Journal of Science Education and Technology (2021) 30:431–446
https://doi.org/10.1007/s10956-020-09886-w

Biology Undergraduate Students' Graphing Practice in Digital Versus Pen and Paper Graphing Environments

Stephanie M. Gardner[1] · Elizabeth Suazo-Flores[2] · Susan Maruca[3] · Joel K. Abraham[4] · Anupriya Karippadath[1] · Eli Meir[3]

Accepted: 19 November 2020 / Published online: 12 January 2021

Abstract

Graphing is an important practice for scientists and in K-16 science curricula. Graphs can be constructed using an array of software packages as well as by hand, with pen-and-paper. However, we have an incomplete understanding of how students' graphing practice vary by graphing environment; differences could affect how best to teach and assess graphing. Here we explore the role of two graphing environments in students' graphing practice. We studied 43 undergraduate biology students' graphing practice using either pen-and-paper (PP) ($n = 21$ students) or a digital graphing tool GraphSmarts (GS) ($n = 22$ students). Participants' graphs and verbal justifications were analyzed to identify features such as the variables plotted, number of graphs created, raw data versus summarized data plotted, and graph types (e.g., scatter plot, line graph, or bar graph) as well as participants' reasoning for their graphing choices. Several aspects of participant graphs were similar regardless of graphing environment, including plotting raw vs. summarized data, graph type, and overall graph quality, while GS participants were more likely to plot the most relevant variables. In GS, participants could easily make more graphs than in PP and this may have helped some participants show latent features of their graphing practice. Those students using PP tended to focus more on ease of constructing the graph than GS. This study illuminates how the different characteristics of the graphing environment have implications for instruction and interpretation of assessments of student graphing practices.

Keywords Improving classroom teaching · Post-secondary education · Pedagogical issues · Teaching/learning strategies · Assessment

Introduction

Revealing and Assessing Student Science Practice Knowledge

Recent discussions of K-16 STEM education have focused on the disciplinary knowledge and practices students should develop and use in novel situations (e.g., Seraphin et al., 2013; Wild et al., 2018; Windschitl et al., 2007). Within science these include conducting experiments, understanding and using models, and organizing, analyzing, and interpreting data (e.g., Kjelvik and Schultheis, 2019; Kuhn, 2010; Lehrer, Schauble and Lucas, 2008; Lehrer, Schauble, and Petrosino, 2001). The move from analog to digital work environments has occurred at a fast rate, but the implications of that move have lagged behind technology advancement and adoption. Some research on learning by environment exists. For instance Cromley et al. (2020) recently published a meta-analysis showing that drawing-to-learn tends to improve learning outcomes only when students drew by hand, but not when using digital drawing tools, and other studies have similarly found differences suggesting the environment in which students work can affect learning (e.g., Mueller and Oppenheimer, 2014; Sinclair and Yurita, 2008). In the specific context of graphing, it has been argued that the benefits of computer tools in graph generation (i.e., high degree of accuracy and ease in generating alternative

✉ Stephanie M. Gardner
sgardne@purdue.edu

[1] Department of Biological Sciences, Purdue University, 915 West State Street, West Lafayette, IN 47907, USA

[2] Center for Advancing the Teaching and Learning of STEM Department of Curriculum and Instruction, BRNG, 4166 Purdue University 100 N. University Street, IN 47907-2098 West Lafayette, USA

[3] SimBiotic Software, Inc, P.O. Box 7158, Missoula, MT 59807, USA

[4] Department of Biological Science, California State University-Fullerton, 800 North State College Blvd, Fullerton, CA 92831, USA

displays for examination) may unintentionally lead students to be less intentional and reflective in their graphing decisions (e.g., selecting appropriate variables) (Tairab and Khalaf Al-Naqbi, 2004). It has been recommended that students approach graph construction in phases which includes exploring the data and planning how to best represent them based on the task or purpose for graphing, sketching graphs out by hand, and finally using technological tools for data visualization and meaning making (Patterson and Leonard, 2005; Tairab and Khalaf Al-Naqbi, 2004). These studies have implications for instruction—much less research examines the impact of environments on assessment features and design.

Assessment is critical for improving STEM teaching (Harwell et al. 2015) but assessing science practices, such as conducting data analysis, is difficult to do with closed form questions like multiple choice (e.g., Garfield, 2003; Zieffler et al., 2008). Instead, those practices are often assessed through open-ended activities such as writing, drawing, or presentations. Such open-ended, manual assessments are difficult to conduct at larger scales such as with the hundreds of students found in many introductory college science classes.

Digital tools are widely used for large-scale assessments of students' procedural knowledge and application of STEM learning (e.g., recognizing and defining terminology, applying knowledge to solve well-defined problems). With notable exceptions (e.g., in biology Urban-Lurain et al., 2013; Beggrow et al., 2014; Weston et al., 2015; Vitale et al., 2015, 2019; Zhai et al., 2020), there are fewer examples of digital tools designed to automatically measure practices in complex, less well-defined problem spaces, in large part because of the difficulty in valid automatic scoring (e.g., Beggrow et al., 2014; Ha and Nehm, 2016). Furthermore, few studies compare assessment between pen and paper vs. digital formats, and those have differing conclusions about student performance across the two environments (e.g., Kumar et al., 1994; Aberg-Bengtsson, 2006; Guimaraes et al., 2018; Oqvist and Nouri, 2018). As development of digital assessment tools increases, the field would benefit from more research on how differences in paper versus digital environment interact with the way students demonstrate understanding through performance tasks.

A key difference in digital assessment tools is the suite of affordances and constraints not present in the blank canvas of a piece of paper and a pen (or generic writing and drawing software). Constraints are almost inevitable with a digital assessment tool when auto-scoring is desired, as the tool will only have a limited set of functionality, and a certain degree of constraint on student actions increases the accuracy and preciseness of scoring algorithms (Scalise and Gifford, 2006; Kim et al., 2017; Meir et al., 2019). However, those constraints can also impact learning to be measured by the assessment (Meir et al., 2019; Cromley et al., 2020). Constraints can preclude measurement of certain aspects of higher-order thinking, but can also help reveal aspects of thinking that were hidden by other confusions in a more lightly constrained pen–and paper assessment. For instance, auto-scored question formats that are intermediately open, between multiple choice and short essays, can help students express their thinking more precisely than in either multiple choice or essay format (Meir et al., 2019). Similarly, affordances of a digital tool may change the graphing practice the students demonstrate due to different or novel aspects of the tool (e.g., Sinclair and Yurita, 2008). Thus, it is important to understand how the affordances and constraints of any digital tool interact with student answers when judging what may be learned from those answers.

In this study, we explore undergraduate biology students' graphing practices in two different environments: pen-and-paper (PP) and a digital assessment tool GraphSmarts (GS).

Graph Construction Consists of Many Graphing Practices and Decision Points

Undergraduate science students are required to interpret and construct graphs (AAAS, 2009; Kjelvik and Schultheis, 2019; Shanahan et al., 2011), yet studies show that students at all levels from K-12 to post-graduate struggle with graphing (e.g., Angra and Gardner, 2017; Chick, 2004; D'Ambrosio et al., 2004; Roth and Hwang, 2006; Tufte, 1983). By the postsecondary level, students know the mechanics of graphing (e.g., placing data points in the cartesian system—Padilla et al. 1986) but still struggle with higher-order skills such as the relationship between variables and graph type based on the goal of the graphing task. The continued struggles may be evidence that graphing is a practice that must be learned through experience rather than as a cognitive skill (Roth and Bowen, 2001; Bowen and Roth, 2005; Roth and McGinn, 1997). Here "poor" graphing is viewed as lack of graphing experience rather than lack of cognitive ability, so that "graph sense develops gradually as a result of one's creating graphs and using already designed graphs in a variety of problem contexts that require making sense of data." (Friel et al., 2001). These practices lead to broadly agreed upon principles of what makes a "good" graph (Kosslyn, 1994; Tufte, 1983,). We build from a body of research on ways in which both developing novices and experienced graphers' graphs differ from the established norms of good graphs (e.g., Angra and Gardner, 2016, 2017; Bowen et al., 1999; Diong et al., 2018; Weissgerber et al., 2015, 2019) and focus here on five central graphing practices that would enhance graphs (Table 1 and described below).

Table 1 Description of graphing practice concepts, and their application, evaluated in the current study

Graphing practice concepts	Application of graphing practice concept
Variable relevance	Did the participant identify appropriate variables for the research question?
Form of the data	Did the participant plot individual data points or summarized values?
Acknowledging variability	Did participants who plotted summarized values show a measure of variation?
Graph type	Did the participant plot data in a line, scatter, or bar graph type?
Graph communication	Did the participant select variables, a graph type, form of data, and display of variability that helps others to see the patterns and trends intended to be communicated?

