# Lessons Learned from Quantitatively Exploring Visualization Rubric Utilization for Peer Feedback

**D. J. Barajas, X. S. Apedoe, D. G. Brizan, A. P. Joshi and S. J. Engle**
University of San Francisco

*Abstract*—We present our experience of adapting a rubric for peer feedback in our data visualization course and exploring the utilization of that rubric by students across two semesters. We first discuss the results of an automatable quantitative analysis of the rubric responses, and then compare those results to a qualitative analysis of summative survey responses from students regarding the rubric and peer feedback process. We conclude with lessons learned about the visualization rubric we used, as well as what we learned more broadly about using quantitative analysis to explore this type of data. These lessons may be useful for other educators wanting to utilize the same data visualization rubric, or wanting to explore the utilization of rubrics already deployed for peer feedback.

**■ BOTH RUBRICS AND PEER FEEDBACK** are widely-established pedagogical techniques [1], both within data visualization [2], [3] and beyond. Rubrics help provide clear evaluation and quality guidelines, and their use has been well-established in higher education for decades [1], [4]. For example, rubrics have been used to clearly communicate grading criteria to students as well as to create grading consistency [5], and for peer review to both improve student work and their understanding of the core concepts [6], [7].

Peer review refers to the process where students provide their peers feedback on their work, and is another long-established pedagogical technique within higher education [8], [9], [10]. This process plays a particularly important role in teaching data visualization [11]. For example, Moxley presents a community-derived rubric and peer review tool for student writing courses [7] and Friedman explores using this rubric and tool

for peer review of visualizations instead [12].

Despite rubrics and peer review being well-established pedagogical tools, peer review in data visualization course syllabi is not yet widely adopted [3]. Prior research also highlights the "importance of adapting a rubric" to different settings [1]. Educators may want to explore how to adapt a rubric for their demographic of students, and explore how well students are utilizing that rubric for peer feedback. However, a detailed qualitative analysis is often prohibitively time-consuming and expensive to employ in this setting.

We help address this gap by presenting our experience adapting a data visualization rubric for peer feedback in our course, and demonstrate what we can learn from an automatable quantitative analysis of how this rubric was utilized by students. We start with the rubric proposed by Friedman and Rosen in 2017 for data visualization feedback [2] and their experience designing and teaching data

visualization coursework. We discuss how we adapted this rubric for our 2019 and 2020 courses in the "Course Details" section.

We then perform a quantitative analysis of 890 peer feedback submissions using that rubric from 54 students across 2 semesters. We compare our quantitative findings to those from a qualitative analysis of 448 text snippets from 47 anonymous end-of-semester survey responses from students about the peer feedback process. Many, but not all, of our findings agree with prior work by Beasley et al. [3], which also integrated the original rubric into their own courses and conducted a quantitative analysis using natural language processing and an end-of-semester survey in 2020. We contrast our findings in the "Discussion" section.

We finally report our lessons learned from this process in two categories:

- Lessons learned specific to the rubric adapted for our data visualization course.
- Lessons learned about quantitatively exploring rubric utilization by students for peer feedback.

In short, we first confirmed that the adapted rubric is useful for peer feedback of data visualizations, but that some adjustments could be made to improve the process. We also demonstrated that automated quantitative analysis can indeed help educators understand how students are utilizing a rubric for the peer feedback process. Most of all, we confirmed how important it is to utilize rubrics and peer feedback (especially verbal) when teaching data visualization.

## COURSE DETAILS

We collected data from our 16 week data visualization course in Spring 2019 and Spring 2020. This course has been taught yearly since 2013 and has co-listed undergraduate and graduate sections taught simultaneously. Both semesters were taught by the same instructor with similar content, except that half of Spring 2020 was remote due to COVID-19 restrictions. The course was completed by 54 students (33% female, 48% graduate).

### Course Assignments

Final grades were determined by a mix of homework, project, and other assignments. See Table 1 for details. The homework programming assignments gave students opportunities to practice im-

**Table 1. Assignments and Modalities**

| Assignment | % | Week | Feedback Modality |
|---|---|---|---|
| Homework 1 (H1) | 5% | 4 | Written (*No Rubric*) |
| Homework 2 (H2) | 5% | 6 | Written |
| Midterm Project (MP) | 25% | 9 | Verbal, Numeric |
| Homework 3 (H3) | 5% | 12 | Numeric, Written |
| Homework 4 (H4) | 5% | 14 | Numeric, Written |
| Final Project (FP) | 30% | Finals | Verbal, Numeric, Written |
| Other Assignments | 25% | – | N/A |

Assignments, the percent of the final grade it is worth, the week it was due, and the modality used for peer feedback. Homework feedback was given asynchronously and due one week after the deadline; project feedback was given synchronously for prototypes two weeks before the project was due.

plementing different visualization and interaction techniques, and thus their grades were determined by the functionality implemented (not whether the implemented visualizations were effective).

The projects were graded both on functionality *and* whether the visualizations were effective. The midterm project was completed in groups of 2 to 3 students. Groups choose from 1 or 2 approved datasets, then decided on their own narrative. Each student worked on an interactive multi-component visualization that supported that narrative, and presented their visualizations together as part of a cohesive group website. The final project was completed individually. Students could propose any dataset and had to provide at least 3 different perspectives of that dataset. At least one visualization had to be highly-interactive and use an advanced visualization technique. For both projects, students had to create prototypes, provide each other feedback on those prototypes, and prepare a website showcasing their visualizations.

### Peer Feedback Process

Students were required to provide peer feedback for homework and projects. Students earned pass/fail participation credit for providing peer feedback, but that feedback did not directly impact the assignment grades. At the end of the semester, students voted on who gave the most thoughtful and constructive peer feedback.

The instructor provided lectures on evaluation principles and conducted an in-class evaluation exercise. The evaluation principles were then reinforced in subsequent lectures when discussing example visualizations. To practice evaluation, students were randomly assigned 2 to 3 homework submissions to provide asynchronous peer

2

feedback. While this process gave students opportunities to practice evaluation, it did not provide students an opportunity to improve their visualizations based on that feedback.

Unlike homework, the peer feedback process for projects gave students an opportunity to improve their visualizations based on that feedback. Students first prepared prototypes of their project visualizations, and then presented and collected synchronous feedback on those prototypes during an in-class exercise.[1] Students had 2 weeks to improve their prototypes based on that feedback.

