material, which is available to authorized users.

✉ Kevin Lai
klai@mills.edu

Julio Cabrera
jcabrera@umn.edu

Jonathan M. Vitale
jonvitale@berkeley.edu

Jacquie Madhok
jjmadhok@berkeley.edu

Robert Tinker
bob@concord.org

Marcia C. Linn
mclinn@berkeley.edu

¹ Mills College, Oakland, CA, USA

² University of Minnesota, Twin Cities, Minneapolis, MN, USA

³ University of California, Berkeley, Berkeley, CA, USA

⁴ Tinker Group, Amherst, MA, USA

Introduction

The need to understand and use graphs in science motivated us to create and test the Graphing Inventory. Our design of items and development of rubrics is guided by the knowledge integration (KI) framework understanding (Kali 2006; Linn and Hsi 2000; Quintana et al. 2004) that emphasizes coherent links between evidence and scientific reasoning to construct arguments. In the case of graph understanding, the knowledge integration rubric rewards students for using evidence from graphs to answer a question about a science topic, such as photosynthesis, plate tectonics, or kinematics. To construct a strong response, students need to make sense of the features of graphs—including labels, scale, shape, noise, and patterns—to describe, depict, and evaluate claims regarding scientific phenomena.

We developed, tested, and refined items to measure graph comprehension, critique, and construction. We took advantage of an interactive, online system to create and deliver the items. We used the knowledge integration (KI)

framework to develop rubrics that captured links between graph knowledge and scientific understanding. We investigated how diverse middle school students perform on these items and identified student characteristics that predicted success.

We focused on middle school because this is when students begin to study complex science phenomena that can be captured in graphs. Graphing is generally taught in mathematics classrooms, where students typically graph only linear functions (Cobb 1999; Watson 2008), and rarely encounter features important in science, such as units, scientific notation, non-integer values, noise, oscillations, and non-linear functions. Graphs with these features are essential in science. Science teachers rarely teach about the graph features needed in science, so students are left to flounder (Gal 2002; Galesic and Garcia-Retamero 2011; Gallimore 1991; Jarman et al. 2012).

Research identifies many weaknesses in student ability to interpret graphs. International comparisons (OECD 2006) and previous research (Chang and Linn 2013) document students' inability to understand more than simple graphs. Many students interpret graphs as pictures, fail to note differences in scale, expect scales to start at zero, and connect dots rather than fit curves to noisy data (Clement 1985).

A predator–prey item illustrates the role of graphs in science learning (Fig. 1). In a simple agent-based model of this system, the time-series graphs of predator and prey populations oscillate irregularly (Wilensky and Reisman 2006). Students experienced in graph interpretation would ignore the irregularities and notice that the repeating patterns of the predator population usually occur just after those in the prey population. Such students would recognize that the predator population increases slowly and is followed by a rapid drop in the prey population. Following the predators' crash, the prey recovers quickly, but is soon

followed by an increase in the predator population. The fourth population maximum in Fig. 1 is a slight exception to this rule that students need to think about. This repeated pattern emerges from the model's rules and can be understood by a convergence of knowledge of graphs and science. Knowledge of cycles and noise in graphs allows students to recognize a pattern. Understanding species relationships suggests possible causes of the graph's oscillations.

To conduct this research, we identified a preliminary set of 14 multiple-choice and explanation questions that require students to link graph knowledge and scientific understanding. We refined these items and administered them to a diverse population of middle school students. In this paper, we investigate the following research questions:

1. What are the characteristics of a valid and reliable Graphing Inventory that measures student ability to link graph knowledge and scientific understanding?
2. How do diverse middle school students perform on comprehension, critique, and construction items and what student characteristics predict success?
3. What types of responses do students give to illustrative items and what are the implications for instruction?

Rationale for Assessing Graph Understanding in Science

Graphs take advantage of the human capacity to visualize large amounts of data in ways that reveal patterns, uncertainty, and critical events (Friel et al. 2001; Wu and Krajcik 2006). Graphs in science, specifically, require students to observe general shapes and patterns that reveal natural processes. Developing expertise with science graphs requires students to make connections between these graphical patterns and the underlying processes that support them (Shah and Hoeffner 2002). Likewise, strong science graph understanding can facilitate learning of new scientific concepts. From this perspective, the relationship between understanding of graphs and scientific concepts is bidirectional and contextualized.

Due to their common usage by practicing scientists, engineers, and mathematicians (Mokros and Tinker 1987) and in daily experiences with media (Gillen 2006; Ozcelik and McDonald 2013), both the American Association for the Advancement of Science (1993) and the National Research Council (1996) emphasize the need for diverse graphing activities (creating, reading, reasoning, and modeling) in science education. Likewise, the recent Next Generation Science Standards (NGSS) call for the use of graphs to convey core disciplinary ideas, such as the relationship between potential and kinetic energy (NGSS Lead States 2013).

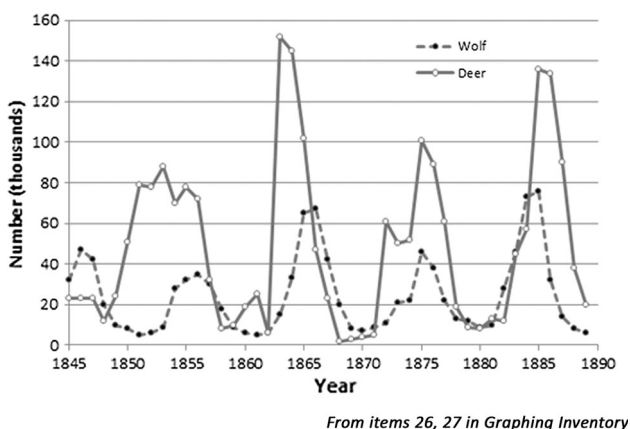


Fig. 1 The prey–predator model illustrates the synergies between interpreting graphs and understanding science

Beyond their ability to convey particular ideas, graphs represent a tool for thinking that can be applied to a wide range of scientific domains and practices. This flexibility enables graphing activities to meet the NGSS emphasis on crosscutting concepts—including “scale, proportion, and quantity”—and science and engineering practices—including “developing and using models” and “obtaining, evaluating, and communicating information.” These standards, while somewhat open-ended, reflect a coherent approach to science education, in which general practices and conceptual themes are recurring. Students are expected to understand not only the procedural aspects of a practice such as graph interpretation, but the potential applications and justification for the practice.

Assessing graph understanding in science therefore requires a focus not only on the technical skills required to successfully interpret or construct a graph, but also on an understanding of how graphs are used to convey evidence and persuade an audience. This multi-level approach to graph understanding requires an appropriate assessment framework that takes into account specific ideas as well as how they are composed into an argument. In prior research the KI framework has successfully guided the design of items and rubrics that require students to link ideas and evidence involving multi-level, cross-cutting concepts, such as students’ ideas about energy flow (Liu et al. 2011). In this research we draw on the KI framework to design measures of graph understanding.

The KI framework emphasizes engaging students in scientific practices that involve combining scientific concepts and evidence into arguments that explain complex phenomena. For example, a KI assessment item may ask students to analyze data showing the role of a plant in converting light energy into chemical energy. Another item might ask students to explain how a “green roof” covered with plants can be used to reduce cooling costs. Design of these items requires trial and refinement starting with a comprehensive review of possible applications of a practice or concept and insight into the potential student ideas (including incorrect ideas) that may arise.

Measures of Graph Understanding

Graph understanding is poorly represented in current standardized tests, which contributes to its neglect in the curriculum. Yeh and McTigue (2009) reviewed 985 released items in state and national tests and found that for late elementary and middle school only 6.7 % of the items included graphs. The few examples of graph comprehension items involved either predator–prey relationships or thermal equilibrium.

Several assessments focused on graph-based instruction and assessment have been developed by researchers. For

the Graphing Inventory, we draw on measures of graph understanding in science developed to assess the impact of instruction using real-time data collection in microcomputer-based laboratory activities (MBL) (Brasell 1987; Linn et al. 1987; Mokros and Tinker 1987).

The 26-item multiple-choice test, called the Test of Graphing in Science (TOGS), measures secondary school students’ graph understanding (McKenzie and Padilla 1986). TOGS items measure students’ ability to interpret graphs, determine the X and Y coordinates of a point, interpolate and extrapolate, state relationships between variables, and interrelate the results of two or more graphs. While TOGS is reliable (McKenzie and Padilla 1986), the multiple-choice format limits its potential to assess how students interpret evidence from graphs to explain science outcomes. Berg and Smith (1994) interviewed students concerning responses to TOGS and found that the multiple-choice items were inadequate for measuring students’ complex ideas about graphs. As a result, Adams and Shrum (1990) created TOGS-R by adding constructed-response explanations to TOGS to assess the impact of MBL on students’ understanding of graphs. Adams and Shrum also created an interview format, I-TOGS, to assess students’ reasoning processes. Similarly, Boote (2014) engaged students in a think-aloud protocol as they engaged in TOGS items, revealing that students have more sophisticated as well as more problematic ideas than are captured in their multiple-choice answers. Several groups have concluded that open-response items are needed to clarify students’ understanding and provide teachers with insight into students’ thinking processes (e.g., Wang et al. 2012). We respond to these insights in developing the Graphing Inventory.