Variable Relevance. When designing an inquiry or experiment, what variables are relevant to the inquiry (Cobb and Moore, 1997; Mayes et al., 2014; Wild and Pfannkuch, 1999). The measure of those relevant variables (i.e., data) is analyzed, for example, by constructing graphs.

Form of the Data. Descriptive statistics are helpful to search for signals and evidence patterns in a data set (Konold and Pollatsek, 2002). By Form of the Data plotted we mean whether students plotted data as collected or given (i.e., every single data point), versus summarized data (i.e., mean or median of the data set) in a graph. Often summarized data such as means are plotted to compare data from different treatments. Yet, researchers in different fields (e.g., Diong et al., 2018; Friel et al., 2006; Shaughnessy, 2006; Weissgerber et al., 2015, 2019) are calling for plotting raw data as a better illustration of data distribution.

Acknowledging Variability. In exploratory data analysis one must acknowledge and account for variability (Cobb and Moore, 1997; Moore, 1997; Shaughnessy, 2006; Watson et al., 2003; Wild and Pfannkuch, 1999). This practice is more indispensable when plotting summarized data such as means, as means could hide important information about the data distribution (Watson and Moritz, 1998; Friel et al., 2006; Shaughnessy, 2006).

Graph Type. Selecting a graph type is a subjective practice involving choosing "a type of graph from the point of view of being able to characterize shape (of the distribution) and then relating the shape to the context being investigated" (Friel et al., 2006, p. 126). That choice may change as inquirers familiarize themselves with the main characteristics of a data set (Friel et al., 2001; Gelman and Unwin, 2013; Kosslyn, 1985; Tukey, 1977; Wild and Pfannkuch, 1999) but there are some well accepted recommendations. For example, in general, scatter graphs are good to explore the distribution of the data set and identify its outliers, spread, and clusters (Friel et al., 2001, 2006; Weissgerber et al., 2015). Bar graphs are good to communicate results of a treatment or experiment, while line graphs are good to describe and explore relationships among quantitative data points where one could reasonably interpolate between points (Friel et al., 2001, 2006; Kosslyn, 1985, 1994).

Graph Communication. Once the inquirer has explored the data set with all its characteristics and wants to communicate to the public, they need to select a graph type and form of the data that easily help others see the patterns and trends they are wanting to highlight (Friel et al., 2001; Kosslyn, 1985; 1994). This often makes graphing an iterative and subjective process where the inquirer tries many graphs until they identify one that best communicates the data in that context (Friel et al., 2006), and Graph Communication is the ability to evaluate the whole graph more holistically.

Research Questions

Given the background above and paucity of studies on the effect of environment on assessment, this study set out to answer the following two research questions:

- What are the similarities and differences in undergraduate students' graphing practices in relation to the graphing environments in which they worked?
- How do the constraints and affordances of the two environments influence undergraduate students' graphing practices?

Context for the Study

The Digital Graphing Tool

The digital graphing tool GraphSmarts (GS), whose properties we explore here, is being developed as part of a broader project whose goals are to build performance-based assessments of graphing practices. Comparing GS with pen-and-paper (PP) graphing was originally conceived as a form of validation of GS, but the data have proven interesting in their own right.

GS and its evolution are more fully described elsewhere (manuscript in prep), but some of its attributes are important to the comparison with a PP graphing environment. GS had a left-hand panel with six buttons that gave users access to all the features they could add to the graph, as well as some graphing process items such as being able to review the

Fig. 1 Screenshot of the digital graphing tool

research question (Fig. 1). On the right was the graph they were constructing. As users selected features on the left, the graph on the right reflected those features. We used features of the graphs produced by participants as evidence of their skill applying the concepts related to graph construction practice (Friel et al., 2001; Roth and McGinn, 1997).

The GS design was guided by the evidence-centered design framework (Mislevy, 2013) with a major design goal for the tool of algorithmically scoring students' performance. We thus included constraints and affordances that we hypothesized would better enable us to accurately capture student graphing practice. For instance, we constrained users to three commonly used graph types: bar, line, and scatter. By limiting the types of graph available, we made it easier to algorithmically determine whether the users' choice of graph type matched well with the form of the data they were attempting to plot. As another example, knowing that most undergraduate students are already competent at plotting data points (e.g., Brasell and Rowe, 1993), we afforded users increased speed and decreased tedium by plotting the points, bars, and/or lines for them, freeing up time to assess other practices more thoroughly.

The digital tool was embedded in several pages of text and images that introduced the context, research question and data, as well as some multiple choice, intermediate constraint (Meir et al., 2019), and essay questions before and after the graphing task. We do not discuss data from the other questions here. We used the SimUText system from SimBio to serve the tool to students and collect the resulting data.

Biological Context of the Task

Calls for teaching of graphing as part of data analysis suggest using data sets embedded in contexts that represent authentic explorations (Friel et al., 2001; Kjelvik and Schultheis, 2019; Moore, 1997; Weiland, 2017). The problem presented to students involves Marine Protected Areas (MPAs) that have been set aside off the coast of Tasmania to protect marine communities from fishing. One of the marine communities is a well-studied food chain where lobsters (which are fished

Study Plot ID	Month Sampled	MPA Status	Lobster Density (#/m^2)	Average Lobster Size (g)	Urchin Density (#/m^2)	Kelp Abundance Score
1	Aug.	YES	1.10	410	9.5	HIGH
2	Sept.	YES	1.55	445	8.5	MED
3	Aug.	NO	1.15	350	12.0	MED
4	Oct.	YES	2.00	435	7.0	MED
5	Aug.	NO	0.75	385	9.5	MED
6	Oct.	NO	1.05	355	11.0	LOW
7	July	YES	0.99	460	6.5	MED
8	Sept.	NO	1.55	355	14.0	LOW
9	July	NO	1.40	370	12.5	MED
10	Aug.	YES	0.80	500	7.5	HIGH

Fig. 2 Variables and a subset of the values given for the graphing task

for human consumption) eat sea urchins, which in turn eat kelp. Much of the marine life in these communities is dependent on "forests" of mature, tall kelp. Too many sea urchins can cut down the kelp in an area, so the presenting research question is "Do Marine Protected Areas (MPAs) in Tasmania succeed in promoting healthy kelp forests?" From this general question, users are asked to examine a specific hypothesis that "Eliminating lobster fishing will result in improved kelp forest health, due to food chain dynamics." They are asked to test this hypothesis by testing a resulting prediction, that "Areas with no lobster fishing (MPAs) have fewer urchins than do areas with lobster fishing."

Users are given a data set with eighteen samples taken from eighteen different locations. Each sample contains three quantitative variables (lobster density, lobster size, urchin density) and four categorical variables (MPA status, kelp abundance, plot ID, time of sample). Users first see this data set in table form (Fig. 2). Of these data, time of sample and plot ID are not relevant to the research question. The rest of the variables are relevant to the hypothesis, but only two are relevant to the prediction the user is asked to test (MPA and urchin density). For this comparison, we decided to focus on criteria related to five practices that we could measure regardless of whether users constructed graphs against the prediction or against the more general hypothesis, including variable relevance to the hypothesis (Table 1).

Materials and Methods

Participants

All work with human subjects was done in accordance with an approved human studies protocol (IRB# 1,706,019,374). Participants in this study were 43 undergraduate biology students from two different Midwestern universities, of which 22 were assigned to GS and 21 to PP. Six students were biology majors at a small regional, private university, and 37 students were biology majors at a large, public research-intensive university. To make the two treatment groups as comparable as possible with respect to participant characteristics, participants were assigned

Table 2 Participants graphing environment and demographic information

Graphing environment	Class standing				Binary gender		% Self-reported non-white*	% First gen. or Pell Grant eligible
	Freshman	Sophomore	Junior	Senior	Female	Male		
GS ($n = 22$)	5	4	5	8	18	4	32%	33%
PP ($n = 21$)	4	5	4	8	14	7	43%	32%

*Self-reported Asian, African American, Hispanic, Latinx (categories defined by the US Census Bureau for race and ethnicity)

to treatments considering both self-reported class standing and race/ethnicity (Table 2) and then randomly within the graphing environment categories. This approach was taken to increase our confidence in the ability to attribute student graphing practices to the graphing environment category and not to participant characteristics. However, we did not have a sufficient sample size to make claims about student graphing in relation to any participant characteristics.

