Spring 2016

Exploratory study of graph drawing on a continuum of expertise

Emily MacLeish
James Madison University

Exploratory Study of Graph Drawing on a Continuum of Expertise

An Honors Program Project Presented to

the Faculty of the Undergraduate

College of Biology

James Madison University

by Emily Rose MacLeish

May 2016

Accepted by the faculty of the Department of Biology and Mathematics & Statistics, James Madison University, in partial fulfillment of the requirements for the Honors Program.

FACULTY COMMITTEE:

Project Advisor: Joseph Harsh, Ph.D.,
Assistant Professor, Biology

Reader: Samantha Prins, Ph.D.,
Assistant Professor, Mathematics & Statistics

Reader: Steve Cresawn, Ph.D.,
Associate Professor, Biology

HONORS PROGRAM APPROVAL:

Bradley R. Newcomer, Ph.D.,
Director, Honors Program

PUBLIC PRESENTATION

This work is accepted for presentation, in part or in full, at James Madison University Biology Symposium on April 14, 2016.

Table of Content

Acknowledgments

I would like to acknowledge the Biology department and faculty of James Madison University for the funds and feedback that made this research possible. I would also like to thank Dr. Steve Cresawn and Dr. Sam Prins for their questions, guidance, and support that forced me to look at each situation from multiple perspectives. Lastly, I would like to acknowledge my advisor, Dr. Joe Harsh. Words cannot express the impact you have had on me throughout this process. Thank you for the patience, wisdom, and drive that was necessary to make this study successful.

Abstract

Graphs are used in our lives daily to communicate information such as political ads or car sales. In the sciences, understanding graphs is essential to effective communication as graphs are often used to report experimental results or observed trends. However, research suggests that college students are not fluent in this form of scientific communication. Additionally, research has also found that standardized assessments of quantitative literacy fail to be clearly defined at the curricular or institutional levels. This research looks at the differences between the cognitive and metacognitive strategies of how individuals along a continuum of biological expertise visually represent data. As a result, an instrument was created from expert feedback and graphing literature to test if differences exist in how individuals transform graph data and if those differences are a function of scientific expertise. The instrument collected data on graph drawing and cognitive interviews (i.e. think-aloud) from 35 participants with varying biology experience, including 13 non-biology majors, 9 non-senior biology majors, 7 senior biology majors and graduate students, and 6 biology faculty. Rubrics were used to evaluate performance in graph drawing and think-aloud components. Although no statistical differences were identified between groups in graph drawing tasks, analysis of specific graph drawing components (e.g., graph type) did reveal variation as a function of expertise. Significant differences were found between expertise groups in the cognitive and metacognitive strategies discussed in the think-aloud data (e.g., why a graph was drawn in that manner). These findings begin to identify differences between experts and novices in Biology, as well as the lack of alignment in one's ability to depict graphical data and actual understanding of graphing practices, which may be used to inform instruction to increase graph literacy. Additionally, the instrument designed for

this study has high face validity, but future work will be needed to establish reliability as only one researcher was able to score data. Increasing reliability will allow this instrument to be an effective tool for faculty interested in assessing their students' data display skills.

Key words: graph, drawing, think-aloud, biology, expertise

Exploratory Study of Graph Drawing on a Continuum of Expertise

1. Introduction

The prevalence of data displays in today's society has led to an increasing need for all students to develop competency in visual data analysis skills. Using visual representations is beneficial because it allows for the rapid perception of linkages and relationships among data which literary language does not. While valued across disciplines, the need for quantitative literacy skills is particularly important in the sciences for the effective communication of varied and complex information. (Kotzebue, Gerstl, & Nerdel, 2015)

The ability to communicate through numerical data is referred to as quantitative literacy (AACU, 2014). Quantitative literacy (QL) is described as "the skill of using simple mathematical thinking to make sense of numerical information" and "refers to the ability to interpret data and to reason with numbers within "real-world" situations" (Bray-Speth, Momsen, Moyerbrailean, Ebert-May, Long, Wyse, & Linton, 2007; pg. 324). The Association of American Colleges and Universities (AACU) has identified QL as one of the key competencies that all students, independent of discipline, should attain throughout the course of their undergraduate educations. Strong QL skills are apparent when an individual can generate and communicate an argument supported by an assortment of quantitative evidence (e.g., graphs, mathematical equations, and tables; AACU, 2014).

Within one's broader QL skill set, the learned cognitive ability that requires the use of mental tools to build and interpret graphical data representations is identified as graph literacy (Duesbery, Werblow, & Yovanoff, 2001). Similar to reading text, graph literacy requires "repeated practice and focus on greater complexity as students develop their skills" (Zucker,

Staudt, & Tinker, 2015; p.20). As the complexity of visual data representations increases, the cognitive demands in attending to this information increase as well. Cognitive demand for a given task can be viewed as the combination of the degree of processing (i.e. the quantity of simultaneously observed information) and depth of knowledge (i.e. familiarity or skills related to the topic) required (Duesbery, Werblow, & Yovanoff, 2001). For example, a graph with one variable would have a lower degree of processing than a graph with two variables, and therefore would require less cognitive processing. The more depth of knowledge tasked, the more cognitive processes required (Duesbery et al., 2001).

In undergraduate biology education, graph literacy has been recognized as a core competency in preparing students for STEM careers and data-based decision making as educated citizens (AAAS, 2011; Woodin, Carter, & Fletcher, 2010). Given the importance of succinctly conveying complex information in the field, scientific communication is often measured based on one's ability to construct, interpret, and apply graphs to various data. Despite this focus, standardized measures for assessing levels of QL largely fail to be clearly defined at the curricular or institutional levels (Bray-Speth, et al., 2007). Many college students are not fluent in scientific communication as measured by their ability to make sense of or construct visual data representations (Glazer, 2011).

To date, few studies have focused on the development of adults' graph literacy in the sciences, and even fewer on graph construction (as reviewed in Glazer, 2011). Recent investigations have begun to explore differences between undergraduate students and scientists' performance in interpreting graphical data (Maltese et al., 2015); however, there has been little research done of the development of the graph drawing skills as a function of expertise. The

purpose of this study is to better understand *how* individuals of varying biology backgrounds represent graph data. To address this, the research questions for this study include: (1) Are there differences in the cognitive and metacognitive strategies used to represent biological data graphically? (2) If differences do exist, are they a function of biological expertise? (3) Is the instrument developed for this study a valid and reliable measurement of one's graph drawing skills?

In line with prior graph research (Maltese et al., 2015 & Harsh, Maltese, and Warner, 2012), it is anticipated that graph construction is a function of scientific expertise. The information collected during this project aims to help instructors educate students on the proper way to address complex graphical information. By examining differences in students and scientists' cognitive and metacognitive processes, it is anticipated that the findings of this study will provide information addressing knowledge gaps and areas most in need of instructional emphasis to foster the development of data skills (Harsh, 2014). Understanding how experts solve problems is an effective means of facilitating the transfer from novice to more expert-like performance (Hmelo-Silver, 2004).

2. Methods

2.1 Performance based assessments (PBAs)

A performance-based instrument was developed to measure the relationship between expertise and graph construction. Performance based assessments (PBAs) are tools used to measure the knowledge utilized during the construction of a response to authentic domain-specific tasks, which provide direct evidence to educational outcomes (Linn, Baker, & Dunbar, 1991). PBAs can have instructional, diagnostic, and monitoring purposes lending to their

effectiveness as means of testing the science knowledge and practices of college students (Linn et al., 1991). This study focuses on the development and implementation of an instrument (i.e. tasks and associated rubrics) to assess individuals' graph construction skills. Similar to previous research that involved the design and testing of performance instruments (e.g., Stein, Haynes, & Redding, 2007), the tasks and associated rubrics were influenced by preexisting relevant instruments (e.g., Bray-Speth et al., 2007, Harsh et al., 2012, and Picone et al., 2007), assessment literature (e.g., Linn et al., 1991; Mehrens, 1992), and recursive feedback from experts in biology and science education.

The first step of the design process was to review publicly available graphs (i.e. those from textbooks, governmental websites, etc.) to identify a set of exemplar data representations that could serve as the basis for the graph drawing tasks. Based on key graphing characteristics identified in the literature (Glazer, 2011), ten graphs were initially selected from a variety of scholarly sources that varied in type, number of variables, specific topic, and other graph components. Each of the potential graphs were described by a number of features, including: general background (i.e. graph focus, title, and citation), graph characteristics (e.g., graph type, unique features), a difficulty score based on graph characteristics identified as being challenging for students (Glazer, 20ll), and a rationale to why the graph data were selected. In addition, a brief (2 to 3 sentence) background describing the nature of the data (e.g., defining general terminology, highlighting topic importance) was included to provide context for the participant. Contextualization is important in PBAs to increase test fairness (Linn, 1991), and has been identified as key feature in making sense of graph data for students and scientists (Roth & McGinn, 1997). The four graphs that serve as the basis of the instrument were selected based on

their design features (e.g., data, graph type) to provide a range in complexities and feedback from faculty in biology, statistics, and biology education at James Madison University (JMU) and Indiana University (IU).

The graph drawing tasks were piloted with an expert faculty member from both the Education and Biology departments at JMU. Pilot testing consisted of the expert constructing graphs that he/she felt "best" represented the provided information (i.e. data table and context background) as well as a discussion about various graphing elements. The discussion included the expert detailing how he/she felt about each task, and drawing predictions regarding the future performances of each expertise group. The feedback was used to refine the graphing tasks and inform rubric development.