Despite these attempts to remedy difficulties in the TOGS, ongoing educational and psychological research as well as new standards emphasizing the use of graphs suggests that graph understanding in science deserves renewed attention. The development of the Graphing Inventory allows us to integrate both the insights from research about graph understanding and the needs of educational practitioners.

Processes of Graph Understanding

At a general level, graph understanding in science is the ability to recognize and communicate about features of graphs, coordinate between multiple representations conveying similar information (e.g., graph and table), and interpret graphs in a science context. This ability is sometimes referred to as “graphicacy” (Wainer 1992) or “graph sense” (Friel et al. 2001). This definition affirms that graph understanding is a multi-level process that spans

from perceptual actions such as observing a point to complex activities involving multiple inferences.

Educational theorists and cognitive psychologists (e.g., Bertin 1983; Kosslyn 1989), Shah and Hoeffner (2002) describe graph understanding in terms of three component processes: encoding visual information, relating visual features to the represented concepts, and understanding the disciplinary context. Specifically, to analyze graphs, students must initially recognize and encode specific visual features of a graph, such as the curvature of a line or the grouping of points. Next, students need to interpret these visual features in terms of the magnitude of relationships between variables, such as how an upward curving parabola in a distance vs. time plot reflects accelerating distance. Finally, students must make connections to the disciplinary context, enabling them to make accurate inferences about the system being graphed.

While these three processes are applicable to all graphs, the specific features and patterns in a graph differ according to graph type. For example, patterns that arise in a line plot are not the same as those that might arise in a scatter or bar graph. Each graph type requires a specific set of knowledge and skills. For example, bar graphs promote discrete comparisons between magnitudes, while line graphs highlight trends in x - y data (Shah and Hoeffner 2002). If a bar graph displays change in quantity over time (i.e., a non-univariate relationship), students are likely to miss important trends. In all types of graphs that depict two dimensions (x , y) students often have difficulty interpreting scale, or depicting information at an appropriate scale (Leinhardt et al. 1990).

Within each graph type, students are prone to specific errors. For graphs of functions, students tend to make errors interpreting the slope and axes in terms of the content (Beichner 1994). In particular, students often treat a graph as a literal picture of a problem scenario, thereby ignoring the meaning of axes (Clement 1985). For example, in depicting the speed over time of a bicycle traveling over a hill, students display graph-as-picture errors and simply draw a picture of a hill. Also, when presented with an inverted U-shaped curve, students are likely to interpret motion as up and over a hill (Leinhardt et al. 1990). Similarly, in a graphing activity of a race car traveling around a track, students confuse oscillations in a speed vs. distance graph for the curvature of the race track referenced in the problem situation (Janvier 1981). In these cases, students fail to recognize the meaning of variables represented on each axis.

In line graphs, students often confuse the meaning of slope and height (Clement 1985; McDermott et al. 1983). Specifically, when asked to compare the speeds of two objects, students often evaluate the height of the graph

representing these objects, rather than their relative slopes (McDermott et al. 1983). While students may understand the concept of slope from everyday experience with covariation (Schlieman et al. 1997), this error reflects a mis-mapping between the slope concept and the graph feature (Clement 1985). For other graphical features, such as the y -intercept, students may not have a clear conceptual understanding from everyday experience to map onto the graph features (Davis 2007).

In addition to the variety of graph types (e.g., line, bar, scatter), graph assessment items may vary how students are prompted to interact with graphs. They may be asked to interpret a valid graph, critique a potentially faulty or misleading graph, or construct a new graph. To measure graph understanding, we designed or selected items that required either *comprehension*, *construction*, or *critique*, across a range of graph types.

Comprehension

A graph can depict a great deal of information efficiently. To comprehend a graph, the student needs to make sense of this information. This requires understanding of the graphs' features and context and the ability to make valid inferences from that information (Preece and Janvier 1992; Roth and Bowen 2001). The ability to make inferences from graphs represents a fundamental graphing skill (Glazer 2011).

Successful graph comprehension requires all three of Shah and Hoeffner's (2002) processes. Individuals must recognize graph features, interpret general relationships, and relate relationships to the disciplinary context. Similarly, Curcio (1987) describes these processes as reading with the data, reading between the data, and reading beyond the data.

We generated items to represent these three aspects of graph comprehension. To measure identification of graph features, some items prompted students to locate the coordinates of a point or distinguish the relative height of bar graphs. To measure the ability to interpret general relationships, some items prompted students to describe shapes, trends, and noise in the depicted graph. To relate relationships to the disciplinary context, some items required students to interpret the quantitative relationships in the science context (see examples in Fig. 2). The first two aspects allow us to evaluate students general graph knowledge, independent of their knowledge of science. The third aspect, however, is intended to directly measure students' ability to integrate science and graph ideas. By focusing on all three aspects, we can investigate relationships between general graph abilities and scientific reasoning.

A hot bowl is left on a cool table. The graph below represents what happens to the temperature of the bowl and the cool table.

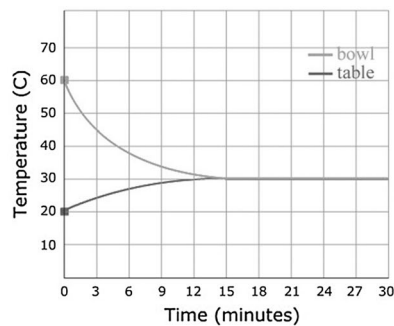


Fig. 2 A thermodynamics item illustrates the three processes of graph comprehension. Questions 47 and 48 ask about specific points on the graph (starting temperatures). Question 49 asks about the relationship between the data points and requires noticing the shapes

47. What is the starting temperature of the bowl?

Ans. 60 °C, Correct: 87%, Incorrect: 13%

48. What is the starting temperature of the table?

Ans. 20 °C, Correct: 91%, Incorrect: 9%

49. Describe the differences in SHAPE between the two graphs.

KI Score 1: 9%, 2: 24%, 3: 29%, 4: 34% and 5: 4%

50. Use evidence from the graph to explain what is happening to the temperature of the bowl and the table.

KI Score 1: 5%, 2: 17% 24%, 3: 18%, 4: 51% and 5: 10%

Critique

One of the central goals of the NGSS is to engage students in the practices of scientists, such as making arguments from evidence. This is consistent with the role that graphs play in conveying scientific arguments as well as their ubiquity in persuasive advertising. Thus, measuring student ability to critique scientific graphs is important for science literacy (Chang and Linn 2013). Many see critique as a key component of metacognition (Bertin 1983; DiSessa and Sherin 2000). Specifically, diSessa and Sherin (2000) describe critique as a main element of meta-representational competency (MRC). MRC involves the ability to judge whether a representation conveys information accurately and effectively. Critical evaluation of representations in science is advocated by many research groups (e.g., Wiley et al. 2009).

While graph critique activities are prone to the types of misinterpretations described for comprehension activity, they also engender unique challenges. In many cases students simply do not know how to use graphs as evidence for a specific argument (Lovett and Chang 2007). In the Graphing Inventory, we measure this dimension by developing items that ask student to critique alternative interpretations of graphs. Consistent with the contextual aspect of graph comprehension items, critique items can ask students to evaluate links between graph and science concepts.

Construction

Graph construction involves illustrating relationships in data sets, experiments, or scientific processes (Barclay 1985; DiSessa et al. 1991; Latour 1990). This process requires students to not only interpret graph features but

of the graphs. Question 50 asks about the relationship between the graph information and science disciplinary knowledge concerning heating and cooling curves in thermodynamics as well as thermal equilibrium

also to use the features to design a graph. Leinhardt et al. (1990) explained, “Construction is quite different from interpretation. Whereas interpretation relies on and requires reaction to a given piece of data (e.g., a graph, an equation, or a data set), construction requires generating new parts that are not given” (p. 12). Graph construction is an important yet neglected skill that is especially crucial today when there are many graph construction tools available on computers.

Mevarech and Kramarsky (1997) documented difficulties students face in graph construction. They noted three main sources of alternative conceptions: constructing an entire graph as one single point, constructing a series of graphs, each representing one factor from the relevant data, and assuming an increasing function under all conditions. Graph construction assessment activities provide an opportunity to measure how students represent scientific information.

