International Journal of Science Education

ISSN: 0950-0693 (Print) 1464-5289 (Online) Journal homepage: http://www.tandfonline.com/loi/tsed20

Measuring Primary Students' Graph Interpretation Skills Via a Performance Assessment: A case study in instrument development

Karen Peterman, Kayla A. Cranston, Marie Pryor & Ruth Kermish-Allen

To cite this article: Karen Peterman, Kayla A. Cranston, Marie Pryor & Ruth Kermish-Allen (2015) Measuring Primary Students' Graph Interpretation Skills Via a Performance Assessment: A case study in instrument development, International Journal of Science Education, 37:17, 2787-2808, DOI: 10.1080/09500693.2015.1105399

To link to this article: http://dx.doi.org/10.1080/09500693.2015.1105399

International Journal of Science Education, 2015
Vol. 37, No. 17, 2787–2808, http://dx.doi.org/10.1080/09500693.2015.1105399

Measuring Primary Students' Graph Interpretation Skills Via a Performance Assessment: A case study in instrument development

Karen Peterman[a*], Kayla A. Cranston[b], Marie Pryor[c] and Ruth Kermish-Allen[d]

[a]*Karen Peterman Consulting Co., 2706 Stuart Drive, Durham, NC 27707, USA;* [b]*Environmental Studies Department, Antioch University New England, 40 Avon Street, Keene, NH 03431, USA;* [c]*College of Health and Public Affairs, University of Central Florida, HPA II Suite 238, 4364 Scorpius Street, Orlando, FL 32816, USA;* [d]*Maine Math and Science Alliance, 219 Capitol St #3, Augusta, ME 04330, USA*

This case study was conducted within the context of a place-based education project that was implemented with primary school students in the USA. The authors and participating teachers created a performance assessment of standards-aligned tasks to examine 6–10-year-old students' graph interpretation skills as part of an exploratory research project. Fifty-five students participated in a performance assessment interview at the beginning and end of a place-based investigation. Two forms of the assessment were created and counterbalanced within class at pre and post. *In situ* scoring was conducted such that responses were scored as correct versus incorrect during the assessment's administration. Criterion validity analysis demonstrated an age-level progression in student scores. Tests of discriminant validity showed that the instrument detected variability in interpretation skills across each of three graph types (line, bar, dot plot). Convergent validity was established by correlating *in situ* scores with those from the Graph Interpretation Scoring Rubric. Students' proficiency with interpreting different types of graphs matched expectations based on age and the standards-based progression of graphs across primary school grades. The assessment tasks were also effective at detecting pre–post gains in students' interpretation of line graphs and dot plots after the place-based project. The results of the case study are discussed in relation to the common challenges associated with performance assessment. Implications are presented in relation to the need for authentic and performance-based instructional and assessment tasks to respond to the Common Core State Standards and the Next Generation Science Standards.

Keywords: *Standards; Performance assessment; Graphs; Place-based education*

*Corresponding author. Karen Peterman Consulting Co., 2706 Stuart Drive, Durham, NC 27707, USA. Email: karenpetermanphd@gmail.com

Introduction

Graph interpretation has traditionally been deemed a math rather than science skill within the context of educational standards in the USA. However, it is also a foundation of effective science communication and thus a skill that spans the full range of science disciplines. As such, graphical interpretation is a skill that generalizes across the traditional boundaries imposed by specific content areas, with the potential to serve as a unifying science practice.

The interpretation of graphs is a challenge for students of all levels (Glazer, 2011). With the introduction of the Next Generation Science Standards (NGSS; National Research Council [NRC], 2014), the opportunity exists to leverage this skill more directly within the context of science classrooms. Bybee (2011, 2012) noted that the NGSS science and engineering practices should be thought of as both instructional strategies and learning outcomes and that the NGSS performance expectations combine practice and content to assess student learning. Whether as part of instructional or assessment practice, standards-aligned graph interpretation tasks hold promise as a strategy for integrating science practice into the classroom.

Furthermore, the focus on performance expectations necessitates assessment methods that extend beyond the constraints of traditional multiple-choice tests. Performance assessments that challenge students to demonstrate their knowledge and skills, and authentic assessments where the tasks mirror real-life problem-solving situations (Rural School & Community Trust, 2001; Wilson & Sloane, 2000), are ideal methods for collecting student proficiency data pertinent to the NGSS.

In relation to the development of graph interpretation skills, it could be argued that place-based education (PBE) offers one of the most effective pedagogies for engaging students with data. PBE has been defined as a holistic approach to education, conservation, and community development that uses project-focused learning and school–community partnerships to boost student achievement (Johnson, Duffin, & Murphy, 2012). Many PBE projects include a data-based investigation of a locally relevant topic that is conducted by students in partnership with community members. PBE often involves data collection, as well as data analysis and the communication of results. Graph interpretation is a key skill in both digesting the information being collected via place-based projects and in communicating the results to others. As such, PBE offers an ideal context from which to study students' graph interpretation skills within the science classroom.

The current case study was conducted as part of an exploratory research grant. It included a partnership with teachers to develop a performance-based instrument of standards-aligned tasks that could be used to document students' graph interpretation skills within the context of a PBE project. The instrument is called the Graphical Interpretation Assessment (GIA). Results from this exploratory study were used to answer two research questions:

> RQ1: What preliminary validation evidence exists to support the use of the GIA?
> RQ2: To what extent does the GIA document change in primary students' abilities to read and interpret graphs within the context of an educational intervention?

Performance Assessment

Progressive educators and evaluators have advocated for alternative assessment-techniques that allow students to demonstrate their ability to use knowledge (Baron, 1991; Becker-Klein, Stylinski, Peterman, & Phillips, 2014; Gardner, 1992; Malcom, 1991; Peterman, 2013; Wiggins, 1993). One example, performance assessments, 'require students to demonstrate their achievements by producing authentic responses to evaluation tasks, such as written or spoken answers … portfolios of work products, or group solutions to defined problems' (Stufflebeam, 2001, p. 25). The benefits of performance assessments seem intuitive. They 'offer the potential for greater ecological validity and relevance [as well as] the assessment of a wider range of skills and knowledge' (Wilson & Sloane, 2000, p. 183) when compared to traditional assessment measures such as multiple choice tests. Performance assessments also demonstrate that students can apply their skills in real-life situations (Stufflebeam, 2001).

In these ways, authentic performance assessments offer an innovative way of capturing student skills that complements traditional methods. Perhaps most importantly for the current educational climate in the USA, they offer a direct response to the NRC's consensus report on how to best assess the performance expectations articulated in the NGSS (NRC, 2014). There, the authors emphasize the importance of creating assessments that can capture three-dimensional science learning to adequately measure students' mastery of performance expectations. The current study explores the use of a series of standards-aligned tasks as a method for the assessment of learning outcomes related to NGSS performance expectations.