The feedback from the pilot studies was one means of establishing the face validity of this instrument. In assessment, *validity* refers to the extent to which a given assessment succeeds in measuring the particular competencies (e.g., graphing drawing) that is was developed to assess (Mehrens, 1992). Along with the pilot study feedback, the face validity of the instrument was strengthened through the use of data displays and associated information drawn from primary literature.

2.2 Rubric development

Elements significant to graph drawing were identified from pre-existing literature to act as the basis of the scoring rubrics (e.g., Bray-Speth et al., 2007, Harsh et al., 2012; Kotzebue et al., 2015; Picone et al., 2007). A modified version of Harsh and others' (2012) graph drawing rubric was used to assess how participants represented the four provided tabular data sets. The rubric focuses on three elements (i.e. framework, content, and labeling) identified by Kosslyn

(2006) as being essential to graph design. To assess the cognitive and metacognitive strategies employed in graph drawing, scoring criteria were developed for this study as no relevant rubrics were identified in a review of the literature. Criteria to measure how and why participants chose to represent data in a given manner were developed based on the author's experience in collecting data for a separate study on graphing (Harsh et al., *in preparation*) and graphing literature (Kotzebue et al., 2015). Both rubrics were developed to be easily modified based on the characteristics of the graphing task (e.g., std. error bars).

Weightings of the scoring criteria (Appendix D-G) are based on prior instruments (Harsh et al., 2012; Kotzebue et al., 2015) and feedback from JMU biology faculty. Expert feedback was collected on what graphing features (i.e. criteria) biology faculty identified as being important, and how they would weight each identified feature relative to others (Appendix A). Feedback was requested from 11 biology professors with seven (64% response rate) providing feedback. The faculty feedback was averaged along with the weights suggested in related literature (Harsh et al., 2012) to generate the weighting of each criterion. The averages were rounded to the nearest half a point to allow for clearer values while scoring (Appendix B). The same procedure was taken to establish weights for think-aloud criteria weights (Appendix C and D).

For the development of effective measures, evidence to the validity and reliability of the scoring criteria is of particular importance (Linn et al., 1991). In respect to validity, which was defined above, the feedback gathered from these experts contributed to the face validity of the rubrics as they identified criteria that they would consider important and how they would weight the respective criteria for evaluating graph drawings in a professional context. Reliability of an instrument is supported through the instrument's reproducibility (Wass et al., 2001; p.946),

which can be improved through training and extensive practice in the consistent use of scoring criteria. In preparation for this study, over a four month period, the primary researcher gained familiarity in the collection and scoring of graph and think-aloud data as part of a prior research project that served as the basis of this work (Harsh et al., *in preparation*).

2.3 Participant Recruitment

Participants were recruited through in-person and electronic solicitation during the Fall of 2015. Social media and departmental listservs were also used to communicate with potential participants. Non-biology majors were recruited to participate and served as a baseline group for comparative purposes. Permission was obtained from the JMU Institutional Review Board (IRB), and participants were provided consent forms detailing the project prior to the voluntary completion of the task. For their efforts, participants were compensated with a small stipend ($5).

2.4 Procedure

Graph drawing and think-aloud data were collected using an electronic tablet and Vittle, a recording application (http://www.qrayon.com/home/vittle/), to allow pen strokes and audio to be synced. Participants were given a brief orientation of the application, then asked to construct the "best" graph for a given data table and context, while verbalizing their reasoning for representing data in such a manner. A cognitive interview (CI; Appendix E) followed each task to further reveal the depth of participants' understanding of data representation. Cognitive interviews are methods for improving information recall that are based on the premise that "retrieval will be enhanced if the context experienced at retrieval matches that experienced during encoding" (Wright & Holiday, 2007; p.20). After the completion of each session,

participants completed a 20 question, online Qualtrics survey to collect data regarding their educational and demographic background as well as experience with graphing.

For each graph task, the primary researcher scored the drawing and think-aloud tasks separately using the graph-specific rubrics included in Appendices F-I. In addition, audio from the CIs were scored as think-aloud data. The scored results from the graph drawing and think-aloud components were analyzed using *IBM SPSS v.23* to examine potential differences as a function of expertise using Kruskal-Wallis H tests, which are the nonparametric analog of ANOVA a used to identify statistically significant differences in the dependent variable across groups defined by at least two independent variables (S. Prins, personal communication, March 4, 2016). This test has four major assumptions: (1) there are two or more independent, categorical, variables, (2) the dependant variable is ordinal or continuous, (3) groups have the same shape and variance, and (4) there is independence of observances.

3. Results

3.1 Participants

Data were collected from 35 participants at James Madison University (JMU) distributed across four levels of biological expertise including non-biology students (n=13), biology undergraduate students (n=8), biology graduate or senior undergraduate students[1] (n=7), and biology faculty (n=6). Due to technical issues in audio recording, cognitive interview data (i.e. think-aloud) from 25 participants (faculty [*n*=3], biology graduate or senior students [*n*=4] biology undergraduate students [*n*=8], and non-biology students [*n*=10]) were collected and

[1] Biology graduate and undergraduate students were grouped together due to the large quantity of upper level biology courses completed by both groups.

analyzed. In addition, two individuals chose not complete graph drawing tasks 2 and 3, because they concluded that the data were best represented in the data table provided. As these participants did not attempt to complete these tasks their abilities could not be measured, therefore, their scores were removed prior to statistical analysis. However, for those few individuals (n=2) who began to construct a graph, but then stopped and stated that the data table was the best representation. These individuals' scores, although outlier, were included in statistical analysis, because the participant had already revealed some of their abilities. It should be noted that these incomplete tasks occurred in the middle of the instrument (i.e. tasks 2 and 3) suggesting that the participants' failure to complete these tasks was not a result of timing (i.e. being rushed to finish) or test fatigue.

The Qualtrics survey responses reported on the average approximate number of college science (i.e. life sciences, physical sciences, applied sciences, and environmental sciences) and math (i.e. mathematics, statistics, and economics) classes taken by each level. Survey data indicated non-biology students averaged 3.4 (±0.85) science and 3 (±0.71) math classes. Biology undergraduates averaged 5.8 (±3.04) science classes and 3.9 (±2.81) math classes. Graduate and senior Biology students averaged 22.7 (±2.18) science classes and 4.7 (±0.64) math classes. Biology faculty averaged 41.3 (±12.05) science classes and 6.8 (±1.45) math classes. Participants also ranked their comfort reading and interpreting graphs on a Likert-type scale from one (*no experience*) to five (*can instruct others how to complete*). The average comfort level for non-Biology students was 4 (±0.17), Biology students was 3.5 (±0.27), Biology graduate and senior undergraduates was 4.4 (±0.20), and Biology faculty was 4.8 (±0.17).

3.2 Kruskal-Wallis H

The results of the Kruskal-Wallis H nonparametric tests (Appendix J) indicated differences across groups. Eleven tests were run including comparisons drawn for each graph independently (i.e graph drawing and think-aloud data separately), the totals (i.e. accumulative score for graph drawing and think-aloud data separately) across the four graphs, and then the graph drawing and think-aloud totals were combined to look at potential differences between expertise groups. Significant differences were identified between expertise groups in seven of the eleven tests, including: the graph drawing of graph task 4 (i.e. Figure 4, Appendix I), the think-aloud data of all graphs including the total, and the combined think-aloud and graph drawing totals. To account for potential multiple comparison effects, significance values were adjusted from a standard 0.05 to 0.0045 (i.e. p=0.05/11 [number of tests]) using the Bonferroni approach (S. Prins, personal communication, March 4, 2016). With this adjustment, significant differences were identified between groups in the think-aloud data for Graphing Task 1 and 2 (Figures 1 & 2) and then across all graphing tasks (Figure 3). In addition, significant differences among groups were also noted in the combined drawing and think-aloud across all tasks (Figure 4).

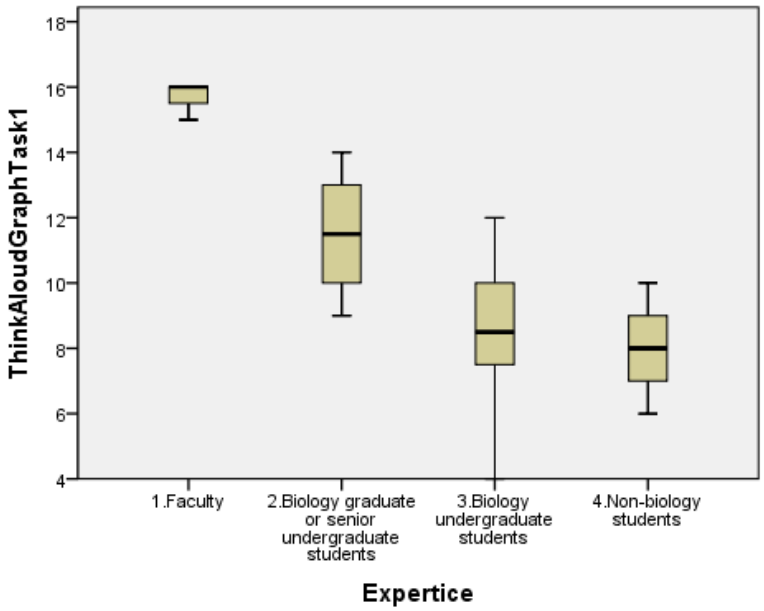


Figure 1: Box and Whiskers Plot displaying participant scores for the think aloud test data by expertise group for Graph Task 1. The numbers displayed on the X-axis are representative of the expertise groups: 1.faculty (n=3), 2.biology graduate or senior students (n=4), 3.biology undergraduate students (n=8), and 4.non-biology students (n=10).