Additionally, graph construction activities can expose students’ non-normative ideas about scientific concepts (Vitale et al. 2015). For example, if students depict cooling as a linear relationship between temperature and time, they demonstrate an incomplete knowledge of the science. Yet, revealing students’ science ideas in graph construction requires activities with the flexibility to accommodate a range of features representing distinct ideas. Plotting values from a table is unlikely to demonstrate students’ scientific understanding. Rather, Hattikudur et al. (2012) recommend the use of qualitative graphs that allow students to depict general relationships without specifying all numerical values. By allowing students to depict qualitative relationships, graph construction can be used to evaluate integration of graph and science knowledge from a different perspective than graph comprehension. We developed a graph construction item that exposes diverse student

ideas (including alternative science conceptions) and differentiates between levels of students' graph understanding.

The Knowledge Integration Framework

The KI framework describes both the developing nature of students' scientific knowledge and an instructional approach to promote conceptual change. The KI framework emphasizes that students generally hold multiple, diverse, and conflicting ideas (Linn and Hsi 2000; Linn et al. 2006). Progressing toward a coherent, normative understanding of a scientific phenomenon requires students to actively reason about and evaluate their own ideas in the light of new evidence. The KI instructional pattern describes successful ways to design instruction that promotes coherent instruction. This involves eliciting students' ideas so they can reconsider them. It also involves adding new normative ideas for students to consider. This is not sufficient, however, since students may simply add the new ideas to their already diverse ideas. A key to KI is encouraging students to distinguish between the multiple ideas and formulate a coherent perspective. This instructional pattern has been used to design successful curriculum units and assessments across a wide variety of scientific disciplines (Lee et al. 2010).

Liu et al. (2011) used the KI framework to design and develop valid, authentic and efficient assessments and rubrics for the Graphing Inventory. The items were designed to measure the process of distinguishing among ideas. Successful answers involve linking the most promising scientific ideas to evidence from the graphs. KI rubrics assess the depth of students' scientific understanding by evaluating alternative and correct ideas, and the extent to which the students make valid links between correct ideas. Liu et al. (2011) showed that the constructed-response (CR) items capture more nuanced and diverse student ideas than multiple-choice (MC) items. Knowledge integration rubric-based scoring of constructed-response items therefore provides a clearer profile of student knowledge than traditional, multiple-choice-based assessments. In the following we describe the design of our instrument and its application in a set of middle school science classrooms.

Methods

Participants

Over 460 students (6th to 8th graders) taught by seven different teachers in 22 classrooms in five different schools participated in this research. Students were 47 % female;

41 % spoke English as a second language; and 52 % reported using computers for homework.

Designing Graph Understanding Items

To measure graph understanding, we reviewed previously tested KI items from the Web-based Inquiry Science Environment (WISE) units (i.e., global climate, seasons, graphing stories) and available items from standardized tests. In addition, we created new KI graphing items designed to measure ability to link scientific ideas and evidence from graphs. We created or selected items where graph and science knowledge intersected. This resulted in a set of 60 items. We pilot tested these items with over 60 students for comprehensibility.

The pilot tested items and student responses were reviewed by a team of expert reviewers consisting of middle school and high school science teachers, professional developers, science education researchers, and science discipline experts. The reviewers examined the questions and student responses. They identified responses they felt reflected misunderstanding of the question and suggested revisions. They reviewed the science behind the questions and identified any items with erroneous or inappropriate content for middle school. By looking at the student responses, they also suggested ways to rephrase confusing wording and improve graph displays. We revised items based on reviewer comments and eliminated items that reviewers found inappropriate.

To compose the final version of the Graphing Inventory, the research team developed a set of item features appropriate for the target population (6th to 8th graders) and used these features to select a representative set of items. We created a set of core items consisting of 14 constructed-response or multiple-choice core items that could be performed in one class period. We assigned additional promising items to Form A or Form B and included them at the end of the core items to be attempted by students who finished the 14 core items.

Administration of the Graphing Inventory

The Graphing Inventory (See supplementary materials) was administered using WISE (<http://wise.berkeley.edu>). WISE is a robust platform that incorporates technological features, such as interactive visualizations, embedded assessments, and data logging to communicate complex scientific phenomena (Linn et al. 2003; Slotta and Linn 2009). We recently added a feature that supports student construction of graphs (Vitale et al. 2014).

All the students ($n = 463$) responded to the core set of items. In addition, 149 took form A with one set of additional items, and 159 took form B with a different set of

additional items. One hundred and eighty-two students completed the construction items.

Rubric Design and Coding

We used the knowledge integration framework to develop rubrics for all the comprehension, critique, and construction items (Linn and Eylon 2011; Liu et al. 2011). We scored multiple-choice questions as correct or incorrect. The constructed-response rubrics were developed by a team of science teachers, professional developers, science education researchers, and expert coders. The rubrics rewarded students for linking evidence from graphs to relevant science disciplinary knowledge. The scores for the constructed-response items ranged from 0 to 5.

In general, the rubric awards a 0 for “no answer” and a 1 for “off task” or irrelevant responses. Students receive a 2 if they display non-normative ideas or links. Students receive a 3 at the “partial link” level, when they generate relevant ideas but do not elaborate on how two ideas are connected. Students receive a 4 at the “full link” level, when they demonstrate one link between two relevant and normative ideas, and receive a 5 for having two or more links (“complex” level).

We sought to make links consistent across items. In the case of graphical representations, we coded a link when students were able to either make a connection between one aspect of a graph to another aspect of the graph or to a piece of scientific information. For example, in an item about a race, one link connected the steepness of the slope of a line in a position and time graph to the speed of the runner in a story. Another link connected the slope of the line to the distance covered by the runner in a specified time (i.e., rise over run).

We used a similar approach to identify links in the critique items. In one critique item, students say which of two graphs best explains global climate change. One link is scored for noting that more data points make the results more interpretable (due to including more variation in temperature, more data across the years). Another link is scored for noting that the data points display a specific trend.

For the construction item (Fig. 3), students created their own graphs by using their mouse to select and plot points on a grid. Students could drag existing points anywhere within the grid. Graphs could be cleared and reset.

To score students’ graph construction responses, the rubric takes into account students’ representation of the features of the graph (e.g., linking the narrative information to represent initial points, slope, general shape in the graph). More specifically, the three features identified in this particular problem are: (1) starts near (0, 20), (2) consists of a positive sloped line and a zero-sloped line,

and (3) consists of a positive sloped line and a negative-sloped line.

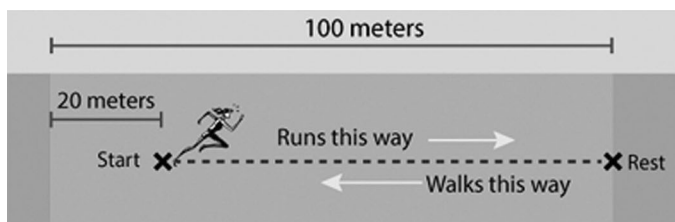
For Fig. 3, the students’ sample responses A are rewarded a 2, because they depict situations with major errors (non-functional graph) and demonstrate students’ non-normative ideas (i.e., graph as picture). Response B is rewarded a 3, because the graph is a functional graph and includes at least one normative feature (a positive sloped line and a negative-sloped line). Response C (the correct answer) is rewarded a 4, because it is a functional graph that consists of at least two of the normative features.

For each item, two independent coders scored 20 % of the responses. When inter-rater reliability reached .80, one coder then scored the remaining items. The results from selected case study items are presented toward the end of the results section.

Data Analysis Procedure

The psychometric properties of the graph understanding test were established using both classical test theory and Item Response Theory (IRT) methods. The reliabilities (internal consistencies) of the test forms were calculated using Cronbach’s alpha. Item-test correlations were calculated to show how well each item correlated with the entire test. For the IRT analyses, a Rasch Partial Credit model (Rasch 1966) and the parameter estimation algorithm EM (Bock and Aitkin 1981) was used to estimate student latent abilities and item difficulty estimates. The ConQuest software (Wu et al. 1998) provided an ability estimate [maximum likelihood estimate (MLE)] for each student and an item difficulty estimate for each item, calibrated on the same logit scale ranging from -3.0 (lower ability/difficulty) to 3.0 (higher ability/difficulty). A Wright map was also created to visually show the distribution of student ability and item difficulty estimates along the same scale. It is ideal that the item difficulty distribution covers approximately the same span of the student ability distribution, thus providing accurate measures of student proficiency (i.e., small errors in ability estimation) over the whole scale.

