Implementing Oral Exams at Scale Using Graduate Student Instructors

Abstract

We won't write anything here until the paper is complete

Introduction

As instructors of statistics, we frequently find ourselves asking, "When a student answers a question, are they parroting what we've already told them, or do they truly understand?" Unfortunately, its hard to answer this question based on written exam grades alone. Addressing this same concern, Theobold (2021) convincingly argued that, in line with the 2016 GAISE standards (Carver et al. 2016), one method that does allow for statistics and data science educators to more deeply assess what students actually know are oral exams.

Oral exams, as the name implies, are assessments where questions and answers are verbally given, with the opportunity for further probing and follow up questions. Oral exams are more commonly used outside of the United States at the undergraduate level (Asklund and Bendix 2003; Ramella 2019) but are still regularly used inside the US at the doctoral level. At the undergraduate level, some research has been done on using oral exams in introductory chemistry classes (Ramella 2019), math classes (Iannone and Simpson 2012), introductory computer science classes (Ohmann 2019), and, as mentioned, statistics classes (Theobold 2021).

While oral exams still act as an assessment in a course, Theobold (2021) and others have argued that their benefits go beyond mere assessment. They develop communication skills (Joughin 2010), are a more authentic way to assess (Wiggins 2019; Beccaria 2013), are powerful tools in gauging a student's understanding through conversation (Iannone and Simpson 2012; Asklund and Bendix 2003; Huxham, Campbell, and Westwood 2012), encourage greater preparation (Ohmann 2019), and (especially pertinent in the age of AI) greatly discourage cheating (Newell 2023; Ramella 2019).

Like other forms of assessment, oral exams are not without their challenges. These challenges range from accommodating ESL learners, mitigating evaluator bias, and managing student

anxiety. However, for larger classes, perhaps the biggest challenge in implementation is scale. Several studies have acknowledged the open problem scale poses to administering high quality oral exams (Asklund and Bendix 2003; Ohmann 2019), and the unknown effect scale would have on reliability, validity, required resources, and bias (Memon, Joughin, and Memon 2010; Kang et al. 2022; Huxham, Campbell, and Westwood 2012). Though many studies have followed the administration of oral exams with an instructor to student ratio ranging from 1:10 to 1:60, what happens when that ratio becomes very large, say 1:100 or more, is still an open question.

In this study, we detail our attempt at implementing oral exams at this type of scale in the setting of multiple section introductory statistical programming classes. Unlike other studies, which generally focus on the student experience, this study follows the graduate teaching assistants that were administering oral exams under the guidance of a course coordinator. Though understanding the student experience is important, the experience of statistics and data science graduate student instructors is of particular importance in this setting since administering and grading oral exams at a large scale is likely impossible without their help. Their experience and voice matters for understanding whether delivering oral exams at scale with novice instructors can be effective and manageable.

In short, this study has two main research questions:

- RQ1: What problems of practice arose for these graduate teaching instructors when attempting to scale oral exams to our large programming classes?
- RQ2: What recommendations can we offer to others as we reflect on our attempt to administer oral exams at scale?

Course Background and Oral Exam Design

Six graduate teaching assistants that were instructors of record for one-credit programming courses (henceforth referred to as graduate student instructors - GSIs) were the subjects of this study. These graduate students were either Master's or PhD students studying statistics. The courses they taught covered either SAS programming or R programming. Each graduate student instructor had three or four sections of one of these courses with about 35 students per section.

SAS Course Information

The SAS course introduces students to programming in SAS through SAS Studio in the SAS OnDemand for Academics browser-based platform. The course covers topics like reading in raw data with `PROC IMPORT`, basic row and column manipulations of SAS datasets through the `DATA step`, summarizing data numerically with `PROC FREQ`, `PROC UNIVARIATE`, and `PROC MEANS`, summarizing data graphically with `PROC SGPLOT` and `PROC SGPANEL`, analysis of means

using `PROC TTEST` and `PROC GLM`, and fitting linear models (one-way ANOVA and Multiple Linear Regression) through `PROC GLM`.

The course has a prerequisite of a business statistics course or a corequisite of a second course in statistics. These restrictions allow the course to cover the creation and interpretation of statistical models without the need to introduce the statistical concepts surrounding these topics.

R Course Information

This course introduces students to programming in the R software through the RStudio Interactive Development Environment. The course covers topics like using and manipulating common R objects (such as lists, data frames, and (atomic) vectors), reading in raw data using the `readr`, `haven`, and `readxl` packages, creating output documents with R Markdown, common row and column manipulations with the `dplyr` and `tidyr` packages, summarizing data numerically, summarizing data graphically with the `ggplot2` package, using vectorized functions and for loops, and writing custom functions.