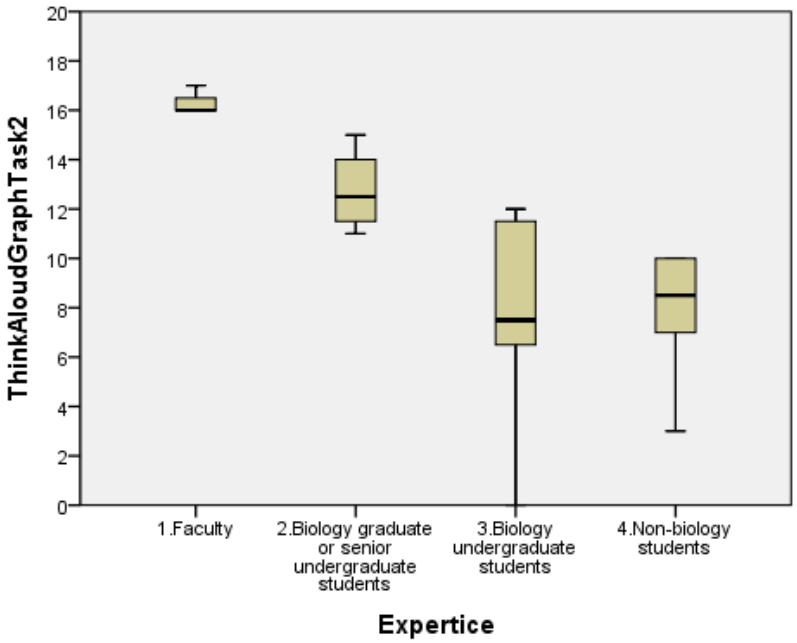


Figure 2: Box and Whiskers Plot displaying participant scores for the think aloud test data by expertise group for Graph Task 2. The numbers displayed on the X-axis are representative of the expertise groups: 1.faculty (n=3), 2.biology graduate or senior students (n=4), 3.biology undergraduate students (n=8), and 4.non-biology students (n=10).

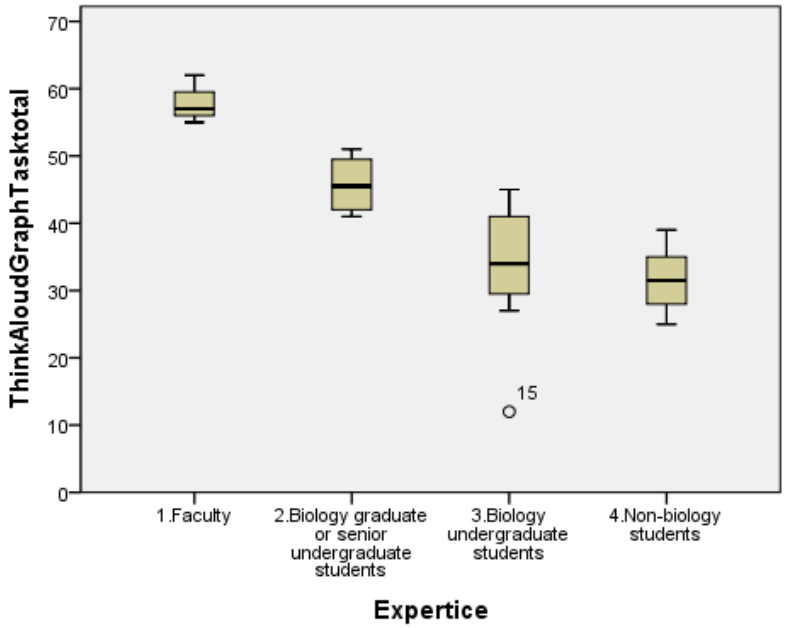


Figure 3: Box and Whiskers Plot displaying participant scores for the think aloud test data by expertise group across all four graphs. The numbers displayed on the X-axis are representative of the expertise groups: 1.faculty (n=3), 2.biology graduate or senior students (n=4), 3.biology undergraduate students (n=8), and 4.non-biology students (n=10).

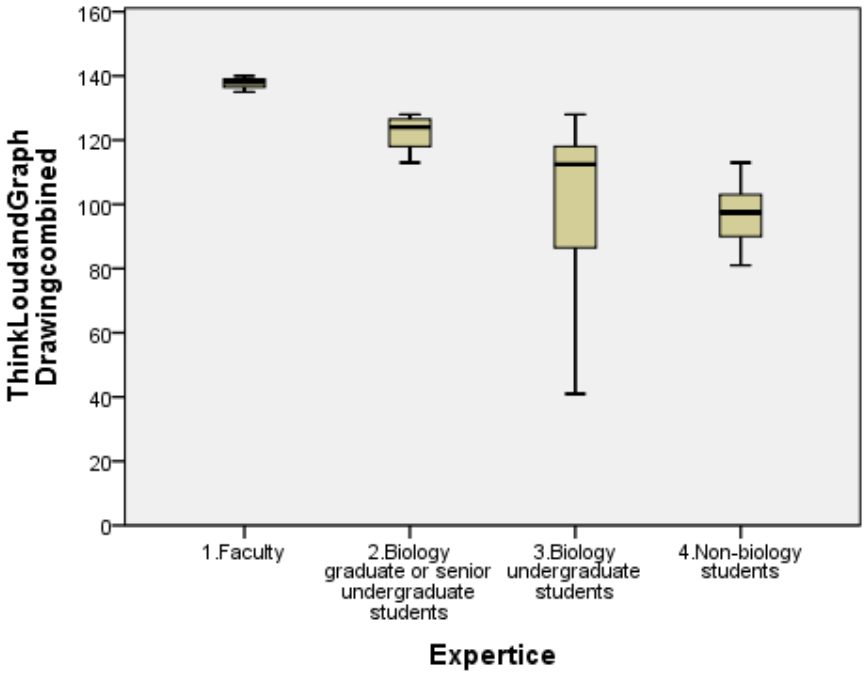


Figure 4: Box and Whiskers Plot displaying participant scores for the think aloud and graph drawing tests data by expertise group all four graphs. The numbers displayed on the X-axis are representative of the expertise groups: 1.faculty (n=3), 2.biology graduate or senior students (n=4), 3.biology undergraduate students (n=8), and 4.non-biology students (n=10).

3.3 Comparisons in Decision Making

A closer look at participant scores within each section of the rubrics was taken to further identify where differences between groups exist. The graph drawing rubric was divided into three subsections including framework (i.e. axes layout, graph type, and variable position), content (i.e. proper data and placement), and labels (i.e. axes labels, color/texture labels, unit labels, proper scaling, grouping labels, key or legend, and title or figure legend) and the mean score for each expertise was determined (Table 1).

Table 1: Means and standard error for the graph drawing subsections by expertise group. "Mean possible score" is the sum of scores possible for each criteria within a subsection, divided by the number of criteria included.

Framework			
Expertise group	**Mean possible score**	**Mean**	**S.E.**
Non-biology undergraduates	1.667	1.236	0.022
Biology undergraduates		1.333	0.027
Biology graduates and seniors		1.472	0.037
Biology faculty		1.5	0.039
Content			
Expertise group	**Mean possible score**	**Mean**	**S.E.**
Non-biology undergraduates	1.444	0.843	0.056
Biology undergraduates		0.826	0.043
Biology graduates and seniors		0.870	0.008
Biology faculty		0.986	0.076
Labels			
Expertise group	**Mean possible score**	**Mean**	**S.E.**
Non-biology undergraduates	1.283	0.699	0.011
Biology undergraduates		0.718	0.018
Biology graduates and seniors		0.801	0.024
Biology faculty		0.859	0.037

As seen in Table 1, average scores across content, framework, and labeling subsections were found to generally increase as a function of expertise to varying degrees.

The think-aloud rubric was also divided into three subsections including framework (i.e. explanation of axes labels and graph type choice), identification (i.e. identification of illustrated relations and mistakes and explanation of scale range, color use, and error bar inclusion), and off-reading (i.e. focus on trend, predictions, and interpretation of graph). The mean score for each expertise was determined (Table 2).

Table 2: Means and standard error for the think-aloud subsections by expertise group. "Mean possible score" is the sum of scores possible for each criteria within a subsection, divided by the number of criteria included.

Framework			
Expertise group	**Mean possible score**	**Mean**	**S.E.**
Non-biology undergraduates	2.333	0.885	0.039
Biology undergraduates		0.947	0.060
Biology graduates and seniors		1.281	0.047
Biology faculty		2.125	0.051
Identification			
Expertise group	**Mean possible score**	**Mean**	**S.E.**
Non-biology undergraduates	1.706	0.806	0.032
Biology undergraduates		0.876	0.055
Biology graduates and seniors		1.171	0.019
Biology faculty		1.372	0.021
Off-reading			
Expertise group	**Mean possible score**	**Mean**	**S.E.**
Non-biology undergraduates	1.833	0.742	0.009
Biology undergraduates		0.964	0.045
Biology graduates and seniors		1.208	0.032
Biology faculty		1.394	0.018

In total, two general patterns were identifiable across the graph drawing and think-aloud data. First, the mean ratings for all graph drawing and think-aloud subsections increased as a function of expertise. Second, in general, there was greater differentiation between groups in the think-aloud data in comparison to the graph drawing data.