Furthermore, Differential Item Functioning (DIF) analyses were used to detect any potential bias by subgroup. An item is labeled with DIF (i.e., biased) if the probability of a correct response to the item is significantly different for members of one subgroup (e.g., females) over another (e.g., males) at the same ability level. We compared male and female subgroups, students who spoke only English at home versus a language other than English, and students who used computer for homework versus those who did not. Previous research has demonstrated that these characteristics frequently impact outcomes in science and technology domains and that assessment items may harbor unintended biases against subgroups represented by these



60. Gizelle trains for a race on a 100 meter long field. To train she repeats the following routine every 2 minutes (120 seconds).
- She begins her sprint 20 meters from the end of the field.
 - She sprints to the other end of the field.
 - She takes a short rest.
 - She walks back slowly to her starting position.
- Draw a graph to represent how Gizelle could complete a 2 minute routine.

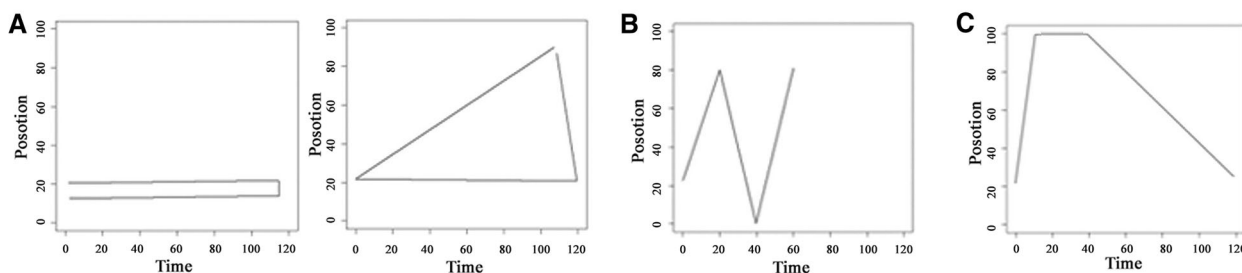


Fig. 3 The Graph construction item “Training for a Race” (Question 60) and student constructed graphs. 65 % of respondents drew pictures (see a), 25 % used incorrect functions (see b), 1 % gave correct answers (see c), and the rest were off-task

variables (i.e., females, non-native English speakers, and students without home access to computers) (Aberg-Bengtsson and Ottosson 2006; Lee et al. 2009; Liu et al. 2011). Our goal in tracking and evaluating these student characteristics was twofold: to ensure that no clear biases were inherent in our instrument and to investigate how these characteristics influence performance in graph-based science activities, which require a complex array of knowledge and experience.

Results and Discussion

To answer our first research question, we assessed the characteristics of the graph understanding test using IRT and classical test theory. To answer our second research question, we analyzed performance of subgroups using indicators of gender, home language, and computer use. To answer our third research question, we analyzed student responses to illustrative comprehension, critique, and construction items.

Properties of the Graphing Inventory

Internal Consistency

Overall, the Graphing Inventory had good internal consistency. The core items had a Cronbach’s alpha of .80. The two longer forms of the test (Form A and B) had Cronbach’s alpha values of .86 and .88, respectively. The Cronbach’s alpha value for the constructed-response core

items was .71; the value for constructed-response items in Form A and B was both equal to .83.

Item Fit

Using the Rasch Partial Credit Model, all item fit statistics were found to be satisfactory, i.e., there were no items showing misfit based on infit and outfit statistics. Specifically, the values are within acceptable values for weighted fit mean square estimates, between 0.75 and 1.30 with *t*-values of less than 2.0 for samples <500 (Bond and Fox 2007).

Wright Map

Results from the Wright map (Fig. 4) show that the design and scoring of the items was satisfactory. The Wright map shows the distributions of student ability and item difficulty estimates on the logit scale -3.0 to 3.0 . On the left side, the distribution of students according to their ability is shown with each *x* representing 3.2 students. The higher on the scale, the more able students are on the construct. On the right side of the figure, the distribution of the item difficulties is shown along with the item thresholds. For example, items 17 and 48 appear lower on the scale, being the easiest items, while item 55 appears highest on the scale, being the hardest item. The relative position of the “*x*” to the item number tells us about the students’ response to each item. If a student estimate and an item estimate are at the same horizontal position, the student has a .50 probability of answering the item correctly. If the student position is above the item, the student is able to answer the item correctly with a

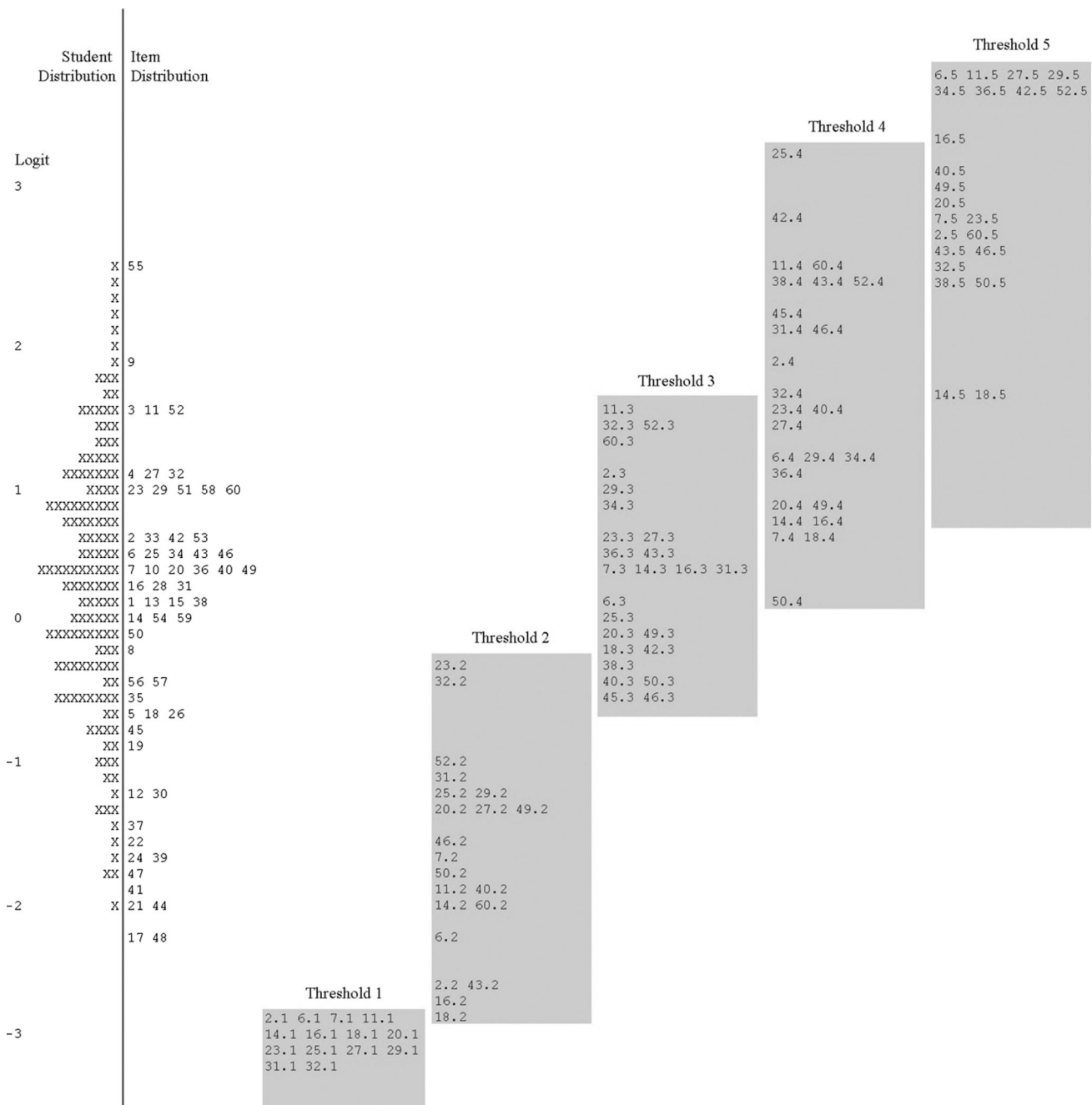


Fig. 4 Wright Map: Distribution of student ability and item difficulty estimates. The thresholds represent the difficulty of scoring $k + 1$ instead of k ($k = 0, 1, 2, 3, 4$). For example, Threshold 4 is the difficulty of scoring 4 instead of 3. Threshold 2.1 stands for Threshold 1 on item 2

probability larger than .50, and if the student position is below the item estimate, the student has <50 % chance of answering it correctly. The Wright map shows that the item distribution covered the span of the student estimates. Most of the item difficulties along with the student ability fell between -2.0 and 2.0 logits. There were also some easy and difficult items at the tail end of the item distribution, providing adequate measures for a range of student abilities.