Performance assessments are a particularly good fit for experiential learning environments. For example, proficiency skill measures have been cited as a valuable strategy to address the challenges associated with assessing lab-based, problem-based, and hands-on learning experiences (Savin-Baden, 2004; Steck, DiBiase, Wang, & Boukhtiarov, 2012). Berg and Smith (1994) demonstrated that performance tasks are crucial for understanding student proficiencies with graph interpretation in particular.

Despite these benefits, there are a number of challenges associated with performance assessments. Stufflebeam (2001) noted that performance assessments require a significant amount of time to develop and implement, and that they are costly to score. Challenges have also been noted in relation to the reliability and validity of performance assessments (Gipps, 1995; Stufflebeam, 2001). Teachers typically do not have the time or training to design performance assessments. It is imperative that educators and researchers who have expertise in developing assessments assist in the creation of tools that both meet the needs of today's classroom and respond to the NRC's assessment recommendations.

This case study explores the use of standards-aligned performance tasks as a strategy to document primary students' interpretation skills in relation to three types of graphs. The tasks were developed as part of an exploratory research grant and were intended to serve as a performance assessment of a PBE project. Johnson,

Penny, and Gordon (2009) state that performance assessments are a system composed of (1) the purpose of assessment, (2) tasks or prompts that elicit the performance, (3) a response demand that focuses the examinee's performance, and (4) systematic methods for rating performances. As applied to this case study, our performance assessment was developed to (1) evaluate a PBE project by (2) using a series of graphs and question prompts to elicit student performance during (3) individual interviews that featured graphical displays. Finally, (4) percent correct scores were generated to document performance.

Theories of and Research about Graphical Interpretation

Two theories underlie our understanding of how students comprehend graphs: Bertin (1983, 2001) and Curcio (Curcio, 1987; Friel, Curcio, & Bright, 2001). Each has been used to guide the study of graph interpretation in the literature, and each served as the foundation for the development of the GIA. According to Bertin, reading a graph happens in three rapid steps. The first is External Identification, in which the reader perceives the external factors of the graphic only (such as the title, axes labels, and the scale of each axis). The next step is Internal Identification, in which the reader perceives the internal factors of the graphic; these details might be the bars, lines or dots that represent the data. In the third step, the Perception of Pertinent Correspondences, the reader combines the details identified via stages one and two to understand the information being displayed through the interaction of the external and internal features.

Once a graphic has been read, Bertin states, it can then be used to answer a number of questions. Elementary-level questions ask the reader to identify a single piece of information from the graph. Intermediate questions require readers to notice patterns among items within the graph to identify trends within the data. The third level, referred to as Overall-level questions, asks readers to draw conclusions about the information presented in the graphic as a whole.

Curcio (1987) asserts that graphs are a type of text and that their interpretation is affected by factors such as a students' prior knowledge about a topic, the mathematical content involved, and the graphical form itself. Curcio's three levels of graph interpretation are: reading the data (answering explicit questions with information that can be found directly in the graph itself), reading between the data (finding relationships in the data presented in a graph), and reading beyond the data (making predictions or inferences based on graph content).

Studies of 5–10-year-olds have focused largely on how young students work with data overall, including how they describe data in graphical form (Konold & Higgins, 2003). In a series of case studies on these topics, Russell, Schifter, Bastable, Konold, and Higgins (2002) found that five-year-olds require a personal connection to data in order to begin to understand it, and that some students can derive information from graphed data even at this young age. Young children typically extract individual values from graphs rather than an overall understanding of the display, and this tendency continues throughout the early school years (Konold & Higgins,

2003). Konold, Higgins, Russell, and Khalil (2014), for example, found that young students focused on either the value for each individual case in a graph or that they focused on classifiers such as the highest or lowest frequency in a graphical display. Friel et al. (2001) suggested that seven-year-olds can begin to understand simple bar and line graphs as abstract representations of data. The use of scale and an understanding of axes begin in the later primary grades (Friel et al., 2001). Even so, 9–10-year-olds still have notable deficits in attending to data labels on axes and interpreting scales and reference markings (Parmer & Singer, 2005). The use of more complicated bar graphs is prevalent in the later primary grades, and understanding of line graphs continues to improve beyond the primary grades as well (Friel et al., 2001; Russell et al., 2002). Konold et al. (2014) concluded that sustained teaching related to data and graph interpretation are still recent additions in education and that there is much to learn about how students in this age range approach and understand data.

The NGSS categorizes a number of graph interpretation skills as Common Core State Standards (CCSS) that are expected to connect across all ages, thus opening the door for additional study of graph interpretation skills within the context of both math and science classrooms. In the early grades, the CCSS in math include a progression from bar graphs, to line graphs, and eventually to dot plots (NRC, 2012). Studies of graph interpretation abilities among young students often use performance assessments to document skills (Curcio, 1987; Glazer, 2011; Konold et al., 2014; Norman, 2012; Parmar & Signer, 2005). Unfortunately, none of these studies offer a practical guide for educators who need to measure these skills for formative or summative purposes.

We have found one example in the literature that studies an assessment method used to document students' graph interpretation skills directly. Boote (2012) collected think-aloud data from 11-year-olds who completed the Test of Graphing in Science (TOGS; McKenzie & Padilla, 1986; Padilla, McKenzie, & Shaw, 1986). This was used to study the Graph Interpretation Scoring Rubric, an instrument created by combining cited factors from Bertin and Curcio to code four types of responses: External Identification, Elementary, Intermediate, and Overall. The results from the study were interpreted in relation to the utility of the rubric itself. Boote concluded that the Graph Interpretation Scoring Rubric provides a common language for the assessment of graph interpretation skills and that its expanded use across age level and graph type warranted additional study.

Together, the education standards and Boote's conclusions illustrate the different approaches and needs of practitioners and researchers. Boote states the need for a common language to describe student skill development and documents the range of existing instruments used by academic researchers to study such skills. By contrast, the CCSS and NGSS provide a common language for the developmental trajectory of students' graph interpretation skills, and yet practical formative and summative instruments do not yet exist to measure those skills. This article is a first step in trying to bridge this gap between research on students' graph interpretation skills and classroom assessment practices.

Method

Setting

This study was conducted within the context of *WeatherBlur*, a two-year project funded by the National Science Foundation. *WeatherBlur* consisted of an online learning community that brought together elementary school teachers and students, fishermen, and scientists to investigate the local impacts of climate change. The project was place-based, in that students worked with community members to investigate topics related to local weather and climate.

Primary data were collected by the students and local fishermen, and via weather stations mounted at school sites. Students used the *WeatherBlur* platform to analyze and graph their results in relation to regional data from the National Oceanic and Atmospheric Administration's 10-, 30-, and 60-year climatologies for precipitation and air temperature. As a final project, students created posters that included graphs and tables to show their results.