The nature of the synchronous in-class exercises and small class size made facilitating anonymous feedback difficult in those scenarios. Knowing that to be the case, the instructor made peer feedback non-anonymous for both synchronous and asynchronous feedback in hopes students would build a cooperative community and grow comfortable providing each other non-anonymous feedback.

### AFaR Rubric Questions

Previous years of this course utilized a similar peer feedback process, but without a specific rubric. For 2019 and 2020, we modified the visualization rubrics by Friedman and Rosen [2] to fit the preexisting course content and peer feedback process. We refer to our modification as the **Adapted Friedman and Rosen** (AFaR) rubric. We removed questions inappropriate for our course, combined some questions, reworded prompts, and added a new "Insight" question. The resulting rubric, shown in Table 2, went from 28 to 14 questions and 3 categories: "**Description**" questions that asked students to *describe* the visualizations, "**Feedback**" questions that asked to *evaluate* the visualizations, and an "**Insight**" question that asked to *interpret* the visualizations.

The "**Description**" questions asked students to describe the visual encodings, interactivity, and the goal or theme of the visualizations. The "**Feedback**" category asked students to evaluate the effectiveness of the color and non-color encodings used, context (e.g. axis, tick labels, legend) provided, lie factor, data-ink ratio, data density, use of Gestalt principles, interaction,

design and aesthetics, and if appropriate, how well the visualization supported the narrative. Both of these categories were derived from the original rubric [2]. The new "**Insight**" category included one question. We felt it important that students attempt to use the visualization under evaluation. It also allowed them to receive findings of their visualizations from others, which might not match those intended. We also replaced the 5 checkboxes of the original rubric with numeric ratings, which allowed us to implement the rubric in our learning management system (LMS).

We only included questions relevant to the assignment. Homework 1 did not use the rubric, as it was assigned before the evaluation content. Homework 2 did not include the interactivity questions. Only projects included the goal and overall effectiveness questions.

We also varied the modality used to provide feedback to gradually introduce the rubric into the course, as well as to study the impact of the different modalities. Homework feedback was given asynchronously and project feedback was given synchronously during class. Homework 1, with no rubric, allowed students to gain familiarity with the feedback process. Homework 2 introduced the AFaR rubric and asked for written feedback. The midterm project introduced verbal feedback and numeric ratings. Afterwards, homework 3 and 4 included both written feedback and numeric ratings. The final project included all modalities: verbal, written, and numeric. Table 1 shows a summary of the feedback modalities.

## METHODOLOGY

For our analysis, we collected and analyzed data relevant to the peer feedback provided using the AFaR rubric for the homework and projects. We also collected and analyzed data from an anonymous end-of-semester survey from students about the feedback process and rubric.

### AFaR Data Analysis

The homework and final project feedback used the "Peer Review" and "Rubric" features in the Instructure Canvas Learning Management System (LMS). Canvas is a cloud-based LMS with several features like gradebooks, assignments, discussions, and quizzes. The "Peer Review" feature can randomly assign students each other's submissions

---

[1]Due to COVID-19 restrictions, the final project verbal feedback exercises were adapted to support remote learning.

3

**Table 2. AFaR Visualization Rubric Questions**

| Type | Label | Example Prompt | Asynchronous H2 | Asynchronous H3/4 | Synchronous 💬 MP | Synchronous 💬 FP |
|---|---|---|---|---|---|---|
| ☰ | Visual Encodings | How are visual channels used to encode data? Visual channels include: position, length or size, shape, color, and other visual attributes. | ✏ × | ✏ × | × × | ✏ × |
| 👍 | Non-Color Encodings | Rate and comment on the non-color encodings on a scale of 5 (highly appropriate) to 1 (inappropriate). Consider both the data type and strength of the visual channel (or preattentive attribute) used to encode that data. | ✏ × | ✏ ★ | × ★ | ✏ ★ |
| 👍 | Color Encodings | Rate and comment on the color encodings on a scale of 5 (excellent use of color) to 1 (inappropriate use of color). Consider whether the type color scheme (diverging, sequential, categorical) is appropriate. For categorical color schemes, also consider the number of distinct colors used and whether the colors are separable. | ✏ × | ✏ ★ | × ★ | ✏ ★ |
| 👍 | Context | Rate and comment on the provided context for interpreting the visualization on a scale of 5 (excellent context provided) to 1 (insufficient context provided). Consider how visual elements such as axis and tick labels, legends, annotations, descriptive text, grid lines, and other non-data visual elements help with providing context and interpretation of the visualization. | ✏ × | ✏ ★ | × ★ | ✏ ★ |
| 👍 | Lie Factor | Rate and comment on the lie factor on a scale of 5 (no lie factor) to 1 (high lie factor). Consider whether the visualization has misleading context (e.g. misleading but properly labeled scales), weak visual encodings (e.g. choosing volume instead of area), exaggerated encodings (e.g. scaling circles by radius instead of area), or unnecessary use of 3D and depth. | ✏ × | ✏ ★ | × ★ | ✏ ★ |
| 👍 | Data Ink Ratio | Rate and comment on the data ink ratio on a scale of 5 (high data ink ratio) to 1 (low data ink ratio). Consider both whether more data ink should be added or non-data ink should be removed. | ✏ × | ✏ ★ | × ★ | ✏ ★ |
| 👍 | Data Density | Rate and comment on the data density on a scale of 5 (high data density) to 1 (low data density). Consider both the amount of data included as well as the overall size of the visualization(s). | ✏ × | ✏ ★ | × ★ | ✏ ★ |
| 👍 | Gestalt Principles | Rate and comment on the Gestalt principles on a scale of 5 (used well) to 1 (used poorly). Consider the principles of background versus foreground, proximity, and similarity. | ✏ × | ✏ ★ | × ★ | ✏ ★ |
| 👍 | Design and Aesthetics | Rate and comment on the design and aesthetics on a scale of 5 (aesthetically pleasing or beautiful) to 1 (not aesthetically pleasing). | ✏ × | ✏ ★ | × ★ | ✏ ★ |
| ☰ | Interactivity | What interaction mechanisms are being used? Consider details-on-demand, highlighting, brushing, filtering, linked views, focus plus context, zooming, panning or translating, rotating, and others. | × × | ✏ × | × × | ✏ × |
| 👍 | Interaction Effectiveness | Rate and comment on the interactivity on a scale of 5 (highly effective) to 1 (ineffective). Consider whether the interactivity improves exploration, search, or engagement. | × × | ✏ ★ | × ★ | ✏ ★ |
| ☰ | Visualization Goal | What is the overall goal of this visualization? Consider both what audiences should learn from the visualization, and how well it fits into the overall narrative. | × × | × × | × × | ✏ × |
| 👍 | Visualization Effectiveness | Rate and comment on the visualization effectiveness on a scale of 5 (highly effective) to 1 (ineffective). Consider given the visualization goal and your understanding of the data from the visualization. | × × | × × | × ★ | ✏ ★ |
| 💡 | Understanding | List 1 to 3 things about the data that you learned or understood by this visualization, or new questions you have about the data as a result of the visualization. | ✏ × | ✏ × | × × | ✏ × |