4. Discussion

The purpose of this study is to better understand how people of varying biology backgrounds draw graphs through the use of performance based data. The statistical analysis through the Kruskal-Wallis H test identifies that significant differences exist in the cognitive and metacognitive strategies of end members to represent biological data graphically. These initial findings extend prior work that has largely focused on the graph drawing difficulties of early science majors (Bray-Speth, 2007; Picone et al., 2007) by identifying that differences exist between how students and scientists represent and think about graph data.

The graphs generated from the Kruskal-Wallis H tests (Figures 1-4) suggest these differences exist as a function of expertise, specifically between members of Biology faculty and non-biology undergraduates. Beyond this, in looking closer at the data collected (Table 1), the means of each rubric subsection increased as a function of expertise. Given the nature of learning, it was anticipated that incremental increases would be seen in performance along the continuum of expertise as one progressed from novice to expert.

In the identification of decision-making differences between expertise groups, the highest variation in graph drawing data skills was found within the framework subsection. Similarly, there was high variation of means within the framework subsection of the think-aloud data. This variability suggests students may lack understanding of *why* to represent data in a given manner (e.g., types of data, how to display data types), and more instructional emphasis should be placed here. Further support for this claim can be heard in participant comments. Participant A, a biology undergraduate, wavered in his selection of graph type (i.e. scatter plot, line graph, and bar graph) and ultimately admitted that he selected a scatter because “a scatter plot is go-to when

I don't know what to do with data." A mistake or misunderstanding in the framework of graph construction can often lead to multiple mistakes. Participant B, a biology graduate or senior undergraduate, experienced this ripple effect when she reversed the variable positions on the basis "of figures I see a lot". Participant B admitted to sacrificing the graph type she thought best fit the data, for the variable position with which she was familiar: "I hadn't looked at the data yet and saw that I wasn't going to be able to draw a line." As a result, participant B selected a scatter plot when a line would have been the "best" selection.

These results also show that there is no a significant difference in the cognitive and metacognitive strategies of undergraduates within the field of biology and undergraduates within other fields of study, which may be due to three reasons. First, this may be attributed to other fields training undergraduates in generalizable strategies for graph drawing or participants may have a background or interest in the field of biology even though it is not their academic major. The lack of difference between groups could also be a result of the non-biology undergraduates including other science majors. The non-biology group included the results of two health science students, a geology student and a geographic sciences student. These non-biology students may have more graphing experience, based on other biology coursework, than first year biology undergraduates which were included in the biology undergraduate group. As a result, the inclusion of these majors may have shifted the means of the non-biology group closer to the mean of biology undergraduates. Similarly, the biology undergraduate group, consisted of primarily freshman biology majors. Although entry level biology classes are expected to focus some attention on the topic of graph drawing, these students have not had repeated exposure to the material in a college setting. Additionally, according to the Qualtrics survey data these

students have not had many math classes, specifically statistics. One participant expressed this concern, suggesting that she would know what to do with standard error once she takes statistics. Second, as noted above, learning occurs as small, gradual changes, and therefore, significant differences should not be expected between close groups. Third, it is possible that the tests or rubrics of this instrument were not sensitive enough to identify differences between groups.

Although not all tasks resulted in a significant difference, the lack of significance may be equally important. The box and whisker analysis (Figures 1-4), as well as the means and standard error (Appendix K) for each task reveal the high range in performance within the biology undergraduates group and non-biology undergraduate groups. Through the collection of further data, the variability of these two groups, and ideally within all groups, would not hide the true means and possibly reveal a significant difference that is currently blocked by outliers. These findings provide evidence to support that more data should be collected to effectively determine if more differences exist and to further evaluate the sensitivity or effectiveness of the tasks to distinguish between groups. The subsection scores analysis begins to identify areas in which differences exist and gaps are present in student learning. Future studies should consider analyzing scores for particular criteria within each subsection in addition to graph drawing as a whole.

The final intention of this study is to determine if the instrument used is a valid and reliable measurement of one's biological graph drawing skills. Expert feedback, pilot studies, and pre-existing literature have begun to successfully establish the face validity of this instrument. At this time, the reliability of the instrument has not been established to the extent of validity. One rater increases the potential of bias in scoring, limiting the validity of scores. Only

having a single rater also affects the reliability of the instrument. The reliability can be strengthened through the future incorporation of a second trained scorer. The inter-rater reliability will assist in determining the instrument's reliability.

4.1 Limitations

There are three limitations that can be noted for this study. First, there is the potential that the participants' graph drawing skills may not be fully tested by the scope of the tasks used here. It is impossible to explore every graph type or subject; however, these tasks are what we expect members of the field to do (i.e. read and draw graphs with data they may or may not be familiar with). This limitation is accounted for by varying the difficulty and topics, as well as providing background information to help participants contextualize, minimizing the likelihood this limitation will occur.

Additionally, this study consisted of small sample sizes. The think-aloud tests also suffered smaller sample sizes due to Vittle's inability to generate the audio for 10 participants. Although notes regarding participant responses were collected, the data would not not have been consistent if these 10 participants were scored according to written notes only and no audio. Therefore the participant pool decreased to 25 participants for the think-aloud component, which limits the generalizability of the findings presented here due to the small sample size. Despite this limitation, this exploratory study has advanced a valid graphing measure that future studies can use to assess graphing skills and further evaluate the instrument's properties.

The third limitation of this work is that participant data was assessed by only one researcher, limiting the reliability of the assessment, as the responses were only scored by a

single rater, estimates to the reliability of the measure could not be drawn. Future work will have multiple scorers to enhance reliability through interrater reliability.

4.2 Implications

The ability to effectively understand increasingly complex data representations is of growing importance and has become a key component of being scientifically literate (Tairab & Khalaf Al-Naqbi, 2004). According to Harsh and Maltese, "[i]nstructors of undergraduate courses should not expect students to come into courses with high proficiency for understanding, interpreting and creating data visualizations" (2012b; pg. 10). Teachers work to develop graphical literacy skills in students at all levels of the education system. At each level, however, misconceptions surface that are not recognized by the instructors and are therefore not corrected for the students. This study has begun to identify general differences in student graph drawing and interpretation performances compared to the performance of experts in biology.

Primary implications of this study are to fill in gaps in the literature by focusing on (a) how adults draw graphs and (b) the differences between novices and experts in graphing performance. This study extends prior research that has narrowly focused on first-year science students' graphing skills based on general descriptions of their graph design (Bray-Speth et al, 2077; Kotzebue et al., 2015; Picone et al., 2007). In addition, verbal components (i.e. participants voicing thoughts during construction and post construction question responses) provide a deeper look into the cognitive and metacognitive processes of students and scientists.

The broader objective of this exploratory study was to design valid and reliable performance-based tasks and scoring criteria that can be adopted by biology faculty to assist in the assessment of their students' graph drawing skills. While further research needs to be

conducted to evaluate the reliability of these measures, components of this project have been incorporated into the 2016 JMU Biology Majors Assessment for graduating seniors and will be used in the assessment of the department's new first year curriculum. Contributing to the development of students' graphing skills has the potential to identify and close gaps in the understanding of graph drawing.

Appendix A: Faculty Graph Drawing Rubric Feedback

	Criteria:	Validate graph type	Explanation of Axis labels	Recognition of illustrated relation	Explanation of scale	Recognition of mistakes	Recognition of the effect of familiarity on construction	Focus on values vs. trends	Offer prediction	Explanation of data from graph	Recognition of difficult elements	Explanation of color/pattern usage	total
Values according to literature		2	3	2	3	2		1.5	2	3		1.5	20
Responder	BIO Faculty 1	2	1	2	1	2		2	2	3	2	1	18
Responder	BIO Faculty 2	1	1	3	1	2	1	4	3	2	1	1	20
Responder	BIO Faculty 3	4	2	3	1	1	1	2	1	3	1	1	20
Responder	BIO Faculty 4	5	5	5	5	0	0	0	0	0	0	0	20
Responder	BIO Faculty 5	3	2	2	1	1	1	2	3	2	2	1	20
Responder	BIO Faculty 6	2	2	2.5	1.5	2	1	1.5	2	2.5	1.5	1.5	20
Responder	BIO Faculty 7	3	3	3	1	1	1	1	1	4	1	1	20
	Average	2.75	2.375	2.8125	1.8125	1.375	0.833333333	1.75	1.75	2.4375	1.214285714	1	

Appendix B: Faculty weights average rounded for written criteria

Criterion	Average feedback score	Assigned weight
Layout of axes	2.3071429	2.5
Graph type	4.2357143	4
Variable position	2.8071429	3
Proper information	3.2357143	3
Data positions	1.8071429	2
Content differentiation (via color, patter, etc.)	1.5928571	1.5
Error bars	1.5928571	1.5
Connected with a line or trend line	1.0214286	1
Calculated mean or average	1.45	1.5
Axes labels	2.8071429	3
Correct units	1.5214286	1.5
Title or figure legend	1.0928571	1
Appropriate scale	1.5928571	1.5
Groupings labeled	1.0928571	1
Legend or caption/key	1.8071429	2
Total	29.964286	30