The Wright map also shows the results for the four levels of the knowledge integration rubric. For example, threshold 1 is the difficulty of scoring 1 instead of zero, and threshold 5 is the difficulty of scoring a 5. As expected, the Wright map shows that the scores of 5 were the most difficult to obtain, followed by scores of 4, 3, 2 and 1. The lower the item threshold on the scale, the less difficult for students to receive that score on the item. This progression confirms the design and scoring of the items.

Item-Test Consistency

We also calculated item-test consistency in order to examine the correlations between the students' responses to each item with their ability estimates measured by the entire test and found that all but two items had an item-total correlation $>.20$. The two items were kept in the analyses because their discrimination values were close to $.20$, and their item fit statistics were also satisfactory.

Differential Item Functioning (DIF) Analysis

Uniform DIF analysis was conducted across the subgroups (gender, ELL, computer use) to determine whether items measuring the graph understanding construct significantly favored students in one subgroup (e.g., males) over students in another subgroup (i.e., females) of equal ability. The DIF value for each item is computed as the difference between the two groups' relative difficulty estimates. An item is flagged as having substantial DIF if the Chi-square test performed on the DIF estimate and the standard errors is significant at a 0.01 level with an absolute DIF value >0.25 logits. DIF analyses were performed on female/male subgroups, students who spoke only English at home versus those who spoke another language at home, and students who used computers for homework versus those who did not.

The results from the DIF analysis show that the instrument provided a satisfactory measure for assessing students' graph understanding. Analyses of the 14 core items in all the three subgroups revealed that only one item had a significant DIF favoring one subgroup over the other, a result that could occur by chance. This suggests that subgroup performance differences mainly reflect graph understanding. DIF analyses on the additional items showed that some items favored one subgroup over the other, but did not reveal a systematic bias. For example, for gender, 5 items (items 21, 30, 39, 47, 48) showed moderate to large DIF in favor of females (8.3 %), while 2 items (9, 27) showed moderate to large DIF in favor of males (3.3 %). Regarding English language speakers, there were 7 items (21, 24, 40, 47, 48, 54, 56) that showed moderate to large DIF in favor of the English speaking group (11.7 %), while 2 items (26, 51) showed moderate to large DIF in favor of the other language group (3.3 %). Regarding computer use for homework, 5 items (21, 22, 24, 41, 55) showed moderate to large DIF in favor of computer users (8.3 %), while 4 items (35, 39, 51, 58) showed moderate to large DIF in favor of non-users (6.7 %). The analysis shows that the items are fairly invariable measures of performance for the gender, language spoken at home, and computer use for homework subgroups, since for the most part the items balance each other out. This group

invariance means that for each item, subgroup performance is consistent with the overall rate of success on the item for the subgroup. This analysis shows that the items are equitable but does not imply that all subgroups perform equally.

Performance by Subgroup

To answer research question two, we conducted *t* tests to determine whether gender, home language, or computer use resulted in performance differences on the test and on individual items. We calculated a total for the core items and only analyzed those who completed the test. We found no differences by gender. Students who speak only English at home ($n = 222$) outperformed those who speak English and another language ($n = 156$) for both multiple choice [M1 = 4.76, out of 9, SD = 1.78, M2 = 3.89, SD = 1.94, $t(376) = 4.51$, $p < .001$] and constructed-response items (CR) [M1 = 14.11, out of 25, SD = 3.34, M2 = 12.72, SD = 3.65, $t(376) = 3.93$, $p < .001$]. Moreover, students with computer-access for homework ($n = 197$) had a significant advantage over those without ($n = 180$) [MC: M1 = 4.72, SD = 1.92, M2 = 4.06, SD = 1.81, $t(375) = 3.425$, $p = .001$; CR: M1 = 14.21, SD = 3.44, M2 = 12.78, SD = 3.51, $t(375) = 4.01$, $p < .001$] (Fig. 5).

We conducted further comparisons (*t* tests) to determine whether subgroups differed by constructed-response item type, including comprehension, critique, and construction. We found no gender differences across the item types. There were significant differences between the language subgroups and the computer-for-homework subgroups for the comprehension and critique items, but not for the construction item (see Table 1).

Access to a computer for homework is probably associated with socioeconomic status and opportunity to learn, especially since the results do not extend to the construction item which is the most technologically demanding of the questions. Since there is no DIF for computer use for homework, these results do not suggest that there is any bias for students who do not use computers for homework. Our results are consistent with those of other assessments (e.g., National Assessment of Educational Progress) showing a relationship between socioeconomic status and achievement (U.S. Department of Education 2000). Socioeconomic status is associated with school quality suggesting that opportunity to learn may be a factor.

Findings for language learners are also consistent with other science assessments showing effects for language learners. For this group as well, there is no DIF. Thus, these results may also be mediated by socioeconomic status and reflect opportunity to learn.

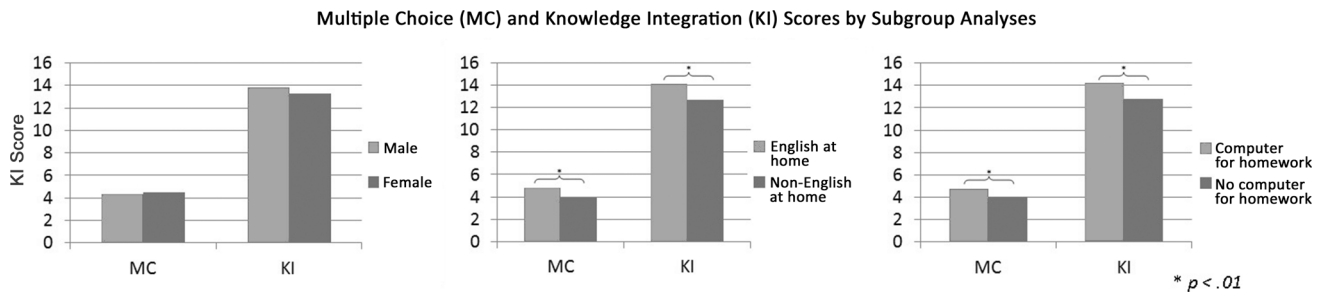


Fig. 5 Students’ performance on multiple-choice items and constructed-response items based on subgroups

Table 1 Comparison of subgroups by item type

	Comprehension (out of 20)				Critique (out of 5)				Construction (out of 5)			
	<i>n</i>	Mean	SSD	<i>t</i>	<i>n</i>	Mean	SSD	<i>t</i>	<i>n</i>	Mean	SSD	<i>t</i>
Male	199	10.41	2.96	−0.95	198	2.86	0.96	−1.22	87	2.11	0.65	−1.35
Female	179	10.68	2.52		178	2.98	0.95		95	2.24	0.64	
English at home	222	10.96	2.58	3.59**	221	3.05	0.93	3.13**	104	2.22	0.67	0.93
Non-english at home	156	9.94	2.90		155	2.74	0.97		78	2.13	0.61	
Computer for HW	197	11.1	2.70	4.20**	196	3.02	0.94	2.13*	101	2.17	0.61	−0.21
No computer for HW	180	9.93	2.71		179	2.81	0.97		81	2.19	0.69	

*** $p < .001$; ** $p < .01$; * $p < .05$

Case Studies of Items

To answer research question three, we clarify understanding of graphing by analyzing performance on illustrative comprehension, critique, and construction items. We explain the aspects of the question that cause confusion, identify consistent patterns of responses across items, and discuss the range of concepts that were tapped in the items.

Comprehension Items

Comprehension items ask students to make sense of a graph to reach conclusions about scientific phenomena. We discuss three items.

The *race* item asks students to interpret a position and time graph representing two runners racing toward a finish line from different starting positions. Students predict the winner of the race and use evidence from the graph to justify their choice. The graph has a nonstandard format, showing the starting position of the two runners at equal and opposite sides of the finish line (at different y-intercepts) and the slopes of their paths. This question measures whether students can read the graph properly and interpret the relationship between the paths of the two runners.

Students had difficulty interpreting the position and time graphs for the two runners. For the runner who started at

zero on the y-intercept, 62 % of the respondents could figure out how many seconds it took to go 10 meters. For the runner who started at a 100 on the y-intercept only 19 % could figure out how many meters were covered in 10 s. Only 13 % of the respondents answered both of these questions correctly. Thus, most of the students were unable to interpret a position and time graph with non-standard starting positions. This is consistent with other research on position and time graphs (e.g., McElhaney and Linn 2011).