☰ Description  👍 Feedback  💡 Insight  ✏ Typed or written  ★ Numeric ratings

Summary of the rubric, including question types (☰ 👍 💡) and feedback modes (✏ ★). Homework feedback (H2, H3, H4) was asynchronous after the deadline. Midterm project (MP) and final project (FP) feedback was synchronous in-class before the deadline.

to provide feedback. If the "Rubric" feature is also used, students can click a button to open a small pop-out widget with specific questions to answer as a part of this process. We used these features to prompt students to provide each other feedback using subsets of the AFaR rubric questions.

We used Jupyter Notebooks [13] and the Instructure Canvas LMS REST API[2] to collect the student, assignment, grade, comment, and rubric data. The API returned data in JSON format, which we converted to CSV. The API is complex with few examples of its use, and currently the only way to export this data from the Canvas system. As such, we provide examples at github.com/djbarajas/canvas-rubrics-api/ to illustrate how to use the API to obtain this data.

The midterm feedback was collected on paper in 2019 and using Google Forms in 2020. The data was converted to CSV and manually cleaned to prepare them for analysis. The CSV files were combined into a SQLite database.

Analysis was performed using the pandas[3] Python library. We selected quantitative metrics that could be quickly calculated from our collected rubric data, starting with median values of our numeric data and word counts of the text data. We also explored metrics utilizing various natural language processing tools, including topic modeling and sentiment analysis. We used the Natural Language Toolkit (NLTK) [14] to tokenize comments to compute word counts and the Gensim [15] library for topic modeling. We used the VADER [16] library to calculate the sentiment of each response to a rubric question.

We then selected the data using SQL queries to explore these metrics grouped by grades, semester, undergraduate versus graduate status, assignment, rating, feedback modality, and whether students were recognized with "best reviewer" awards. We visualized these metrics and subsets as swarm plots and bar plots using Vega [17]. We used these visualizations to explore how students utilized the rubrics for peer feedback. We looked at word count to determine whether certain questions were underutilized, median rating and sentiment to determine whether feedback was positive or negative, and topic modeling to determine whether

feedback used relevant terms to each question.

## Survey Data Analysis

We asked students to comment on the feedback process and AFaR rubric at the end of the semester. The instructor stepped out of the room and a third party led the discussion. Then, students were asked to complete an anonymous survey. The survey included 6 freeform text questions that asked students: (1) the feedback mode they preferred (written, ratings, numeric, verbal) and why, (2) the quality of the feedback received, (3) whether the feedback received was helpful for future assignments, (4) whether giving feedback was helpful for understanding the assignment or material, (5) if they ever found themselves detecting, diagnosing or solving problems during the review process, and (6) any improvements or changes they recommended for the course.

Those responses were downloaded as a CSV file. Of the 54 students that completed the course, 47 (87%) responded. We used Taguette[4] to manually tag the responses into 29 hierarchical tags. For example, the tag "Helpful/Perspectives/Others Viewing" captured responses that mentioned it was helpful to have others view and interpret your visualizations. The tag "Helpful/Perspectives/Viewing Others" captured responses that mentioned it was helpful to view other visualizations to get new ideas. Both fall under the "Helpful/Perspectives" and "Helpful" tags as well. Each response was tagged by two individuals from the research team; tags were refined until there was at least $85\%$ agreement on the tags applied.

## FINDINGS

We visually explored the data looking for trends to learn more about how students utilized the rubrics for peer feedback and present our findings here.

## Written Feedback Breakdown

We explored whether students used relevant terminology for the written feedback using Latent Dirichlet Allocation (LDA) topic modeling [18]. After filtering for stop words, we manually explored the most salient terms per AFaR rubric question. We found the terms used were relevant to the question. See Figure 1 for an example for the

---

[2]See https://canvas.instructure.com/doc/api/ for details.
[3]See pandas.pydata.org and DOI 10.5281/zenodo.3509134.

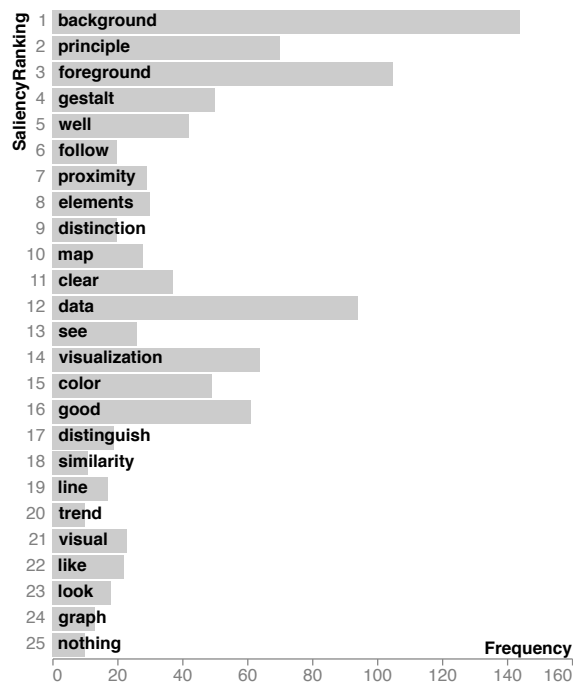[4]See taguette.org and DOI 10.5281/zenodo.4560784.