Appendix C: Faculty Think-Aloud Rubric Feedback

	Criteria:	Layout of axes	Graph type	Variable position	Proper information	Data positions	Content differentiation (via color, patter, etc.)	Error bars	Connected with a line or trend line	Calculated mean or average	Axes labels	Correct units	Title or figure legend	Appropria te scale	Groupings labeled	Legend or caption/key	total
Values according to literature		1.65	3.65	1.65	1.65	2.65	1.65	1.65	2.15	1.65	2.65	1.65	1.65	2.65	1.65	1.65	30.25
Responder	BIO Faculty 1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	30
Responder	BIO Faculty 2	4	5	2	3	2	3	0.5	0.5	0.5	2	1	2	0.5	0.5	3	29.5
Responder	BIO Faculty 3	1.5	4	4	4	3	0.5	1	0.5	1	4	2	0	2	0.5	2	30
Responder	BIO Faculty 4	5	10	5	5	0	0	0	0	0	5	0	0	0	0	0	30
Responder	BIO Faculty 5	2	2	2	3	2	1	3	2	2	3	2	2	2	1	1	30
Responder	BIO Faculty 6	1	2	3	3	2	2	2	1	3	2	2	2	1	3	1	30
Responder	BIO Faculty 7	1	3	2	3	1	3	3	1	2	1	1	2	1	1	3	30
	Average	2.26875	3.95625	2.70625	3.08125	1.83125	1.64375	1.64375	1.14375	1.51875	2.70625	1.58125	1.20625	1.64375	1.20625	1.83125	

Appendix D: Faculty weights average rounded for verbal criteria

Criterion	Average feedback score	Assigned weight
Validate graph type	2.857143	3
Explanation of Axis labels	2.571429	2.5
Recognition of illustrated relation	2.928571	3
Explanation of scale	1.928571	2
Recognition of mistakes	1.285714	1
Recognition of the effect of familiarity on construction	0.833333	1
Focus on values vs. trends	1.714286	1.5
Offer prediction	1.714286	1.5
Explanation of data from graph	2.357143	2.5
Recognition of difficult elements	1.083333	1
Explanation of color/pattern usage	1	1
Total	20.27381	20

Appendix E: Post Graph Drawing Questions

Script

1. Is there a reason you choose to represent the data in this way?
    a. Why did you choose to use a _____ graph instead of other options?
2. Is there a reason you chose to position the variables as you did on the graph?
    a. Why did you use the scale that you did?
3. After drawing the graph is there any modifications or changes you would make to your graph?
4. How would you interpret the data in the graph?
5. Have you been asked to draw a graph like this in the past?
6. Do you have any background knowledge, or wish you had some background knowledge, regarding these topics that might have helped you?
7. What were some of the difficult components you encountered in these graphs?

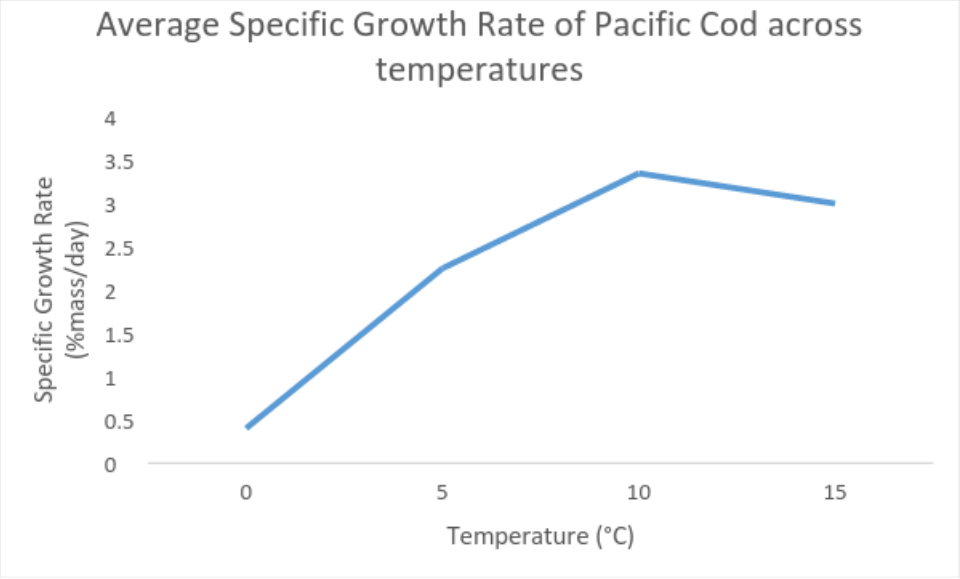
Appendix F: Graph One Task and Associated Rubrics

Task

Temperature (°C)	Specific Growth Rate (%mass/day)		
	Trial 1	Trial 2	Trial 3
0	0.40	0.55	0.25
5	2.45	2.25	2.05
10	3.55	3.35	3.15
15	3.00	3.15	2.85

Historically, the Pacific Cod (*Gadus macrocephalus)* is an important commercial food species. As a result of intensive fishing practices globally reducing the number of wild Pacific Cod, studies have been undertaken to assess the fishes' growth dynamics. These are assessed through feeding experiments at different temperatures in artificial environments for conservation and food production purposes. The data above represent the specific growth rate (measured in percent change of body mass per day) for Pacific Cod grown at four different incubation temperatures in three replicate tanks per treatment. Generate a graph(s) that best represents the data provided.

Anticipated result from: Helser, T. E., Colman, J. R., Anderl, D. M., Kastelle, C. R. 2016. Growth dynamics of Saffron cod (*Eleginus gracilis*) and Arctic cod (*Boreogadus saida*) in the Northern Bering and Chukchi Seas. US Dept. of the Interior, Bureau of Ocean Energy Management, Alaska OCS Region. OCS Study BOEM 2011-AK-11-08 a/b. 50 pp.



Graph 1 Written		
Participant #:	1	2
Framework – indicates what kinds of measurements are being used and what things are being measured		
· Is the layout of the graph axes being used the most effectively to represent the data? (0 = no axes, 1=layout not effective, 2 = layout affective but not best 2.5=layout is extremely effective)		
· Is the proper type of graph being used to most effectively represent the data? (0 = nothing drawn, 1 = graph being used is not appropriate, 2 = a graph that can be used but is not possibly the best means to represent the data, 3 = the best means to represent the data, 4=a graph that exceeds the best)		
· If applicable, are the variables properly positioned on the X & Y axis? (0 = the independent and dependent variables are not correctly placed on the X & Y, 1.5 = one variable is correctly placed, 3= both variables are correctly placed) [Is IV on X and DY on Y?]		
Content – lines, bars, point symbols, or other marks that specify particular relations among the things represented by the framework		
· Is the proper data or information being plotted (0 = no, 3 = yes)		
· Are the relative positions of the data plotted on the Y axis properly paired with the values along the X axis? (i.e. proper relationship between X & Y) (0 = none or few, 1 = roughly 50% accurate, 2 = nearly 100%; given a +/-1 on either scale)		
Labels – indicates the variables, the value along the measurement scale, the particular entities that were measured, and the title of the graph		
· Are the independent and dependent variables properly labeled? (0 = lacking labels 1=incorrect labeling, 2= one correct label, 3= two correct labels)		
• Are data points connected with a line or is a trend line graphed? (0= no line, 0.5=connecting line, 1=trend line) [when 'best graph' is a line]		
• Are data manipulated by the participant to show mean/average? (0=no, 1=partial 1.5=yes, all)		
• If multiple content elements are being graphed, is the content (lines, points, bars) represented via color, texture, and so on to allow the reader to readily read and interpret the data being presented? (0 = no, 1 = yes)		
· If appropriate, are the correct respective units for each variable properly labeled? (0 = no, 0.75=1/2 correct, 1.5=all correct)		
· Given the provided data, does the scaling for each axis seem appropriate to construct an effective graph? (0= no scale 0.5= both axes with improper scale, 1 = one axis with proper scale, 1.5 = both axes with proper/best scale)		
· If values along an axis fall into secondary groups are they labeled correctly? (0 = no, 0.5=some, 1 = yes)		
· If necessary based on the means of graphing, is a key or legend used to clarify the meanings of symbols, patterns, color, etc. used in the graph? (0 = no, 1 = yes, but not clear 2=yes, very clear)		
· Was the graph properly labeled with a title or figure legend/caption? (0 = no, 1 = yes)		
Total possible = 28 Participant:	1	
Total		

