The Graph Rubric: Development of a Teaching, Learning, and Research Tool

Aakanksha Angra[†] and Stephanie M. Gardner[‡]*
[†]Department of Biology, Georgia State University, Atlanta, GA 30303; [‡]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907

ABSTRACT

As undergraduate biology curricula increasingly aim to provide students with access to courses and experiences that engage them in the practices of science, tools are needed for instruction, evaluation, and research around student learning. One of the important skills for undergraduate biology students to master is the selection and creation of appropriate graphs to summarize data they acquire through investigations in their course work and research experiences. Graphing is a complex skill, and there are few, discipline-informed tools available for instructors, students, and researchers to use. Here, we describe the development of a graph rubric informed by literature from the learning sciences, statistics, representations literature, and feedback and use of the rubric by a variety of users. The result is an evidence-based, analytic rubric that consists of categories essential for graph choice and construction: graph mechanics, graph communication, and graph choice. Each category of the rubric can be evaluated at three levels of achievement. Our analysis demonstrates the potential for the rubric to provide formative feedback to students and allow instructors to gauge and guide learning and instruction. We further discuss and identify potentially interesting research targets for science education researchers.

INTRODUCTION

Reforms to biology education from K–12 through undergraduate levels call for students taking part in the practices of science, including inquiry and quantitative data analysis, interpretation, and decision making (American Association for the Advancement of Science [AAAS], 2011; College Board, 2011; National Research Council, 2011; Common Core State Standards Initiative, 2012; Next Generation Science Standards Lead States [NGSS], 2013). Furthermore, there are calls for all undergraduate students to participate in research (AAAS, 2011; President's Council of Advisors on Science and Technology, 2012; Howard Hughes Medical Institute, 2013), an experience that will ultimately engage them in data analysis and communication of their findings. As part of these reforms, students will need to develop quantitative literacy skills, such as graphing, to enable them to solve problems and ask questions using quantitative evidence and methods (American Association of Colleges and Universities, 2010). Therefore, graphical literacy and competence is essential for undergraduate biology students and an important life skill for non–science majors, as well.

Graphing skills can be broadly separated into graph interpretation and graph construction. The interpretation of graphs requires cognitive engagement in statistical, experimental, and proportional reasoning in addition to visuospatial skills (Shah *et al.*, 1999; Garfield, 2003; Garfield *et al.*, 2007; Bengtsson and Ottosson, 2006). The graph must be decoded to allow for the extraction of information to make inferences (Friel and Bright, 1996; Shah *et al.*, 1999). Graph construction is a more complex, generative task involving the integration of knowledge, skills, and reasoning from many content areas and incorporating broader thinking about experiments and/or inquiry. The graph constructor needs to draw on knowledge of graphical representations (i.e., representational competence), spatial and proportional reasoning, and statistical and quantitative

Jennifer Knight, *Monitoring Editor*
Submitted Jan 8, 2018; Revised Aug 13, 2018; Accepted Sep 11, 2018
CBE Life Sci Educ December 1, 2018 17:ar65
DOI:10.1187/cbe.18-01-0007
*Address correspondence to: Stephanie M. Gardner (sgardne@purdue.edu).

skills (Tufte, 1983; Mathewson, 1999; Garfield, 2003; Garfield *et al.*, 2007; Bengtsson and Ottosson, 2006). For example, choosing an appropriate graph type to represent data requires an understanding of the variable types to be plotted (e.g., categorical vs. continuous), a consideration of the purpose for graphing the data (e.g., the research question), knowledge of different ways to statistically summarize data, and a basic knowledge of graph types that exist for the type of data to be plotted (diSessa and Sherin, 2000; diSessa, 2004; Grawemeyer and Cox, 2004; Novick, 2004). Further, the nature of the variables and the approaches and measurements used to acquire them play a role in graph construction. Another important feature of a well-constructed graph is its visual appearance. Aesthetic and spatial aspects of graphs impact visual processing and interpretation and need to be considered to ensure clear communication of the data (Tufte, 1983; reviewed in Montello *et al*., 2014). A well-constructed graph will not only be aesthetically pleasing but will also leverage Gestalt principles (e.g., proximity and continuity; Kellman, 2000; reviewed by Hegarty, 2011), which facilitate the global and local visual analysis that is a natural feature of human visual processing (Franconeri *et al.*, 2012). Finally, the construction of high-quality and meaningful graphs also requires reflective processes that ensure that the form of the data plotted (Konold *et al.*, 2015) and graph chosen are aligned with its purpose for the creator and readers of the graph (Angra and Gardner, 2016, 2017). This reflective piece extends the representational competence needed for graph choice and construction to metarepresentational competence (diSessa and Sherin, 2000; diSessa, 2004), which is implicitly practiced by experts (Angra and Gardner, 2017).

Students (Bray-Speth *et al.*, 2010; McFarland, 2010; Gormally *et al.*, 2012) and even experts (Roth and Bowen, 2001; Rougier *et al.*, 2014; Weissgerber *et al.*, 2015) struggle when choosing appropriate graphs to display their data. Indeed, reviews of primary literature articles published in science, technology, engineering, and mathematics journals have documented the overuse and/or the inappropriate use of certain graph types and data representations (Cooper *et al.*, 2001, 2003; Puhan *et al.*, 2006). Further, several journals have featured articles with guidelines for scientists on graph construction in an effort to improve the clarity of data communication (*PLoS Biology*, *PLoS Computational Biology*, *BioMed Central*). Related to the graph type is the form in which the data are plotted. There is currently a backlash against the overreliance and overinterpretation of descriptive and inferential statistics in the scientific community (e.g., Klaus, 2015, 2016; Saxon, 2015; Weissgerber *et al.*, 2015). While experts have room for improvement in this area, graph creation is far easier for them, given their knowledge of the system under study, data analysis and statistics, and the research question they are addressing (Konold *et al.*, 2015). However, students may lack a sufficient understanding of data (Dasgupta *et al.*, 2014) and statistical techniques, or the proper context given details of the study system or audience (Lovett and Chang, 2007), leading to differences in graph construction decision making between novices and experts (Konold and Lehrer, 2008; Konold *et al.*, 2015; Angra and Gardner, 2016, 2017).

Despite the ubiquity and importance of graphs in science, instructors do not regularly use time in class to engage and enculturate students into the norms and behaviors of experts in graph construction and interpretation (Bowen and Roth, 1998). Further, instructors tend to use oversimplified graphs and fail to deconstruct and analyze figures with students (Bowen and Roth, 1998). This lack of graphical enrichment limits students' experiences of dealing with the "messiness" that comes with data from biological experiments and the statistical and quantitative techniques used to summarize, analyze, and interpret those data. Commonly used software with graphing features can exacerbate the problems by facilitating quick decision making without thoughtful reflection on the multidisciplinary concepts that are part of data representation. Previous work within the statistics and science education communities, including our own, has revealed some of the basic areas in which undergraduate students have difficulty (Cobb *et al.*, 2003; Lehrer and Schauble, 2004; Novick, 2004; McFarland, 2010; Angra and Gardner, 2016, 2017).

Resources exist to help both students and practitioners increase their competence with graph format selection and construction. These include instructional books (Bertin, 1983; Tufte, 1983; Kosslyn, 1994; Few, 2004) and Web-based interactive tools and modules such as TinkerPlots, BeSocratic (graphical thinking), and CODAP (Common Online Data Analysis Platform; Concord Consortium). As mentioned previously, current recommendations from research journals aim *generally* to promote the creation of better data visualizations, including graphs (Cooper *et al.*, 2001, 2002; Puhan *et al.*, 2006; Rougier *et al.*, 2014; Slutsky, 2014; Saxon, 2015; Weissgerber *et al.*, 2015; Klaus, 2016; Nuzzo, 2016). As such, these resources rarely focus on the complex reasoning behind graph choice and construction, nor are they grounded in the concepts and measures of a particular discipline. It is therefore difficult to choose an appropriate graph for data (e.g., bar graph for summarized categorical data), without evaluating the advantages and disadvantages of using a particular graph within the context of a given scientific discipline or audience.

The multifaceted and complex nature of graphing makes it difficult for instructors to diagnose student difficulties and for students to master the skill of graphing. There have been scattered efforts to identify and address student difficulties with graphing. For example, Vitale and colleagues (2015) have developed an automated digital tool to evaluate line graphs created by middle and high school students in chemistry and physics classrooms. Their tool can provide quick feedback to researchers and instructors about difficulties that students have based on the slope and trajectories of the lines graphed. However, the tool is limited by graph types and the scientific concepts they model, which are distinct from graphing data from experiments, predictions, or explanations. For example, the data structure in data summary graphs (e.g., bars, points, box and whisker) is an abstraction and distinct from the identity of the data and system in which they were generated. Scaffolded instruction at the undergraduate level has been somewhat successful in increasing the graph interpretation and construction competence of students during part of (Bray-Speth *et al.*, 2010; McFarland, 2010) or an entire (Harsh and Schmitt-Harsh, 2016) semester. However, continued guided and reflective practice over a longer period of time has been recommended (Roth *et al.*, 1999; diSessa, 2004). Therefore, there is a need for additional tools to aid in graphing instruction and research that have broad applicability.

Designing tools for instruction can be done easily through the rubric format, which is commonly used in diverse settings and by a variety of users. Rubrics are commonly used in the classroom by both instructors and students (Jonsson and Svingby, 2007; Panadero and Jonsson, 2013; Brookhart and Chen, 2015), but can be used for program evaluation (e.g., PULSE rubrics; Aguirre *et al.*, 2013; Brancaccio-Taras *et al.*, 2016) and in research (Dasgupta *et al.*, 2014; Ashley *et al.*, 2017), as examples. Rubrics allow for transparency of expectations and a level of objectivity in evaluation, and they require minimal time on the part of the user (Allen and Tanner, 2006; Dawson, 2017). Regardless of the specific type of rubric (i.e., analytic or holistic), the purpose and context of its use, or the user, rubrics typically have the following design features: an articulation of the categories under which something will be evaluated, a definition of the quality of different levels of achievement, and a scoring strategy (Popham, 1997; Mertler, 2001; Allen and Tanner 2006; Allen and Knight, 2009; Jonsson and Svingby, 2007; Panadero and Jonsson, 2013; Brookhart and Chen, 2015; Dawson 2017).

We developed an analytic graph rubric with three levels of achievement for each of the graphing subcategories. The objectives for the design of the rubric were to create a tool that would 1) facilitate the teaching and evaluation of data summary graphs, 2) provide undergraduate students with formative and summative feedback on their graphs, and 3) allow education researchers to evaluate graphing artifacts to assess experimental and quantitative skills. In this article, we describe the rubric development process, the sources of validity and reliability evidence we gathered, and the insights we gained related to the scope of use and potential of the rubric to improve student competence with graph construction.