**Figure 1.** The term frequency for the top 25 most salient terms for the "Gestalt Principles" question. Terms are ordered by saliency as determined by LDA *not* frequency; "background" was the most salient but other terms such as "proximity" and "similarity" are mentioned too. This analysis showed students were using appropriate terminology in their feedback.

"Gestalt Principles" question. The topic models for the other questions also had relevant terms listed.

We also explored the ratings, word counts, and sentiment of each question, as illustrated by Table 3. Students primarily used the highest 5 rating, making it the median for every question and assignment. The average rating was 4.5 with a 0.9 standard deviation. The "Context" and "Lie Factor" questions had the lowest average rating of 4.4. The "Visualization Effectiveness" question had the highest 4.6 average rating.

Students' responses had similar word count per question; the word count ranged from 0 to 133, had a 20 word average, 17 median, and standard deviation of 13. Three "Feedback" questions had the lowest median and the new "Insight" question had the highest. From the topic modeling and median word counts, students seemed to respond earnestly to each question.

We expected a neutral sentiment for the "Description" or "Insight" categories, and a wide range

for the "Feedback" questions. However, the sentiment had a 0.3 average and 0.4 median, indicating a positive sentiment for feedback overall. The "Design and Aesthetics" question consistently had a higher median, which may indicate students were reluctant to add subjective versus objective criticism. The "Lie Factor" question had the lowest median at 0.0 (neutral), but this may be due to the terminology used for that question. For example, a student commented "no lie factor" with the highest 5 rating. This comment has a negative $-0.3$ sentiment despite being positive feedback.

## Assignment Breakdown

Table 3 provides a breakdown by assignment. We take a closer look at these metrics in Figure 2. Consider the difference in word counts for homework 2 and 3, near the top middle column in Figure 2. Homework 2 had only written comments, no ratings or verbal feedback. After that, the word count drops and never recovers. That also corresponds with when ratings were added to homework 3, homework 4, and the final project. It is possible that students felt less need to provide as much written feedback when the ratings were present. Interestingly, adding verbal feedback did *not* impact the word count in the same way.

The median sentiment remained consistent across assignments, with a slight dip for homework 4. In the survey, students reported receiving more critical verbal feedback for projects, but that did not impact the sentiment of the written feedback. However, references to the different datasets used by each assignment did influence sentiment. Consider the "Lie Factor" feedback, "I am confused by the inclusion of the California average for the violent crime rate in a graph that shows the rate for specific crime types in a specific county." This had a positive 4 rating, but one of the lowest sentiments in our dataset at $-0.9$. Removing the dataset references "violent" and "crime" results in a less negative $-0.3$ sentiment.

## Numeric Rating Breakdown

We explored how the ratings were used in Figure 3. This figure illustrates how many times students rated a question highly: the "Rating 5" category includes $67\%$ of the responses. Another $21\%$ fall into the "Rating 4" category. The other 3 categories combined have only $12\%$ of responses.

6

**Table 3. Calculated Metrics per AFaR Rubric Question**

| Question Type and Label | Average (Mean) Rating ☆ | | | | | Median Word Count ✎ | | | | | Median Sentiment ✎ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H3 | H4 | MP | FP | All | H2 | H3 | H4 | FP | All | H2 | H3 | H4 | FP | All |
| ☰ Description | – | – | – | – | – | 22 | 19 | 15 | 18 | **18** | 0.4 | 0.4 | 0.2 | 0.4 | **0.4** |
| ☰ Visual Encodings | – | – | – | – | – | 22 | 18 | 19 | 19 | **20** | 0.4 | 0.4 | 0.3 | 0.4 | **0.4** |
| ☰ Interactivity | – | – | – | – | – | – | 19 | 13 | 18 | **17** | – | 0.3 | 0.0 | 0.4 | **0.3** |
| ☰ Visualization Goal | – | – | – | – | – | – | – | – | 17 | **17** | – | – | – | 0.4 | **0.4** |
| 👍 Feedback | 4.4 | 4.6 | 4.5 | 4.6 | **4.5** | 23 | 15 | 14 | 15 | **16** | 0.4 | 0.4 | 0.4 | 0.4 | **0.4** |
| 👍 Non-Color Encodings | 4.5 | 4.8 | 4.5 | 4.7 | **4.6** | 23 | 15 | 15 | 18 | **18** | 0.4 | 0.4 | 0.4 | 0.4 | **0.4** |
| 👍 Color Encodings | 4.3 | 4.6 | 4.4 | 4.6 | **4.5** | 23 | 20 | 16 | 18 | **19** | 0.4 | 0.4 | 0.2 | 0.4 | **0.4** |
| 👍 Context | 4.2 | 4.4 | 4.4 | 4.5 | **4.4** | 24 | 16 | 16 | 17 | **17** | 0.5 | 0.4 | 0.4 | 0.4 | **0.4** |
| 👍 Lie Factor | 4.4 | 4.5 | 4.3 | 4.5 | **4.4** | 23 | 15 | 10 | 13 | **16** | 0.0 | 0.0 | 0.0 | 0.0 | **0.0** |
| 👍 Data Ink Ratio | 4.4 | 4.7 | 4.4 | 4.7 | **4.5** | 20 | 12 | 11 | 13 | **15** | 0.4 | 0.2 | 0.0 | 0.4 | **0.4** |
| 👍 Data Density | 4.4 | 4.6 | 4.4 | 4.6 | **4.5** | 23 | 16 | 11 | 14 | **16** | 0.4 | 0.3 | 0.0 | 0.3 | **0.3** |
| 👍 Gestalt Principles | 4.4 | 4.6 | 4.5 | 4.7 | **4.6** | 23 | 12 | 11 | 09 | **14** | 0.5 | 0.4 | 0.4 | 0.4 | **0.4** |
| 👍 Design and Aesthetics | 4.3 | 4.5 | 4.5 | 4.6 | **4.5** | 20 | 13 | 13 | 13 | **14** | 0.6 | 0.5 | 0.6 | 0.5 | **0.5** |
| 👍 Interaction Effectiveness | 4.3 | 4.5 | 4.5 | 4.7 | **4.6** | – | 16 | 15 | 15 | **15** | – | 0.4 | 0.4 | 0.5 | **0.5** |
| 👍 Visualization Effectiveness | – | – | 4.6 | 4.6 | **4.6** | – | – | – | 14 | **14** | – | – | – | 0.5 | **0.5** |
| 🔍 Insight/Understanding | – | – | – | – | – | 30 | 30 | 19 | 17 | **22** | 0.2 | 0.0 | 0.1 | 0.3 | **0.2** |
| **Overall by Assignment** | 4.4 | 4.6 | 4.5 | 4.6 | **4.5** | 23 | 17 | 14 | 16 | **17** | 0.4 | 0.4 | 0.3 | 0.4 | **0.4** |

The ratings range from 1 to 5 and are provided for homework 3, homework 4, the midterm project, and the final project (not homework 2). The sentiment ranges from -1 to 1 and the per-question word counts range from 0 to 90. Those per-question metrics are provided for homework 2, homework 3, homework 4, and the final project (not the midterm). The questions and assignments with the lowest and highest average rating, median word count, and median sentiment are highlighted. See Table 2 for icon legend.