Through these experiences, *WeatherBlur* provided a range of opportunities for students to use the graph interpretation skills featured in the math CCSS. For example, students created scaled bar graphs to represent a data set with several categories (CCSS.MATH.CONTENT.3.MD.B.3) and used line plots to display a set of measurements (CCSS.MATH.CONTENT.4.MD.B.4). Students also graphed points on the coordinate plane to solve real-world and mathematical problems (CCSS.MATH.CONTENT.5.G.A.1). Graphs were also used to reinforce NGSS Earth Systems standards. Students observed, recorded, and shared representations of local weather conditions to describe changes over time and identify patterns (K-ESS2-a); organized simple weather data sets to identify both day-to-day variations and long-term patterns of weather (NGSS.3-ESS2-a); and displayed simple data sets in tables and graphs to describe typical weather conditions expected during a particular season and variations over years (NGSS.3-ESS2-b).

Participants

The students in this study were from five island-based schools in Maine, USA. Seventy-one students participated in the *WeatherBlur* project in autumn, 2013. Matching pre–post data were collected from 55 students (77%). See Table 1 for a demographic description of participants.

Instruments

The GIA was created jointly by two groups of stakeholders: educators from the Island Institute and *WeatherBlur* teachers. The goal was to create an authentic assessment of standards-aligned tasks that could be used to document change in students' graph interpretation skills as the result of the project. The *WeatherBlur* platform and classroom activities were designed to mirror the kinds of graphical displays featured in the CCSS and thus included bar graphs, line graphs, and dot plots.

Table 1. Demographic description of participants ($N = 55$)

	n	%
Gender		
Male	27	49
Female	27	49
Missing	1	2
Age		
6-year-olds	9	16
7-year-olds	3	5
8-year-olds	16	29
9-year-olds	15	27
10-year-olds	12	22
Ethnicity		
Asian	2	4
White	52	94
Missing	1	2

As a first step, the stakeholders agreed that the GIA would include each of these graphical displays. Next, the CCSS and NGSS were reviewed to identify the performance expectations related to 6–10-year-olds' graphing and interpretation skills (see previous section). Finally, teachers described the kinds of skills they hoped their students would gain from the project. Specifically, teachers hoped that *WeatherBlur* projects based on real-world scenarios would help students (1) draw and interpret line and bar graphs, and (2) understand the difference between weather and climate.

At the conclusion of these steps, a draft assessment was shared with the *WeatherBlur* stakeholders who then provided feedback. Revisions were made to create the final version of the assessment.

The GIA was developed in two versions for use in a pre–post design (see Appendix for Version 1). Each form included a bar graph, line graph, and dot plot. Boote (2012) recommended that future studies investigate students' discussion of graphs through open-ended rather than multiple-choice questions; the GIA was a hybrid of these approaches. Discussion of each graph began with an open-ended question that challenged students to describe the information conveyed. A series of specific prompts then followed to document students' understanding of the graph, its features, and the information it was meant to communicate. Some questions were designed for all 6–10-year-olds, while others were designed for the older students only.

Three GIA items were used to document students' line graph skills. Nine items were used to assess bar graph skills (6 that were appropriate for all students, one for 8–10-year-olds only, and two for 10-year-olds only). Seven items were used to measure students' graphing and interpretation skills in relation to dot plots (five for 8–10-year-olds only and 2 for 10-year-olds only). To combat a challenge associated with performance assessment—specifically, that it is time consuming to administer and score—*in situ* scoring was an essential design constraint. *In situ* is defined as 'in the natural or original

position or place' (www.merriam-webster.com/dictionary/in%20situ). For the purposes of this study, *in situ* scoring was defined as tabulating students' correct responses on individual items during the natural context of the interview's administration.

Procedure

Students were interviewed individually by one of two researchers prior to and soon after completing the *WeatherBlur* investigation. One researcher conducted all the interviews at a given school. At baseline, the researcher alternated between Version 1 and Version 2 of the GIA in each class in order to counterbalance the assessment; students took the opposite version at post. Interviews, all audio recorded, took place in a quiet area of the classroom and lasted approximately 10 minutes.

In situ scoring was conducted to document responses during the interview. More specifically, students' responses to both the open-ended and prompted items were documented on the interview protocol. In cases where students provided information related to one of the prompt questions as part of their initial open-ended description of the graph, the researcher noted those responses with the corresponding prompt on the protocol, then skipped those items during the prompt portion of the interview. Some questions were scored simply as correct versus incorrect. Questions that required discussion or interpretation often resulted in multiple correct responses for one prompt; the student received credit for each piece of information provided.

Responses were entered into a Survey Monkey form to create the database needed for additional coding and analysis. Inter-rater reliability of *in situ* scores was established for each field researcher. The lead author scored a representative sample of approximately one-fifth of the interviews to calculate inter-rater reliability of the *in situ* scores; Krippendorf alpha was .71.

Coding Method

Students' responses were totaled for each graph. Correct responses were coded as 1 and incorrect responses were coded as 0. (See Table 2 for the total possible score for each graph by age.) Raw scores were converted to percent correct to adjust for the different number of items assessed across age. Percent correct scores were calculated by dividing the number of correct items by the total number of items asked of the student. A total of four scores were compiled per student (an overall score and one for each of the three graph types).

Table 2. Total possible scores on the GIA, by graph type and age band

	Ages 6–7	Ages 8–9	Age 10
Line graph	6	6	6
Bar graph	7	9	11
Dot plot	–	10	12

The Graph Interpretation Scoring Rubric (Boote, 2012; referred to hereon as the Rubric) was used to create a second set of scores that were used to document the convergent validity of the GIA. Recall that the Rubric can be used to code student responses related to four literature-based levels of graph interpretation. Students' responses to the GIA were coded for each level. Verbatim transcripts of all intelligible utterances were created from the interview audio files. Student transcripts were then parsed by utterance, which was defined as a unique idea. Each sentence received at least one code. Complex sentences that included multiple ideas were parsed to ensure that each unique utterance was coded. Duplicate codes were not permitted within a response. Once parsed, the transcripts were coded using the Rubric.

Inter-rater reliability was established by two members of the research team. Each coded 31 transcripts, approximately one-fifth of the open-ended responses. The 31 transcripts were selected to reflect the overall sample as much as possible based on grade level, school, and time (baseline, post). The sample used for reliability resulted in a total of 1,100 coded utterances (Cohen's kappa = .79).