METHODS

All work with human subjects, as appropriate, was performed under approved protocols (IRB#1210012775 and #1803020378). During the process of designing the graph rubric, we gathered validity and reliability evidence so that we, and others, could use the graph rubric in teaching and research. Validity in our study is "the relationship between the content of the test and the construct it is intended to measure" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME], 2014, p. 14). In the context of our work, we sought validity evidence to ensure that the graph rubric has appropriate categories, descriptions, and guidelines that can be used to measure and assess student understanding and application of concepts and skills relevant to graph choice and construction. To this end, our design process involved establishing construct validity, which refers to the claim that the content and features of the instrument (i.e., the graph rubric) are well supported with evidence (Benson, 1998; AERA, APA, and NCME, 2014, p. 11). In support of our overall claim of construct validity for the graph rubric as a tool to evaluate graphs, we gathered evidence for content and face validity. Establishing content validity involves gathering data in support of the claim that the instrument includes all relevant features of the subject under examination (Benson, 1998). In our case, we consulted diverse sources to ensure that the graph rubric encompasses appropriate criteria or content used to evaluate graphs (Table 1). We also approached diverse users to gather evidence of face validity, which is the ability to conclude that an instrument (i.e., the graph rubric) is appropriate and effective in achieving its aims (Holden, 2010). While the rubric is not a test instrument, our design and construct validation process was informed by the instrument design literature and its application in discipline-based education research (Benson, 1998; Corwin *et al.*, 2015) and consisted of three stages: 1) substantive, 2) structural, and 3) external. Although this process generally follows a linear path, there were cycles of revision and repetition of some stages. These design stages, our activities, and the types of validity evidence they contribute to are summarized in Table 1. As part of the evaluation of the construct validity of the rubric, we used interrater reliability (IRR) with a diverse group of users to understand consistency in judgment and scoring of graphs using the rubric (Holsti, 1969; Jonsson and Svingby, 2007; see *Data Analysis* below).

Stage 1. Substantive Stage: Identifying Graphing Elements by Consulting the Literature and Ongoing Research

The substantive stage led to the initial draft of the graph rubric with its categories, subcategories, and definitions. Three sources of information contributed to this stage and supplied content validity evidence for the concepts within the rubric (Table 1). We consulted the graphing and visual representations literature, student-generated graphs and reflections from a classroom study (Angra and Gardner, 2015; Angra, 2016), and graphs and the articulated reasoning constructed by students and professors in a think-aloud clinical graphing interview (Angra and Gardner, 2017).

We began the process of rubric development by consulting books and primary literature that discuss appropriate graphing practices. Because graphs are ubiquitous in many fields, we did not restrict our literature search to biology at this stage. When doing our literature search for articles on graphing, we consulted Google Scholar and the university's online library for article recommendations. We searched broadly for articles using keywords including "graph," "construction," "choice," "presentation," "science," and "practices." We then extended our research by consulting the reference sections in the articles. We read each reference, made notes on the authors' recommendations on proper graph choice and construction practices, and grouped similar recommendations together. As graphs are visual representations of data, we consulted select seminal work in the visual representations literature to identify theory and best practices (e.g., Tufte, 1983; diSessa, 2004).

To supplement the literature review and aid in rubric development, we used data from two ongoing graphing studies (Angra, 2016; Angra and Gardner, 2017). Briefly, the first graphing study took place in a physiology laboratory in which students produced graphs from their experimental data. Specifically, we were interested in the general qualities of the graphs produced (graph type, data plotted, overall appearance, understanding of the take-home message) and student reasoning for graphs they produced (Angra, 2016). The second graphing study was an expert–novice analysis conducted to understand how professors and students constructed and reflected on their graphs in a think-aloud interview setting (Angra and Gardner, 2017).

TABLE 1. Process of graph rubric design and construct validation with the three stages for graph rubric construct validation defined, the associated steps taken for each stage presented, and places in which evidence for content and face validity were obtained in support of the construct validation indicated

Stage and purpose	Type of validity	Activities and sources of evidence
1. Substantive Review literature and data to establish the graph rubric categories, subcategories, and definitions.	1. Content validity: Assurance from diverse sources that the graph rubric encompasses appropriate criteria or content used to evaluate graphs	1. Review of graphing and visualizations literature formed the initial basis of the rubric 2. Mine classroom data: Graph artifacts and student reflections on graph choice (Angra, 2016) 3. Mine clinical interview data: Graph artifacts and themes from student and professor graphing interviews (Angra and Gardner, 2017)
2. Structural Solicit feedback from diverse audiences. Revise rubric categories and descriptions, as needed.	1. Face validity: The quality enabling diverse users to conclude that the purpose of the rubric is to evaluate graphs 2. Content validity: Assurance from diverse sources that the graph rubric encompasses appropriate criteria or content used to evaluate graphs	Solicit input to establish content and face validity from: 1. Science education scholar feedback from rubric use 2. Non-education graduate student feedback from rubric use 3. Undergraduate student feedback from use of the rubric in the classroom (Spring 2015) and on the graph rubric categories, usability, and utility 4. Biology instructor feedback on the graph rubric categories, usability, and utility
3. External Evaluate the rubric by using it to assess a diversity of graphs. Confirm the features and structure of the graph rubric as appropriate and useful for evaluating graphs.	1. Face validity: The quality enabling diverse users to conclude that the purpose of the rubric is to evaluate graphs 2. Content validity: Assurance from diverse sources that the graph rubric encompasses appropriate criteria or content used to evaluate graphs	Rubric use by different stakeholders and to evaluate diverse graphs: 1. Undergraduate student evaluation of graphs generated in a class they had taken previously 2. Biology instructor evaluation of student-generated graphs from their courses 3. Evaluation of graphs from selected chapters from various introductory biology texts

Stage 2. Structural Stage: Soliciting Feedback to Establish Content and Face Validity

During this stage, we sought content and face validity evidence to convince us that the rubric contents and structure were appropriate and relevant for evaluating graphs in biology (Table 1). We accomplished this by seeking feedback on the rubric from four different groups of people: 1) science education scholars, 2) non–education research biology graduate students who were actively pursuing either a master's or doctoral degree, 3) undergraduate biology students enrolled in an upper-level physiology laboratory course, and 4) biology instructors. Incorporating feedback from participants at various levels of education and with expertise in various fields allowed us to check the learning goals and usability of the rubric. Feedback from students allowed us to make sure that the language in the rubric was clear and easy to understand.

Science Education Scholars. Drafts of the rubric were presented to an interdepartmental biology education research group of science education scholars (Table 1) that includes chemistry and biology education graduate students and postdoctoral fellows and instructors from the department of curriculum and instruction, biology, and chemistry. The reason for sharing the graph rubric with science education scholars was to obtain feedback from people with pedagogical expertise. The objective of the first meeting with this group was to obtain targeted feedback on the first draft of the graph rubric. In the first draft, we used a binary scale (i.e., present/not present) for the mechanics category and three levels of achievement for the other categories. We presented two de-identified student graphs (Graphs 1 and 2 in Appendix C, Supplemental Material) produced by different student groups in a physiology laboratory course along with a brief overview of the students' experimental designs and variables associated with that particular laboratory context. Each science education scholar was instructed to independently use the graph rubric to evaluate both student graphs, then pair and discuss their ratings with a partner; this was followed by a group discussion guided by A.A. and S.M.G. The guided group discussion began with broad questions to solicit feedback from the participants about rubric use, appropriateness, and descriptions of the rubric categories and subcategories. Percent agreement as an estimate of IRR between the science education scholars and authors was calculated after the meeting to gauge consistency in rubric scoring across the categories (see *Results*). IRR scores from the first meeting were low and are not reported in this article, but conversations about rubric scoring are provided in the *Results* section, as they were fruitful for rubric revisions.

After the initial round of feedback, the rubric categories and subcategories were expanded and refined based on comments from science education scholar group, further literature review, and ongoing graphing research (Table 1). We standardized the levels of achievement to three categories: "present/appropriate," "needs improvement," and "unsatisfactory." In addition, we adjusted the weighting of the scoring of the subcategories across the three main categories of the rubric to reflect the level of cognitive demand; scoring of items in the "mechanics" category is weighted less than scoring of items in the "communication" and "graph choice" categories (Figure 1). Using similar protocols but at a later time, the science education scholars were asked to use the revised rubric to evaluate Graph 3 (Appendix C, Supplemental Material).

Research Question:
Hypothesis:

	Present/Appropriate (P/A) = 0.5 pts Present but Needs Improvement (NI) =0.25 pts Unsatisfactory (U)= 0 pts	GRAPH 1			GRAPH 2		
		P/A	NI	U	P/A	NI	U
Graph Mechanics	**Descriptive title** • P/A-Should be: a) in the form of a statement, b) mention the subject, c) appropriate variables, and d) include relevant details about the experiment that help understand the take home message. • NI- If the title is missing any one of the four points mentioned above. • U-The graph does not have a title.						
	Label for the X axis (e.g. time) • P/A- Should be appropriate and descriptive for the experiment. For graphs with categorical independent variables, there needs to be a label under each set of data and a larger label under all data plotted. • NI- If the label is missing any one of the points mentioned above. • U- The graph does not have an x-axis label.						
	Label for the Y axis (e.g. heart rate) • P/A- Should be appropriate and descriptive for the experiment. If the data is manipulated (average, change, percentage, etc.), then it should be indicated on the y axis. • NI- If the label is missing any one of the points mentioned above. • U- The graph does not have a y-axis label.						
	Units for the X axis (e.g. seconds) • P/A- Should be appropriate and descriptive for the data displayed. • NI- If the units are not appropriate or descriptive. • U- The graph does not have units for the x-axis.						
	Units for the Y axis (e.g. average beats per minute) • P/A- Should be appropriate and descriptive for the data displayed. • NI- If the units are not appropriate or descriptive. • U- The graph does not have units for the y-axis.						
	Scale (appropriate intervals and range for data) • P/A- Should be appropriate for the data displayed with appropriate significant figures. If the scale is discontinuous or doesn't start at the origin, it should be indicated by a break in the axis. • NI- If the scale is not appropriate for the data and/or if the scale does not indicate axis break. • U- The graph does not have a scale.						
	Key (defines different data sets that are plotted) • P/A- Should be appropriate and descriptive for the data displayed. It should include: a) descriptions of different colors (if applicable), b) the sample size and c) the number of trials. • NI- If the key is not descriptive and does not indicate the sample size. • U- A key is necessary but it not shown on the graph.						
	Total Points for Mechanics: /3.5pts						