Despite being underutilized, rating appears related to word count. The word count is lowest for 1 ratings, likely for non-functional or missing visualizations. The highest counts are from ratings 2, 3, and 4, where we would expect there to be issues requiring more feedback than the extremes.

The sentiments per rating somewhat match expectations as well. The responses with a 3, 4, and 5 rating have increasingly higher median sentiments, indicating that students are in fact more positive in higher rated responses. The sentiment for ratings 1 and 2 surprisingly increase, but many of these appear to be mistakes. For example, the comment "I love the design of the website" has a positive 0.64 sentiment, but was assigned the lowest 1 rating. With how few responses had 1 and 2 ratings, it is possible this finding is due to noise.

### Student and Grade Breakdown

We explored whether any of the metrics for numeric or written feedback reflected the assignment grades or final course letter grades assigned, however we did not find any patterns. We also did not find any patterns between the different semesters, despite the transition to remote learning due to the COVID-19 pandemic. We found no patterns when comparing undergraduate versus graduate students. Whether a student received a "Reviewer Award" also appears unrelated to the feedback metrics and grades received.

### Survey Tag Breakdown

We finally explored the anonymous survey feedback provided by students at the end of the semester. All 100% of responses mentioned **peer feedback was helpful** in some way, however 68% also had one or more negative comments.

**Verbal Feedback**   We explored the survey responses in Figure 4 by modality: numeric, written, versus verbal feedback. There was a clear favorite: 79% of all responses mentioned verbal feedback positively and 0% mentioned it negatively. Some were specific about why: 32% of the positive responses mentioned preferring the interaction between those giving and receiving feedback, and 32% mentioned verbal feedback was higher quality. One student wrote, "As for the verbal feedback, everyone was a lot more comfortable giving constructive criticism and suggestions for changes that could improve the visualization." Additionally, 13% of all responses suggested adding more verbal feedback opportunities.
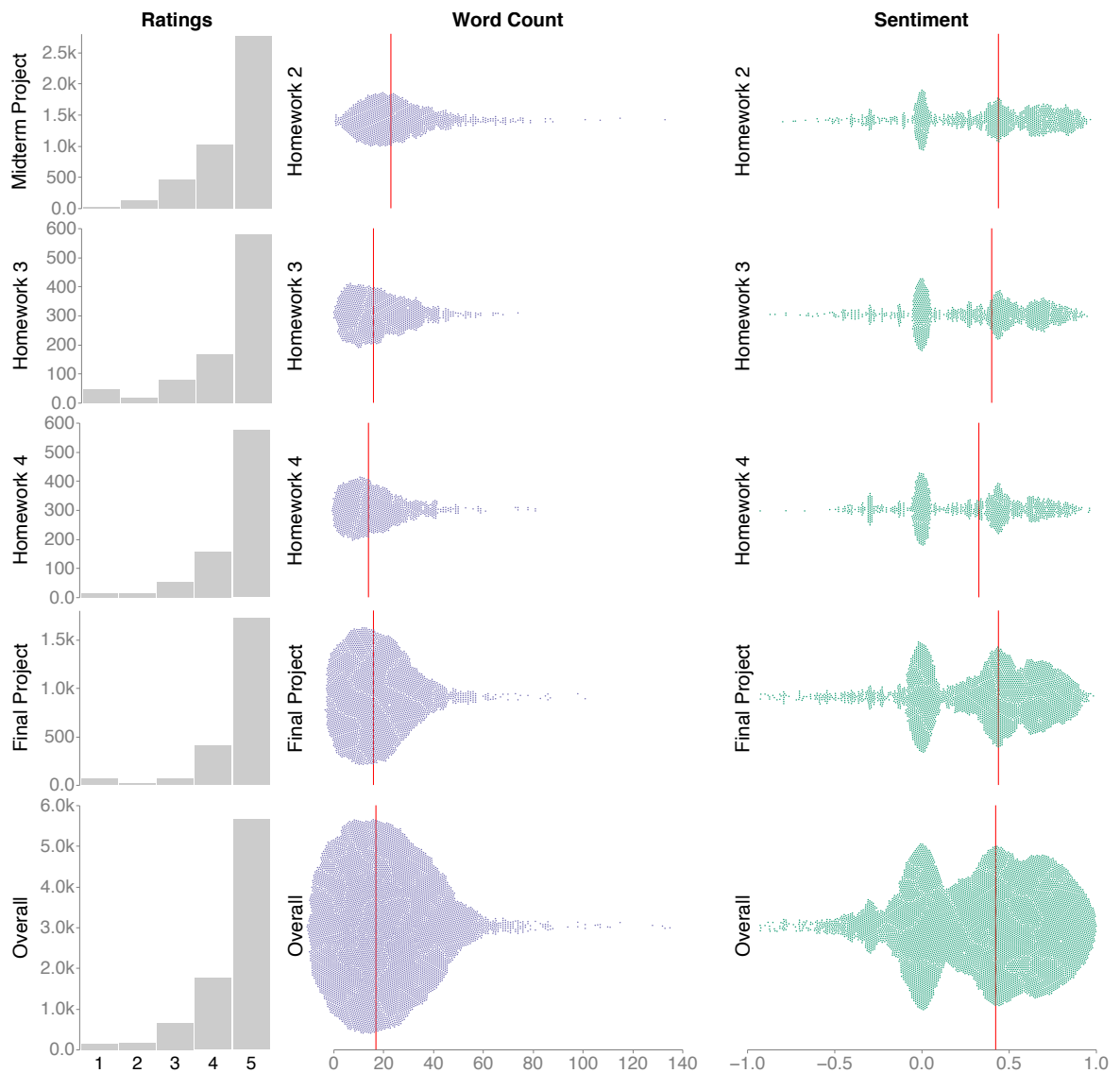
7

**Figure 2.** Overview of per question metrics by assignment. The left column shows rating histograms for the midterm project, homework 3 and 4, and final project. The x-axis is the rating value from worst (1) to best (5), and the y-axis is the number of times students used that rating when providing feedback. The other columns show swarm plots of the word count (middle, purple) and sentiment distributions (right, green) for homework 2, 3, and 4, and the final project. Each circle is one response to one rubric question. The red line provides the median value for each swarm plot. The bottom row shows the metrics overall, not broken down by assignment.