Two Graphing Environments: Pen-and-Paper Versus Digital Tool

Participants in both graphing environments read the same on-screen three-page introduction to the biology context described above. As part of this introduction, participants were shown a table of the data set from which they could select variables to be graphed (Fig. 2). Following the on-screen introduction, PP participants received the data set on a piece of paper, a blank paper, and a pen that recorded their pen strokes as well as audio (LiveScribe pen ™). Participants in PP were also informed that they could use their cell phone or a calculator to make any calculations they wanted to perform with the data. GS participants went on to a short video introducing the digital graphing tool, and then to a screen with the GS. Each participant was asked to make a graph to address the research question. Participants could make as many exploratory graphs as they wished before submitting a "final graph." During the graph construction process, participants were asked to "think-aloud." After making the final graph, participants in both treatments answered some additional verbal interview questions (i.e., What type of graph did you make? Why did you decide to create the graph that you did? How does graphing in this interface compare to other ways you have built graphs?) that had as a main goal to retrieve participants' justifications for some of their graphing decisions. One researcher witnessed the whole process and conducted the post-graphing semi-structured interview.

In deciding on graphing practices for which to analyze the data (Table 1), we chose only those for which both PP and GS participants were able to use similar feature sets. As an example, participants in PP had a blank piece of paper with a pen, which means they could plot data on any scale they wanted, but needed to draw all elements themselves (i.e., axes and their units, axis breaks if they wanted to have them). Meanwhile, participants in GS were provided with a screen that already had, among other features, the axes. While GS participants did need to assign minimum and maximum values for each axis plotting a quantitative variable, GS evenly spaced the data on the axis, and axis breaks were not available. Thus, because some students in PP used scaling features not available in GS, we did not specifically compare PP and GS participants on those features. We discuss some graphing practices we saw used in one environment but not the other in a qualitative sense only.

Data Collected

This study aimed to understand biology students' graphing practices in two different environments through a mixed-methods approach (Patton, 2015). The data for this study comprise (1) all participants' graph information (variable plotted, graph type, whether the data plotted was raw or summarized data, the total number of graphs created) and (2) audio records of participants' answers to set questions during a semi-structured interview (see Sec 2.2, above). The combination of these different data sources resulted in a rich supply of information that provided a complete picture than would have been possible with a single method of data collection. We analyzed these data in two ways, as described below.

Evaluating Participants' Graphs

The first four graph construction practice concepts in Table 1 were simple to score. *Form of the Data*, *Graph Type*, and *Acknowledging Variability* are read directly off each graph. To score *Selection of Variables* we noted that a graph that did not include the variable MPA Status would do a poor job of addressing the research question, so we scored whether that variable was included on either one of the axes, or a distinct feature of points plotted (e.g., color). In PP participants were able to select a subset of the data to be plotted (e.g., only MPA yes), an action that was not available for GS participants. Therefore, to be consistent across the two environments, we did not differentiate whether participants plotted the complete data set (MPA yes and no) or a subset (only MPA yes).

For the more synthetic concept of *Graph Communication* we scored each participant's graph being of high, medium, or low quality. As shown in Fig. 3, we based the graph quality determination on the four more atomic criteria we scored but considered whether they were combined in ways that did a good job of communicating the data. For this data set, we judged that these criteria reflect the ease with which a viewer can discern patterns in the data.

Figure 4 shows examples of how we scored the quality of several graphs made in both PP and GS graphing environments according to the rules in Fig. 3.

To provide quantitative insight in answering our research questions, where possible, we compared the proportion of

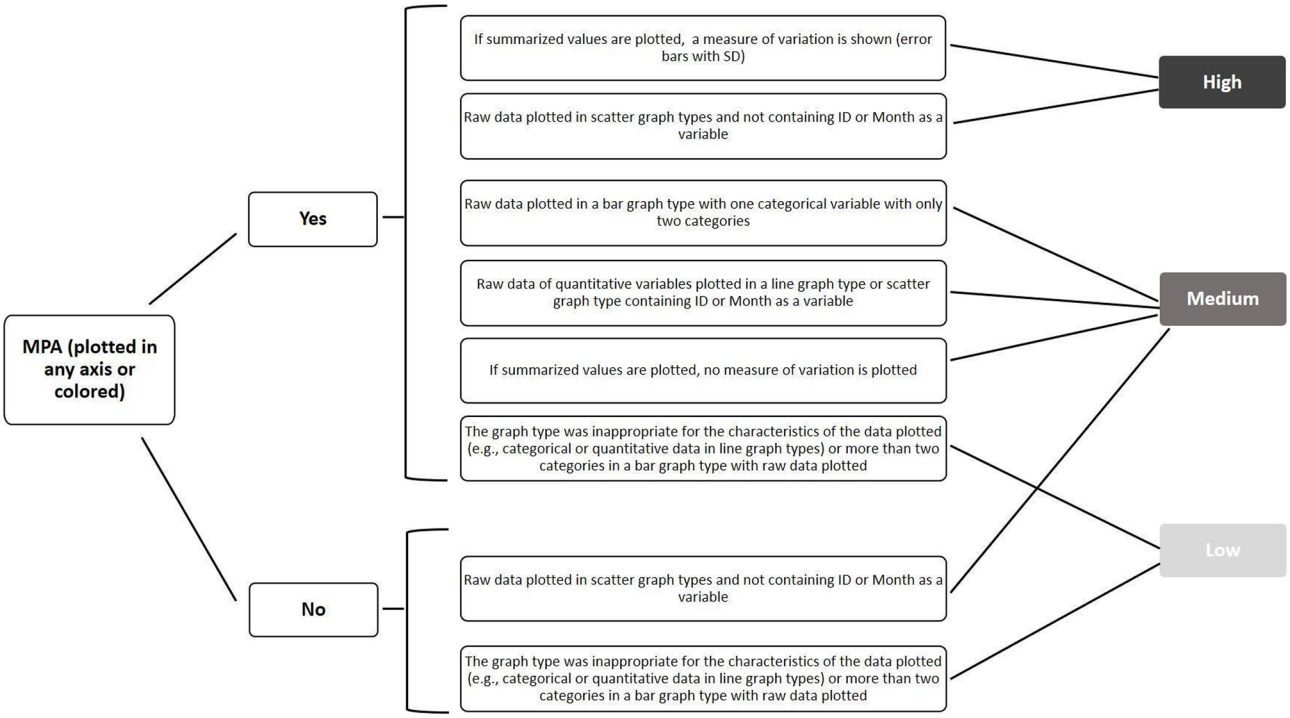

Fig. 3 The rules used to score graph quality. The segments show how a graph was evaluated as a high-, medium-, or low-level

participant scores between the two graphing environments using Fisher's Exact tests in R 3.5.0 (R Core Team, 2018). We chose this test because of the structure of our data and analyses (categorical levels), our research questions, and its appropriateness for small sample sizes in the categories being compared. Finally, to account for the multiple comparisons made, while reducing the risk of Type II errors, we applied a Holm-Bonferroni correction to all tests; this takes a sequential approach to applying Bonferroni corrections such that the likelihood of Type II error is not increased (Holm 1979). We draw inferences based on this correction throughout.

Analyzing Participants' Justifications

In addition to considering how the visible products participants produced might be affected by graphing environment, we also wished to address whether the graphing environment affected participants' graph construction thought process. As a window to their thinking, we analyzed participants' justifications for their graph type as expressed in the interviews after they completed their final graph in response to the set questions from the interviewer (see "Two Graphing Environments: Pen-and-paper Versus Digital Tool") We used the constant comparative method (Strauss and Corbin, 1994) to build categories of participants' justification, and then three researchers coded to consensus the transcript of participants' justifications. Our coding process led to the following four intertwined categories of justification, which were not mutually exclusive: data visualization, data exploration, data characteristics, and ease of use (Table 3).

Once participants' justifications were analyzed using the described codes, we explored patterns related to participants' graph levels and justifications.

Results

Participants' Graphs Show a Mix of Similarities and Differences Between Environments

We scored participants' final graphs (those they reported as final to the interviewer) along the five criteria we defined above (Table 1) for evaluating participants' graphs. Participants using the two graphing formats produced similar graphs on three criteria, and different graphs on two other criteria (Fig. 5; Table 4). We discuss each criterion in turn below.

Selection of variables. As described in the methods, we looked at whether one relevant variable, called MPA, was plotted on one of the axes or was used to color data points. A much higher percentage of participants using GS included MPA on their graph then those using PP (Fig. 5a), and this was a significant difference ($p = 0.007$).