Four scores were created from the Rubric coding to document the depth and breadth of students' responses. Depth was defined as the total number of Rubric levels a student used in response to the GIA (out of a total of four possible levels). This score represents the range of literature-based perspectives that students used to interpret the graphs. Breadth was defined as the total number of code-able utterances provided by a student according to the Rubric. This score reflects the frequency of unique literature-based interpretations that each student used to describe the graph. Depth and Breadth scores were created to document both students' responses to the initial open-ended prompt and their responses to the prompt questions.

Results

Validity Evidence

Three types of validity evidence were explored as part of this study (RQ1). To investigate the criterion validity of the GIA, student scores were analyzed by age group. The literature on graph interpretation indicates that students become more sophisticated in their understanding of data and graphs by ages 9–10. In addition, the CCSS in math provide students with more opportunity to use graphs as they progress through primary school. Preliminary criterion validity evidence of the GIA would be demonstrated then if GIA scores increased by age group. A one-way analysis of variance (ANOVA) was conducted to evaluate the relationship between age and students' overall percent correct scores at baseline. The independent variable, age, included four levels: 6-, 8-, 9-, and 10-year-olds. Given the small sample of 7-year-old students, these data were omitted from this analysis. The ANOVA was significant with an effect size in the large range, $F(3, 48) = 15.50$, $p < .001$, $\eta^2 = .49$. Follow-up Tukey tests were conducted to evaluate pairwise differences among the means. The 95% confidence intervals for the pairwise differences, as well as the means and standard deviations for each age, are presented in Table 3.

Table 3. 95% confidence intervals of pairwise differences in baseline GIA scores by age ($N = 52$)

	M	SD	Age 6	Age 8	Age 9
Age 6	.27	0.18			
Age 8	.55	0.18	(−.47, −.10*)		
Age 9	.66	0.17	(−.58, −.21*)	(−.27, .05)	
Age 10	.74	0.12	(−.67, −.28*)	(−.35, −.02*)	(−.25, .09)

*$p < .05$.

Students' overall GIA scores increased with each age at baseline, with significant increases in scores between ages 6 and 8 and ages 8 and 10. The range of students' scores showed a seeming lack of familiarity with graph interpretation among 6-year-olds and a growing proficiency across older ages. This age-level progression matches expectations based on the academic literature and the progression of the math CCSS, providing initial confirmation of the criterion validity of the GIA.

Next, discriminant validity was explored. The literature states that student understanding of bar graphs usually precedes that of line graphs (Friel et al., 2001; Russell et al., 2002), and the math CCSS introduce bar graphs before line graphs. Dot plots are not included explicitly in the CCSS for primary grades but are sometimes used with students of this age to create an opportunity to graph X and Y on a coordinate plane (CCSS.MATH.CONTENT.5.G.G.1). Preliminary discriminant validity evidence for the GIA would be established then if students' scores were highest for bar graphs and if scores were higher for line graphs compared to dot plots.

Two analyses investigated whether the GIA differentiated student scores based on graph type. A one-way repeated-measures ANOVA was conducted to evaluate the relationship between graph type and percent correct scores at baseline for all students. The independent variable, graph type, included two levels: line graph and bar graph. The results for the ANOVA indicated a significant main effect with a moderate effect size, *Wilk's* $\Lambda = .78$, $F(1, 54) = 15.31$, $p < .001$, $\eta^2 = .22$. Students scored significantly higher on the bar graph compared to the line graph ($M = 78\%$ and 40%, respectively). This pattern of results establishes initial discriminant validity evidence to support the use of the GIA; the findings match expectations based on the academic literature and the progression of the math CCSS which introduces students to bar graphs in the early primary grades and line graphs in later years of primary schooling.

Students aged 8–10 years completed the GIA for a third graph, the dot plot. A second one-way repeated-measures ANOVA was conducted for this subset of students to compare their scores across all three GIA graphs assessed. Again, a significant main effect was found, *Wilk's* $\Lambda = .21$, $F(2, 41) = 76.43$, $p < .001$, $\eta^2 = .79$. Three pairwise comparisons were conducted among mean baseline scores (line graph, bar graph, dot plot). Two of the three were significant controlling for familywise error across the three tests at the .05 level using the Holm's sequential Bonferroni procedure. These included a statistically significant difference between students' baseline bar graph

and line graph scores ($p < .001$) and between baseline bar graph and dot plot scores ($p < .001$). Students' scores on the line graph and dot plot did not differ at baseline ($p = .15$). (See Table 4 for means and standard deviations by graph type.)

As with the previous results, students' proficiency with bar graphs is expected, given the order in which graphs are introduced to students across grade level. Students' dot plot scores were higher than those for line graphs, though not statistically so. Given that line graphs are included in the CCSS for this age group and dot plots are not, this pattern of results is the opposite of that expected. Figures 1–3 present pre and post scores by age and graph type to offer a descriptive summary of the results presented thus far.

A final analysis was conducted to establish convergent validity of the GIA by comparing students' scores on the assessment to scores from the Graph Interpretation Scoring Rubric (Boote, 2012). Table 5 presents the mean number of unique utterances provided by students in response to both open-ended and prompted items from the GIA. The pattern of results demonstrates that students provided a broader range of graph interpretations, as defined by the four levels of the Rubric, and that they discussed the graphs in more depth when prompted by the GIA questions. The standard deviations indicate a range of variability in both the Breadth and Depth with which students discussed the graphs in relation to both question types.

To explore the convergent validity of the GIA, students' Rubric-based Breadth and Depth scores were correlated with their scores from the GIA. Breadth and Depth scores were correlated separately based on responses to open-ended and prompted GIA items. The full correlation matrix is presented in Table 6.

GIA scores correlated with students' Rubric-based interpretation scores for both open-ended and prompted questions. Correlations between the GIA and students' open-ended responses were moderate for Breadth scores and the large range for Depth scores. Correlations between the GIA and students' responses to the prompted items were in the large range for both Depth and Breadth. Collectively, these data demonstrate that the GIA is positively associated with established measures of graph interpretation in the anticipated direction, thus providing evidence to support the use of the GIA as a measure of students' graph interpretation skills.