**Written Feedback** The results were more mixed for numeric and written feedback. Written feedback was the next highest, with 64% of responses mentioning it positively and 45% negatively. Of the negative responses, 76% mentioned written feedback was too time consuming. One student wrote, "Due to the amount of time spent just filling out the form, I did not go into detail about how to solve a issue."

**Numeric Feedback** Numeric feedback was the least preferred, mentioned positively in only 47% of all responses. One student wrote, "No ratings because I don't like being rigid with classmates and having bad ratings on things they can easily improve on." However, numeric ratings had fewer negative mentions (19% of all responses) than written. Of the positive responses, 41% were because students preferred the midterm
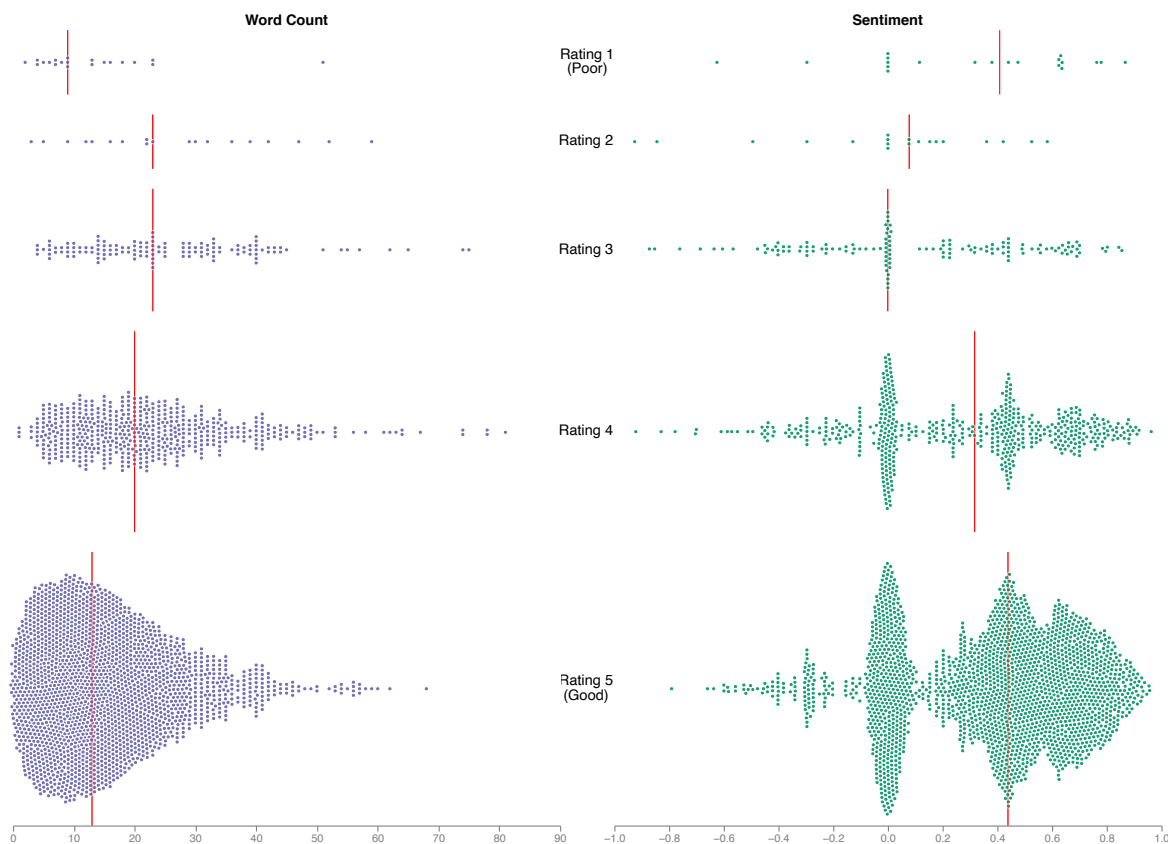
8

**Word Count** **Sentiment**

**Figure 3.** Beeswarm plots illustrating the distribution of word count (left, purple) and sentiment (right, green) of the written feedback broken down by rating on a scale of 1 (poor) to 5 (good). Each circle is one response to a rubric question. The red line provides the median value for each subplot.

feedback, where ratings were given quickly and anonymously. These responses could reflect less of a preference for ratings as much as a preference for quicker and anonymous feedback. Looking at just negative responses, $44\%$ indicated the ratings were unhelpful and $22\%$ indicated that ratings were too positive.

**Feedback Quality** Feedback being too positive was another general theme. When asked about the feedback quality, $77\%$ of all responses indicated at least some of the feedback was constructive. For example, one student wrote, "I think that most of the feedback was very constructive and actionable criticism. There were some general and some more specific comments, but it was largely helpful." However, $60\%$ of all responses felt feedback was mostly positive or praise. This was an issue for some: $30\%$ of responses mentioned issues with feedback quality,

with $50\%$ of those responses stating the feedback was too positive to be helpful and $50\%$ of those responses indicating the feedback was too general. One student wrote, "The majority of it was praise in text. I think people were too cautious to offend anyone, so the majority of feedback wouldn't really be actionable."

**Assignment Type** We also explored responses that mentioned homework versus projects feedback. Overall, $79\%$ of all responses mentioned feedback was helpful for improving future submissions. However, $30\%$ mentioned homework feedback negatively, and $93\%$ of those negative responses were due to homework feedback being completed after the deadline and not directly applicable to other homework.