Continued

	Present/Appropriate (P/A) = 2 pts Present but Needs Improvement (NI) =1 pt Unsatisfactory (U)= 0 pts	GRAPH 1			GRAPH 2		
		P/A	NI	U	P/A	NI	U
Communication	**Ease of Understanding-Aesthetics** • P/A- If the graph is aesthetically pleasing, meaning that: a) the data plotted takes up sufficient room in the Cartesian plane, b) makes use of legible size font, c) the x and y axis lines are clear and legible, d) the graph displays data in an appropriate number of bars and lines, and e) is devoid of chart junk elements such as: distracting background colors, patterns, and dark gridlines • NI- If the graph has one of the following flaws: a) the graph displays too much white space, b) the font size is too small, c) the x and y axis lines are not clear and legible, d) the graph shows too many bars or lines OR e) elements of chart junk are clouding interpretation of data. • U- If the graph has multiple flaws, which interfere with the understanding and interpretation of data.						
	Ease of Understanding-Take home message • P/A- If the graph is constructed in a way that is: a) clear to sort trends and b) easy to note the take home message. • NI- If data trends are difficult to observe or it is difficult to formulate a proper take home message. • U- If the graph is ineffective at communicating data trends and take home message.						
Graph Choice	**Graph Type (Bar, line, scatter, dot, box and whisker)** • P/A- If data displayed in a graph is appropriate for both independent and dependent experimental variables (i.e. categorical and continuous) and data. (*Referring to the data form) • NI- If data displayed in a graph is a) not suitable for either the dependent or independent experimental variables OR b) there is a better way to present data. • U- If the graph type is not suitable for both experimental variables.						
	Data Displayed (Raw, Averages, Changes, Percentage) • P/A - If the graph indicates the type of data (ex. Raw, averages, etc.) that are plotted. There should be a clear distinction between raw data and manipulated data based on the information presented in the key (ie. sample size and number of trials) and axis label. If the graph is showing averages, then it should also be accompanied with STDEV or error bars. • NI- If the graph is missing one of points mentioned above. • U- If data type is inappropriate for the graph type						
	Alignment *(at least one of the graphs presented should align with the research question and hypothesis. Other graphs can be exploratory.)* • P/A - If the graph is completely aligned with the research question and/or hypothesis. In other words, the independent, dependent variables, and information about the experiment are explicit. • NI- If the graph is partially aligned with the research question and/or hypothesis. In other words, the graph is missing information about either the independent, dependent, or details about the experiment. • U- If the graph is not aligned with the research question and/or hypothesis.						
Total Points for Communication and Graph Choice		**/10pts**					
Total Score		**/13.5pts**					

FIGURE 1. The graph rubric. Final version of the analytic graph rubric with three levels of achievement. There are three broad categories: graph mechanics, communication, and graph choice. Within graph mechanics are seven subcategories: title, x-axis and y-axis labels and units, scale, and key. Within communication are two subcategories: aesthetics and take-home message. Within graph choice are three subcategories: graph type, data displayed, and alignment. We suggest weighting the graph mechanics lower than the other two categories, as indicated by the scoring criteria.

Biology Graduate Students. We obtained feedback from 10 biology graduate students present at a biweekly graduate seminar (Table 1), using the revised version of the graph rubric (Figure 1). Feedback from this group is important because of the role they play as teaching assistants in assisting the main instructor to deliver knowledge and/or provide feedback to students, usually with a specific rubric or answer key. We gave the biology graduate students a copy of the graph rubric and a student-generated graph (Graph 3 in Appendix C, Supplemental Material) with the corresponding research question and hypothesis to review independently; this was followed by a think–pair–share and a general discussion. IRR was calculated after the meeting to gauge consistency of rubric scoring across the graph rubric categories.

Undergraduate Students. We tested the utility of the graph rubric in an upper-level physiology laboratory classroom with undergraduate students to 1) provide instructor feedback on graphs they constructed as a group and 2) have them use the graph rubric to provide peer feedback. Briefly, students worked in teams to design original experiments, collect data, and display findings in graphs. In conjunction with previously published graph tools (Angra and Gardner, 2016), students used the graph rubric to guide their graph construction and to inform their anonymous graph peer review, which occurred four times during the semester. At the end of the semester, students were prompted to anonymously fill out a survey and provide feedback on the usability of the rubric and the appropriateness of the rubric for the task and to offer suggestions for improving the rubric.

Biology Instructors. We recruited four research-active biology instructors from diverse biology subdisciplines to gather face and content validity. Instructors were shown a copy of the graph rubric (Figure 1) and were asked for feedback regarding the appropriateness of the rubric categories, its potential usability in the classroom and helpfulness to students, and the scoring features of the rubric.

Stage 3. External Stage: Usage of the Graph Rubric in Different Contexts and by Diverse Users

This stage consisted of using the final rubric (Figure 1) to evaluate graphs from different sources and by users from diverse external stakeholder groups to provide us with additional content and face validity evidence. The sources of evidence were derived from evaluation of 1) student-generated graphs from an upper-level undergraduate physiology class; 2) student-generated graphs from a biology instructor's class; and 3) graphs from selected chapters from five introductory biology textbooks. To standardize and guide independent users' scoring of graphs with the rubric, we constructed graph rubric training materials (Appendix B, Supplemental Material). These materials define and explain the features of the rubric and include example scoring of five graphs, each from the three levels of achievement, as shown on the final version of the graph rubric. IRR was calculated for each external user and an expert rater.

Feedback from Undergraduate Biology Majors. We gathered feedback on an independent graph evaluation task from undergraduate students (n = 7) who had successfully completed an upper-level physiology course. We provided the participants with the graph rubric training materials (Appendix B, Supplemental Material) and five, de-identified student-generated graphs to evaluate with the rubric (Appendix D, Supplemental Material). Graphs chosen represented typical graph types and displayed some common undesirable attributes such as plots of all raw data when a descriptive statistic would be appropriate; the use of dark backgrounds and gridlines, which deflect attention from the data displayed; plots of averages without error bars; and misalignment of the graph with the research question and/or hypothesis. Students were encouraged to comment and explain their reasoning for their scoring in each of the graph rubric subcategories.

Feedback from Biology Instructors. To gather feedback and evaluate the rubric as a teaching tool within the context undergraduate biology courses, we recruited biology instructors who have students create or interpret graphs as part of their normal classroom instruction. We purposely recruited instructors who teach courses ranging from the introductory levels to advanced undergraduate and graduate levels. The four faculty instructors taught a range of courses: a course-based undergraduate research experience (CURE) introductory biology laboratory; intermediate-level physiology and cell biology courses; and upper-level field ecology, conservation biology, and neurobiology courses. We provided each instructor with the graph rubric and rubric training materials (Appendix B, Supplemental Material) and asked them to select and evaluate between five and 10 student graphs (with accompanying research question and/or hypothesis statements) with the graph rubric (see Appendix E in the Supplemental Material for descriptions). The graphs were returned to the research team for "expert" scoring with the graph rubric for comparison of scoring with each instructor. In addition, each instructor completed a brief survey to provide feedback on the clarity, usability, and appropriateness of the rubric for evaluating student graphs in their courses.

Evaluation of Biology Textbook Graphs. Because undergraduate students may encounter graphs in their textbooks as part of their course work, we evaluated graphs from five introductory biology textbooks to augment our content validity evidence (see Table 7 later in this article and Appendix H, Supplemental Material). We chose four textbooks (Raven *et al.*, 2008; Sadava *et al.*, 2009; Singh-Cundy and Shin, 2010; Urry *et al.*, 2014) based on the undergraduate curriculum for biology students at a large midwestern university. The fifth textbook (Campbell *et al.*, 2014) was chosen because it integrates the recommendations put forth by *Vision and Change* to incorporate more quantitative thinking in biology (AAAS, 2011). Our selection criteria and graph analysis followed that of Rybarczyk (2011) and Hoskins *et al.* (2007). We randomly selected 10 chapters from each textbook and analyzed pages with graphs as stand-alone artifacts using the graph rubric. The definition that we use for a graph is taken from Kosslyn's (1994, p. 2) work: "a visual display that illustrates one or more relationships among numbers." We expanded this definition and analyzed graphs that were in a Cartesian coordinate system, framed with x- and y-axes, and found in the main chapter or in the side-panel chapter exercises (see Appendix G in the Supplemental Material for a list of graphs on which evaluation was performed). We excluded

interactive graphs, graphs found in videos, and graphs found in the end-of-chapter exercises. Because the graphs in textbooks were rarely directly derived from or presented as related to experiments, we did not include evaluation of the "alignment" subcategory of the rubric.

Data Analysis

We used IRR to quickly identify and refine areas of the rubric during the structural stages of rubric design and to provide us with feedback on the broad use and scope of the rubric during the external stage (Table 1). In this way, the IRR analysis contributed to both content and face validity evidence. We were able to identify areas in which the content and the structure of the rubric were well understood and relevant to users. In addition, IRR provided us with insight into how different raters at various skill levels use the rubric and how they rate graphs that they are most likely to encounter in their own contexts. We first calculated IRR in the form of percent agreement between raters to quantify reliability between expert raters (A.A. and S.M.G.) and each individual population that was asked to use the graph rubric for the structural stage (McHugh, 2012). Because the percent agreement between the two expert raters was high (>90%), percent agreement between other raters (e.g., students or instructors) and either expert rater is used for the values presented here (Stemler, 2004). In qualitative research, an IRR agreement of 80% or higher is considered acceptable (Holsti, 1969). This will inform limitations and usage of the rubric and suggest possible avenues of implementation in the classroom.

RESULTS

Rubric Content and Structure

A critical feature of analytic rubrics is the clear articulation of areas for evaluation with clear explanations for the evaluative criteria (i.e., categories) that users of the rubric need to complete the task (Dawson, 2017). To construct our rubric for graph construction, we began by seeking appropriate evaluative criteria that are characteristic of graphs during both the construction and interpretation processes. For general criteria regarding data presentation and visualizations, we used five books that include guidance on data visualizations (Tufte, 1983; Kosslyn, 1994; Few, 2004; Evergreen, 2014, 2018), and we also consulted 26 primary literature sources on topics that ranged from graph construction with middle school students to evaluation of graphs constructed by physicians for medical journals (Table 2). Additionally, we contributed findings from our ongoing research toward the graph rubric (Angra 2016; Angra and Gardner, 2016, 2017; Table 2). We found that books on data presentation and visualizations heavily emphasize the importance of aesthetics and considering the ink–data ratio that exists in each graph representation. The books also emphasize: descriptive labels on the x- and y-axes and a title to frame the message that is being conveyed by the graph; a key to show the various colors used in the graph; appropriate axis scaling to show proper intervals conveyed by the data represented in the graph; and thinking about the appropriateness of the graph representation for the data to be displayed. Given the general target audience for the books, the graph criteria were illustrated in multiple non-science examples with pros and cons of each. However, our graphing literature review further supported the importance of the categories mentioned above and expanded support for axis units, data displayed, and aligning the graph to its original intended purpose (e.g., question to be answered; Table 2).