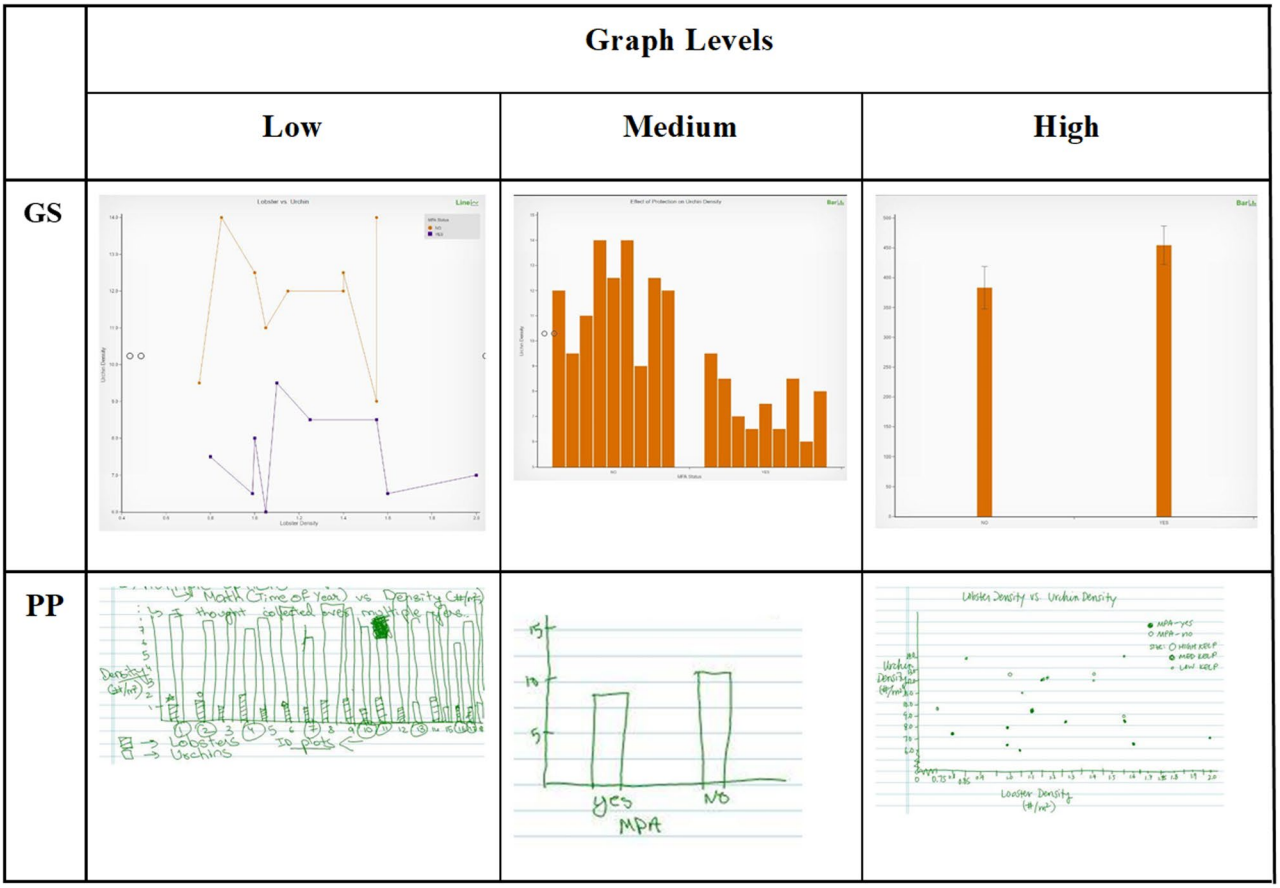

Fig. 4 Examples of graphs made in PP and GS grouped by the graph quality scores they received

Form of the data. Regardless of graphing environment, the majority of the participants (84%) plotted raw data instead of summarized data (Fig. 5b), and we saw no statistical difference between participants using GS vs PP ($p = 0.412$).

Acknowledging variability. Few students in either graphing environment plotted summarized data in their final graphs (7 of 43), but of the five participants who plotted summarized data in GS, four added error bars, while neither of the two participants who plotted summarized data in PP added error bars.

Graph type. Participants using both graphing environments tended to choose either scatter or bar graphs over line graphs, as was appropriate for the data and prediction being tested (Fig. 5c). We saw no significant difference in graph type choice between environments ($p = 0.559$).

Table 3 Categories of participant graph type and variable selection justifications

Justification category	Definition
Data visualization	Participants' justifications were coded as *data visualization* if they referred to being able to better see the data or interpret the information with a particular graph type
Data exploration	Participants mentioned selecting a type of graph because they wanted to explore possible relationships or patterns without knowing or having in mind a pre-determined idea of the characteristics, or shape, of the data to be plotted
Data characteristics	Participants selected a graph type because of the characteristics of the data (i.e., categorical or numerical), form of the data (i.e., raw or summarized), and the size of the data set. For instance, if they computed means, they referred to the bar graph type as the option to represent the mean of the data set. If they considered that the number of points involving the data set was manageable, they plotted raw data in a scatter graph type
Ease of use	Participants referred to selecting a graph type because it was easy to draw the signifiers of that graph type. For instance, scatter graph types were easier to draw than bar graphs because their signifiers are dots rather than bars, and dots are easier to draw than rectangles

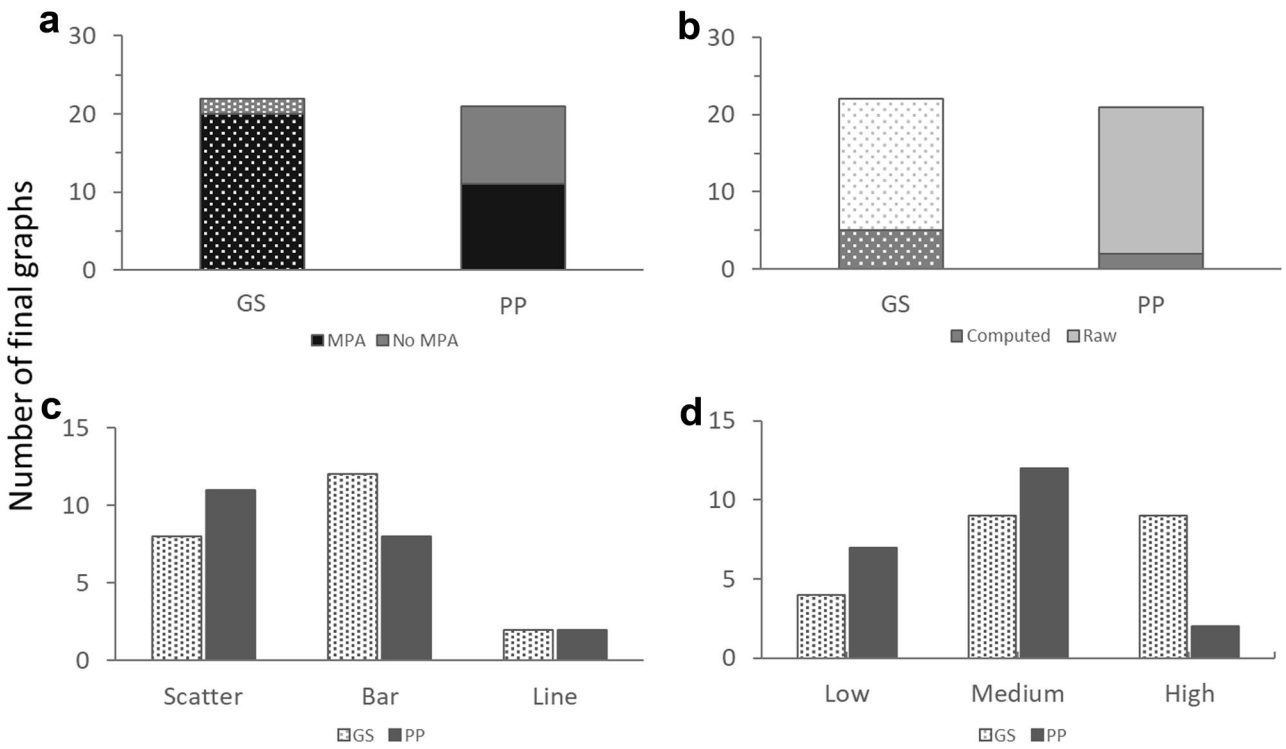

Fig. 5 Summary of graphs between the GS and PP environments. Displayed are the number of final graphs with the relevant variable, MPA (**a**), displaying raw or computed data (**b**) of different graph types (**c**), and that were in the low, medium, or high graph communication quality level (**d**). All stippled bars are GS and solid bars are PP

Graph communication. Using the graph quality rules in Fig. 3, most graphs in both graphing environments were categorized as medium-level (Fig. 5d). Although there is a visually apparent shift towards higher level graphs among GS participants compared with PP, this was not significant ($p = 0.0723$).