Projects were mentioned positively in $45\%$ of all responses. Of those, $76\%$ mentioned that the feedback was helpful as there was enough

9

time to incorporate the suggested changes into the final project version. One student wrote, "Honestly I wouldn't review much of my peer feedback for any of the homeworks. Once we were done with an assignment, we were done. Therefore it didn't incentivize me to read what others wrote. With the midterm and the final it has been much different though. It encourages me to listen to what they have to share and actually implement some changes unlike with the homeworks."

**Other Benefits** There were other benefits highlighted by the responses. For example, $43\%$ of all responses mentioned the feedback process helped them better understand the course concepts and another $40\%$ mentioned it lead to them diagnosing or solving problems in their peers' work. One student wrote, "The feedback process makes me to think of data visualizations in those detailed aspects, and that definitely helped me to understand the inner core of data visualization, that everything we care about is with reason and conveys information from data."

Approximately $62\%$ of all responses mentioned feedback was helpful for gaining perspective. Of those responses, $90\%$ mentioned it was helpful to see submissions of their peers. This was often either to compare their work, get new ideas or inspiration, and/or seeing other approaches to visualizing the same data. Another $31\%$ of those responses on perspective mentioned it was helpful to receive multiple peer perspectives and interpretations of their visualizations. For example, a student wrote that feedback "helped me to see how an audience can interpret my visualization and check whether I am able to convey the intended message." This highlights the usefulness of the new "Insight" question added to the rubric.

**Suggestions** Finally, $38\%$ of all responses made suggestions to improve the feedback process. Most of these responses indicated a desire for more verbal feedback, an alternative to Canvas, shorter rubrics or fewer feedback assignments, or more emphasis in terms of grades or class time on feedback. Another $9\%$ of all responses, however, included suggestions that hinted at why written feedback might have been so positive: concern that non-anonymous negative feedback would impact grades and may be perceived as being more harsh
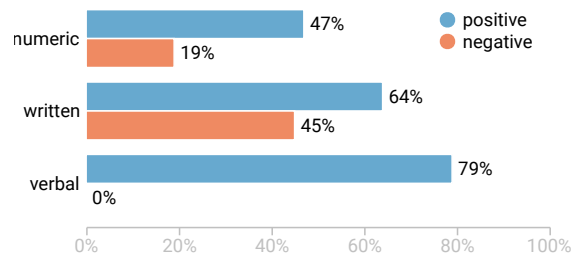


**Figure 4.** Percent of survey responses that mentioned a feedback modality positively or negatively at least once. Since some responses mentioned modalities both positively and negatively, percentages will not add up to 100% within categories.

in written versus verbal form and could even impact in-class friendships.

## DISCUSSION

Overall, we found that students found the peer feedback rubric helpful, but struggled to provide critical feedback. These are common findings for these pedagogical techniques within higher education [8]. However, it shows that our efforts paid off even with the shorter rubric.

We were also able to supplement our quantitative findings via the qualitative analysis of the end-of-semester surveys. For example, topic modeling showed that students were using relevant terminology and word counts showed the students were providing considerable feedback per question. However, those quantitative analyses could not show whether the terms were used correctly to provide helpful feedback. The survey analysis, thankfully, confirmed the feedback was helpful.

The median ratings and sentiment analysis showed the feedback was too positive, which was again confirmed by the survey results. These results show promise that by utilizing the "Rubric," "Peer Review," and API features of the Canvas LMS, we may be able to automate these analyses and track utilization real-time. This could allow us to make rubric modifications throughout the semester, rather than conducting a time-consuming qualitative analysis after the course ends.

However, we were disappointed that sentiment analysis was problematic due to visualization terminology and the often negative underlying datasets. This issue persisted even when we switched from VADER [16], trained for "senti-

ments expressed in social media," to Stanford CoreNLP. These issues might be addressed by using a model tailored for data visualization feedback, or via other NLP techniques.

We were also unable to find any relation between feedback and grades. One factor could be the high $A-$ average final grade, leaving little room for variability. Another factor might be the preexisting peer feedback structure of the course, which had an intentional disconnect between grades (based on functionality) versus feedback (based on effectiveness) for the homework.

Our findings highlight that providing peer feedback constructively and clearly is challenging for students. Min [19] found that while peer feedback is beneficial, it is frequently criticized due to the students' "inability to provide concrete and useful feedback." Lam [20] recommends a "peer review workshop" approach where students are given training to giving and receiving peer feedback. They found that through training, students can learn to provide and evaluate the received feedback for subsequent revisions in the context of an ESL/EFL writing course. We may be able to replicate this in our course as well.

Finally, most of our findings confirm those of Beasley et al. on the original rubric [3]. Their work also shows students were using appropriate terminology for the questions, preferred written to numeric responses, valued the peer feedback process, and that some even wanted a greater importance on peer feedback in the course. However, we also discovered a strong preference for verbal and formative feedback amongst our students. And, despite having reduced the questions by half, our students still felt the rubric was too long.

Based on our findings and those of prior work, we recommend the following to other educators considering a similar approach:

- **Keep all of the rubric questions, including the new "Insight" category.** However, let students choose a smaller subset of "Feedback" questions appropriate for the submission. Require students to provide *critical* feedback for at least one "Feedback" question.
- **Allow the feedback to appear anonymous between students, but not for the instructors.** This allows instructors to grade feedback and enforce a code of conduct, while encouraging

students to provide critical feedback. This functionality is supported in Canvas.

- **Provide incentives for students to utilize feedback.** This could be a feedback reflection exercise, or a small extra credit assignment to improve submissions based on the suggestions.
- **Keep written feedback, but eliminate numeric ratings.** Students underutilized the rating scale, preferred written feedback over ratings, found ratings unhelpful, and ratings may result in a drop in word count for written feedback.
- **Facilitate synchronous, formative verbal feedback opportunities.** While logistically difficult to manage, the end of semester survey indicates these opportunities were the most helpful and preferred form of feedback.
- **Use word count and topic modeling to track how students are utilizing the rubric for peer feedback**, but be wary of simple sentiment analysis of the feedback.

However, we have one recommendation that rises above the rest: **utilize rubrics and peer feedback when teaching data visualization**.