Our preliminary evaluative criteria consisted of the categories emphasized by the existing literature but were further refined from our ongoing research (Angra and Gardner, 2016, 2017). Think-aloud interviews conducted to understand how graphs are constructed by experts and novices revealed that students titled their graphs with the subject and variables, a detail that is important for the graph reader to see when interpreting the graph (Angra and Gardner, 2017). We also noted that professors verbally articulated experimental details that belong in the key, such as sample size and number of trials. Finally, an important characteristic of data summary graphs in biology or sciences, in contrast to conceptual graphs or data exploration graphs, is the alignment of the graph with its intended purpose, such as the research question and hypothesis. Experts do this routinely, while students do not (Angra 2016; Angra and Gardner, 2016, 2017). Additionally, two articles from our literature search (Konold and Higgins, 2003; Rougier *et al.*, 2014; Table 2) mention making a graph so it has a purpose, but do not explicitly state the alignment of the graph to the research question or hypothesis.

On the basis of the literature and research review, we created a list of 12 graph construction categories with definitions (Table 2). To organize and characterize the evaluative criteria for the 12 categories, we aggregated them into three broader categories: graph mechanics, communication, and graph choice. Graph mechanics includes the title, axis labels, axis units, axis scaling, and a key. Communication consists of aesthetics and take-home message. In our literature search, we noticed a high emphasis on communication (Table 2 and Figure 1), which is why we decided to create this separate category. Finally, graph choice includes tasks like choosing a graph type, thinking about the data displayed, and alignment of the graph with its intended purpose (Table 2 and Figure 1).

An important feature of any rubric is the quality levels or numeric criteria that tell students how they will be graded. We chose to use three quality levels and express them in statements of student performance that are used to distinguish specific graph construction elements as "present/appropriate," "needs improvement," or "unsatisfactory" (Dawson, 2017), each with associated point values (Figure 1). Feedback, testing, and use of the rubric by diverse users suggests that using three levels of achievement works well for most users (see below).

Rubric Testing and Implementation

We share here our findings in the form of conversations and IRR with science education research scholars (graduate students, postdoctoral fellows, and faculty), non-education research biology graduate students, undergraduate biology students enrolled in an upper-level physiology laboratory course, and biology instructors. These data contribute to the evidence in support of the content and face validity of the graph rubric (Table 1).

As the first step in the structural stage of the rubric design, we used the first draft of the graph rubric to gather feedback from science education scholars. Three important outcomes resulted from the first round of structural stage of the rubric design. First,

TABLE 2. Graph rubric elements compiled in the substantive stage of rubric design and preliminary construct validation

Graph rubric elements	Sources that informed descriptions[a]
Descriptive title Should 1) be in the form of a statement, 2) mention the subject, 3) include appropriate variables, and 4) include relevant details about the experiment that will help readers understand the take-home message.	**Kosslyn, 1994**; **Evergreen, 2018**; *Puhan et al., 2006*; Angra, 2016; Angra and Gardner, 2017
Label for the x-axis (e.g., time) Should be appropriate and descriptive for the experiment. For graphs with categorical independent variables, there needs to be a label under each set of data and a larger label under all data plotted.	**Kosslyn, 1994**; **Few, 2004**; *Federico et al., 2012*; *Elliott et al., 2006*; *Puhan et al., 2006*; Angra, 2016; Angra and Gardner, 2017
Label for the y-axis (e.g., heart rate) Should be appropriate and descriptive for the experiment. If the data are manipulated (average, change, percentage, etc.), then that should be indicated on the y-axis.	**Kosslyn, 1994**; **Few, 2004**; *Federico et al., 2012*; *Elliott et al., 2006*; *Puhan et al., 2006*; Angra, 2016; Angra and Gardner, 2017
Units for the x-axis (e.g., seconds) Should be appropriate and descriptive for the data displayed.	*Leinhardt et al., 1990*; *Puhan et al., 2006*; Angra, 2016
Units for the y-axis (e.g., average beats per minute) Should be appropriate and descriptive for the data displayed.	*Leinhardt et al., 1990*; *Puhan et al., 2006*; Angra, 2016
Scale (appropriate intervals and range for data) Should be appropriate for the data displayed such that the increments are clear and without clutter and should include appropriate significant figures. If the scale is discontinuous or does not start at the origin, it should be indicated by a break in the axis.	**Tufte, 1983; Kosslyn, 1994**; **Few, 2004**; **Evergreen, 2018**; *Leinhardt et al., 1990*; *Cleveland, 1984*; *Duke et al., 2015*; Angra, 2016
Key (defines different data sets that are plotted) Should be appropriate and descriptive for the data displayed. It should include: 1) descriptions of different colors (if applicable), 2) the sample size, and 3) the number of trials.	**Few, 2004**; **Evergreen, 2018**; Angra, 2016; Angra and Gardner, 2017
Ease of understanding—aesthetics[b] The graph is aesthetically pleasing if 1) the data plotted take up sufficient room in the Cartesian plane, 2) a legible size font is used, 3) the lines of the x- and y-axes are clear and legible, 4) data are displayed in an appropriate number of bars and lines, and 5) there are no "junk" elements such as distracting background colors, patterns, and dark gridlines.	**Tufte, 1983**; **Kosslyn, 1994**; **Evergreen, 2014, 2018**; *Cooper et al., 2003*; *Puhan et al., 2006*; *Stengel et al., 2008*; *Federico et al., 2012*; *Rougier et al., 2014*; *Duke et al., 2015*; Angra, 2016
Ease of understanding—take-home message[b] If the graph has sound construction and mechanics that allow for clear sorting of trends and take-home message.	**Evergreen, 2018**; *Cooper et al., 2003*; *Federico et al., 2012*; *Rougier et al., 2014*; *Duke et al., 2015*
Graph type (bar, line, scatter, dot, box and whisker) If data displayed in a graph are appropriate for both independent and dependent experimental variables (i.e., categorical and continuous) and data. (Referring to the data form.)	**Kosslyn, 1994**; **Evergreen, 2018**; *Padilla et al., 1986*; *Cleveland, 1984*; *Schriger and Cooper, 2001*; *Few, 2004*; *Leonard and Patterson, 2004*; *Patterson and Leonard, 2005*; *Drummond and Tom, 2011*; *Drummond and Vowler, 2011*; *Franzblau and Chung, 2012*; *Humphrey et al., 2014*; *Duke et al., 2015*; *Saxon, 2015*; *Weissgerber et al., 2015*; *Klaus, 2016*; Angra, 2016; Angra and Gardner, 2016, 2017
Data displayed (raw, averages, changes, percentage) If the graph indicates the type of data (e.g., raw, averages, etc.) that are plotted. There should be a clear distinction between raw data and manipulated data based on the information presented in the key (i.e., sample size and number of trials) and axis label. If the graph is showing averages, then these should also be accompanied with SD or error bars.	*Wild and Pfannkuch, 1999*; *Friel and Bright, 1996*; *Konold and Higgins, 2003*; *Konold et al., 2015*; Angra, 2016
Alignment (at least one of the graphs presented should align with the research question and hypothesis. Other graphs can be exploratory.)[b] If the graph is completely aligned with the research question and/or hypothesis. In other words, the independent and dependent variables and information about the experiment are explicit.	*Konold and Higgins, 2003*; *Rougier et al., 2014*; Angra, 2016; Angra and Gardner, 2016, 2017

[a]Sources in bold are from graphing books, sources in italics are from primary literature, sources that are underlined are from our own graphing research projects.
[b]Graph rubric elements not included in the first draft of the graph rubric. The "ease of understanding" category consisted of combined descriptions of both aesthetics and take-home message.

all participants approved of the categories, subcategories, and descriptions within the rubric, but suggested that changes be made to the levels of achievement for the subcategories within graph mechanics by weighting the individual criteria for the subcategories to be less than those in the graph choice and communication categories. Participants felt that the cognitive difficulty of the mechanics category was lower compared with the other categories and should be weighted accordingly.

Second, during the conversations, a science education research scholar from the College of Education suggested that we might want to consider not just evaluating one graph, but a set of graphs with the graph rubric to determine a more accurate take-home message. Although we agree that looking at multiple but related graphs like those found in science articles first before formulating a take-home message is helpful, it was our purpose to produce a graph rubric that evaluates one graph at a time, because this is a common graph construction practice in the classroom. Further, the spirit of this subcategory of the rubric was to capture whether the graph was constructed in a way in which one could discern whether or not there were trends in the data and not necessarily the specific type of conclusion, which requires direct knowledge of the discipline or experiments.

Third, a graduate student in chemistry suggested that a subcategory on figure legends be considered for the graph rubric. Although we agreed that figure legends provide helpful information when encountered in a science paper, they are not universal; for example, they are not found in oral presentations when the graph is a stand-alone item. Furthermore, different sets of skills are required when writing a figure legend, and these fall outside of the scope of our current work (Angra 2016; Angra and Gardner, 2017).

These types of conversations were vital during this first round of the structural stage of the rubric design process (Table 1), providing us with feedback on the structural components and content within the rubric. Revisions to the rubric from this round included refining and clarifying the subcategory definitions and adjusting the point values assigned to parts of the rubric to reflect the cognitive difficulty of the items (i.e., graph mechanics scoring was decreased in weight). We also separated the communication category into two subcategories, "aesthetics" and "take-home message," and added the "alignment" subcategory within the graph choice category (Figure 1).

The revised rubric was presented later to the science education scholars for further feedback. Science education scholars were asked to evaluate a student-generated graph (Graph 3, Appendix C, Supplemental Material) and then engaged in a discussion. Percent agreement with the ratings of an expert rater was calculated for the attendees for each category within the graph rubric. The overall percent agreement with attendees was 82%, which is considered excellent (Holsti, 1969). There was greater than 80% agreement on all subcategories of graph mechanics, except for the scale, which scored 33% before the general discussion (Table 3). Lower percent agreement was also observed for the take-home message subcategory of communication and the graph type subcategory of graph choice. The last category in which there was a low percent agreement and raters tended to underscore was the take-home message subcategory under communication. On the basis of this analysis, we realized that we needed to increase the clarity of our definitions for elements within the rubric (Table 4), and that led us to develop training materials for new users (see Appendix B in the Supplemental Materials).

Feedback from science education research scholars provided us with valuable pedagogical feedback, but we also wanted to solicit feedback from users who grade student assignments. We sought feedback from 10 biology graduate students who were shown the same graph as the science education scholar group (Graph 3 in Appendix C, Supplemental Material) and asked them to score the graph independently, after which there was discussion and feedback. The graduate students reported that the rubric was clear and easy to use (Table 4). While the rubric was used easily by the graduate students, compared with the science education scholars, the graduate students had more low percent agreements with the expert rater. The lowest level of agreement was noticed in the graph mechanics category "key." This resulted from the graduate students deviating from the definition on the rubric and underscoring the category, because they said the key was vague and some graduate students did not like where it was placed in relation to the data on the graph. This is an element that was not explicitly articulated in the rubric but would fall into the subcategory of aesthetics.