Graph 1 Verbal		
Participant #:	1	2
Framework		
• Why did they choose this graph type? (0=no mention 1=poor explanation with low understanding of graphing conventions 2=solid description based on normal graph conventions 3=solid description that talks about type of data being displayed)		
• Can they explain why they labeled the axes as they did (IV & DV)? (why they positioned the variables on the axes) (0=no explanation, 0.5=low understanding 1.5=relies on normal conventions of graphing, 2.5=high understanding with through explanation)		
Identification		
· Are they able to recognize the illustrated relation? (0=no, 1=yes, but no/limited explanation, 2=yes, with an explanation that relies on normal conventions of graphing, 3=yes and provides an explanation using higher thought)		
• Are they able to note and explain their scale range? (0=no mention or explanation, 0.5=mentioned but no explanation 1=mentions and explains with low understanding 1.5=mentions and explains but relies on normal conventions of graphing, 2= mentions and explains with high understanding with through explanation)		
• Are they able to note and explain use of color? 0=no mention or explanation, 0.25=mentioned but no explanation 0.5=mentions and explains with low understanding 0.75=mentions and explains but relies on normal conventions of graphing, 1= mentions and explains with high understanding with through explanation)		
• Are they able to recognize mistakes (metacognition) (0=no, 0.5=yes, but does not note how to correct, 1=yes, and provides an explanation of how to correct or better their current graph)		
Off-reading		
• Does the participant focus on numeric value, two value comparison with trend recognition, or multiple values/multiple trends compared? (0=no focus on numbers, 0.5=focus on numeric value, 1=focus on 2 values compared/trend, 1.5=multiple values/trends compared)		
• Does the participant offer a prediction as to how the data would continue over time? (0=no, 1=yes with a focus on the topic only, 1.5=yes with a focus on other elements effected as well)		
• Can he/she explain, using their graph, what the data show? (0=no explanation, 1=low understanding 2=relies on normal conventions of graphing, 2.5=high understanding with through explanation)		
Total possible = 20 Participant:	1	2
Total	0	0

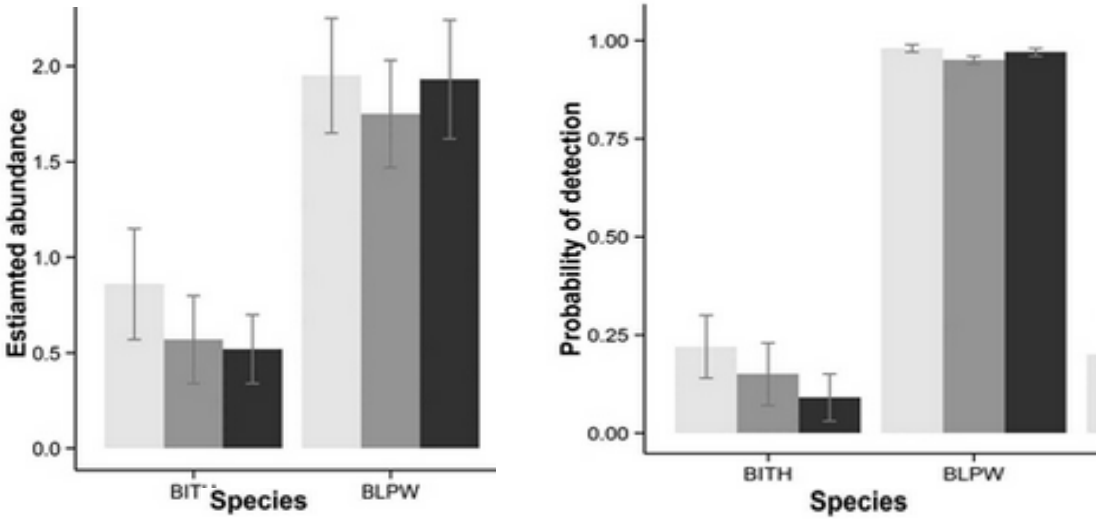
Appendix G: Graph Two Task and Associated Rubrics

Task

Species	Distance (meters)	Abundance (std. error)	Probability of Detection (std. error)
Bicknell's Thrush (BITH)	On the trail	0.9 (±0.35)	.20 (±0.09)
	200 M	0.6 (±0.25)	.15 (±0.09)
	400 M	0.5 (±0.15)	.10 (±0.07)
Blackpoll Warbler (BLPW)	On the trail	1.9 (±0.4)	.95 (±0.025)
	200 M	1.75 (±0.35)	.92 (±0.025)
	400 M	1.85 (±0.4)	.93 (±0.025)

The data provided are from an ecological study on the influence of recreational hiking trails on the abundance of different montane bird populations. To measure the potential effect of hiking trails, the authors examined the probability of detecting (or seeing) a species while on the trails and the estimated abundance (or number) of a bird species in the area. Generate a graph(s) that best represents the data.

Anticipated results from: Deluca, W. V., & King, D. I. (2014). Influence of hiking trails on montane birds. *Journal Of Wildlife Management*,*78*(3), 494-502. doi:10.1002/jwmg.675



Graph 2 Written		
Participant #:	1	2
Framework – indicates what kinds of measurements are being used and what things are being measured		
· Is the layout of the graph axes being used the most effectively to represent the data? (0 = no axes, 1=layout not effective, 2 = layout affective but not best 2.5=layout is extremely effective)		
· Is the proper type of graph being used to most effectively represent the data? (0 = nothing drawn, 1 = graph being used is not appropriate, 2 = a graph that can be used but is not possibly the best means to represent the data, 3 = the best means to represent the data, 4=a graph that exceeds the best)		
· If applicable, are the variables properly positioned on the X & Y axis? (0 = the independent and dependent variables are not correctly placed on the X & Y, 1.5 = one variable is correctly placed, 3= both variables are correctly placed) [Is IV on X and DY on Y?]		
Content – lines, bars, point symbols, or other marks that specify particular relations among the things represented by the framework		
· Is the proper data or information being plotted (0 = no, 3 = yes)		
• Are error bars plotting when std. error numbers are provided (0=no, 0.5=plotted, but incorrectly , 1=yes, and partially correct, 1.5=yes and entirely correct)		
· Are the relative positions of the data plotted on the Y axis properly paired with the values along the X axis? (i.e. proper relationship between X & Y) (0 = none or few, 1 = roughly 50% accurate, 2 = nearly 100%; given a +/-1 on either scale)		
Labels – indicates the variables, the value along the measurement scale, the particular entities that were measured, and the title of the graph		
Are the independent and dependent variables properly labeled? (0 = lacking labels 1=incorrect labeling, 2= one correct label, 3= two correct labels)		
• If multiple content elements are being graphed, is the content (lines, points, bars) represented via color, texture, and so on to allow the reader to readily read and interpret the data being presented? (0 = no, 1 = yes)		
· If appropriate, are the correct respective units for each variable properly labeled? (0 = no, 0.75=1/2 correct, 1.5=all correct)		
· Given the provided data, does the scaling for each axis seem appropriate to construct an effective graph? (0= no scale 0.5= both axes with improper scale, 1 = one axis with proper scale, 1.5 = both axes with proper/best scale)		
· If values along an axis fall into secondary groups are they labeled correctly? (0 = no, 0.5=some, 1 = yes)		
· If necessary based on the means of graphing, is a key or legend used to clarify the meanings of symbols, patterns, color, etc. used in the graph? (0 = no, 1 = yes, but not clear 2=yes, very clear)		
· Was the graph properly labeled with a title or figure legend/caption? (0 = no, 1 = yes)		
Total possible = 27 Participant:	1	2
Total		

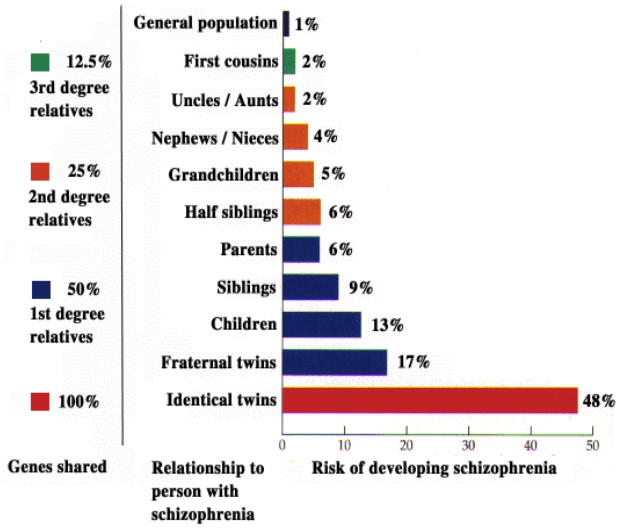
Graph 2 Verbal		
Participant #:	1	2
Framework		
• Why did they choose this graph type? (0=no mention 1=poor explanation with low understanding of graphing conventions 2=solid description based on normal graph conventions 3=solid description that talks about type of data being displayed)		
• If they participant drew multiple graphs, can they explain why? (0=no explanation, 0.5=yes, but for ease/low understanding 0.75=yes, recognizes that two dependent variables exist, 1=yes, recognizes that two dependent and a single independent variable exist and provides through explanation)		
• Can they explain why they labeled the axes as they did (IV & DV)? (why they positioned the variables on the axes) (0=no explanation, 0.5=low understanding 1.5=relies on normal conventions of graphing, 2.5=high understanding with through explanation)		
Identification		
· Are they able to recognize the illustrated relation? (0=no, 1=yes, but no/limited explanation, 2=yes, with an explanation that relies on normal conventions of graphing, 3=yes and provides an explanation using higher thought)		
• Are they able to note and explain their scale range? (0=no mention or explanation, 0.5=mentioned but no explanation 1=mentions and explains with low understanding 1.5=mentions and explains but relies on normal conventions of graphing, 2= mentions and explains with high understanding with through explanation)		
• Are they able to note and explain use of color? 0=no mention or explanation, 0.25=mentioned but no explanation 0.5=mentions and explains with low understanding 0.75=mentions and explains but relies on normal conventions of graphing, 1= mentions and explains with high understanding with through explanation)		
• Are they able to recognize mistakes (metacognition) (0=no, 0.5=yes, but does not note how to correct, 1=yes, and provides an explanation of how to correct or better their current graph)		
• Does the participant explain the use of error bars (0=no mention, 0.5=mentions, but does not explain, 1=yes, and provides an explanation with relation to the data)		
Off-reading		
• Does the participant focus on numeric value, two value comparison with trend recognition, or multiple values/multiple trends compared? (0=no focus on numbers, 0.5=focus on numeric value, 1=focus on 2 values compared/trend, 1.5=multiple values/trends compared)		
• Does the participant offer a prediction as to how the data would continue over time? (0=no, 1=yes with a focus on the topic only, 1.5=yes with a focus on other elements effected as well)		
• Can he/she explain, using their graph, what the data show? (0=no explanation, 1=low understanding 2=relies on normal conventions of graphing, 2.5=high understanding with through explanation)		
Total possible = 22 Participant:	1	2
Total		

Appendix H: Graph Three Task and Associated Rubrics

Genes shared	Relationship to person with Schizophrenia	Risk of developing Schizophrenia (%)
100%	Identical twins	44
50%	Fraternal Twins	16
	Children	11
	Siblings	7
	Parents	5
25%	Half siblings	5
	Grandchildren	4
	Nephews/ nieces	3
	Uncles/ Aunts	2
12.5%	First Cousins	2
None	General population	1

The degree of family relatedness, which influences the percent of genes shared between two family members, can be used in medicine to predict one's likelihood of having a disease. Draw a graph(s) that you believe best represents the data.