## CONCLUSION

We adapted and simplified a preexisting visualization rubric for peer feedback designed for a different program into our own classroom. We then analyzed the utilization of that rubric across 2 semesters. Our quantitative and qualitative analyses confirm well-established findings that rubrics and peer feedback are helpful, but also that we can learn how students are utilizing the rubric via quantitative analysis, and that this specific rubric was utilized well by students in our course.

In addition to the 6 recommendations we make based on our findings, this project also provides the following artifacts for other educators:

- The AFaR Rubric, an adapted (and shorter) data visualization rubric for peer feedback based on the work of Friedman and Rosen [2].
- Results from our analyses of the feedback provided using the AFaR rubric. These results are helpful for educators deciding whether to use the same rubric in their own courses.
- Jupyter notebooks showing how to access rubric data from the Canvas API at github.com/djbarajas/canvas-rubrics-api/. These

**11**

notebooks will be helpful for educators wanting to measure utilization of this or other rubrics within the Canvas LMS for their own courses.

● Additional supplementary material in the appendix, including what the rubrics looked like for different assignments, anonymized rubric responses from students, the tag dictionary used for the end-of-semester survey analysis, and example course materials.

Given our findings and those of prior work [3], we strongly encourage other educators to integrate rubrics for peer feedback (especially verbal feedback) in similar data visualization courses.

## ACKNOWLEDGMENT

## ◼ REFERENCES

1. Y. M. Reddy and H. Andrade, "A review of rubric use in higher education," *Assessment & Evaluation in Higher Education*, vol. 35, no. 4, pp. 435–448, 2010. [Online]. Available: https://doi.org/10.1080/02602930902862859

2. A. Friedman and P. Rosen, "Leveraging peer review in visualization education: A proposal for a new model," Accepted Abstracts, Pedagogy of Data Visualization Workshop (PDVW '17), Phoenix, AZ, USA, 2017.

3. Z. Beasley, A. Friedman, L. Pieg, and P. Rosen, "Leveraging peer feedback to improve visualization education," in *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2020, pp. 146–155.

4. E. Panadero and A. Jonsson, "The use of scoring rubrics for formative assessment purposes revisited: A review," *Educational Research Review*, vol. 9, pp. 129–144, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1747938X13000109

5. D. D. Stevens and A. J. Levi, *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Stylus Publishing, LLC, 2013.

6. S. P. Balfour, "Assessing writing in MOOCs: Automated essay scoring and calibrated peer review," *Research & Practice in Assessment*, vol. 8, pp. 40–48, 2013.

7. J. Moxley, "Big data, learning analytics, and social assessment," *The Journal of Writing Assessment*, vol. 6, no. 1, pp. 1–10, 2013.

8. J. McGourty, P. Dominick, and R. R. Reilly, "Incorporating student peer review and feedback into the assessment process," in *Proceedings of the 28th Annual Frontiers in Education - Volume 01*, ser. FIE '98. USA: IEEE Computer Society, 1998, p. 14–18.

9. C. Moore and S. Teather, "Engaging students in peer review: Feedback as learning," *eCULTURE*, vol. 5, no. 1, 2012. [Online]. Available: https://ro.ecu.edu.au/eculture/vol5/iss1/4

10. D. Nicol, A. Thomson, and C. Breslin, "Rethinking feedback practices in higher education: A peer review perspective," *Assessment & Evaluation in Higher Education*, vol. 39, no. 1, pp. 102–122, 2014. [Online]. Available: https://doi.org/10.1080/02602938.2013.795518

11. A. Kerren, J. T. Stasko, and J. Dykes, *Teaching Information Visualization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 65–91. [Online]. Available: https://doi.org/10.1007/978-3-540-70956-5_4

12. A. Friedman, "Toward peer-review software and a rubric application in visual analytics classes: A case study," *Education for Information*, vol. 35, no. 3, pp. 241–249, 2019.

13. T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing, and J. development team, "Jupyter notebooks - a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Scmidt, Eds. Netherlands: IOS Press, 2016, pp. 87–90. [Online]. Available: https://eprints.soton.ac.uk/403913/

14. S. Bird, E. Loper, and E. Klein, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

15. R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

16. C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text,"

in *Proceedings of the Eighth International Conference on Weblogs and Social Media*, ser. ICWSM-14. Palo Alto, CA: Association for the Advancement of Artificial Intelligence (AAAI), June 2014, pp. 216–225. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109

17. A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer, "Reactive vega: A streaming dataflow architecture for declarative interactive visualization," *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2016. [Online]. Available: http://idl.cs.washington.edu/papers/reactive-vega-architecture

18. H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, 2019.

19. H.-T. Min, "Training students to become successful peer reviewers," *System*, vol. 33, no. 2, pp. 293–308, 2005.

20. R. Lam, "A peer review training workshop: Coaching students to give and evaluate peer feedback," *TESL Canada Journal*, pp. 114–114, 2010.

**Daniel J. Barajas** is a Software Engineer at Rescale. He received a B.S. in Computer Science from the University of San Francisco. His research interests include visualizing machine learning models and natural language processing. He co-authored a paper, "AGAMI: Scalable Visual Analytics over Multidimensional Data Streams," published at BDCAT by IEEE in 2020. Contact him at danielbarajas21@gmail.com.

**Xornam S. Apedoe** is an Associate Professor in the Department of Learning and Instruction at the University of San Francisco. Dr. Apedoe received her Ph.D. in Instructional Technology from the University of Georgia. Her recent research interests include examining how STEM learning experiences influence identity development, and strategies for addressing the inequities of who participates in STEM+C (computer science) learning environments. Contact her at xapedoe@usfca.edu.

**David G. Brizan** is an Assistant Professor in the Department of Computer Science at the University of San Francisco. Dr. Brizan received his Ph.D. in Computer Science from the Graduate Center, City University of New York (CUNY). His research focuses on natural language processing and machine learning. Contact him at dgbrizan@usfca.edu.

**Alark P. Joshi** is an Associate Professor in the Department of Computer Science at the University of San Francisco. Dr. Joshi received his Ph.D. in Computer Science from the University of Maryland Baltimore County. His research has focused on data visualization and computer science education. Contact him at apjoshi@usfca.edu.

**Sophie J. Engle** is an Associate Professor in the Department of Computer Science at the University of San Francisco. Dr. Engle received her Ph.D. in Computer Science from the University of California, Davis. Her research focuses on data visualization and computer science education. Contact her at sjengle@usfca.edu.