Potential Impact of GS Affordances on Data Exploration and Graph Communication

The most dramatic affordance of GS Potential Impact we observed was enabling students to explore the data by making many more graphs on average than in PP (Fig. 6). While no participant in PP made more than two graphs, and the vast majority (17/21) made but a single graph, almost half (10/22) of participants in GS made more than two graphs.

Table 4 Differences in graphing practices between formats

Graph practice	Comparison (PP vs GS)
Selection of variables	GS better
Form of the data	Similar
Acknowledging variability	GS better? (low N, no statistics)
Graph type	Similar
Graph communication	Similar (with GS better in some subsets of data)

Student comments supported the idea that digital tools facilitate exploration. For instance, one student in the PP treatment commented that they found it harder than making graphs by computer because with a program like Excel "you don't have to worry about putting dots in the right place and you can do like calculations easy and make your lines." (ID 7711, PP, Final graph = high). Several students in the Potential Impact treatment compared the tool with using Excel, with mixed thoughts on which was preferred but almost universally using the word "easy" in their comments. For instance, "[Excel] was a lot easier to play around with it and see which one you like" (ID 5507, GS, Final graph = high) from a student that preferred Excel, or "[GS] is easier [than Excel] because there are a lot of less noise there is just the buttons that you really need and that makes it very easy" (ID 9112, GS, Final graph = high). Regardless of which interface they preferred, students seem to associate digital graphing tools with ease of use.

It is a trickier question to assess whether the ease of making multiple graphs affected the participants' graphing practices. As reported above, although there were more graphs of the highest quality in GS, the differences in graph

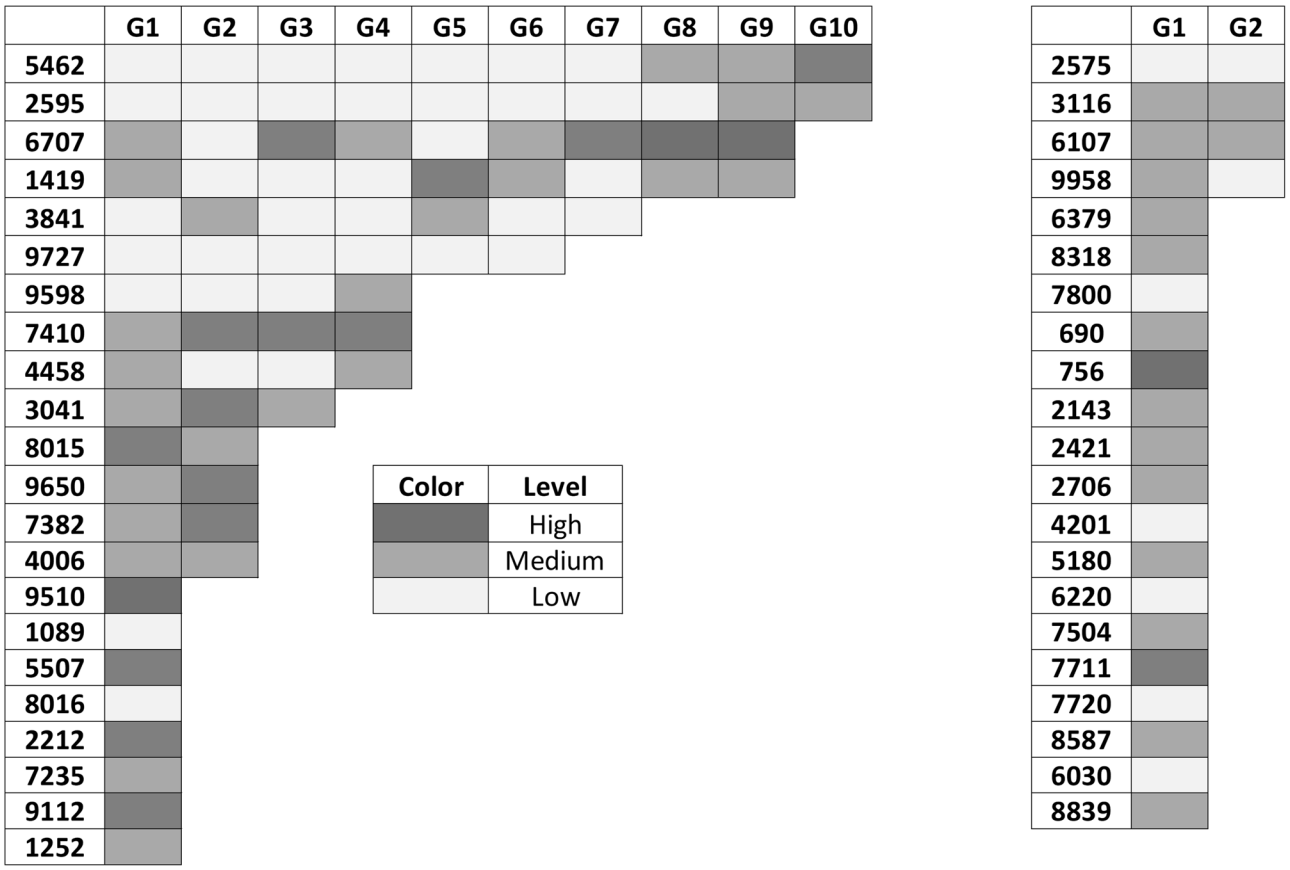

Fig. 6 Graph communication quality for succession of graphs made by each participant. GS participants on the left, and PP participants on the right. For each numbered participant, the rectangles each represent one graph made, in the order they were made from left to right. Shading represents the graph quality according to the rules shown in Fig. 3

communication quality between formats was not significant, overall. However, among GS participants who made more than two graphs, there appears to be a trend from lower quality initial graphs to higher quality final graphs (Fig. 6). In particular, none of those participants started with a high-quality graph, but 3/10 had a high-quality final graph.

Because the use of a digital tool enables quick data exploration and graphing, we explored the relationship between the time to create the first graph and the graphing environment. We did not see a significant difference between the two ($p > 0.05$). We further explored any relationship between graph quality level and time to complete the first graph and again saw no trend notable enough to comment on.

Graph Type Justifications Differed Between PP and GS Participants

We asked each participant to describe why they had selected a particular graph type (bar, line, scatter, etc.) in their final graph, and then classified these into four types of answers (see Methods). Many students from both PP and GS environments mentioned good data visualization and the characteristics of the data as rationales for the type of graph they chose, and we did not see much difference in the frequency of those responses between groups (Fig. 7). For instance, students made statements such as "A bar graph allowed me to see the average instead of all the points" (ID 7382, GS). The most common answer for participants using the PP format, though, related to how easy it was to draw a particular graph. PP participants picked graphs that they could draw with less effort, something that was much less common among GS participants. For instance, PP participants made statements such as "I ended up making a line graph to save time" (ID 6030, PP).

There were also more PP than GS students that mentioned being able to explore the data as a justification for the graph type they chose, but the difference between groups was smaller than with ease of use (Fig. 7). We did not see differences in graph type justifications in relation to graph quality that were large enough to comment on.

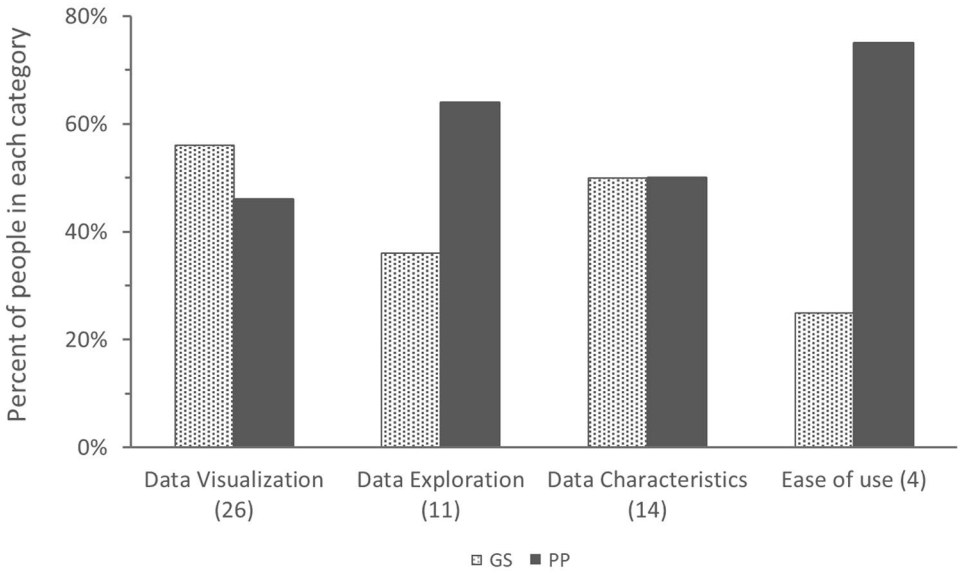

Fig. 7 Participants graph type justifications compared between graphing environments. Shown are the percent of people between the two graphing environments who stated a given justification for the graph type chosen. In parentheses are the total numbers of participants stating the justification. A given participant may have stated more than one justification category. Stippled bars are GS and solid bars are PP