When asked who would win the race, 49 % predicted correctly. However, only 7.5 % of the respondents made a link between evidence in the graph and their prediction. Many students misinterpreted the slope of the lines as actual paths to the finish. One respondent said that the winner of the race “is running in a straighter line to the finish than” the loser of the race. Some students focused on the slope of the lines but conflated distance and time saying, the loser is going farther because she is “running at a longer angle.” Others focused on only part of the information, saying for example, that the actual loser “will finish first because she is closer to the finish line” but ignoring the time dimension. These responses show that the students were struggling to interpret the starting points and slopes of the lines. They could draw on ideas about graphs but were applying them inaccurately. Students need instruction to sort out and refine their graphing ideas. It

may be especially difficult to extend graphing ideas from mathematics, such as slope, to a novel problem, such as a race. Instructional studies show that well-designed KI instruction can improve success (Vitale et al. 2015).

The *thermodynamics* item asked students to interpret the temperature changes when a hot bowl is placed on a cool table (Fig. 2). Overall 86 % of the students could identify the starting temperature of the bowl and table, consistent with other research showing that students can read values on graphs (Shah and Hoeffner 2002). When asked to describe the differences in the shape of the graphs for the bowl and table, many students noted that the starting temperatures were different. Those who discussed the shape generally gave superficial answers such as noting that one line was purple and the other was blue. A typical response was to say that the purple line was lower than the blue line. Very few students noted that the graphs had opposite shapes: the counter was warming, while the bowl was cooling. Even fewer students directly compared the curvature of the graphs: the counter warmed more slowly than the bowl cooled. When asked to use evidence from the graph to explain what was happening to the temperature of the bowl and the table, many students correctly noted that the bowl was cooling but few mentioned that heat from the bowl was transferred to the table. Essentially, students were able to describe a common phenomenon: that hot things cool, but did not connect this observation to the specific situation or to the graph. Research shows that KI instruction emphasizing rate of heat flow and thermal equilibrium improves performance on this item (Chang and Linn 2013; Clark et al. 2009).

The *population growth* item asked students to compare the population increase for two types of birds over 10 years. The graphs had different scales on the y-axis so it appeared that the slope for the population that increased more was actually flatter. Most, 46 % of the students, compared the slopes directly without accounting for differences in scale. Students justified their choice by saying that the smaller population growth condition graph had a steeper incline. Inspection of responses revealed that a few students interpreted the question as asking which population grew more relative to its original size over the 10-year time span. This response required a successful interpretation of the relationship between y-axis scale and slope and was therefore awarded a full score. The opportunity for multiple paths to a reasonable response has advantages and also introduces ambiguities since the selection of either population could be justified.

In summary, graph comprehension items revealed that students could locate points on graphs and had some ideas about slope. Students also often focused on superficial features of the graph rather than using either ideas about graphs or ideas about the science context. Some students

interpreted graphs more as pictures, consistent with other research (e.g., Clement 1985). Few students combined their ideas or used multiple sources of information to debug their responses. A few instructional studies show that students' ability to integrate graph and science ideas responds to instruction based on the KI framework. These results contribute evidence for the validity of the items since they can detect the impact of relevant instruction. The results also attest to the importance of strengthening emphasis on graph understanding in science courses.

Critique Items

Graph critique items ask students to analyze a conjecture about one or several graphs. For example the *climate change over time* item asked students which of two depictions of climate change data better explains the phenomena. One graph included data from 1940 to 1960. The other graph included data from 1880 to 2010. 30 % of the students selected the graph with fewer data points and a smaller time frame. Most students explained their choice based on readability saying, "There are less dots on the right one so I thought it was easier to read/use [the graph with fewer points]" or "You can't really see the information on the left graph because it's small and it's all bunched up together. The bigger one would really make you see and understand." These students prioritize readability over valid depiction of the phenomena. Students who selected the graph with data from 1880 to 2010 generally argued that the graph had more information. Only a very few students explained that the graph with more data also communicated a change in slope for average temperature that was not represented in the graph with less data. Thus, for critique as for comprehension many students drew on superficial criteria (such as readability) and failed to incorporate their knowledge of the science (such as climate trends) into their response. Instructional studies that emphasize knowledge integration show that it is possible to improve the ability to critique graphs (Chang and Linn 2013; Svihla and Linn 2012; Vitale et al. 2015).

Graph Construction Items

Graph construction items require students to draw a graph that captures scientific information. To implement this type of item, we took advantage of the WISE graphing tool that provided students with a predefined set of axes and an intuitive point-plotting interface as well as a method for designing automated scoring of the resulting graph (Vitale et al. 2015). Since the axes were predefined for the students, the automatic scoring only scored the actual lines the students constructed, and not the other basic graph elements (e.g., axes, axes titles).

The *training for a race* item asked students to construct a graph based on a narrative that described a training routine. The narrative specified the training sequence with many constraints but left some details to the designer of the graph (such as the length of a short rest). We were able to develop an automated scoring rubric that was based on detecting specific features of the student responses.

Consistent with responses to comprehension and critique items, 65 % of the students focused on superficial features of the narrative and drew a picture of the routine rather than a graph. In addition, 25 % of students graphed incorrect functions, 9 % provided off-task responses (e.g., drawing a star), and only 1 % gave fully correct answers. The tendency for students to draw pictures, consistent with other research (Clement 1985), often led to the production of non-functional graphs (e.g., a rectangle), representing physically impossible situations (such as a reversal in time). Students appeared to implement an intuitive idea about the narrative, rather than focusing on elements of the graph, such as the meaning of the axes. Instructional studies demonstrate that students can learn to integrate their intuitions about the narrative and their knowledge of graphs to represent a complex sequence of actions and incorporate a series of constraints (Vitale et al. 2015).

Taken together these case studies provide an answer to research question three. They illustrate the confusions and difficulties students face when responding to items in the Graphing Inventory. Most students give superficial answers and many students interpret graphs as pictures rather than as capturing relationships. A few instructional studies support the claim that the graphing items are valid by detected the impact of instruction. These cases demonstrate the need for more comprehensive instruction that emphasizes graphing of scientific phenomena.

Conclusions

We developed a Graph Inventory that included items measuring ability to comprehend, critique, and construct graphs involving scientific concepts. In answer to our first research question, we were able to create items and rubrics based on KI that resulted in an inventory with good psychometric properties. In answer to our second question, we found no DIF for subgroups, demonstrating that the items were unbiased for gender, access to computers for homework, and home language. We did find subgroup performance differences for students who lack access to computers for homework and for those who speak a language other than English at home. We hypothesize that these results reflect differences in opportunity to learn.

In response to our third research question, we found that students have great difficulty with graphs in science. These

results were consistent with the expectations of the teachers and with other research showing that most people have difficulties in understanding graphs (OECD 2006). These findings signal the need for additional research to improve instruction incorporating graph understanding into science.

Inventory Properties

Results show that items measuring graph understanding in science and scored using a knowledge integration rubric have good internal consistency (Cronbach's alpha) and form an acceptable IRT scale. The items have satisfactory item-total correlations, and the Wright map shows good alignment of item scores.

The items have face validity as established by expert review. Instructional studies show that the items tested can detect the impact of curriculum materials designed to improve graph understanding (Chang and Linn 2013; Svihla and Linn 2012; Vitale et al. 2015). These studies demonstrate instructional validity for the items tested and suggest that the remaining items falling on the same IRT scale are likely to have instructional validity as well.

Overall students struggle to interpret graphs embedded in science contexts. In many cases students pay attention to superficial characteristics of graphs, consistent with previous studies (Shah and Hoeffner 2002). Students draw on ideas from mathematics (slope, y-intercept) and intuitions about the situation (often reflected in graph-as-picture responses). They especially have difficulty integrating multiple features of graphs, such as axis labels; slope; cyclical patterns; and noisy data. Additional research could help clarify the difficulties students have with graphs and test promising remedies.

Subgroup Performance

Consistent with other research, we found no bias or performance differences by gender (Aberg-Bengtsson and Ottosson 2006). We found no bias for computer use for homework or for home language. As expected, students who use computers for homework outperformed non-users, a finding we hypothesize reflects opportunity to learn. Future research might clarify this finding by documenting opportunity to learn and looking for specific ways that student computer use for homework might improve graph understanding. This is important because most standardized tests are now being administered by computer.

Consistent with other research, we found that English-only speakers outperformed students who spoke another language at home on both the core multiple-choice and constructed-response items (Aberg-Bengtsson and Ottosson 2006). Since DIF analyses for the core items showed no significant bias toward one group, the finding reflect

factors other than item design. We hypothesize that these findings reflect opportunity to learn. Further research is needed to understand the role of home language in graph understanding in science.

Interestingly, for the construction item, there were no subgroup differences. It is worth exploring whether the reduced language demands neutralize the advantage for English-only speakers. Our interactive, online system makes it feasible for researchers to use more of these items. Since language and graphical knowledge are important in many science practices such as making sense of data, critique, and argumentation, the relationship among these skills warrants further research.