Next, we sought informal written feedback from students enrolled in the Spring 2015 semester (Table 1) of a physiology laboratory course. These students used the rubric multiple times over the semester to inform their graph construction, critique peer graphs, and interpret feedback from the instructors. Graphing and presenting data were important components of the course, and students readily used the graph rubric and found it to be a valuable resource. Comments from students are displayed in Table 4.

Finally, we showed the graph rubric to biology instructors and asked them for written feedback regarding the appropriateness of the rubric categories, usability in the classroom, helpfulness to students, and scoring. All instructors agreed with the content and structure of the rubric, including the division and distribution of elements within categories and weighting of the scoring, and felt it could be useful in their classrooms (Table 4).

Application of the Graph Rubric to Diverse Contexts in Undergraduate Biology Instruction

We wanted to explore the broad utility of the rubric by having a diverse set of rubric users from a variety of classroom contexts evaluate student-generated graphs and also by characterizing the features of graphs found in textbooks, which is one of the ways undergraduate students are exposed to graphs and provides a potentially strong model of graphs for students. Here, we report results from 1) undergraduate student evaluation of student-generated graphs from a classroom, 2) instructor use of the graph rubric to score graphs produced by their students, and 3) analysis of textbook graphs. These data contributed to further content and face validity evidence for the graph rubric (Table 1).

Undergraduate Student Evaluation of Graphs. We gave undergraduate students who had some previous experience with the graph rubric from their physiology course a variety of student-generated graphs to score, ranging from a more unfamiliar graph type like a box-and-whisker plot (Graph 1, Appendix D, Supplemental Material) to more familiar graph types like line graphs (Graphs 2 and 4, Appendix D, Supplemental Material), scatter plots (Graph 3, Appendix D, Supplemental Material),

TABLE 3. Graph rubric use during the structural stage of rubric design and construct validation

Graph rubric category		IRR (% agreement)[a]	
		Science education scholars (n = 6)	Biology graduate students (n = 10)
Graph mechanics	Descriptive title	83	50
	Label for the x-axis	100	50
	Label for the y-axis	83	100
	Units for the x-axis	100	90
	Units for the y-axis	100	90
	Scale	33	70
	Key	100	20
Communication	Ease of understanding—aesthetics	100	60
	Ease of understanding—take-home message	50	80
Graph choice	Graph type	67	80
	Data displayed	83	70
	Alignment	83	100
	Average task IRR	82 ± 22	72 ± 24

[a]IRR (% agreement) with science education scholars (n = 6) and biology graduate students (n = 10) before graph rubric discussion is shown. All members who participated evaluated Graph 3 in Appendix C in the Supplemental Material.

and a bar graph (Graph 5, Appendix D, Supplemental Material). Overall, the graph rubric ratings of student-generated graphs by students (n = 7) were consistent with those of the expert rater with an overall average percent agreement of ≥ 71% (Table 5). However, student scoring of graphs using the rubric revealed several things. One interesting finding is that almost all students scored Graph 1, the box-and-whisker plot, as "needs improvement" instead of "present/appropriate" for the data-displayed category. Student reasoning for underscoring the graph was that it was not explicit to the type of data plotted. This hints at

TABLE 4. Feedback in the form of quotes from the users who were asked to provide feedback on the content, structure, and use of the graph rubric during validation

Science education scholars feedback use
"Do all graphs have to be hypothesis driven?"
"Label for the x-axis, what about categorical data?"
"Add more detail to the graph type category. Tease out and define words like appropriate."
Graduate student feedback from use of the rubric
"What about figure legends?"
"The language in the rubric was easy to understand but it was a lot to read."
"I didn't encounter problems using the rubric."
Biology instructor feedback on the graph rubric categories, usability, and utility
"I have been wanting to improve the way I teach graphing in my classes and this seems like a useful tool."
"In some points, I felt that having three categories was too restrictive and figures that were really different in that feature ended up together in the middle category (Present but needs Improvement)."
"I feel that students who are given this rubric, in courses where graphs are used to present data and are graded using this rubric, would quickly conform to consistently produce high quality, informative graphs."
"Rubric items were well constructed and clear to apply."
"I think this rubric is useful as long as I prepare the students well before they make the graph. I could imagine needing to spend a fair amount of time in class going over how to make graphs in order to get students aligned with this rubric. I'm willing to do this because the rubric provides good guidelines and teaching students how to do science is part of the classes I teach."
Undergraduate student feedback from use of the rubric in the classroom, Spring 2015
"It took me awhile to understand all the necessary components in graphs. This class helped me understand when certain graph types are relevant and also what to include on each graph type."
"Provide more feedback on the level of detail you are looking for in graphs eg. titles, etc."
Undergraduate student feedback on the graph rubric categories, usability, and utility
"The rubric was very clear."
"No problems were encountered."
"Specific points detailing what should be present in the graph were helpful."
"This rubric is useful as a guideline for creating effective graphs."
"The rubric is both comprehensive and flexible enough to be used in other scientific courses."

TABLE 5. Graph rubric use by undergraduate students during the external stage of rubric validation[a]

	Graph rubric category	Graph 1	Graph 2	Graph 3	Graph 4	Graph 5
Graph mechanics	Descriptive title	29	100	71	57	71
	Label for the x-axis	100	29	86	100	100
	Label for the y-axis	71	86	43	100	86
	Units for the x-axis	86	100	100	100	86
	Units for the y-axis	100	71	100	100	100
	Scale	100	71	86	86	100
	Key	14	57	57	71	86
Communication	Ease of understanding—aesthetics	86	86	71	43	14
	Ease of understanding—take-home message	86	71	0	71	57
Graph choice	Graph type	100	14	100	29	14
	Data displayed	14	86	57	86	86
	Alignment	43	100	86	43	57
	Average (%) task IRR[b]	69 ± 9	73 ± 12	71 ± 14	74 ± 9	71 ± 12

[a]IRR (% agreement) with seven undergraduate students who evaluated five graphs from a physiology lab that they successfully completed is shown.
[b]Overall average is among all seven undergraduate students for each individual graph critiqued

student difficulty in interpreting box-and-whisker plots, which display data and descriptive statistics in a way that is challenging to novices (Bakker *et al.*, 2004). We also observed that students did not object to the dark background in Graph 5, clashing with Tufte's rule to maximize data–ink ratio.

Instructor Use of Graphs in the Classroom. To evaluate the potential of the graph rubric to be used in diverse classrooms and by diverse instructors, we recruited four undergraduate biology instructors from different biological subdisciplines and course contexts. Instructors were provided with the graph training materials (Appendix B, Supplemental Material) and were asked to thoroughly study them before proceeding to evaluate student-generated graphs from their classrooms with the graph rubric. None of the graphs submitted by the instructors were accompanied by a figure legend and all were therefore scored as stand-alone artifacts by all raters (instructors and expert). Overall, the graph rubric ratings of student-generated graphs by the instructors were consistent with those of the expert rater, with an overall average percent agreement of ≥ 72% (Table 6 and Appendix F, Supplemental Material), which is good, given that no other training on the rubric had been provided. There were three graph rubric subcategories that had the highest numbers of differences in ratings: title and key (mechanics) and aesthetics (communication). Instructors 2 and 3 consistently rated titles as "present/appropriate," while the expert rater rated the titles as "needs improvement." Examining the survey feedback revealed that Instructor 3 realized that the titles were not fully complete, but felt they were close enough to warrant full credit. Full credit was given for the keys on graphs from the classrooms of Instructors 2 and 3 more often than by the expert rater, even when elements such as the sample size were not indicated on the graph. Instructors 2, 3, and 4 consistently rated aesthetics as "present/appropriate" instead of "needs improvement" for graphs that contained unnecessary grid lines in the background or lacked y-axis lines. All instructors felt that the rubric categories and definitions were appropriate and that the rubric itself was easy to use and would be a valuable addition to their introductory and upper-division biology classrooms (see Table 4).

Analysis of Textbook Graphs. Because students also encounter graphs in their assigned readings for their courses, which includes textbooks, we wanted to determine how well the graph rubric captured features of those graphs. The analysis of textbook graphs supports our first objective for the development of the rubric, which is to facilitate the teaching and evaluation of data summary graphs. The graph rubric was generally useful and appropriate for evaluating graphs from introductory biology textbooks. Because the purpose of the Campbell *et al.* (2014) textbook is to incorporate more experiments, data, and quantification in biology, we noticed that, compared with the other four textbooks analyzed, there were approximately seven times more graphs present in this online textbook on average. Bar and line graphs were the most common type across all five textbooks, and scatter, dot, and box-and-whisker plots were the least common (Figure 2).

TABLE 6. Graph rubric use by biology instructors during the external stage of rubric validation[a]

Graph rubric category	Instructor 1 (n = 8)	Instructor 2 (n = 10)	Instructor 3 (n = 5)	Instructor 4 (n = 12)
Course type	Introductory laboratory and upper-level ecology	Introductory cell biology and upper-level neurobiology	Upper-level ecology	Upper-level physiology
Mechanics	86 ± 18	73 ± 24	71 ± 22	83 ± 11
Communication	75 ± 0	70 ± 28	70 ± 42	67 ± 0
Choice	83 ± 19	80 ± 10	73 ± 12	72 ± 10
Average (%) task IRR[b]	83 ± 9	74 ± 4	72 ± 15	78 ± 7

[a]IRR (% agreement) with four instructors who evaluated 5–10 graphs constructed by students from their respective courses is shown. n = number of graphs evaluated.
[b]Average from the overall rubric scoring across graphs for each instructor.

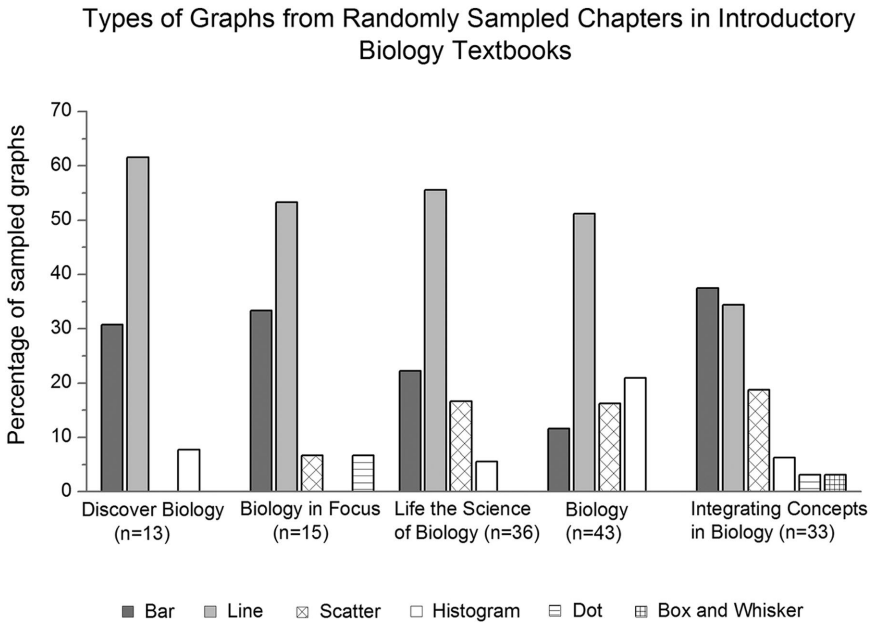

FIGURE 2. Types of graphs in introductory biology textbooks: *Discover Biology* (Singh-Cundy and Shin, 2010); *Campbell Biology in Focus* (Urry *et al.*, 2014); *Life: The Science of Biology* (Sadava *et al.*, 2009); *Biology* (Raven *et al.*, 2008); *Integrating Concepts in Biology for Introductory College Biology* (Campbell *et al.*, 2014). Summary of the types of graphs evaluated from randomly sampled chapters in introductory biology textbooks. Data are expressed as the percentage of all graphs in the sampled chapters. Number of graphs for each book is indicated in parentheses.