Using the GIA to Detect Change within the Context of an Educational Intervention

The GIA was originally developed as a way to use standards-aligned tasks as a performance assessment for the *WeatherBlur* project. A series of one-way

Table 4. Means and standard deviations for baseline scores by graph type ($N = 43$)

	M	SD
Line graph	.47	0.24
Bar graph	.87	0.15
Dot plot	.53	0.27

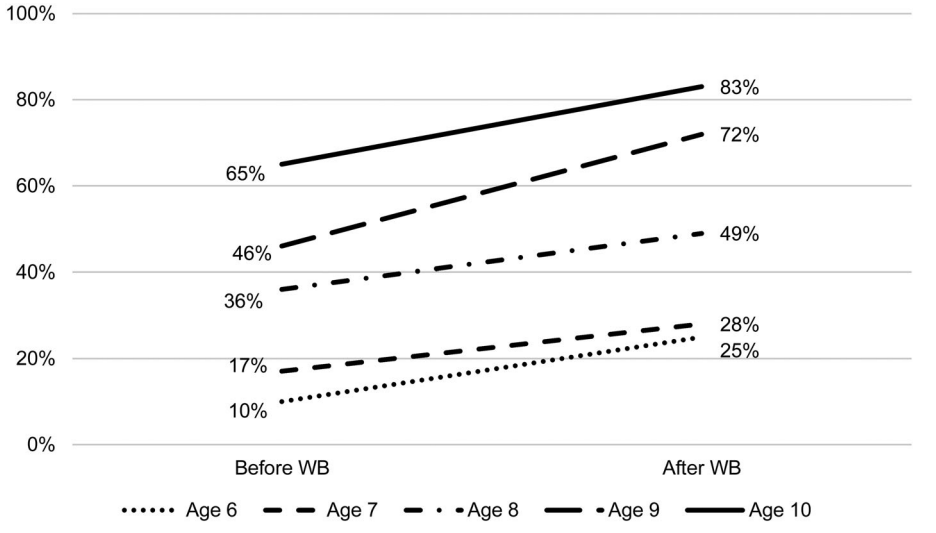


Figure 1. Percent correct scores for the GIA line graph, by age ($N = 55$)

repeated-measures ANOVAs was conducted to evaluate whether the GIA could be used to detect changes in students' graph interpretation skills after participating in *WeatherBlur* (RQ2). Given the program-focused nature of this research question, the small sample size, and the consistent rise in scores across grade levels reported above, it seemed most conservative to conduct this analysis for the entire sample rather than by age group. The within-subjects factor for each ANOVA was time at two levels: pre and post. The dependent variable was percent correct scores for the line, bar, and dot plot, respectively.

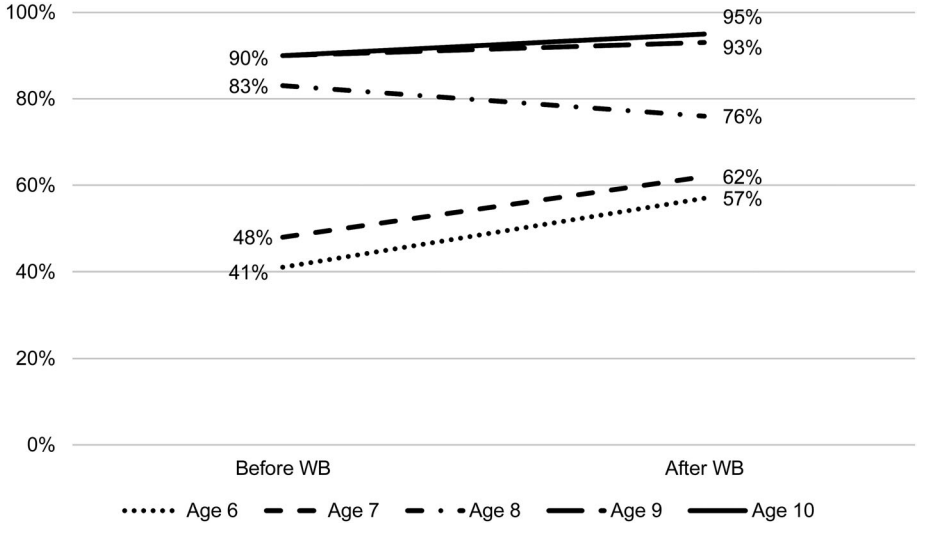


Figure 2. Percent correct scores for the GIA bar graph, by age ($N = 55$)

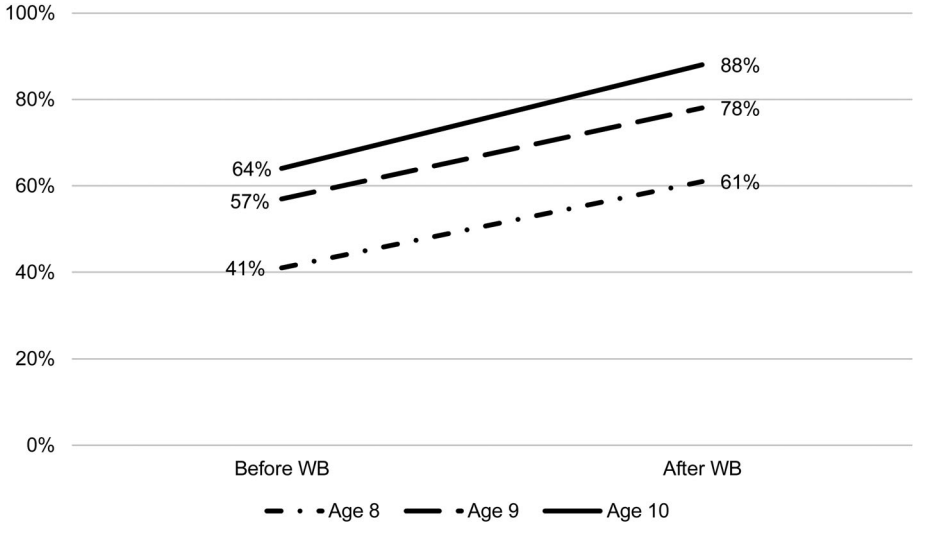


Figure 3. Graphical interpretation scores for common dot plot items, by age ($N = 55$)

Table 5. Means and standard deviations for breadth and depth rubric scores, by GIA question type ($N = 55$)

	M	SD
Breadth		
Open-ended	1.00	0.75
Prompted	2.62	0.83
Depth		
Open-ended	1.58	1.44
Prompted	8.38	4.39

Table 6. Correlation matrix of GIA and rubric scores question type ($N = 55$)

		1	2	3	4
1	GIA score				
2	Breadth-open	.40**			
3	Breadth-prompted	.72***	.33*		
4	Depth-open	.52***	.76***	.38**	
5	Depth-prompted	.84***	.40**	.70***	.49***

*$p < .05$.

**$p < .01$.

***$p < .001$.

Table 7. Changes in percent correct GIA scores after the *WeatherBlur* investigation, by graph type

	N	*M*	SD	Wilk's Λ	*F*	*p*	η^2
Line graph							
Pretest	55	.40	0.27	.67	26.04	<.001	.33
Posttest		.57	0.30				
Bar graph							
Pretest	55	.78	0.26	.99	0.84	.37	.02
Posttest		.81	0.27				
Dot plot							
Pretest	43	.53	0.27	.52	40.69	<.001	.49
Posttest		.74	0.26				

Significant gains with large effect sizes were found in two of three instances (Table 7). Students demonstrated improved graph interpretation skill in relation to line graphs and dot plots, but not bar graphs, after the *WeatherBlur* investigation. Students' pre–post change in line graph and dot plot scores increased at similar levels (17 and 19 percentage points, respectively). Interpreting line graphs remained a challenge for students even after the *WeatherBlur* investigation, with average post scores at 57% correct. Dot plot scores at the end of the investigation showed moderate proficiency, with students answering 78% of the items correctly on average. Bar graph scores remained higher than those for the other two graph types at post, but also demonstrated the potential for additional gains in student proficiency. Collectively, these results provide initial evidence to support the use of the GIA as a pre–post measure to document change in students' graph interpretation skills across time.