This course does not have a prequisite or corequisite. This causes the course to cover more programming-centric topics and less statistical topics when compared to the SAS programming course.

Student Information

Both courses serve three major audiences: statistics majors, statistics minors, and business majors.

The statistics majors generally have taken an introductory statistics course and a three-credit introductory programming course in python prior to enrolling in these courses. Statistics majors are generally taking additional statistical coursework concurrently with either programming course.

The statistics minors are similar to the majors but do not generally have the introductory programming course.

The business majors usually enroll in the course having taken a business statistics course at some point earlier in their academic careers.

For all three audiences, the broad purpose of the courses from the program level is to provide the students with an introduction to using SAS or R in order to facilitate the use of these languages in their upper level statistics or business courses.

Course Structure

Both programming courses have a flipped structure in which students receive the course learning materials prior to the in-class session. This is done in order to spend most of the in-class time on activities involving active participation. Each class meets for 50 minutes once a week. Prior to class, students are expected to watch one or two 10-20 minute videos created by the course coordinator. In completing these videos, students take two to three quizzes embedded in the content. During most class sections, instructors briefly recap material for the week, take questions, and then introduce the in-class activity. The students are then given the bulk of the time to work through the material while the GSI can check in on students and answer questions. The last five minutes of the class are used to recap important parts of the activity. The students are then formally assessed on the material by taking an asynchronous quiz due a few days later.

For this iteration of the courses, there were four weeks that did not follow this structure. Two of those weeks replaced the usual in-class period with a short paper-based quiz (10 minutes) followed by time to work on a homework assignment. These assignments involved creating their own original program to answer questions of interest along with finding a data set of their own and applying class concepts. The remaining two weeks were dedicated to administering oral exams. These weeks involved no new content and no class was held. Each oral exam accounted for eight percent of the students' overall course grade.

Week	SAS Course Topic	SAS Course Assignment	R Course Topic	R Course Assignment
1	Basics of SAS, SAS Studio	LMS Quiz 1	Basics of R Programming	LMS Quiz 1
2	Libraries & Reading Data	LMS Quiz 2	Common Data Objects	LMS Quiz 2
3	Reading Data	LMS Quiz 3	R Packages & readr	LMS Quiz 3
4	Column Manipulations	In-Class Quiz 1 & Homework 1	R Markdown	In-Class Quiz 1 & Homework 1
5	Creating New Variables	LMS Quiz 4	No Class	
6	Row Manipulations	LMS Quiz 5	Oral Discussion 1	Oral Discussion 1
7	Oral Discussion 1	Oral Discussion 1	Row and Column Manipulations	LMS Quiz 4
8	Contingency Tables & SAS Options	LMS Quiz 6	No Class	
9	Numeric Summaries	LMS Quiz 7	Creating New Variables & Reshaping Data	LMS Quiz 5
10	Plotting	In-Class Quiz 2 & Homework 2	Numeric Summaries	In-Class Quiz 2 & Homework 2

Week	SAS Course Topic	SAS Course Assignment	R Course Topic	R Course Assignment
11	Statistical Concepts & Correlation	LMS Quiz 8	ggplot2	LMS Quiz 6
12	Oral Discussion 2	Oral Discussion 2	Oral Discussion 2	Oral Discussion 2
13	Linear Regression	LMS Quiz 9	Loops & Vectorized Functions	LMS Quiz 7
14	No Class		No Class	
15	Analysis of Means	LMS Quiz 10	Writing Functions	LMS Quiz 8
16	No Class	Final Project	No Class	Final Project

Table 1 - SAS and R Course Schedules. LMS = Learning Management System

The course format allows for GSIs, even without a robust understanding of the material or effective pedagogical practices, to run the classes. During in class activities, these instructors mainly use prior programming experience and problem solving skills to appropriately guide students to successfully solve their problems. In order to manage the instructor load, all sections are capped at 40 students. Most sections had between 35 and 40 students, with a few having much lower enrollment. In total, there were approximately 700 students across all of the sections of the two courses.

Designing the Oral Discussions

As oral exams are not common for courses taken by these students, they were branded as an 'oral discussion' to hopefully lessen the anxiety students felt with the term 'exam.'