Constraints of GS Can Both Reveal and Obscure Participants' Graphing Practices

As part of the design of the GS graphing tool to facilitate auto-scoring and to decrease the number of dimensions to be scored, we implemented a number of constraints in the GS environment compared with the PP environment (see Introduction), which had effects on student graphing practices. We saw this in the type and nature of the graphs some participants made in the PP environment which were not possible in the GS environment (11/21 participant graphs), such as graph types other than bar/scatter/line, graphs with multiple *y*-axes, multi-panel graphs comprised subsets of the data (e.g., one graph for MPA and another for non-MPA), trendlines and axis breaks, and plotting all of the variables (Fig. 8). The most common graphing difference was the inclusion of multiple variables (e.g., urchin and lobster density) plotted using the same *y*-axis (4/21 participant graphs) or by adding a second *y*-axis (1/21 participant graphs) (Fig. 8; 6220 and 8587). Of particular interest, the ability of PP environment graphers to plot more than three variables reveals more about their variable relevance graphing practice than might be captured in the constrained GS environment (4/21 participant graphs) (Fig. 8; 7711, 6220, 8587).

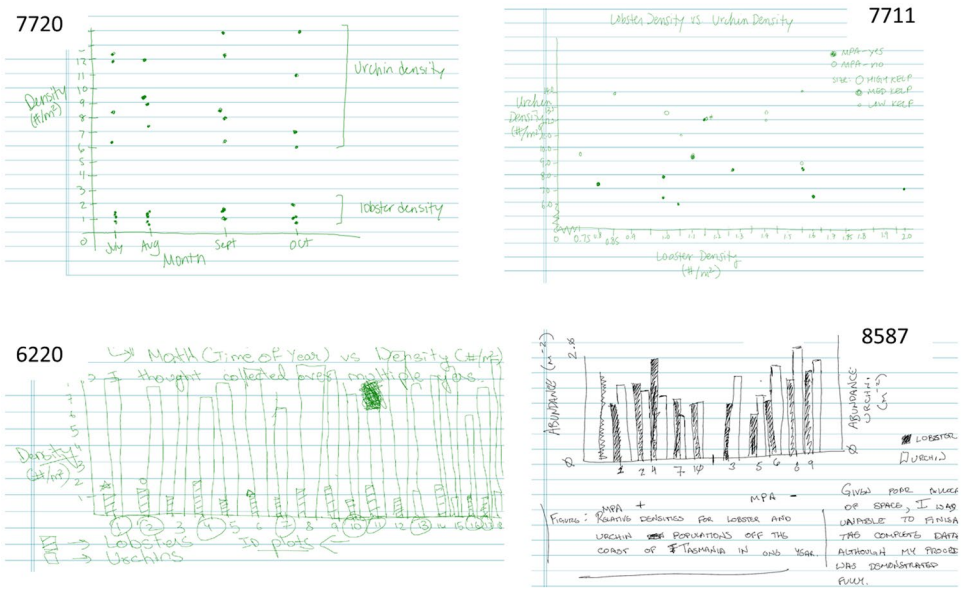

Fig. 8 Examples of graphs with features not possible in GS

Discussion

Undergraduate Students Performed Similarly on Several, But Not All, Criteria Between Graphing Environments

To address our first research question, asking about similarities and differences of student graphing practices in relation to the graphing environment, we compared the graphs students made in PP versus those in GS on five criteria. On three of the criteria, graph type, form of the data, and graph quality, there were no large differences in participant performance between environments. Plotting raw data rather than summarized data conforms with the findings of previous studies on student graph construction as well (Angra and Gardner, 2017). It could also be that participants considered the size of the data set (18 points per variable) small enough to plot every single data point. Future scenarios could include a larger data set to explore this graphing practice further.

Virtually all participants in GS plotted MPA status; while the majority of those in PP plotted MPA status as well, significantly more PP than GS participants chose not to plot MPA status. That most participants plotted MPA status indicates that there was high understanding among participants of the biological scenario and the question being posed, and it seems unlikely to us that more PP participants, uniquely, did not understand the scenario of data since both treatments read the background material in the identical digital environment. Although we speculate below on how the affordances and constraints of GS may have led to other differences, no obvious hypothesis comes to mind as to why the GS as opposed to the PP environment would have guided students towards selecting MPA status.

The other criteria where we noted a difference, albeit with a very small sample size, is in accounting for variability. Here, as we discuss more below, we speculate that features of GS may have led to or made it easier for students to plot error bars.

To generalize, it appears that many aspects of students' graphing practices can be captured similarly in a digital assessment as with pen-and-paper.

The Constraints and Affordances of Digital Assessments May Activate Latent Knowledge or Lead to Just-in-Time Learning

To address our second research question, asking how constraints and affordances of graphing environments may affect student graphing practices, we used both records of students graphing exploration as well as qualitative data from think-aloud interviews. Students in PP commented on how ease of drawing was a factor in the choices they made in constructing their graphs (see quote from ID 2143), and only in GS did students make more than two graphs, in many cases improving their graphs in the process. Thus, the affordance of GS in allowing quick construction of graphs may allow students to focus more on higher-order graph construction practices such as the shape of the distribution and the context of the inquiry (Friel et al., 2006; Aberg-Bengtsson, 2006), providing a better picture of their abilities and understanding.

As previous studies have reported (de Freitas and Sinclair, 2012, Sinclair and de Freitas, 2013; de Freitas and Sinclair 2014; Sinclair and Yurita, 2008) the constraints of a digital tool like GS may also affect the skill students demonstrate. Only certain options are available in GS, and the visibility of those options may remind students of those possibilities in constructing their graphs—both possibilities that lead to better graphs, and those that lead to worse graphs. There are indications that, on average, the constraints in GS prevented poor graphing or activated latent knowledge of graph construction that helped participants produce what we evaluated as high level graphs. Several PP participants (6/21 participants) may have benefited from the GS constraints in not being allowed to plot extra, irrelevant variables (4/21 participants) or a subset of the data (i.e., only MPA yes data, 2/21 participants).

The affordance of easier graphing and constraint options may also lead to learning through the process of taking the assessment. Of particular interest is that participants in GS who made more than two graphs tended to start out with worse graphs than those who made just one or two graphs (Fig. 5), and further, that those multiple graph participants tended to improve between their first and final graph. This may be an indication of activating latent skill with graphing practice concepts and/or a demonstration of learning within the assessment. We speculate that students who had more experiences graphing to begin with knew exactly what an appropriate graph would look like and went straight to constructing that graph. Those with less experience may have needed to play around more, but in the process of playing found their way (in many cases) to a higher quality graph. This interpretation is supported by previous studies involving the role of the graphing environment (Sinclair and de Freitas 2013), graphing experience (Roth and McGinn, 1997), work on other assessments showing such learning through intermediate constraint question formats (Meir et al., 2019) and in a constrained experimental design task (Meir, in preparation). While prior studies explicitly comparing science skill assessment by environment are rare, at least one prior study in a similar context also found that students performed better on digital assessment than on pen-and-paper, and

similarly speculated that the constraints and affordances aided students in drawing on latent knowledge (Kumar et al., 1994).

To generalize, the constraints and affordances in a digital assessment may aid students to show higher skill than in a pen-and-paper environment, and this may, on balance, provide a better representation of the students' true level of skill.

Limitations of the Study

The results we discuss above are general in that we demonstrate the existence of these similarities (such as similar assessments of certain practices) and differences (such as the affordance of easier graph creation) as a possibility. However, we only present data for a single, idiosyncratic digital tool, and a single data set and hypothesis against which students were asked to graph. Other tools and presenting stories may lead to different results. In particular, we could imagine different results if students were presented with larger data sets, or with different levels of constraint in the digital tool.