Conclusions

This case study explored the validity of the GIA as a measure of primary students' proficiency interpreting multiple kinds of graphical displays. As with any case study, the purpose of this research was to offer a detailed description of a phenomenon with the hope of identifying hypotheses that can be explored in the future.

The GIA was effective at differentiating primary students' graph interpretation skills by both age and graph type, providing preliminary evidence of criterion and discriminant validity. Furthermore, preliminary evidence of the GIA's concurrent validity was established by correlating GIA scores with scores from a rubric of literature-based graph interpretation skills. The GIA also documented improved student scores after the *WeatherBlur* investigation, suggesting that the instrument is sensitive enough to detect short-term gains associated with an educational unit or intervention. The counterbalanced design used for the data collection increases our confidence that the gains found were not a simple test–retest effect.

This evidence suggests that the GIA, or other assessments of standards-aligned tasks, holds promise as a method for assessing performance expectations. The NGSS categorizes a number of graph interpretation skills as CCSS that are expected

to connect across all ages (NRC, 2012). These skills are likely to be the focus of future instruction and assessment for many teachers. Unfortunately, the vast majority of graph interpretation research to date has focused on instructional techniques rather than on assessment itself, providing educators with scant literature to inform the development and utility of tasks that measure performance expectations in the classroom.

This case study was born out of the need to conduct evaluation of and research on *WeatherBlur*. The performance expectations provided via the CCSS and NGSS provided the language needed to study the development of these skills within the context of the project. Specifically, the GIA was designed to include mathematics standards for students aged 8–10, as well as science standards for students aged 5–8. The interpretation of graphs is a challenge for students of all levels (Glazer, 2011). There is much to be learned about the developmental progression of students' skills via the context of the new standards. The findings of this study provided meaningful information to document students' graphical interpretation skills at the beginning of the school year, and change in those skills as the result of the *WeatherBlur* project.

Though the need for assessments like the GIA is clear, the challenges of using performance assessments are not insignificant. The current study reiterates the utility of this method, and takes a first step toward addressing the concerns associated with performance assessments. A lack of reliability and validity evidence, for example, is a general criticism of performance assessments (Gipps, 1995; Johnson et al., 2009; Stufflebeam, 2001). Indeed, many performance assessments in the existing literature on graph interpretation lack validation evidence. In this study, inter-rater reliability for GIA scores was quite high, and preliminary evidence was provided to begin establishing the GIA as a valid measure of students' graph interpretation skills.

Another criticism of performance assessments is the time that they take to develop, implement, and score (Stufflebeam, 2001). The GIA begins to respond to two of these three challenges by combining the time devoted to implementation and scoring. The GIA's administration and scoring were completed in 10 minutes per student. It is our hope that *in situ* scoring will make it feasible for some teachers to use the GIA or similar standards-aligned assessment tasks for formative or summative purposes.

Implications

The GIA was designed to be project-focused and thus the instrument itself may not generalize to other contexts. This fact is a potential limitation of the GIA that serves to highlight the need for additional research in this area. Context-relevant and context-specific tasks are ideals for both performance assessments and assessments of graph interpretation skills (Berg & Smith, 1994; Johnson et al., 2009). While these characteristics offer optimal support to students, they present a considerable challenge to establishing a generalizable performance measure. A scan of the current literature supports this notion. While multiple-choice assessments of graph interpretation skills such as the TOGS have been used in many studies and contexts (Boote, 2012; McKenzie & Padilla, 1986; Padilla et al., 1986), we are unaware of performance-based measures that have been used in these ways. Creating both generic and

customizable versions of the GIA (or other performance measures of graph interpretations skills) is a crucial next steps in understanding whether and how these types of performance measures can serve as a valuable resource to classroom teachers. A generalized version of the GIA is currently being developed for these purposes, and instrument development work in the context of citizen science has demonstrated the potential to create validated scales that can be customized to the context of individual projects (http://www.birds.cornell.edu/citscitoolkit/evaluation/instruments).

Generalized and customizable standards-aligned tasks are each promising avenues of further research. Given the heightened role that performance tasks are likely to play in the coming years, it is imperative that studies focus on practical solutions that can equip teachers to utilize performance tasks in their classrooms as both instructional and assessment techniques.

We believe that the results from the current study warrant replication via a more comprehensive validation effort that includes a larger sample and a more comprehensive set of standards-aligned tasks. Future work should include both generic and context-specific options that can be used to harness the potential of performance assessments, while striving to meet the practical needs of today's classroom.

Additional research is also needed to document strategies that are effective at fulfilling the NRC's vision for parallel instructional and assessment practices in the classroom, and we believe that the GIA tasks hold promise in this area as well. Though used for assessment purposes, standards-aligned tasks like those on the GIA tasks could easily be used as instructional practices. Indeed, research suggests that when students work with graphs that feature real data they can relate to, their graph interpretation skills are enhanced (Konold et al., 2014).

In sum, we believe, the results from this case study provide compelling evidence that can be used to generate new hypotheses for future study of practical solutions that can help teachers respond to the instructional and assessment challenges related to the CCSS and NGSS. Focusing these efforts on the math and science standards that inform teaching practice, the formative and summative assessment needs of teachers, and the cited factors from the literature has the potential to bridge research and practice. These efforts can provide teachers with the information they need about their students' skill and education researchers with new evidence to continue to inform theory.

Acknowledgements

We would like to thank Martin Lodish, Nancy Miles, Jane Robertson, and Michael Weatherwax for assisting in data collection and coding for this study. We would also like to thank Jan Makros and Julie Johnson for reviewing early versions of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Science Foundation under Grant numbers DRL#1217247 and DRL#1451315.

References

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education*, *4*(4), 305–318. doi:10.1207/s15324818ame0404_4

Becker-Klein, R., Stylinski, C., Peterman, K., & Phillips, T. (2014). *Using embedded assessment to measure science skills within STEM education*. Paper presented at the meeting of the American Evaluation Association, Denver, Colorado.

Berg, C. A., & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple-choice and free-response instruments. *Science Education*, *78*(6), 527–554. issn: 0036-8326

Bertin, J. (1983). *Semiology of graphics: Diagrams, networks, maps* (W. J. Berg, Trans.). Madison, WI: The University of Wisconsin Press.