The design of the oral discussions roughly followed that of a structured interview protocol ("Structured Interviews" 2024). In this format, subject matter experts create relevant questions and a rating scale for responses. The interviewers systematically ask these questions of the interviewees in order to be as fair as possible while ensuring valid and reliable ratings. Follow up or 'probe' questions are also created to help elicit responses at the appropriate level desired by the interviewer.

The first oral discussion occurred in week six of the R course and week seven of the SAS course. For the first oral discussion, the course coordinator developed an example program and a script of candidate questions to ask about the program for each course. They also created a list of probable student responses with subsequent follow up questions to gain the appropriate clarity of responses. The candidate questions were split into two categories: higher-level questions involving more detailed explanations and lower-level questions involving mostly recall. As the oral discussions for a given student were slated to be taken in a five minute window, two of each type of question were developed for the first oral discussion. The grading scales were 0, 1, 2, or 3 points, and 0, 1, or 2 points for the higher-level and lower-level questions, respectively.

During a weekly meeting of all instructors, the instructors and course coordinator went through the example programs (one SAS program and one R program) and questions written by the course coordinator. The GSIs then split into pairs (one a SAS instructor and one an R instructor) and practiced administering the discussion to each other. Using this as feedback, the questions and follow ups were modified as needed.

For the second oral discussion, which occurred in week 12 for both courses, the course coordinator developed an example program to share with students and a similar program to use for the actual oral discussion. This time a weekly meeting was devoted to having the GSIs develop questions and the corresponding follow ups. The GSIs were again split into pairs and practiced administering the discussions.

After the first oral discussions, the GSIs felt strongly that they would like to have a finer scale for their grading rubric. Whereas before the lower-level questions were out of 0, 1, or 2 scale, these were now out of a 0, 1, ..., 4 scale. Similarly, the higher-level questions were now graded on a scale of 0, 1, ..., 6.

The oral discussion scripts used by the graduate student instructors and SAS or R programs are available in the appendix.

Administering the Oral Discussions

The oral discussions were administered through zoom. This decision was made for the ease of scheduling for both the instructors and the students, since both groups each had their own class schedules and finding a time to schedule out rooms that would work for everyone wasn't feasible. The discussions were set to five minutes in length with the intention that the actual time for the discussions would be closer to three-four minutes, allowing for a buffer between appointments. The discussions were closed notes and did not involve students coding. The students were given an example program and instructed to be prepared to discuss the purpose and syntax of the code. Instructors were encouraged to build in down time for their appointments to deal with any appointments that went long or had technical difficulty. For instance, only making appointments available from the start of the hour until 50 minutes past. This would leave 10 minutes to account for issues that may arise.

The weeks prior to the discussions students were able to use an online scheduler to sign up for a five minute time slot. They were instructed to sign up for a time slot by the Wednesday prior to the discussion week and contact their instructor if none of the designated time slots worked for them. If this wasn't followed, a 25% deduction to their score could be given (this did not end up being used for any students). Students with accommodations were to contact their instructor to make sure they were able to have their accommodations met, which often involved booking back to back time slots.

The week prior to the oral discussion, students were provided an example program, the rubric used for grading the discussions, and the instructions for how the oral discussion would

progress. For the first oral discussion, we chose to use the same program as both the example program and the actual program in order to ease student anxiety. As students were more comfortable in the second round of oral discussions, different programs were used. The rubric used was based on that given by (Theobold 2021), and was modified between the first discussion and second discussion based on instructor feedback. The instructions for the discussion itself are given in the appendix.

Students were instructed to log into the zoom meeting a few minutes prior to their designated time slot. The zoom meetings themselves utilized the waiting room feature. This ensured that students would not pop into the meeting and interrupt an ongoing discussion. Students were required to have their camera on during the discussions. Students are not required to have a laptop at this institution but no instructors reported issues with students not being able to accommodate virtual meetings with a webcam (this may be due to the excellent library facilities on campus).

The rubric for the oral discussion was set up on the learning management system (LMS) allowing the GSIs to provide their graded feedback as the discussions were taking place. These grades were not released until all students had completed their discussions. Along with this, the discussions themselves were recorded so that any disagreements about grading between the GSIs and students could be mediated by the course coordinator. Only two disagreements required mediation across all of the administered oral discussions.

Lastly, some GSIs found that they needed to have extra 'hidden' times on the last day of the week to account for students that had missed their appointments due to personal or technological reasons. The number of students needing this extra period was pretty low (five to ten) for each GSI.