Anticipated results from: Debby Tsuang, M.D., M.Sc., University of Washington/VAPSHCS, Special thanks to Dr. Kristin Cadenhead, UCSD



Graph 3 Written		
Participant #:	1	2
Framework – indicates what kinds of measurements are being used and what things are being measured		
· Is the layout of the graph axes being used the most effectively to represent the data? (0 = no axes, 1=layout not effective, 2 = layout affective but not best 2.5=layout is extremely effective)		
· Is the proper type of graph being used to most effectively represent the data? (0 = nothing drawn, 1 = graph being used is not appropriate, 2 = a graph that can be used but is not possibly the best means to represent the data, 3 = the best means to represent the data, 4=a graph that exceeds the best)		
· If applicable, are the variables properly positioned on the X & Y axis? (0 = the independent and dependent variables are not correctly placed on the X & Y, 1.5 = one variable is correctly placed, 3= both variables are correctly placed) [Is IV on X and DY on Y?]		
Content – lines, bars, point symbols, or other marks that specify particular relations among the things represented by the framework		
· Is the proper data or information being plotted (0 = no, 3 = yes)		
· Are the relative positions of the data plotted on the Y axis properly paired with the values along the X axis? (i.e. proper relationship between X & Y) (0 = none or few, 1 = roughly 50% accurate, 2 = nearly 100%; given a +/-1 on either scale)		
Labels – indicates the variables, the value along the measurement scale, the particular entities that were measured, and the title of the graph		
· Are the independent and dependent variables properly labeled? (0 = lacking labels 1=incorrect labeling, 2= one correct label, 3= two correct labels)		
• If multiple content elements are being graphed, is the content (lines, points, bars) represented via color, texture, and so on to allow the reader to readily read and interpret the data being presented? (0 = no, 1 = yes)		
· If appropriate, are the correct respective units for each variable properly labeled? (0 = no, 0.75=1/2 correct, 1.5=all correct)		
· Given the provided data, does the scaling for each axis seem appropriate to construct an effective graph? (0= no scale 0.5= both axes with improper scale, 1 = one axis with proper scale, 1.5 = both axes with proper/best scale)		
· If values along an axis fall into secondary groups are they labeled correctly? (0 = no, 0.5=some, 1 = yes)		
· If necessary based on the means of graphing, is a key or legend used to clarify the meanings of symbols, patterns, color, etc. used in the graph? (0 = no, 1 = yes, but not clear 2=yes, very clear)		
· Was the graph properly labeled with a title or figure legend/caption? (0 = no, 1 = yes)		
Total possible = 25.5 Participant:	1	2
Total	0	0

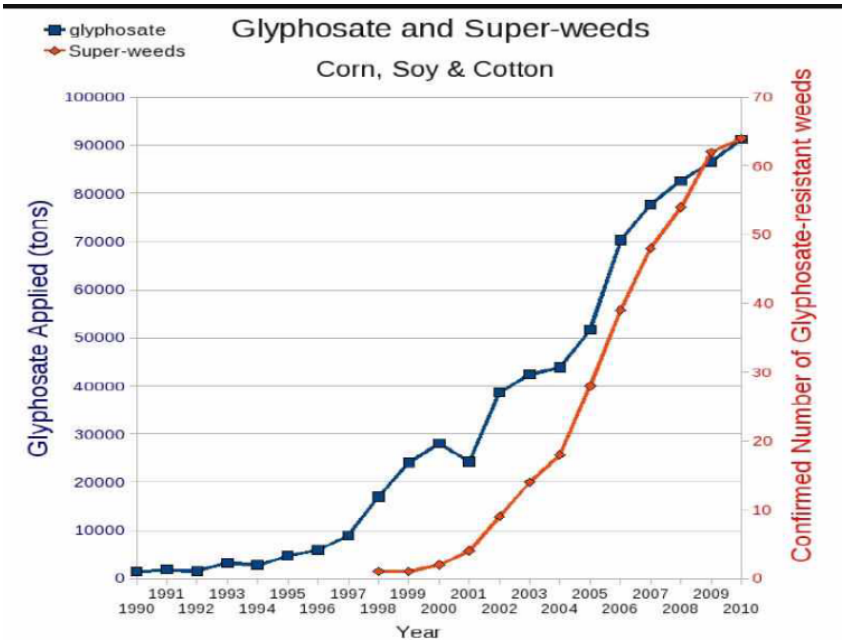
Graph 3 Verbal		
Participant #:	1	2
Framework		
• Why did they choose this graph type? (0=no mention 1=poor explanation with low understanding of graphing conventions 2=solid description that talks about type of data being displayed)		
• Can they explain why they labeled the axes as they did (IV & DV)? (why they position the variables on the axes) (0=no explanation, 1=low understanding 2=relies on normal conventions of graphing, 3=high understanding with through explanation)		
Identification		
· Are they able to recognize the illustrated relation? (0=no, 1=yes, but no/limited explanation, 2=yes, with an explanation that relies on normal conventions of graphing, 3=yes and provides an explanation using higher thought)		
• Are they able to note and explain their scale range? (0=no mention or explanation, 0.5=mentioned but no explanation 1=mentions and explains with low understanding 1.5=mentions and explains but relies on normal conventions of graphing, 2= mentions and explains with high understanding with through explanation)		
• Are they able to note and explain use of color? 0=no mention or explanation, 0.25=mentioned but no explanation 0.5=mentions and explains with low understanding 0.75=mentions and explains but relies on normal conventions of graphing, 1= mentions and explains with high understanding with through explanation)		
• Are they able to recognize mistakes (metacognition) (0=no, 0.5=yes, but does not note how to correct, 1=yes, and provides an explanation of how to correct or better their current graph)		
Off-reading		
• Does the participant focus on numeric value, two value comparison with trend recognition, or multiple values/multiple trends compared? (0=no focus on numbers, 0.5=focus on numeric value, 1=focus on 2 values compared/trend, 1.5=multiple values/trends compared)		
• Does the participant offer a prediction as to how the data would continue over time? (0=no, 1=yes with a focus on the topic only, 1.5=yes with a focus on other elements effected as well)		
• Can he/she explain, using their graph, what the data show? (0=no explanation, 1=low understanding 2=relies on normal conventions of graphing, 2.5=high understanding with through explanation)		
Total possible = 20 Participant:	1	2
Total		

Appendix I: Graph Four Task and Associated Rubrics

Year	Amount of Glyphosate Applied (tons)	Confirmed # of Glyphosate-resistant weeds
1990	1,000	0
1992	1,000	0
1994	4,000	0
1996	8,000	0
1998	18,000	1
2000	29,000	2
2002	40,000	10
2004	42,000	17
2006	70,000	49
2008	82,000	54
2010	90,000	65

Glyphosate is a broad-spectrum systemic herbicide that is commonly used to kill a wide range of weeds in agricultural systems. Over the last decade, there has been growing concern to the use of systemic weed killers (i.e. chemicals that are applied to leaves or foliage to kill the weed). Studies have suggested a side effect to using systemic weed killers is glyphosate resistant weeds – or super weeds. Generate a graph(s) that best represents the data.