Bertin, J. (2001). Matix theory of graphics. *Information Design Journal*, *10*(1), 5–19. issn: 1876-486X

Boote, S. K. (2012). Assessing and understanding line graph interpretations using a scoring rubric of organized cited factors. *Journal of Science Teacher Education*, *25*(3), 333–354. doi:10.1007/s10972-012-9318-8

Bybee, R. W. (2011). Scientific and engineering practices in K-12 classrooms. *Science Teacher*, *78*, 34–40. doi:10.1525/abt.2012.74.8.3

Bybee, R. W. (2012). The next generation of science standards: Implications for biology education. *The American Biology Teacher*, *74*(8), 542–549. doi:10.1525/abt.2012.74.8.3

Curcio, F. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education*, *18*(5), 382–393. doi:10.2307/749086

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124–158. doi:10.2307/749671

Gardner, H. (1992). Assessment in context: The alternative to standardized testing. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 77–119). Netherlands: Springer. doi:10.1007/978-94-011-2968-8_4

Gipps, C. (1995). What do we mean by equity in relation to assessment? *Assessment in Education: Principles, Policy & Practice*, *2*(3), 271–281. doi:10.1080/0969595950020303

Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, *47*(2), 183–210. doi:10.1080/03057267.2011.605307

Johnson, B., Duffin, M., & Murphy, M. (2012). Quantifying a relationship between place-based learning and environmental quality. *Environmental Education Research*, *18*(5), 609–624. doi:10.1080/13504622.2011.640748

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: Guilford Press.

Konold, C., & Higgins, T. L. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics.

Konold, C., Higgins, T., & Russell, S. J., & Khalil, K. (2014). Data seen through different lenses. *Educational Studies in Mathematics*, *88*(3), 305–325. doi:10.1007/s10649-013-9529-8

Malcom, S. M. (1991). Equity and excellence through authentic science assessment. In G. Kulm & S. Malcom (Eds.), *Science assessment in the service of reform* (pp. 313–330). Washington, DC: American Association for the Advancement of Science.

McKenzie, D. L., & Padilla, M. J. (1986). The construction and validation of the test of graphing in science (TOGS). *Journal of Research in Science Teaching*, *23*(7), 571–579. Retrieved from http://onlinelibrary.wiley.com.proxy.antioch.edu/doi/10.1002/tea.3660230702/abstract

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Committee on a Conceptual Framework for the New K-12 Science Education Standards. Washington, DC: National Academies Press.

National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Committee on Developing Assessments of Science Proficiency in K-12. Board on Testing and Assessment and Board on Science Education. In J.W. Pellegrino, M.R. Wilson, J.A. Koenig, & A.S. Beatty (Eds.), *Division of behavioral and social sciences and education* (pp. 1–270). Washington, DC: The National Academies Press.

Norman, R. R. (2012). Reading the graphics: What is the relationship between graphical reading processes and student comprehension. *Reading and Writing*, *25*(3), 739–774. doi:10.1007/s11145-011-9298-7

Padilla, M. J., McKenzie, D. L., & Shaw, E. L. J. (1986). An examination of the line graphing ability of students in grades seven through twelve. *School Science and Mathematics*, *86*(1), 20–26. issn: 144263551

Parmar, R. S., & Signer, B. R. (2005). Sources of error in constructing and interpreting graphs: A study of fourth- and fifth-grade students with LD. *Journal of Learning Disabilities*, *38*(3), 250–261. issn: 00222194

Peterman, K. (2013). Show me what you can do: Performance-based assessments as a measure of scientific practice. A panel presented at the meeting of the American Evaluation Association, Washington, DC.

Rural School & Community Trust. (2001). *Assessing Student Work*. Retrieved from http://www.ruraledu.org/user_uploads/file/Assessing_Student_Work.pdf

Russell, S.J., Schifter, D., Bastable, V., Konold, C., & Higgins, T. (2002) *Working with data, casebook*. Parsippany, NJ: Dale Seymour.

Savin-Baden, M., & Major, C. H. (2004). *Foundations of problem-based learning*. Buckingham: SRHE/Open University Press.

Steck, T. R., DiBiase, W., Wang, C., & Boukhtiarov, A. (2012). The use of open-ended problem-based learning scenarios in an interdisciplinary biotechnology class: Evaluation of a problem-based learning course across three years. *Journal of Microbiology & Biology Education*, *13*(1), 2–10. doi:http://dx.doi.org/10.1128/jmbe.v13i1.389

Stufflebeam, D. (2001). The meta-evaluation imperative. *American Journal of Evaluation*, *22*(2), 183–209. Retrieved from http://www.wmich.edu/evalphd/wp-content/uploads/2011/02/The_Metaevaluation_Imperative.pdf

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappa International*, *75*, 200–214. Retrieved from http://www.jstor.org/stable/20405066

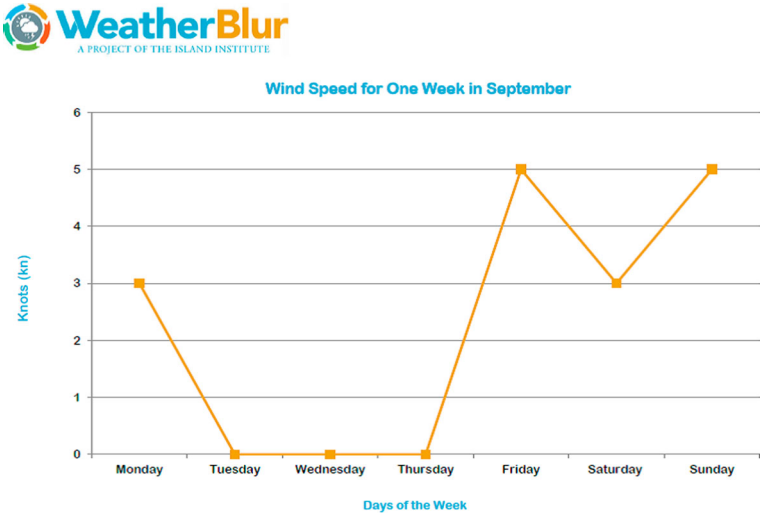
Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181–208. issn: 0895-7347

Appendix

Graphical Interpretation Assessment Version 1

Today we are going to look at several graphs together and then talk about what we see. All of the graphs show information that you might want to know about the Weather-Blur project or that you might learn through the project.