Limitations

The implications of this study are limited by the selection of graphing items, the size and representativeness of the sample, and the use of knowledge integration scoring rubrics. Because we did not include a representative samples at each grade level we could not report on trends across grades. There are other aspects of graph understanding in science that should be included in future assessments. Our items revealed some connections between graphing and science but the limitations of student understanding of graphs limited the evidence for this phenomenon.

In addition, there are other relationships that could be explored in future studies. In particular, to align our inventory with the type of relationships found in common natural processes and the scientific practices used to measure and convey these processes, we recommend increased focus on curvilinear graphs and noisy data.

Many common natural phenomena exhibit nonlinear changes over time (Linn and Hsi 2000). For example, population growth may be modeled with an accelerating exponential curve, cooling or heating with a decelerating exponential (logarithmic) curve, and projectile motion with a quadratic curve. Recognizing the visuospatial differences between these trends, as well as understanding their underlying scientific significance represents an important challenge for students. For example, the decelerating trend of the cooling curve represents a system coming to equilibrium—a pattern that repeats in many other circumstances. While our inventory does incorporate several curvilinear graphs, for the most part they do not focus on comparing specific types of non-linear relationships. For example, the “Hot Bowl on a Table” item (Fig. 2) displays two curves reaching an equal temperature from different directions. Although students were often capable of recognizing how the direction of change indicated cooling or heating, they rarely described how the curvature represented a changing rate of cooling (or heating), or that the

curvature differed for the two graphs. To some extent this may represent vagueness in the prompt, rather than a specific lack of knowledge; however, by asking students to compare or construct different curves and link them to underlying principles, we may target these ideas explicitly.

Additionally, reflecting science and engineering practice standards emphasizing students’ ability to use and understand data to communicate ideas (NGSS Lead States 2013), more emphasis on critiquing potentially misleading or noisy graphs is appropriate. The climate change item demonstrates a successful format for engaging students in critical thinking about graphs by asking them to consider a faulty argument based on a challenging graph. This approach could be applied to additional graph features such as outliers, non-standard scales, and measurement error. Furthermore, asking students to construct graphs with noisy data and draw curves of best fit represents an important real world practice and an insight into students understanding of noise and curve types.

In parallel with the results described here, we expect that new items focusing on curvilinear relationships and noisy data would underscore the limited understanding of students. Because the students completed these items within the first 2 months of the school year, prior to graphing instruction, our results are unlikely to represent students’ full potential. Additional studies demonstrating the measures of graph understanding in science following instruction are needed to demonstrate the validity of the items. In addition, instructional studies are essential to understand how best to improve graph understanding in science. In future work we plan to investigate the changes that occur across the school year as students engage in graph-related learning activities as well as to test specific approaches to graph instruction.

The results seem to indicate that students have particular difficulty with the relationships between two graphs in the same grid, as illustrated in Figs. 1 and 2. This seems reasonable, since understanding the differences between graphs presupposes the ability to understand individual graphs. Additional research is needed to explore the cognitive issues that are raised by two and three graphs on the same grid.

A further limitation of the study is that we did not test the items in other contexts (e.g., international contexts). We hope that others will be motivated to do so, and additional research would provide insight into the generalizability of the results.

Instructional Implications

The difficulties students faced in responding to Graph Inventory items highlight the need for instruction that emphasizes graph understanding in science. To improve

graph understanding, students need to use graphs as a communication tool, not just view graphs in a passive manner (Fry 1981). Kramarski (2004) support this view, noting that students improve their graphical knowledge through communication and self-explanation activities, such as explaining their reasoning, challenging their peers' views, and clarifying their own thinking.

We anticipate that effective instruction and teacher support can help mitigate subgroup differences due to limited English language proficiency (Davis and Krajcik 2005; Lee et al. 2009). In many cases poor performance by students with difficulty reading English can mask critical deficiencies in content knowledge, while in other cases these students can perform as well as their peers if given appropriate guidance and response formats. It is therefore important that assessments help teachers diagnose the primary source of difficulty. In particular, adapting materials for English learners is essential (Lee 2005; Lee et al. 2009).

This study revealed that middle school students could interpret straightforward but not complex graphs consistent with other research (Edelson et al. 1999; Hadjidemetriou and Williams 2002). Students were challenged by critique items, consistent with other research (Chang and Linn 2013; OECD 2006). Students lacked the ability to construct accurate graphs from narrative accounts of scientific phenomena. Students rarely used their science knowledge to interpret graphs (of predator–prey relationships or the day night cycle) and demonstrated limited understanding of graph ideas (such as noise, curve shapes) in their interpretations of science graphs of global climate change or of trends in growth curves. These results show that experience with critique and construction of graphs would strengthen science instruction. Such instruction is also essential for everyday interpretation of persuasive messages. Communication of scientific knowledge increasingly depends on the ability to construct graphs to capture insights.

Although the Graphing Inventory will require further study in a wider range of contexts, these results demonstrate that students' difficulty applying graph knowledge to science concepts (and vice versa) is an important topic for instructional research and development. Graphs, such as those depicting rising global temperatures, can reveal and underscore natural phenomena that may impact the lives of students. Facilitating students' graph literacy is therefore a necessity and requires frequent and consistent attention throughout the science curriculum.

References

- Aberg-Bengtsson L, Ottosson T (2006) What lies behind graphicacy? Relating students' results on a test of graphically represented quantitative information to formal academic achievement. *J Res Sci Teach* 43:43–62
- Adams DD, Shrum JW (1990) The effects of microcomputer-based laboratory exercises on the acquisition of line graph construction and interpretation skills by high school biology students. *J Res Sci Teach* 27:777–787. doi:10.1002/tea.3660270807
- American Association for the Advancement of Science (1993) Benchmarks for science literacy. University Press, Oxford
- Barclay WL (1985) Graphing misconceptions and possible remedies using microcomputer-based labs (Technical report number TERC-TR-85-5). Technical Education Research Center, Cambridge
- Beichner RJ (1994) Testing student interpretation of kinematics graphs. *Am J Phys* 62:750–762
- Berg CA, Smith P (1994) Assessing students' abilities to construct and interpret line graphs: disparities between multiple-choice and free-response instruments. *Sci Educ* 78:527–554
- Bertin J (1983). In: Berg W (ed), *Semiology of graphics: diagrams networks maps* (trans: Berg W). The University of Wisconsin Press, Madison
- Bock RD, Aitkin M (1981) Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika* 46:443–459
- Bond TG, Fox CM (2007) *Applying the Rasch model: Fundamental measurement in the human sciences*, 2nd edn. Lawrence Erlbaum Associates Inc, Mahwah
- Boote SK (2014) Assessing and understanding line graph interpretations using a scoring rubric of organized cited factors. *J Sci Teach Educ* 25:333–354
- Brasell H (1987) The effect of real-time laboratory graphing on learning graphic representations of distance and velocity. *J Res Sci Teach* 24:385–395
- Chang HY, Linn MC (2013) Scaffolding learning from molecular visualizations. *J Res Sci Teach* 50(7):858–886
- Clark DB, D'Angelo CM, Menekse M (2009) Initial structuring of online discussions to improve learning and argumentation: incorporating students' own explanations as seed comments versus an augmented-preset approach to seeding discussions. *J Sci Educ Technol* 18:321–333
- Clement J (1985) Misconceptions in graphing. In: *Proceedings of the ninth international group for the psychology of mathematics education*. Noordwijkerhout
- Cobb P (1999) Individual and collective mathematical development: the case of statistical data analysis. *Math Think Learn* 1:5–44
- Curcio FR (1987) Comprehension of mathematical relationships expressed in graphs. *J Res Math Educ* 18:382–393. doi:10.2307/749086
- Davis JD (2007) Real-world contexts, multiple representations, student-invented terminology, and y-intercept. *Math Thin Learn* 9:387–418
- Davis EA, Krajcik JS (2005) Designing educative curriculum materials to promote teacher learning. *Educ Res* 34:3–14
- DiSessa A, Sherin BL (2000) Meta-representation: an introduction. *J Math Behav* 19:385–398
- DiSessa AA, Hammer D, Sherin B, Kolpakowski T (1991) Inventing graphing: metarepresentational expertise in children. *J Math Behav* 10:117–160
- Edelson DC, Gordin DN, Pea RD (1999) Addressing the challenges of inquiry-based learning through technology and curriculum design. *J Learn Sci* 8:391–450
- Friel SN, Curcio FR, Bright GW (2001) Making sense of graphs: critical factors influencing comprehension and instructional implications. *J Res Math Educ* 32(2):124–158. doi:10.2307/749671
- Fry E (1981) Graphical literacy. *J Read* 25:383–390