The average percentage of graphs that received a "present/appropriate" rating from the graph rubric in the graph mechanics, communication, and graph choice categories is displayed in Table 7 and Appendix H in the Supplemental Material. Looking broadly across the graph rubric categories, we see that there was variability *within* textbooks for good graph design, and no one textbook received a perfect score. For example, we noticed variation across the subcategories of graph mechanics for graphs within any given textbook. There was also variability *across* textbooks. There are clear differences between the books for the attributes and quality of graphs displayed as captured by the rubric. For example, graph choice showed a large range of scores between the textbooks.

DISCUSSION

A Tool for Evaluating and Teaching Graphing in Undergraduate Biology

In this article, we aimed to present the rigorous and systematic development of an evidence-based rubric for teaching and evaluating graphs. The graph rubric is a tool designed within the context of undergraduate biology to 1) facilitate the teaching and evaluation of data summary graphs, 2) provide undergraduate students with formative and summative feedback on their graphs, and 3) allow education researchers to evaluate graphing artifacts to assess students' experimental and quantitative skills. As undergraduate biology students are increasingly engaged in the practice of science as part of their undergraduate curricula, more tools for research and instruction on graphing are needed. Specifically, there is a need for resources that are not generic but are contextualized within the discipline.

The three broad categories of the rubric, and the subcategories within them, allow the rubric user to create and evaluate graphs that are constructed in a manner that is complete (graph mechanics), appropriate for the data and purpose (graph choice), and clear and easy to interpret (graph communication).

TABLE 7. Graph rubric used to evaluate graphs from introductory biology textbooks[a]

Graph rubric category	Introductory biology textbooks[b]	Present/appropriate	Needs improvement	Unsatisfactory
Mechanics	Singh-Cundy and Shin, 2010 (n = 13)	70 ± 6	12 ± 5	18 ± 6
	Urry *et al.*, 2014 (n = 15)	61 ± 7	24 ± 7	15 ± 5
	Sadava *et al.*, 2009 (n = 36)	75 ± 3	13 ± 3	12 ± 2
	Raven *et al.*, 2008 (n = 43)	76 ± 3	19 ± 4	5 ± 1
	Campbell *et al.*, 2014 (n = 33)	69 ± 4	26 ± 5	4 ± 1
Communication	Singh-Cundy and Shin, 2010 (n = 13)	42	58	0
	Urry *et al.*, 2014 (n = 15)	80 ± 5	20 ± 5	0
	Sadava *et al.*, 2009 (n = 36)	85 ± 3	15 ± 3	0
	Raven *et al.*, 2008 (n = 43)	85 ± 2	15 ± 2	0
	Campbell *et al.*, 2014 (n = 33)	88	12	0
Choice	Singh-Cundy and Shin, 2010 (n = 13)	63 ± 5	38 ± 5	0
	Urry *et al.*, 2014 (n = 15)	60 ± 12	40 ± 12	0
	Sadava *et al.*, 2009 (n = 36)	81 ± 1	19 ± 1	0
	Raven *et al.*, 2008 (n = 43)	73	27	0
	Campbell *et al.*, 2014 (n = 33)	83 ± 1	17 ± 1	0

[a]Data displayed are the average scores ± SE received by the graphs for each textbook for each graph rubric category. The average percentage with SE from the overall rubric scoring across graphs for each textbook is shown.
[b]n = number of graphs evaluated.

The graph rubric complements and extends from existing guidebooks and other resources (Table 2) by explicitly incorporating important concepts and skills needed for graph choice and construction in the context of biology. We incorporated expert-like, reflective practices such as the checking the alignment of the graph with its purpose (e.g., evaluating a hypothesis; Angra and Gardner, 2017).

Throughout the rubric design process, we gathered content and face validity evidence to support our claim that the rubric is an appropriate and usable tool to evaluate graphs in the undergraduate biology context (i.e., construct validity; Table 1). We are confident that the evidence we gathered is sufficient to support this claim in our context. As part of the design process, we consulted existing resources (e.g., instructional books and literature) and three important stakeholder groups (i.e., students, instructors, and science education scholars) for the rubric. The face and content validity evidence we gathered and used during the substantive and structural stages allowed us to be confident that the rubric was capturing important and relevant elements of strong graph design. Particularly valuable was the feedback collected from undergraduate biology students, biology instructors, and science education scholars during the structural stage (Table 4). We were able to clarify and refine terms and definitions and organize the rubric in a manner that was understandable to all user groups. We adjusted the weighting of the scoring points to reflect the cognitive difficulty within the three broad categories of the rubric, with graph mechanics weighted less than graph communication and choice. Finally, on the basis of feedback during these stages, we emphasized that the inclusion of this subcategory is meant to emphasize a reflection on the purpose of the investigations that generated the data in the first place, which is something that students do not consistently do (Angra and Gardner, 2017). This is not meant to preclude the creation of graphs to explore the data, however.

The external stage of the graph rubric development provided us with additional content and face validity evidence. We gained important insight into the scope of the appropriateness and utility of the graph rubric and some interesting observations about graphs in different contexts that students may encounter. During this stage, we conducted user testing of the rubric by having students and biology instructors evaluate graphs generated in the classroom, and we used the rubric to evaluate graphs in introductory biology textbooks (Table 1). The graphs that the students and biology instructors evaluated were single graphs extracted from class assignments, which included oral presentations and written work (e.g., research posters and lab reports). Textbook graphs, as previously described (Hoskins *et al.*, 2007; Rybarczyk, 2011), were often stylized representations of data embedded in multimedia figures, also with figure legends. While the two graph contexts (i.e., classroom vs. textbook) were different, the attributes of the graphs aligned with the typical communication purpose of the context: graphs in oral presentations are accompanied by real-time verbal narrations of the graph, while graphs in textbooks are embedded in descriptive text and their purpose is often to summarize data trends, albeit often in an oversimplified manner not true to the natural "messiness" of the actual data (Hoskins *et al.*, 2007; Rybarczyk, 2011).

Limitations of the Graph Rubric

The graph rubric is designed to assist in the creation and/or evaluation of graphs as a stand-alone piece of communication, similar to what would be seen in oral presentations of data. Because of the limited amount of space allotted for figures by research journals, graphs are usually small and do not have titles, labels, and keys. Instead, this information is found in the figure legend, a category not present in our graph rubric. The absence of the figure legend from the rubric was noted by individuals in the science education scholar and graduate student groups. However, while figure legends are informative accompaniments to a graph, we feel they are beyond the scope of the graph rubric for two reasons. First, we want to promote the creation of clear representations of data. Because graphs are meant to be stand-alone representations with the purpose of conveying complex data in a quick and efficient manner, we and others recommend that graphs should be labeled in a descriptive manner and include a key, if necessary (Mack, 2013). Second, writing figure legends is a related but distinct skill that requires knowledge regarding which methods and results to include and a succinct description of what is plotted, with trends noted (Rodrigues, 2013). Users of the graph rubric may modify the rubric to include figure legends and define a set of criteria at each level of achievement to communicate expectations to students.

Although we provided the rubric users with training materials to consult independently in the external stage of the rubric design, there were some instances of low agreement in the three graph rubric categories (Tables 5 and 6). This is likely because the users were not trained on rubric use in collaboration with the expert raters. In addition to the effects of minimal rubric training, low consensus in any of the three graph rubric categories could be affected by the different number subcategories within each and the level of subjectivity potentially used in the evaluation (see Figure 1). For example, we observed the most deviation from the expert rater within the mechanics category, which has seven subcategories compared with two and three subcategories in the communication and graph choice categories, respectively. Therefore, while a well-designed rubric should be clear to any user, in theory (Dawson, 2017), we recommend training, practice, and feedback to ensure rigorous and consistent grading in the classroom or within a research project that uses a rubric for evaluation of research artifacts.

The purpose of the IRR consensus estimates was to help us understand the graph rubric use by different people and with graphs from different contexts (Dawson, 2017) and to highlight areas in which things might not be clear or consistently interpreted by users. As such, it is interesting to note that the areas in which the students' ratings differed from the expert raters are consistent with their status as developing graph makers. For example, students exhibited a tolerance for extraneous features and colors in the graphs (see Graphs 4 and 5, Appendix D, Supplemental Material), which led to differences in scoring in the graph communication category. In contrast, while the biology instructor group was small (n = 4), there was little variability across the four instructors.

The final source of content and face validity evidence we sought was from an evaluation of graphs from introductory biology textbooks. The purpose of our evaluation of textbook

graphs was not to criticize textbooks, but rather to examine a source that is potentially a strong model to students for what constitutes data and graphs, given the presence of these books in the 100-level in university curricula. In general, we found that the graph rubric was able to capture and describe the elements of the graphs in textbooks. However, consistent with their typical purpose, graphs in textbooks rely even more heavily on the figure legend and surrounding text for complete understanding, as reflected in lower scores in the scale and key subcategories (Table 7 and Appendix H, Supplemental Material). In addition, we omitted the alignment subcategory, as the display of data explicitly resulting from experiments was rare, and the data displayed were stylized summaries of trends, not quantitative in nature (see data displayed and y-axis units, Appendix H, Supplemental Material). An interesting observation we made in our textbook graph analysis was that not only did graphs vary *across* the introductory biology textbooks we examined but were variable *within* a given textbook. This general observation echoes the inconsistent use of arrows within textbooks (Wright *et al.*, 2018). This inconsistent graph model could impair student learning of what constitutes a high-quality graph and also impede their understanding and learning of important biological concepts.

While the basic features of sound graph design are discipline neutral, the norms for graph choice and construction may have some variation across the biological subdisciplines that can be perpetuated by the degree to which research journals guide authors. Interestingly, during and since the work in the external stage of the rubric design, there have been calls for improved data representations in research articles by a variety of journals with broad readership (e.g., Rougier *et al.*, 2014; Slutsky, 2014; Saxon, 2015; Weissgerber *et al.*, 2015; Klaus, 2016; Nuzzo, 2016; Boers, 2018; Hertel, 2018). In addition, more textbooks have begun to incorporate the description of experiments and "messy" data displays (Campbell *et al.*, 2014), which will provide students with a more realistic perspective of scientific data.