In total there were six graduate student instructors. Three had four sections of a course and administered oral exams to roughly 140 students each. Three had three sections of a course with sections that tended to have slightly lower enrollment. Two of these GSIs administered oral exams to roughly 90 students. One GSI had a teaching assistant appointment for less time so only administered exams to two of their sections (roughly 70 students). The course coordinator administered oral exams to the remaining section (roughly 30 students).

Methods

Participants and Data Collection

Six GSIs participated in the study and consented to have their data used. Initial thoughts about giving, and experience with, oral discussions were requested on a Google form prior to administering the oral discussions. The R course had its first oral discussion one week prior to the SAS course. Reflections and advice were discussed in the weekly meeting and recorded for future reference. The meeting and conversations after the first SAS course oral discussion were likewise recorded. Another Google form was administered at this point as well. The

second oral discussions occurred during the same week for both courses. The meeting and conversations the week after administration were recorded. Lastly, a Google form was given to record their final thoughts on administering oral discussions in the course. All of the forms are available in the appendix. *Figure 1* shows when surveys and meetings were held in relation to each other and to the oral exams.

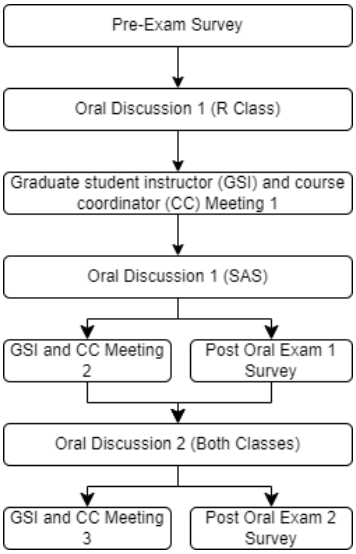

Figure 1: Timeline of oral exam administration and data collection.

Data Analysis

For the open-ended responses on surveys and whole instructor discussions, an iterative thematic analysis using different forms of inductive coding, such as in vivo and descriptive, was used. In the first stage of analysis, one author passed through the data coding first based on the occurrence of particular words such as "time" or "technology". Initial themes were created based on the intial clusters of codes. After a series of discussions about the strengths and weaknesses of the first round of analysis, the authors passed through the data again in order to include more moments and better ground the themes in the context of the data, for example noting why certain conversations were occuring (e.g., a question being asked, it being a certain part of the semester) not just what was being said. After this second round of coding, the authors developed a final set of 4 themes, including 6 sub-themes, that better captured the thoughts, feelings, and experiences of the GSIs. Our codebook can be found in the appendix. Along with the themes generated from the open-ended responses on surveys and whole instructor discussions, closed-ended responses from the surveys were also tabulated.

Results

The closed-ended responses are discussed first, followed by the themes.

Closed-ended Responses

Closed-ended responses were requested at the beginning of the study (Pre-Oral Exam Survey), the middle of the study (Post Oral Exam 1 Survey), and the end of the study (Post Oral Exam Survey).

Pre-Oral Exam Survey

Prior to giving the oral discussions, the course coordinator wanted to know more about the GSIs previous experiences and feelings about oral exams. Based on the pre exam survey, our GSIs had limited experience. None had ever administered an oral exam and only two of the six had taken one. Not surprisingly, there was a range of comfort levels, with half of the GSIs feeling fairly comfortable and the other half feeling fairly uncomfortable.

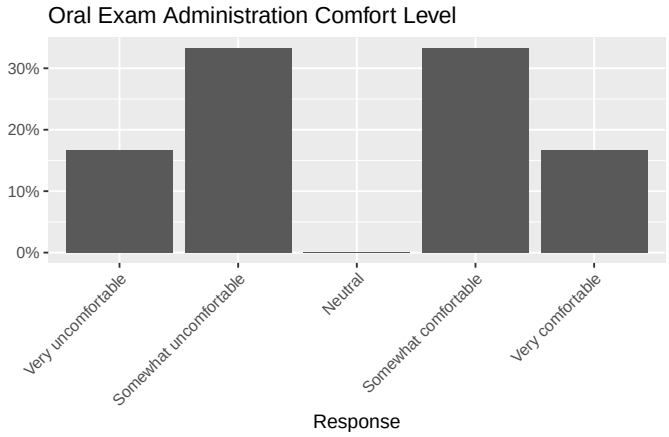

Figure 2: Self-reported comfort with administering an oral exam prior to the start of the semester.