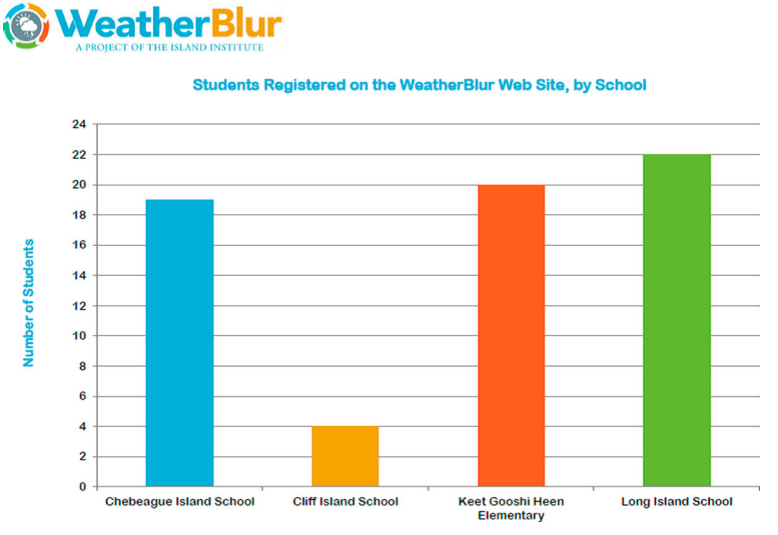
Our first graph is a line graph. Take a look at this graph and tell me what you see. You can point to different parts of the graph and tell me what they mean or just talk about it.



That was great—thank you. Now I'm going to ask a few more questions to see if we can figure out what some of the other pieces of this graph mean. You can answer by pointing, showing me with your fingers, or telling me your answer. [throughout, remind students that they can point our show you their answer with their fingers if/as needed; if they point or use their fingers, narrate back to them what they are doing so we have it on the audio record]

Age range	Question	Correct answer [score as you go]
Ages 6–10	Let's start by talking about the kinds of information we see in this graph. What is this graph really trying to tell us about? [If student answers, repeat answer back to him/her and then ask, "Is there anything else you can tell me about the information in this graph?"]	____ Wind speed ____ Knots (kn) ____ Days of week ____ Change over time
	How many data points do you see in this graph?	____ 7
	What do we think this graph tells us about the wind during the week when these data were collected? [If student answers but doesn't give all possible options then ask, "Is there anything else the graph tells us about the wind?"]	____ How the wind changes ____ Drops on Tuesday ____ No wind mid-week ____ Windy weekend ____ Most wind on Fri and Sun ____ Least wind on Mon-Wed

Now let's talk about a bar graph. Tell me what this one is about—remember you can point to the different parts of the graph and then tell me about them or just start talking.



Now let's see if there is anything else we can figure out about this graph. I'm going to ask you some questions about the graph—you can answer by pointing, showing me with your fingers, or telling me your answer.

Grade band	Question	Correct answer
Ages 6–10	How many schools are shown in this graph?	____ 4
	So what do you think the bars mean?	____ # of students
	Which school has 19 students? You can tell me the name or point.	____ Chebeague
	What do you think it means if one bar is really tall like this one [point to Chebeague] and another one is really short [point to Cliff]?	____ School with a tall bar has more ____ School with a short bar has less
	How many more students does Long Island School have compared to Keet Gooshi Heen?	____ 2
	Let's see if we can figure out the school with the most and least students registered. Let's start with the most first—which school has the most students? Now how about the least?	____ Most – Long ____ Least – Cliff

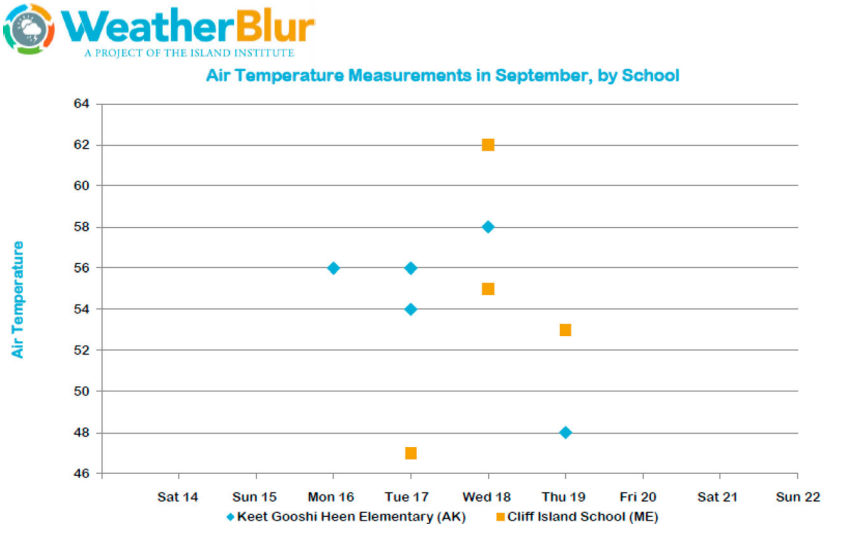
(*Continued*)

Continued

Grade band	Question	Correct answer
Ages 8–10	Let's say that we realized that there was a mistake in this graph and that Cliff island actually had 14 students registered instead of 4. Show me how high their bar would be.	____ Points to the right column for Cliff ____ Points to the line for 14 on the graph
Age 10	What information is shown on the X axis (that is the horizontal axis)?	____ Schools
	What information is shown on the Y axis (or the vertical axis)?	_____ # of students

[Students age 6–7 stop here; students age 8—10 continue]

Okay, this is our last graph. This is a graph that you might make for some of you WeatherBlur data—it shows what graphs actually look like on the WeatherBlur site. Take a look at the graph and tell me what you see. You can point to different parts of the graph and tell me what they mean or just talk.



That was great—thank you. Now I'm going to ask a few more questions to see if we can figure out what some of the other pieces of this graph mean.

Age range	Question	Correct answer
Ages 8–10	Which WeatherBlur schools are shown in this graph?	____ Cliff ____ Keet Gooshi Heen
	How many days were data collected for this graph?	____ 4

(*Continued*)

Continued

Age range	Question	Correct answer
	So what do you think these dots and squares really mean—what does this one mean for example? [point to top Cliff Island data point for Wed 18] [If student answers but doesn't give all possible options then repeat the answer back and ask, "Is there anything else you can tell me about this data point [point]?"]	____ Shows temp ____ for Cliff ____ on Wed ____ It was 62 degrees.
	Alright. We have figured out that we have data in this graph for two schools and that we are looking at air temperature data. Which school has the most data?	____ Keet Gooshi Heen
	Let's see if we can figure out the highest and the lowest air temperatures on this graph. Let's start with the highest first—what is the highest temperature you see? Now how about the lowest?	____ Highest – 62 degrees ____ Lowest – 47 degrees
Age 10	What information is shown on the X axis (that is the horizontal axis)?	____ Days
	What information is shown on the Y axis (or the vertical axis)?	____ Air temperature