As is the case for any assessment and research tool, the external stage is never truly complete, and with each new user and context, validity and reliability evidence need to be gathered to establish the scope of use and inference to help interpret the findings from use of the graph rubric. We chose the elements of the external stage of the rubric development to provide us with initial evidence for the broad usability (e.g., not within a single biological subdiscipline and a single user type) and utility of the rubric (Dawson, 2017) within our institutional context. Our sample of raters for the external stage was limited based on opportunity and volunteer bias. Students were recruited from a physiology course, and the biology instructors who chose to participate in our study were the few who deliberately integrate graphing into their courses. Therefore, we cannot definitively claim that the rubric is universally applicable "as is" for each context, and we encourage users of the rubric to reflect on its appropriateness and utility for their specific use.

Implications for Instruction and Future Research

The graph rubric is a valuable, evidence-based assessment tool for biology instructors, students, and science education researchers, because it provides quick, systematic, and targeted evaluation of essential features of effective graphs. Frequent use of this rubric in the classroom not only communicates the learning objectives for data communication, but also the expectations of a well-constructed graph. The graph rubric has three different levels of achievement (present/appropriate, needs improvement, and unsatisfactory) and provides the instructor a transparent and objective means to evaluate student graphs. Given the diversity of graphs, their contexts, and personal or disciplinary preferences for data representations, the rubric can be used as an instructional catalyst. For example, an instructor can have students evaluate a set of graphs with the rubric and use the similarities and differences in scoring between the students and the instructor in a guided instruction. This activity would serve two important purposes. First, it would allow the instructor to communicate expectations for the attributes of high-quality graphs, and second, it would facilitate a classroom discussion about data and data representations. This discussion would provide the instructor the opportunity to provide guidance with and model reflective graph design and could include a comparison of the affordances and limitations of the form of data to plot (e.g., raw vs. summarized data) or appropriate graph types to use (e.g., categorical dot plot vs. bar graph) appropriate to the data type and purpose of the graph (see "Guide to Data Displays" in Angra and Gardner [2016]).

Current recommendations for undergraduate biology curricula include increasing students' access to experiences that involve them in the practices of science, including designing investigations, grappling with data, and summarizing their findings (Auchincloss *et al.*, 2014; Harsh and Schmitt-Harsh, 2016). These experiences will include work within lectures, laboratory courses, CUREs, or research apprenticeships. The graph rubric can be a valuable tool to guide students in creating effective data representations that will allow them to explore and summarize their data in a reflective manner. The consistent incorporation of the rubric into lab manuals for courses throughout a course sequence would be a valuable way to provide students with the repeated guidance and practice that is needed to aid in mastery of the skill of graphing.

As instructors are changing their instruction to respond to the recommendations for undergraduate biology curricula and funding agencies such as the National Science Foundation and Howard Hughes Medical Institute are committing monies to support the adoption of experiences such as CUREs, tools to guide and improve student learning are needed (Auchincloss *et al.*, 2014; Corwin *et al.*, 2015; Shortlidge and Brownell, 2016; National Academies of Sciences, Engineering, and Medicine, 2017). Until recently, much of the evaluation of student experiences in course-based or other research experiences was predominantly on student perceptions and attitudes. The graph rubric is a valuable addition to the growing list of research and evaluation tools of cognitive gains and learning (summarized in Shortlidge and Brownell, 2016). For example, the rubric can be used by education researchers and evaluators to monitor student learning and reveal persistent difficulties as students progress through a course, program, or curriculum. These data will be valuable in evaluating and refining instruction within student learning experiences.

While the graph rubric is a valuable, evidence-based tool for instruction and research, there are many opportunities for refining and expanding it. As noted, we did not include figure legends in the rubric. There is an opportunity to potentially include figure legends as part of the rubric or to design another

tool to guide students in that part of their science writing and data presentation, as appropriate to the communication medium (e.g., lab reports). In addition, as part of the external stage of the design of the graph rubric, we chose to have a variety of people use the rubric to evaluate graphs in their areas of expertise. For students, these were graphs from the laboratory portion of a course they had taken, and for biology instructors, these were graphs generated by students in their classrooms. Exploring student use of the rubric with graphs in other course contexts or with graphs from the primary literature could reveal common areas of competence and difficulty regardless of graphing context or context-specific difficulties. This knowledge would provide valuable insight for research and instruction.

ACKNOWLEDGMENTS

Thanks to the members of the Purdue International Biology Education Research Group for thoughtful feedback on this work. We are indebted to the late Dr. Aaron Rogat for his tremendous insights, tenacious demand for clarity, and constant support of our work. Many thanks to Mikhail Melomed and Drs. Elizabeth Suazo-Flores and Maurina Aranda for feedback on the article. This work benefited from ideas initiated within the Biology Scholars Research Residency program (S.M.G.). The interpretation of this work benefited from the ACE-Bio Network (National Science Foundation RCN-UBE 1346567).

REFERENCES

Aguirre, K. M., Balser, T. C., Jack, T., Marley, K. E., Miller, K. G., Osgood, M. P., … Romano, S. L. (2013). PULSE vision and change rubrics. *CBE—Life Sciences Education*, *12*(4), 579–581.

Allen, S., & Knight, J. (2009). A method for collaboratively developing and validating a rubric. *International Journal for the Scholarship of Teaching and Learning*, *3*(2), 10.

Allen, D., & Tanner, K. (2006). Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. *Cell Biology Education*, *5*(3), 197–203.

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC. Retrieved December 27, 2017, from https://visionandchange.org/files/2013/11/aaas-VISchange-web1113.pdf

American Association of Colleges and Universities. (2010). *Quantitative Literacy VALUE Rubric. Assessing Outcomes and improving Achievement: Tips and Tools for Using*. Retrieved November 3, 2018, from www.aacu.org/value/rubrics/quantitative-literacy

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC.

Angra, A. (2016). Understanding, evaluating, and diagnosing undergraduate student difficulties with graph choice and construction. ProQuest Dissertations and Thesis A&I (2002580772). Retrieved November 3, 2018, from https://ezproxy.gsu.edu/login?url=https://search.proquest.com/docview/2002580772?accountid=11226

Angra, A., & Gardner, S. M. (2015). Development of an analytic rubric to evaluate undergraduate student graphs and diagnose difficulties in graph choice and construction. *Poster presented at Society for the Advancement of Biology Education Research Meeting (Minneapolis, MN)*.

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.

Ashley, M., Cooper, K. M., Cala, J. M., & Brownell, S. E. (2017). Building better bridges into STEM: A synthesis of 25 years of literature on STEM Summer Bridge Programs. *CBE—Life Sciences Education*, *16*(4), es3.

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., … Dolan E.L. (2014). Assessment of course-based undergraduate research experiences: A meeting report. *CBE—Life Sciences Education*, *13*(1), 29–40. doi: 10.1187/cbe.14-01-0004

Bakker, A., Biehler, R., & Konold, C. (2004). Should Young Students Learn About Box Plots? Curricular Development in Statistics Education, Sweden, 2004. Retrieved November 4, 2018, from https://iase-web.org/documents/papers/rt2004/4.2_Bakker_etal.pdf

Bengtsson, L. A., & Ottosson, T. (2006). What lies behind graphicacy? Relating students' results on a test of graphically represented quantitative information to formal academic achievement. *Journal of Research in Science Teaching*, *43*(1), 43–62.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, *17*(1), 10–17.

Bertin, J. (1983). *Semiology of graphics: Diagrams, networks, maps*. Madison: University of Wisconsin Press.

Boers, M. (2018). Designing effective graphs to get your message across. *Annals of the Rheumatic Diseases*, *77*(6), 833–839.

Bowen, G. M., & Roth W. M. (1998). Lecturing graphing: What features of lectures contribute to student difficulties in learning to interpret graphs? *Research in Science Education*, *28*(1), 77–90.

Brancaccio-Taras, L., Pape-Lindstrom, P., Peteroy-Kelly, M., Aguirre, K., Awong-Taylor, J., Balser, T., … Zhao J. (2016). The PULSE Vision and Change Rubrics, Version 1.0: A valid and equitable tool to measure transformation of life sciences departments at all institution types. *CBE—Life Sciences Education*, *15*(4), ar60.

Bray-Speth, E., Momsen, J. L., Moyerbrailean, G. A., Ebert-May, D., Long, T. M., Wyse, S., & Linton, D. (2010). 1, 2, 3, 4: Infusing quantitative literacy into introductory biology. *CBE—Life Sciences Education*, *9*(3), 323–332.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*(3), 343–368.

Campbell, A. M., Heyer, L. J., & Paradise, C. J. (2014, August 14). *Integrating concepts in biology. Portsmouth, NH: Trunity.net. eTextbook for introductory college biology and high school AP biology*. Retrieved November 3, 2018, from www.trunity.com/trubook-integrating-concepts-in-biology-by-campbell-heyer-paradise.html

Cleveland, W. (1994). *The elements of graphing data*. Belmont, CA: Wadsworth.

Cleveland, W. S. (1984). Graphs in scientific publications. *American Statistician*, *38*(4), 261–269.

Cobb, P., McClain, K., & Grawemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction*, *21*, 1, 1–78. https://doi.org/10.1207/S1532690XCI2101_1

College Board (2011). *AP Biology: Curriculum Framework 2013–2014*. Retrieved November 3, 2018, from http://media.collegeboard.com/digitalServices/pdf/ap/10b_2727_AP_Biology_CF_WEB_110128.pdf

Common Core State Standards Initiative. (2012). *Preparing America's Students for College and Career*. Retrieved December 27, 2017, from www.corestandards.org

Cooper, R. J., Schriger, D. L., & Close, R. J. (2002). Graphical literacy: The quality of graphs in a large-circulation journal. *Annals of Emergency Medicine*, *40*(3), 317–322.

Cooper, R. J., Schriger, D. L., & Tashman, D. A. (2001). An evaluation of the graphical literacy of Annals of Emergency Medicine. *Annals of Emergency Medicine*, *37*(1), 13–19.

Cooper, R. J., Schriger, D. L., Wallace, R. C., Mikulich, V. J., & Wilkes, M. S. (2003). The quantity and quality of scientific graphs in pharmaceutical advertisements. *Journal of General Internal Medicine*, *18*(4), 294–297.

Corwin, L. A., Graham, M. J., & Dolan, E. L. (2015). Modeling course-based undergraduate research experiences: An agenda for future research and evaluation. *CBE—Life Sciences Education*, *14*(1), es1.

Dasgupta, A. P., Anderson, T. R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE—Life Sciences Education*, *13*(2), 265–284.

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment and Evaluation in Higher Education*, *42*(3), 347–360.

diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction*, *22*(3), 293–331.

diSessa, A. A., & Sherin, B. L. (2000). Meta-representation: An introduction. *Journal of Mathematical Behavior*, *19*(4), 385–398.