When specifically asked about what they saw as potential logistical and grading issues, GSI responses aligned with future points of discussion during the semester, including concerns about time, scheduling, and bias. Some GSIs worried about their lack of memory and being biased due to familiarity with students. They also worried about getting students in the right time slots and their own ability to administer exams over a large time span.

Post Oral Exam 1 Survey

Three close-ended questions were asked in this survey. They were "How do you think the administration of the exam went?", "How well did the exam act as an assessment of the material in the course?", and "How would you rate the exam generally?". The GSIs were asked to give a rating from one (Very Poorly) to ten (Very Smoothly).

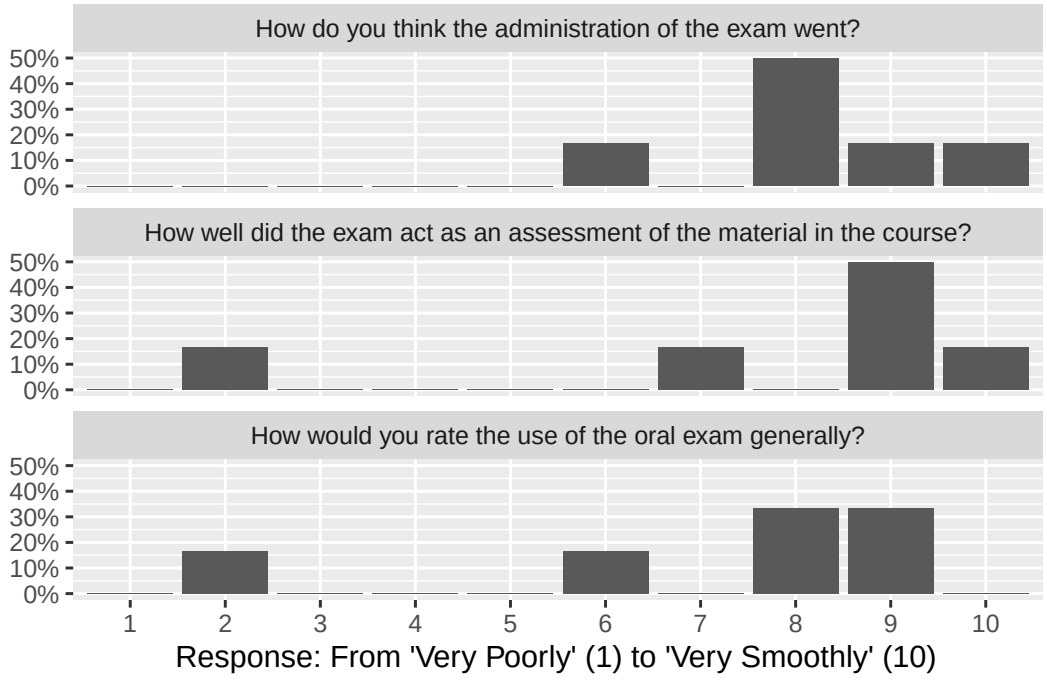

Figure 3: Responses to Questions After Administration of Oral Discussion 1.

As can be seen in Figure 3, generally these questions were answered positively. Further elaboration of these questions is given in the thematic analysis below.

Post Oral Exam Survey

Part of the Post Oral Exam Survey involved a closed-ended question with a follow up response explaining their choice. During the course of the semester, a GSI mentioned that they weren't sure what the oral discussions would accomplish beyond what a usual LMS quiz. To determine the perceived worth of the oral discussion to the GSIs, a survey was given after the final oral discussions were administered where they were asked, "After the first oral exam, a few people commented that a [LMS] quiz would have accomplished the same thing. Do you agree or disagree with this?" The potential answers were "Agree", "It's Complicated", or "Disagree".

Half of the GSIs chose "It's Complicated" and half chose "Disagree". This indicates that they found the oral discussions served a different purpose that a standard LMS quiz.

A follow up question asked them to explain their response. For two of the GSIs that answered "It's Complicated," exam integrity was discussed. For the other GSI, they noted that the oral discussions caused them much more 'suffering and misery' than an LMS quiz. For the GSIs that chose "Disagree," one noted that the oral discussions made it easier to know how deep students understand questions, one noted that it allowed us to evaluate how well student can produce knowledge "on cammand," and one noted that allows us to test how well someone can explain the topics we cover in class and is more lenient than an open-ended quiz question because, if a student says something confusing, they can still get points for correcting course.