Anticipated result from: Glyphosate data from USDA:NASS; Super weed data cited from Charles Benbrook (an American agricultural economist and former research professor at the Center for Sustaining Agriculture and Natural Resources at Washington State University)



Graph 4 Written		
Participant #:	1	2
Framework – indicates what kinds of measurements are being used and what things are being measured		
· Is the layout of the graph axes being used the most effectively to represent the data? (0 = no axes, 1=layout not effective, 2 = layout affective but not best 2.5=layout is extremely effective)		
· Is the proper type of graph being used to most effectively represent the data? (0 = nothing drawn, 1 = graph being used is not appropriate, 2 = a graph that can be used but is not possibly the best means to represent the data, 3 = the best means to represent the data, 4=a graph that exceeds the best)		
· If applicable, are the variables properly positioned on the X & Y axis? (0 = the independent and dependent variables are not correctly placed on the X & Y, 1.5 = one variable is correctly placed, 3= both variables are correctly placed) [Is IV on X and DY on Y?]		
Content – lines, bars, point symbols, or other marks that specify particular relations among the things represented by the framework		
· Is the proper data or information being plotted (0 = no, 3 = yes)		
· Are the relative positions of the data plotted on the Y axis properly paired with the values along the X axis? (i.e. proper relationship between X & Y) (0 = none or few, 1 = roughly 50% accurate, 2 = nearly 100%; given a +/-1 on either scale)		
Labels – indicates the variables, the value along the measurement scale, the particular entities that were measured, and the title of the graph		
· Are the independent and dependent variables properly labeled? (0 = lacking labels 1=incorrect labeling, 2= one correct label, 3= two correct labels)		
• If multiple content elements are being graphed, is the content (lines, points, bars) represented via color, texture, and so on to allow the reader to readily read and interpret the data being presented? (0 = no, 1 = yes)		
· If appropriate, are the correct respective units for each variable properly labeled? (0 = no, 0.75=1/2 correct, 1.5=all correct)		
· Given the provided data, does the scaling for each axis seem appropriate to construct an effective graph? (0= no scale 0.5= both axes with improper scale, 1 = one axis with proper scale, 1.5 = both axes with proper/best scale)		
· If values along an axis fall into secondary groups are they labeled correctly? (0 = no, 0.5=some, 1 = yes)		
· If necessary based on the means of graphing, is a key or legend used to clarify the meanings of symbols, patterns, color, etc. used in the graph? (0 = no, 1 = yes, but not clear 2=yes, very clear)		
· Was the graph properly labeled with a title or figure legend/caption? (0 = no, 1 = yes)		
Total possible = 25.5 Participant:	1	2
Total		

Graph 4 Verbal		
Participant #:	1	2
Framework		
• Why did they choose this graph type? (0=no mention 1=poor explanation with low understanding of graphing conventions 2=solid description based on normal graph conventions 3=solid description that talks about type of data being displayed)		
• Can they explain why they labeled the axes as they did (IV & DV)? (why they positioned the variables on the axes) (0=no explanation, 0.5=low understanding 1.5=relies on normal conventions of graphing, 2.5=high understanding with through explanation)		
Identification		
· Are they able to recognize the illustrated relation? (0=no, 1=yes, but no/limited explanation, 2=yes, with an explanation that relies on normal conventions of graphing, 3=yes and provides an explanation using higher thought)		
• Are they able to note and explain their scale range? (0=no mention or explanation, 0.5=mentioned but no explanation 1=mentions and explains with low understanding 1.5=mentions and explains but relies on normal conventions of graphing, 2= mentions and explains with high understanding with through explanation)		
• Are they able to note and explain use of color? 0=no mention or explanation, 0.25=mentioned but no explanation 0.5=mentions and explains with low understanding 0.75=mentions and explains but relies on normal conventions of graphing, 1= mentions and explains with high understanding with through explanation)		
• Are they able to recognize mistakes (metacognition) (0=no, 0.5=yes, but does not note how to correct, 1=yes, and provides an explanation of how to correct or better their current graph)		
Off-reading		
• Does the participant focus on numeric value, two value comparison with trend recognition, or multiple values/multiple trends compared? (0=no focus on numbers, 0.5=focus on numeric value, 1=focus on 2 values compared/trend, 1.5=multiple values/trends compared)		
• Does the participant offer a prediction as to how the data would continue over time? (0=no, 1=yes with a focus on the topic only, 1.5=yes with a focus on other elements effected as well)		
• Can he/she explain, using their graph, what the data show? (0=no explanation, 1=low understanding 2=relies on normal conventions of graphing, 2.5=high understanding with through explanation)		
Total possible = 20 Participant:	1	2
Total		

Appendix J: Kruskal-Wallis Non-Parametric Test Results

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of VG1 is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.003	Reject the null hypothesis.
2	The distribution of VG2 is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.003	Reject the null hypothesis.
3	The distribution of VG3 is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.012	Reject the null hypothesis.
4	The distribution of VG4 is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.016	Reject the null hypothesis.
5	The distribution of VGtotal is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.003	Reject the null hypothesis.
6	The distribution of CG1 is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.057	Retain the null hypothesis.
7	The distribution of CG2 is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.051	Retain the null hypothesis.
8	The distribution of CG3 is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.272	Retain the null hypothesis.
9	The distribution of CG4 is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.012	Reject the null hypothesis.
10	The distribution of CGtotal is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.010	Reject the null hypothesis.
11	The distribution of TOTAL is the same across categories of Category.	Independent-Samples Kruskal-Wallis Test	.003	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Appendix K: Means and Standard deviation of participant scores by group and graph type

	Biology faculty		Graduate/senior Biology students		Biology undergraduates		Non-Biology undergraduates	
Think-Aloud								
	M	SE	M	SE	M	SE	M	SE
Graph 1	15.58	0.36	19.13	0.60	8.47	1.27	7.7	0.78
Graph 2	16.25	0.33	12.5	0.91	8.86	1.21	7.62	1.16
Graph 3	11.42	0.96	10.31	0.55	8.31	1.03	7.18	0.49
Graph 4	13.83	0.73	10.63	0.33	8.31	0.98	8.25	0.56
Total	57.08	1.91	45.06	3.32	32.84	3.39	30.75	1.37
Graph Drawing								
	M	SE	M	SE	M	SE	M	SE
Graph 1	19.33	0.29	22.08	0.37	18.79	0.62	19	0.35
Graph 2	19.12	0.58	20.58	0.56	18.16	0.83	17.79	0.44
Graph 3	18.34	2.24	17.96	1.60	17.39	2.18	17.65	1.54
Graph 4	17.73	0.69	17.76	2.00	13.64	2.63	15.58	1.12
Total	79.21	0.88	75.67	0.86	72.36	7.17	63.46	4.82
Graph Drawing & Think-Aloud combined								
	M	SE	M	SE	M	SE	M	SE
Total	136.58	1.91	121.31	2.42	108.14	6.30	95.58	3.06

Reference

AACU (2014) Quantitative literacy value rubric.Retrieved from https://www.aacu.org/value/rubrics/quantitative-literacy

Bray-Speth, E. B., Momsen, J. L., Moyerbrailean, G. A., Ebert-May, D., Long, T. M., Wyse, S., & Linton, D. (2007). 1, 2, 3, 4: Infusing quantitative literacy into introductory biology. *CBE Life Sciences Education*, *9*(3), 323–332. doi:10.1187/cbe.10-03-0033

Duesbery, L., Werblow, J., Yovanoff, P., (2011). Grphical Literacy Moderates the interaction of decorative dimensionally and cognitive demand in computer-based graph comprehension. *Journal of Educational Computing Research*. 45(1). 75-93.

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

Harsh, J. (2014). Assessing the effects of undergraduate research on the development of scientific thinking skills as measured by student performance. Indiana: Indiana University. Dissertation.

Harsh, J. A. & Maltese, A. V. (2012). Data interpretation along the novice-expert continuum. Paper presented at the National Association for Research in Science Teaching Annual Meeting, Indianapolis, IN.

Harsh, J., Maltese, A., Warner, J., MacLeish, E. (in preparation). Data Representation Along a Continuum of Expertise in the Sciences.

Hmelo-Silver, C. E. (2004) Problem-based learning: what and how do students learn? Educational Psychology Review, 16, 235–266. doi: 10.1023/B:EDPR.0000034022.16470.f3

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. Oxford: Oxford University Press.

Kotzebue, L., Gerstl, M., & Nerdel, C. (2015). Common mistakes in the construction of diagrams in biological contexts. *Research in Science Education*. 45(2), 193-213. Doi:10.1007/s11165-014-9419-9

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

Maltese, Adam V., Joseph A. Harsh, and Dubravka Svetina. "Data visualization literacy: investigating data interpretation along the novice–expert continuum." Journal of College Science Teaching 45.1 (2015): 84.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. Educational Measurement: Issues and Practice, 11(1), 3-9, 20.

Picone, C., Rhode, J., Hyatt, L., & Parshall, (2007) Assessing gains in undergraduate students' abilities to analyze graphical data. *Teaching Issues and Experiments in Ecology*, Vol. 5: Research #1 [online].

Roth, W. M. and M. K. McGinn. (1997). Graphing: Cognitive ability or practice? Science Education 81: 91-106

Stein, B., Haynes, A., & Redding, M. (2007). Project CAT: Assessing critical thinking skills. In Proceedings of the 2006 National STEM Assessment Conference, Deeds, D, and Callen, B.(eds) Springfiled, MO: Drury University

Tairab, H. H. & Khalaf Al-Naqbi, A. K. (2004). How do secondary school science students interpret and construct scientific graphs. *Journal of Biological Education, 38*(3), 127-132

Wass V, Van der Vleuten C, Shatzer J, Jones R. (2001). Assessment of clinical competence. Lancet, 357, 945-949.

Woodin, T., Carter, V. C., & Fletcher, L. (2010). Vision and Change in Biology Undergraduate Education, A Call for Action—Initial Responses. CBE Life Sciences Education, 9(2), 71–73. http://doi.org/10.1187/cbe.10-03-0044

Wright, A. M., & Holliday, R. E. (2007). Enhancing the recall of young, young–old and old–old adults with cognitive interviews. *Applied Cognitive Psychology*, *21*(1), 19-43. doi:10.1002/acp.1260

Zucker, A., Staudt, C., & Tinker, R., (2015). Teaching graph literacy across the curriculum. *Science Scope*, 38(6), 19-24