- Gal I (2002) Adults' statistical literacy: meaning, components, responsibilities. *Int Stat Rev* 70:1–25
- Galesic M, Garcia-Retamero R (2011) Graph literacy: a cross-cultural comparison. *Med Decis Making* 31:444–457
- Gallimore M (1991) Graphicacy in the primary curriculum. In: Vere-Jones D (ed) *Proceedings of the third international conference on teaching statistics*. Otago University Press, New Zealand
- Gillen CM (2006) Criticism and interpretation: teaching the persuasive aspects of research articles. *CBE Life Sci Educ* 5:34–38. doi:10.1187/cbe.05-08-0101
- Glazer N (2011) Challenges with graph interpretation: a review of the literature. *Stud Sci Educ* 47(2):183–210. doi:10.1080/03057267.2011.605307
- Hadjidemetriou C, Williams J (2002) Children's graphical conceptions. *Res Math Educ* 4:69–87
- Hattikudur S, Prather RW, Asquith P, Alibali MW, Knuth EJ, Nathan MJ (2012) Constructing graphical representations: middle schoolers' intuitions and developing knowledge about slope and y-intercept. *School Sci Math* 112(4):230–240
- Janvier C (1981) Use of situations in mathematics education. *Educ Stud Math* 12:113–122
- Jarman R, McClune B, Pyle E, Braband G (2012) The Critical reading of the images associated with science-related news reports: establishing a knowledge, skills, and attitudes framework. *Int J Sci Educ Part B*. doi:10.1080/21548455.2011.559961
- Kali Y (2006) Collaborative knowledge-building using the design principles database. *Int J Comput Support Collab Learn* 1:187–201
- Kosslyn SM (1989) Understanding charts and graphs. *Appl Cognit Psychol* 3(3):185–225
- Kramarski B (2004) Making sense of graphs: does metacognitive instruction make a difference on students' mathematical conceptions and alternative conceptions? *Learn Instr* 14:593–619
- Latour B (1990) Drawing things together. In: Lynch M, Woolgar S (eds) *Representation in scientific practice*. MIT Press, Cambridge, pp 19–68
- Lee O (2005) Science education and English language learners: synthesis and research agenda. *Rev Educ Res* 75:491–530
- Lee O, Penfield R, Maerten-Rivera J (2009) Effects of fidelity of implementation on science achievement gains among English language learners. *J Res Sci Teach* 46:836–859
- Lee HS, Linn MC, Varna K, Liu OL (2010) How do technology-enhanced inquiry science units impact classroom learning? *J Res Sci Teach* 47(1):71–90
- Leinhardt G, Zaslavsky O, Stein MM (1990) Functions, graphs, and graphing: tasks, learning and teaching. *Rev Edu Res* 60:1–64
- Linn MC, Eylon BS (2011) *Science learning and instruction: taking advantage of technology to promote knowledge integration*. Routledge, New York
- Linn MC, Hsi S (2000) *Computers, teachers, peers: science learning partners*. Lawrence Earlbaum Associates, Mahwah, NJ
- Linn MC, Layman JW, Nachmias R (1987) Cognitive consequences of microcomputer-based laboratories: graphing skills development. *Contemp Educ Psychol* 12:244–253
- Linn MC, Clark D, Slotta JD (2003) WISE design for knowledge integration. *Sci Educ* 87(4):517–538
- Linn MC, Lee HS, Tinker R, Husic F, Chiu J (2006) Teaching and assessing knowledge integration in science. *Science* 313:1049–1050
- Liu OL, Lee HS, Linn MC (2011) Measuring knowledge integration: validation of four-year assessments. *J Res Sci Teach* 48(9):1079–1107
- Lovett MC, Chang NM (2007) Data analysis skills: What and how are students learning? In: Lovett MC, Shah P (eds) *Thinking with data*. 33rd carnegie symposium on cognition. Erlbaum, Mahwah, pp 293–318
- McDermott L, Rosenquist M, Popp B, van Zee E (1983) Student difficulties in connecting graphs, concepts and physical phenomena. In: Paper presented at the annual meeting of the American educational research association. Montreal
- McElhaney KW, Linn MC (2011) Investigations of a complex, realistic task: intentional, unsystematic, and exhaustive experimenters. *J Res Sci Teach* 48(7):745–770. doi:10.1002/tea.20423
- McKenzie DL, Padilla MJ (1986) The construction and validation of the test of graphing in science (TOGS). *J Res Sci Teach* 23:571–579
- Mevarech ZA, Kramarsky B (1997) From verbal descriptions to graphic representations: stability and change in students' alternative conceptions. *Educ Stud Math* 32:229–263
- Mokros J, Tinker R (1987) The impact of microcomputer-based labs on children's ability to interpret graphs. *J Res Sci Teach* 24(4):369–383
- National Research Council (1996) *National science education standards*. National Academy Press, Washington
- NGSS Lead States (2013) *Next generation science standards*. National Academies Press, Washington, DC. <http://www.nextgenscience.org/next-generation-science-standards>
- Organization for Economic Cooperation and Development (2006) *PISA 2006 science competencies for tomorrow's world*. Accessed 19 Nov, 2013, from <http://www.oecd.org/fr/education/scolaire/programmeinternationalpourlesuividesacquisdeselevespisa/pisa2006results.htm>
- Ozcelik AT, McDonald S (2013) Preservice science teachers' uses of inscriptions in science teaching. *J Sci Teach Educ* 24:1103–1132. doi:10.1007/s10972-013-9352-1
- Preece J, Janvier C (1992) A study of the interpretation of trends in multiple curve graphs of ecological situations. *School Sci Math* 92(6):299–306
- Quintana C, Reiser BJ, Davis EA, Krajcik J, Fretz E, Duncan RG et al (2004) A scaffolding design framework for software to support science inquiry. *J Learn Sci* 13:337–386
- Rasch G (1966) An item analysis which takes individual differences into account. *Br J Math Stat Psychol* 19:49
- Roth W-M, Bowen GM (2001) Professionals read graphs: a semiotic analysis. *J Res Math Educ* 32:159–194
- Schlieman A, Carraher D, Ceci SJ (1997) Everyday cognition. In: Berry JW, Dasen PR, Saraswathi TS (eds) *Handbook of cross-cultural psychology: basic processes and human development*, 2nd edn. Allyn & Bacon, Needham Heights, pp 177–215
- Shah P, Hoeffner J (2002) Review of graph comprehension research: implications for instruction. *Educ Psychol Rev* 14:47–69. doi:10.1023/A:1013180410169
- Slotta JD, Linn MC (2009) *WISE science: inquiry and the internet in the science classroom*. Teachers College Press
- Svihla V, Linn MC (2012) A design-based approach to fostering understanding of global climate change. *Int J Sci Educ* 34(5):651–676
- U.S. Department of Education, National Center for Education Statistics (2000) *The condition of education 2000*, NCES 2000-602. Government Printing Office, Washington, DC
- Vitale JM, Lai K, Linn MC (2014) Dynamic visualization of motion for student-generated graphs. In: *Proceedings of the international conference of the learning sciences (ICLS)*. Boulder, CO
- Vitale JM, Lai K, Linn MC (2015) Taking advantage of automated assessment of student-constructed graphs in science. *J Res Sci Teach* 52(10):1426–1450
- Wainer H (1992) Understanding graphs and tables. *Educ Res* 21:14–23
- Wang ZH, Wei S, Ding W, Chen X, Wang X, Hu K (2012) Students' cognitive reasoning of graphs: characteristics and progression. *Int J Sci Educ* 34:2015–2041. doi:10.1080/09500693.2012.709333

- Watson JM (2008) Exploring beginning inference with novice grade 7 students. *Stat Educ Res J* 7:59–82
- Wilensky U, Reisman K (2006) Thinking like a wolf, a sheep, or a firefly: learning biology through constructing and testing computational theories—an embodied modeling approach. *Cognit Instr* 24:171–209
- Wiley J, Goldman SR, Graesser AC, Sanchez CA, Ash IK, Hemmerich JA (2009) Source evaluation, comprehension, and learning in Internet science inquiry tasks. *Am Educ Res J* 27:255–265
- Wu HK, Krajcik JS (2006) Inscriptional practices in two inquiry based classrooms: a case study of seventh graders' use of data tables and graphs. *J Res Sci Teach* 43:63–95
- Wu ML, Adams RJ, Wilson MR (1998) ACER ConQuest: generalised item response modelling software. ACER press, New York
- Yeh Y-FY, McTigue EM (2009) The frequency, variation, and function of graphical representations within standardized state science tests. *School Sci Math* 109:435–449. doi:[10.1111/j.1949-8594.2009.tb18291.x](https://doi.org/10.1111/j.1949-8594.2009.tb18291.x)