The comparison between the GS and pen-and-paper graphing environments was not meant to be a one-to-one comparison with graphing environment as the only variable. Rather, some of the features of the GS were based on our previous work (Angra and Gardner, 2016, 2017) and were meant to represent evidence-based improvements for instruction and assessment, as mentioned in the section above. Further, we narrowed the range of possible graphing decisions and actions in the GS graphing environment in order to interpret our data and discern patterns. Thus, it is likely that the differences between the two environments had additional indirect effects on the student graphing practices we observed. For example, while the ease of switching between graph types in GS may have promoted an exploration of the data for some participants, this may have led students to focus more on aesthetics rather than selecting a graph that was appropriate for the data and purpose (Tairab and Khalaf Al-Naqbi, 2004). The ease of calculating averages in the GS condition compared with the pen-and-paper condition may have affected the way in which students thought about the data and variation (Konold et al., 2015; Lehrer and Schauble, 2007). As a final example, the ability of students in the pen-and-paper condition to plot all the data and multiple *y*-axes could have affected their focus on carefully selecting the most relevant variables to plot and affected their ability to evaluate the prediction (Tairab and Khalaf Al-Naqbi, 2004; Patterson and Leonard, 2005). These examples highlight the importance of comparative research when introducing new technological approaches to student practices.

Finally, some of our conclusions are based on small numbers of students, and thus, the statistical results we present could change in a larger study. Power analysis indicates that we could detect moderate to larger treatment effects, so smaller differences in student behaviors between environments may not have been captured. But due to alignment of quantitative results here with the qualitative data we collected and the previous work of our own and others (Angra and Gardner, 2017; Aberg-Bengtsson, 2006), we think it likely that alternate effects which showed up in larger studies would be relatively small in magnitude.

Digital Performance-Based Assessments Have Potential to Algorithmically Assess Higher-Order Skills and Improve Evidence-Based Teaching

Our data are promising both for research on student understanding of higher-order skills and for helping instructors improve their classes using just-in-time data. On three criteria, student work was similar in the two environments. We also have hints that student work was a more authentic representation of student graphing practices in GS than PP, as evidenced by the number of graphs made, their justifications for their final graph, and the quality of graphs for those participants that only made one or two. Students who may have had the experience with graphing practices to make higher quality graphs did not do so in PP because of the effort involved. In GS, where effort was lower, such students were able to demonstrate their graphing practices more fully. Thus, GS may more authentically represent student graphing practice in some respects. The light constraints enable auto-scoring of the results (Meir, in preparation), which make such assessments feasible to use on a larger scale and on the faster timelines needed for instructors to modify their teaching on the fly. We expect that these advantages of GS are likely to be broader than our particular tool and these results should encourage the development of other digital performance-based assessments of higher-order skills using similar design criteria.

Our results also support using similar features in teaching tools, and in fact an assessment tool such as GS, with small modifications, might additionally serve as a teaching tool. The constraints serve as a form of scaffolding for student thinking about data and graphing and may address previous concerns related to students too rapidly producing graphs without reflection and thoughtful deliberation (Patterson and Leonard, 2005; Tairab and Khalaf Al-Naqbi, 2004).

Acknowledgments We thank numerous people at SimBiotic Software who helped write and release GraphSmarts and provided support for users, especially Steve Alison-Bunnell who did the bulk of the coding. Thanks to Isobel Buck for help with creating the GraphSmarts tool figure (Fig. 1). This work also benefited from collaborations initiated within the ACE-Bio Network (NSF RCN-UBE 1346567).

Author's Contributions CRediT: Stephanie M. Gardner: conceptualization, methodology, validation, formal analysis, resources, writing review and editing, visualization, supervision, project administration, funding acquisition; Elizabeth Suazo-Flores: conceptualization, methodology, validation, formal analysis, investigation, data curation, writing original draft, visualization; Susan Maruca: conceptualization, software; Joel K. Abraham: conceptualization, methodology, writing review and editing, funding acquisition; Anupriya Karippadath: formal analysis; Eli Meir: conceptualization, methodology, software, resources, data curation, writing original draft and review and editing, funding acquisition.

Funding This work was supported by the National Science Foundation under Grant #1726180.

Data Availability All data and materials used on this study are available from the authors, subject to IRB approval. Please write to sgardne@purdue.edu.

Code Availability All software discussed in this paper is available for evaluation purposes from the authors. Please email info@simbio.com to arrange for access.

Compliance with Ethical Standards

Conflicts of Interest SimBiotic Software, who employs authors SM and EM, may eventually develop a product for sale from some of the software discussed in this paper. No other authors have any conflicts of interest.

Ethics Approval All work was done accordance with an approved human studies protocole (IRB # 1706019374).

Consent to Participate All research participants were over 18 years of age, were not from vulnerable populations, and signed the informed consent form prior to participation ((IRB # 1706019374).

Consent for Publication Included in the written and signed informed consent forms (IRB # 1706019374).

References

Aberg-Bengtsson, L. (2006). "Then You Can Take Half... Almost"—elementary students learning bar graphs and pie charts in a computer-based context. *Journal of Mathematical Behavior, 25,* 116–135.

Angra, A., & Gardner, S. M. (2016). Development of a framework for graph choice and construction. Advances in Physiology Education, 40(1), 123-128.

Angra, A., & Gardner, S. M. (2017). Reflecting on graphs: Attributes of graph choice and construction practices in biology. CBE-Life Sciences Education, 16(3), ar53.

American Association for the Advancement of Science (AAAS) (2009). Conference Homepage. *Vision and change in undergraduate biology education: A view for the 21st century*. (accessed July 7, 2019).https://www.visionandchange.org

Beggrow, E.P, Ha, M., Nehm, R.H., Pearl D, & Boone, W.J. (2014) Assessing scientific practices using machine-learning methods: how closely do they match clinical interview performance? *Journal of Science Education and Technology*, v23 n1 p160–182 Feb 2014.

Bowen, G. M., & Roth, W. M. (2005). Data and graph interpretation practices among preservice science teachers. *Journal of Research in Science Teaching, 42*(10), 1063–1088. https://doi.org/10.1002/tea.20086.

Bowen, G. M., Roth, W. M., & McGinn, M. K. (1999). Interpretations of graphs by university biology students and practicing scientists: toward a social practice view of scientific representation practices. *Journal of Research in Science Teaching, 36*(9), 1020–1043. https://doi.org/10.1002/(sici)1098-2736(199911)36:9%3c1020::aid-tea4%3e3.0.co;2-#.

Brasell, H., & Rowe, M. (1993). Graphing skills among high school physics students. *School Science and Mathematics, 93*(2), 62–70.

Chick, H. (2004). Tools for transnumeration: early stages in the art of data representation. In Putt, I., Faragher, R., & McLean, M. (Eds.), *Mathematics Education for the Third Millennium: Towards 2010. Proceedings of the Twenty-seventh Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 167–174). Sydney, Australia: MERGA.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly, 104*(9), 801–823. https://doi.org/10.1080/00029890.1997.11990723.

Cromley, G. J., Du, Y., & Dane, A. P. (2020). Drawing-to-learn: does meta-analysis show differences between technology-based drawing and paper-and-pencil drawing? *Journal of Science Education and Technology, 29,* 216–229. https://doi.org/10.1007/s10956-019-09807-6.

D'Ambrosio, B., Kastberg, S. E., McDermott, G., & Saada, N. (2004). Beyond reading graphs: student reasoning with data. In P. Kloosterman & F. Lester (Eds.), *Results and interpretations of the 1990–2000 mathematics assessments of the National Assessment of Educational Progress* (pp. 363–381). Reston, VA: NCTM.

Diong, J., Butler, A. A., Gandevia, S. C., & Héroux, M. E. (2018). Poor statistical reporting, inadequate data presentation and spin persist despite editorial advice. *PLoS ONE, 13*(8), e0202121. https://doi.org/10.1371/journal.pone.0202121.

de Freitas, E., & Sinclair, N. (2012). Diagram, gesture, agency: theorizing embodiment in the mathematics classroom. *An International Journal, 80*(1), 133–152. https://doi.org/10.1007/s10649-011-9364-8.

de Freitas, E. D., & Sinclair, N. (2014). *Mathematics and the body: material entanglements in the classroom*. Cambridge University Press.

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

Friel, S. N., O'Connor, W., & Mamer, J. D. (2006). More than "Meanmedianmode" and a bar graph: what's needed to have a statistical conversation. In Burrill, G., & Portia C. E. (Eds.), *Thinking and reasoning with data and chance, 68th. Yearbook* (pp. 117–137). Reston, VA: NCTM.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 22–38.

Gelman, A., & Unwin, A. (2013). Infovis and statistical graphics: different goals, different looks. *Journal of Computational and Graphical Statistics, 22*(1), 2–28. https://doi.org/10.1080/10618600.2012.761137.

Guimaraes, B., Ribeiro, J., Cruz, B., Ferreira, A., Alves, H., Cruz-Correia, R., et al. (2018). Performance equivalency between computer-based and traditional pen-and-paper assessment: a case study in clinical anatomy. *Anatomical Sciences Education, 11*(2), 124–136.