Drummond, G. B., & Tom, B. D. (2011). Presenting data: Can you follow a recipe? *Clinical and Experimental Pharmacology and Physiology*, *38*(12), 787–790.

Drummond, G. B., & Vowler, S. L. (2011). Show the data, don't conceal them. *Advances in Physiology Education*, *35*(2), 130–132.

Duke, S. P., Bancken, F., Crowe, B., Soukup, M., Botsis, T., & Forshee, R. (2015). Seeing is believing: Good graphic design principles for medical research. *Statistics in Medicine*, *34*(22), 3040–3059.

Elliott, A. C., Hynan, L. S., Reisch, J. S., & Smith, J. P. (2006). Preparing data for analysis using Microsoft Excel. *Journal of Investigative Medicine*, *54*(6), 334–341.

Evergreen, S. (2014). *Presenting data effectively*. Los Angeles, CA: Sage.

Evergreen, S. D. (2018). *Presenting data effectively: Communicating your findings for maximum impact*. Los Angeles, CA: Sage.

Federico, B., Damiani, G., Scopelliti, L., Venditti, A., Ronconi, A., Errico, A., ... Ricciardi, W. (2012). Can public health researchers effectively communicate their findings? An evaluation of graph use at the 2006 European Conference on Public Health. *Journal of Public Health*, *20*(3), 213–218.

Few, S. (2004). *Show me the numbers: Designing tables and graphs to enlighten*. Oakland, CA: Analytics Press.

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227. doi: 10.1016/j.cognition.2011.11.002

Franzblau, L. E., & Chung, K. C. (2012). Graphs, tables, and figures in scientific publications: The good, the bad, and how not to be the latter. *Journal of Hand Surgery*, *37*(3), 591–596.

Friel, S. N., & Bright, G. W. (1996). Building a Theory of Graphicacy: How Do Students Read Graphs? *Paper presented at the Annual Meeting of the American Educational Research Association held April 8–12, 1996, in New York, NY*.

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

Garfield, J. B., delMas, R. C., & Chance, B. (2007). Using students' informal notions of variability to develop an understanding of formal measures of variability. In Lovett, M.C., and Shah, P. (Eds.), *Thinking with data*. New York: Taylor and Francis.

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a Test of Scientific Literacy Skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE—Life Science Education*, *11*, 364–377.

Grawemeyer, B., & Cox, R. (2004). The effects of knowledge of external representations and display selection on database query performance. *Diagrammatic Representation and Inference Lecture Notes in Computer Science*, *2980*, 351–354.

Harsh, J. A., & Schmitt-Harsh, M. (2016). Instructional strategies to develop graphing skills in the college science classroom. *American Biology Teacher*, *78*(1), 49–56.

Hegarty M. (2011). *The role of spatial thinking in undergraduate science education (Third Committee Meeting on Status, Contributions, and Future Directions of Discipline-Based Education Research)*. Retrieved November 3, 2018, from http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072586.pdf

Hertel, J. (2018). A picture tells 1000 words (but most results graphs do not): 21 alternatives to simple bar and line graphs. *Clinics in Sports Medicine*, *37*(3), 441–462.

Holden, R. R. (2010). Face validity. *Corsini Encyclopedia of Psychology*, 1–2.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Addison-Wesley Pub. Co.

Hoskins, S. G., Stevens, L. M., & Nehm, R. H. (2007). Selective use of the primary literature transforms the classroom into a virtual laboratory. *Genetics*, *176*(3), 1381–1389.

Howard Hughes Medical Institute. (2013). *Sustaining Excellence: New Awards for Science Education to Research Universities, 2014 Competition*.

Humphrey, P. B., Taylor, S., & Mittag, K. C. (2014). Developing consistency in the terminology and display of bar graphs and histograms. *Teaching Statistics*, *36*(3), 70–75.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144.

Kellman, P. J. (2000). An update on gestalt psychology. In Landau, B., Sabini, J., Jonides, J., & Newport, E. (Eds.), *Perception, cognition, and language: Essays in honor of Henry and Lila Gleitman*. Cambridge, MA: MIT Press.

Klaus, B. (2015). Statistical relevance—relevant statistics, part I. *EMBO Journal*, *34*(22), 2727–2730.

Klaus, B. (2016). Statistical relevance—relevant statistics, part II: Presenting experimental data. *EMBO Journal*, *35*, 1726–1729.

Konold, C., & Higgins, T. (2003). Reasoning about data. In Kilpatrick, J., Martin, W. G., & Schifter, D. (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 193–215). Reston, VA: National Council of Teachers of Mathematics. Retrieved November 3, 2018, from https://www.srri.umass.edu/publications/konold-2003rad/

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

Konold, C., & Lehrer, R. (2008). Technology and mathematics education: An essay in honor of Jim Kaput. In English, L. D. (Ed.), *Handbook of international research in mathematics education* (2nd ed.) (pp. 49–72). New York: Routledge.

Kosslyn, S. M. (1994). *Elements of graph design*. New York: Freeman.

Lehrer, R., & Schauble, L. (2004). Modeling natural variation through distribution. *American Educational Research Journal*, *41*(3), 635–679.

Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, graphs, and graphing: Tasks, learning, and teaching. *Review of Educational Research*, *60*(1), 1–64.

Leonard, J. G., & Patterson, T. F. (2004). Simple computer graphing assignment becomes a lesson in critical thinking. *NACTA Journal*, *48*(2), 17–21.

Lovett, M.C., & Chang, C. (2007). Data analysis skills: What and how are students learning? In Lovett, M., & Shah, P. (Eds.), *Thinking with data*. Mahwah, NJ: Erlbaum.

Mack, C. (2013). How to write a good scientific paper: Figures, part 1. *Journal of Micro/Nanolithography, MEMS, and MOEMS*, *12*(4), 040101–040101.

Mathewson, J. H. (1999). Visual-spatial thinking: An aspect of science overlooked by educators. *Science Education*, *83*, 33–54.

McFarland, J. (2010). Teaching and assessing graphing using active learning. *MathAMATYC Educator*, *1*(2), 32–39.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation*, *7*(25), 1–10.

Montello, D. R., Grossner, K. E., & Janelle, D. G. (Eds.), (2014). *Space in mind: Concepts for spatial learning and education*. Cambridge, MA: MIT Press.

National Academies of Sciences, Engineering, and Medicine. (2017). *Undergraduate research experiences for STEM students: Successes, challenges, and opportunities*. Washington, DC: National Academies Press. Retrieved November 3, 2018, from https://doi.org/10.17226/24622

National Research Council. (2011). *Successful K-12 STEM education: Identifying effective approaches in science, technology, engineering, and mathematics*. Washington, DC: National Academies Press.

Next Generation Science Standards Lead States. (2013). *Next Generation Science Standards: For States, by States*. Retrieved December 27, 2017, from www.nextgenscience.org

Novick, L. R. (2004). Diagram literacy in preservice math teachers, computer science majors, and typical undergraduates: The case of matrices, networks, and hierarchies. *Mathematical Thinking and Learning*, *6*(3), 307–342.

Nuzzo, R. L. (2016). The box plots alternative for visualizing quantitative data. *PMandR*, *8*(3), 268–272.

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*(1), 20–26.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, *9*, 129–144.

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*(1), 95–115.

Popham, W. J. (1997). What's wrong-and what's right-with rubrics. *Educational Leadership*, *55*, 72–75.

President's Council of Advisors on Science and Technology. (2012). *Transformation and opportunity: The future of the U.S. research enterprise*. Washington, DC. Retrieved December 27, 2017, from www.whitehouse.gov/sites/default/files/microsites/ostp/pcast_future_research_enterprise_20121130.pdf

Puhan, M. A., Riet, G., Eichler, K., Steurer, J., & Bachmann, L. M. (2006). More medical journals should inform their contributors about three key principles of graph construction. *Journal of Clinical Epidemiology*, *59*(10), 1017–e1.

Raven, P. H., Johnson, G. B., Losos, J. B., Mason, K. A., & Singer, S. R. (2008). *Biology* (8th ed.). Boston: McGraw-Hill.

Rodrigues, V. (2013, April 11). Tips on effective use of tables and figures in research papers. *Editage Insights*.

Roth, W. M., & Bowen, G. M. (2001). Professionals read graphs: A semiotic analysis. *Journal of Research in Mathematics Education*, *32*(2), 159–194.

Roth, W. M., Bowen, G. M., & McGinn, M.K. (1999). Differences in graph-related practices between high school biology textbooks and scientific ecology journals. *Journal of Research in Science Teaching*, *36*(9), 977–1019.

Rougier, N. P., Droettboo, M., & Bourne, P. E. (2014). Ten simple rules for better figures. *PLoS Computational Biology*, *10*, 1–7.

Rybarczyk, B. (2011). Visual literacy in biology: A comparison of visual representations in textbooks and journal articles. *Journal of College Science Teaching*, *41*(1), 106.

Sadava, D. E., Hillis, D. M., Heller, H. C., & Berenbaum, M. (2009). *Life: The science of biology* (Vol. 2). New York: Macmillan.

Saxon, E. (2015). Beyond bar charts. *BMC Biology*, *13*(1), 60.

Schriger, D. L., & Cooper, R. J. (2001). Achieving graphical excellence: Suggestions and methods for creating high-quality visual displays of experimental data. *Annals of Emergency Medicine*, *37*(1), 75–87.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction. *Journal of Educational Psychology*, *91*, 690–702.

Shortlidge, E. E., & Brownell, S. E. (2016). How to assess your CURE: A practical guide for instructors of course-based undergraduate research experiences. *Journal of Microbiology and Biology Education*, *17*(3), 399.

Singh-Cundy, A., & Shin, G. (2010). *Discover biology* (6th ed.). Sunderland, MA: Sinauer Associates.

Slutsky, D. J. (2014). The effective use of graphs. *Journal of Wrist Surgery*, *3*(2), 067–068.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation*, *9*(4), 1–19.

Stengel, D., Calori, G. M., & Giannoudis, P. V. (2008). Graphical data presentation. *Injury*, *39*(6), 659–665.

Tufte, E. R. (1983). *Visual display of quantitative information*. Cheshire, CT: Graphic Press.

Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., Jackson, R. B., & Reece, J. B. (2014). *Campbell biology in focus*. San Francisco, CA: Pearson.

Vitale, J. M., Lai,K., & Linn, M. C. (2015). Taking advantage of automated assessment of student-constructed graphs in science. *Journal of Research in Science Teaching*, *52*(10), 1426–1450.

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 Biology*, *13*(4), e1002128.

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

Wright, L. K., Cardenas, J. J., Liang, P., & Newman DL (2018). Arrows in biology: Lack of clarity and consistency points to confusion for learners. *CBE—Life Sciences Education*, *17*(1), ar6. doi: 10.1187/cbe.17-04-0069