Harwell, M., Moreno, M., Phillips, A., Guzey, S. S., Moore, T. J., & Roehrig, G. H. (2015). A study of STEM assessments in engineering, science, and mathematics for elementary and middle school students. *School Science and Mathematics, 115*(2), 66–74.

Ha, M. & Nehm, R.H. (2016) The impact of misspelled words on automated computer scoring: a case study of scientific explanations. *Journal of Science Education and Technology*, v25 n3 p358–374 Jun 2016

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Kjelvik, M. K., & Schultheis, E. H. (2019). Getting messy with authentic data: exploring the potential of using data from scientific research to support student data literacy. *CBE - Life Sciences Education, 18*(2), es2. doi:https://doi.org/10.1187/cbe.18-02-0023

Kim, K. J., Pope, D. S., Wendel, D., & Meir, E. (2017). WordBytes: Exploring an intermediate constraint format for rapid classification of student answers on constructed response assessments. *Journal of Educational Data Mining, 9*(2), 45–71.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33*(4), 259–289.

Konold, C., Higgins, T., Russell, S. J., & Khalil, K. (2015). Data seen through different lenses. *Educational Studies in Mathematics, 88*(3), 305–325.

Kosslyn, S. M. (1985). Graphics and human information processing: a review of 5 books. *Journal of the American Statistical Association, 80,* 499.

Kosslyn SM. Elements of Graph Design. New York: WH Freeman, 1994.

Kuhn, D. (2010). Teaching and learning science as argument. *Science Education, 94*(5), 810–824.

Kumar, D. D., White, A. L., & Helgeson, S. L. (1994). A study of the effect of hypercard and pen-paper performance assessment methods on expert-novice chemistry problem solving. *Journal of Science Education and Technology, 3*(3), 187–200.

Lehrer, R., Schauble, L., & Petrosino, A. J. (2001). Reconsidering the role of experiment in science education. In K. Crowley, C. Schunn, & T. Okada (Eds.), *Designing for science: implications from everyday, classroom, and professional settings* (pp. 251–278). Mahwah, N.J.: Lawrence Erlbaum Associates.

Lehrer, R., & Schauble (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In: Thinking with data Lovett, M. (Ed.), Shah, P. (Ed.). New York: Psychology Press. https://doi.org/10.4324/9780203810057

Lehrer, R., Schauble, L., & Lucas, D. (2008). Supporting development of the epistemology of inquiry. *Cognitive development, 23*(4), 512–529.

Mayes, R. L., Forrester, J. H., Christus, J. S., Peterson, F. I., Bonilla, R., & Yestness, N. (2014). Quantitative reasoning in environmental science: a learning progression. *International Journal of Science Education, 36*(4), 635–658. https://doi.org/10.1080/09500693.2013.819534.

Mislevy, R. J. (2013). Evidence-centered design for simulation-based assessment. *Military medicine*, *178*(suppl_10), 107–114.

Moore, D. S. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review, 65*(2), 123–137. https://doi.org/10.1111/j.1751-5823.1997.tb00390.x.

Meir, E., Wendel, D., Pope, D. S., Hsiao, L., Chen, D., Kim, K. J. (2019). Are intermediate constraint question formats useful for evaluating student thinking and promoting learning in formative assessments? Computers & Education, 141, 1-21

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: advantages of longhand over laptop note taking. *Psychological Science, 25*(6), 1159–1168. Retrieved from: https://doi.org/10.1177/0956797614524581

Oqvist, M., & Nouri, J. (2018). Coding by hand or on the computer? Evaluating the effect of assessment mode on performance of students learning programming. *Journal of Computers in Education, 5,* 199–219.

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

Patterson, T. F., & Leonard, J. G. (2005). Turning spreadsheets into graphs: an information technology lesson in whole brain thinking. *Journal of Computing in Higher Education, 17,* 95–115.

Patton, M. Q. (2015). Qualitative research & evaluation methods: integrating theory and practice (Fourth edition. ed.). SAGE Publications, Inc.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Roth, W.-M., & Bowen, G. (2001). Professionals read graphs: a semiotic analysis. *Journal for Research in Mathematics Education, 32*(2), 159. https://doi.org/10.2307/749672.

Roth, W.-M., & Hwang, S. (2006). On the relation of abstract and concrete in scientists' graph interpretations: a case study. *Journal of Mathematical Behavior, 25*(4), 318–333. https://doi.org/10.1016/j.jmathb.2006.11.005.

Roth, W. M., & McGinn, M. K. (1997). Graphing: cognitive ability or practice? *Science Education, 81*(1), 91–106. https://doi.org/10.1002/(SICI)1098-237X(199701)81:1%3c91::AID-SCE5%3e3.0.CO;2-X.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: aa framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6), 3–44.

Seraphin, K. D., Philippoff, J., Parisky, A., Degnan, K., & Warren, D. P. (2013). Teaching energy science as inquiry: reflections on professional development as a tool to build inquiry teaching skills for middle and high school teachers. *Journal of Science Education and Technology, 22*(3), 235–251.

Shaughnessy, J. M. (2006) Research on students' understanding of some big concepts in statistics. In Burrill, G., & Portia C. E. (Eds.), *Thinking and reasoning with data and chance, 68th. Yearbook,* (pp. 77–98). Reston, VA: NCTM.

Shanahan, C., Shanahan, T., & Misischia, C. (2011). Analysis of expert readers in three disciplines: history, mathematics, and chemistry. *Journal of Literacy Research, 43*(4), 393–429. https://doi.org/10.1177/1086296X11424071.

Sinclair, N. & de Freitas, E. (2013). The virtual curriculum: New ontologies for a mobile mathematics. Mathematics Curriculum in School Education. New York: Springer.

Sinclair, N., & Yurita, V. (2008). To be or to become: How dynamic geometry changes discourse. Research in Mathematics Education, 10(2), 135-150. https://doi.org/10.1080/14794800802233670

Strauss, A., & Corbin, J. (1994). Grounded theory methodology: an overview. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 273–285). Thousand Oaks, CA: Sage.

Tairab, H. H., & Al-Naqbi, A. K. (2004). How do secondary school science students interpret and construct scientific graphs? *Journal of Biological Education, 38,* 127–132.

Tufte, E. (1983). *The visual display of quantitative information*. Connecticut: Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.: Addison-Wesley Pub. Co.

Urban-Lurain, M., Prevost, L., Haudek, K. C., Henry, E. N., Berry, M., & Merrill, J. E. (2013). Using computerized lexical analysis of student writing to support Just-in-Time teaching in large enrollment STEM courses, *2013 IEEE Frontiers in Education Conference (FIE). Oklahoma City, OK, 2013,* 1709–1715.

Vitale J.M., Lai K & Linn M.C. (2015) Taking advantage of automated assessment of student-constructed graphs in science: AUTO ASSESSMENT OF STUDENT GRAPHS. *Journal of Research in Science Teaching* 52(10) https://doi.org/10.1002/tea.21241

Vitale, J. M., Applebaum, L., & Linn, M. C. (2019). Coordinating between graphs and science concepts: density and buoyancy. *Cognition and Instruction, 37*(1), 38–72. https://doi.org/10.1080/07370008.2018.1539736.

Watson, J. M., & Moritz, J. B. (1998). The beginning of statistical inference: comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145–168.

Weston, M., Haudek, K. C., Prevost, L., Urban-Lurain, M., & Merrill, J. (2015). Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. *CBE life sciences education*, *14*(2), ar19. https://doi.org/10.1187/cbe.14-07-0110

Watson, J., Kelly, B., Callingham, R., & Shaughnessy, M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematics Education in Science and Technology, 34*(1), 1–29.

Weiland, T. (2017). The importance of context in task selection. *Teaching Statistics, 39*(1), 20–25. https://doi.org/10.1111/test.12116.

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Bio, 13*(4), e1002128. https://doi.org/10.1371/journal.pbio.1002128.

Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., et al. (2019). Milic NM (2019) reveal, don't conceal: transforming data visualization to improve transparency. *Circulation, 140,* 1506–1518. https://doi.org/10.1161/CIRCULATIONAHA.118.037777.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248.

Wild, C., Utts, J., & Horton, N. (2018) What is statistics? In D. Ben-Zvi et al. (Eds.), *International handbook of research in statistics education,* (pp. 5–35). Springer International Handbooks of Education. Retrieved from: https://doi.org/10.1007/978-3-319-66195-7_1

Windschitl, M., Dvornich, K., Ryken, A. E., Tudor, M., & Koehler, G. (2007). A comparative model of field investigations: aligning school science inquiry with the practices of contemporary science. *School Science and Mathematics, 107*(1), 382–390.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757.

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the college level? A review of the literature, *Journal of Statistics Education, 16*(2). https://doi.org/10.1080/10691898.2008.11889566