Themes

Theme 1: Challenges with the Oral Exam Process

Scheduling

A main challenge that GSIs discussed was scheduling and time management, with conversations about scheduling centered on two types: exam scheduling, which was talk centered on when the exam itself was to be scheduled in relation to other class and GSI activities, and student scheduling, or talk centered on the process by which students scheduled themselves to take the exam and how GSIs managed that process.

Starting with exam scheduling, after the classes had administered both their oral discussions, a small conversation in the instructor meeting centered on finding the best way to place the oral discussion in relation to course homework assignments and school breaks. This conversation brought up questions about what was best for students such as minimizing instructional time lost due to taking an entire week for each oral discussion and placing the oral discussion so students could get feedback on homework before the corresponding oral discussion (see Table 1 for the course schedules - the GSIs were instructed to grade and return homework assignments within five business days when possible but, as this guide was not always met, this often gave little time for students to review their feedback).

On the final post oral discussion survey a few GSIs also commented on when they thought would be the best time to schedule the oral discussions for them. They noted that having their own exam preparations and homework on top of the oral discussions "drained" them, and suggested solutions such as having more graders and shifting office hours later in the semester. One GSI noted that "if the oral exam format (talking about the scheduling) goes unchanged, I am unsure if I am fit to be an instructor next semester" due to being a first year PhD student and their heavy workload at the end of the semester. Thus exam scheduling talk centered on how to make the oral discussions work best for both student and GSI schedules and, ultimately, reduce the amount of work.

Another issue related to exam scheduling was the overall time commitment needed to administer the oral discussions by the GSIs. During the first instructor meeting, one GSI was asked how long it took them to administer their discussion. They estimated around 12 hours. Though they didn't indicate how they felt about it, other GSIs in the first post oral discussion survey made their feelings very explicit, one saying that it "took way too long" and, for the final oral discussion, one said they were concerned that "15 hours on Zoom will drive me insane." Three GSIs commented on this topic in the final post oral discussion survey. One thought the oral discussions were "extremely time inefficient" and one noted that they were "very time consuming." The third GSI noted that, although the time it took to grade a homework assignment was similar overall, it was "incredibly grueling" to do the oral exams because they had to be engaged most of the time. One GSI, during the final instructor meeting, commented "If it were just me with 120 people, I would not do this to myself [talking explicitly about the time commitment]." Overall, GSIs found the time commitment to be a burden, but voiced this concern much more toward the end of the semester when they too were busier.

Talk about student scheduling came up during the first instructor meeting after the SAS course GSIs had given their discussions but prior to the first R course oral discussions. When the SAS course GSIs were asked an open-ended question about what advice they'd give to the R course GSIs, they immediately began talking about student scheduling. They noted that one needed to watch out for student lateness, making sure space out slots with some extra time, and being aware that students might leave if they were required to wait for too long. After both groups had given their first oral discussions, only one GSI commented on student scheduling on the first post oral discussion survey ("How [do we] make everyone register and show up on time?"). However, on the final survey and in the final instructor meeting, almost all GSIs had something to say about student scheduling. For example, when asked again what advice they'd give to future GSIs, one GSI said, "Scheduling, scheduling, scheduling". They offered strategies on how to deal with students missing their time slots, including building in more breaks or making Friday more open. During the instructor meeting similar sentiments were expressed, where multiple GSIs talked about students needing to make up exams and the logistics that accompanied that. Student scheduling was a topic that was brought up both times GSIs were asked to give advice.

Technology

After scheduling, the second challenge that was frequently mentioned was related to technology use. For example, GSIs mentioned a few moments when technology did not do what it was supposed to. During the first instructor meeting after the SAS GSIs had administered their discussions, they talked about the effects of bad wifi, mix ups with links, and students also experiencing similar technological issues on their discussion experience. One GSI noted in the first post oral discussion survey, after both groups had given their first oral discussions, that the Zoom waiting rooms had not worked properly for them. However, after these initial difficulties, no other technological difficulties were discussed for the second oral discussions.

Distinct from technological difficulties were also challenges with learning to use technology in new ways. The GSIs were instructed to record the oral discussions in case a grade dispute occurred. During the first meeting, several GSIs talked among themselves about different ways to record the discussions and download the videos. One pair had questions about inputting oral discussion feedback into the LMS. Interestingly, no one talked about the challenges of learning how to use technology differently after both groups administered their first oral discussion. It appears that the GSIs were able to get a better grasp on using Zoom and the LMS in different ways after the initial burn-in period.

Student Interactions

Another set of challenges came when GSIs had to interact with students in new ways. One of these new interactions was navigating how to accommodate students with different needs. For example, after both groups had given their oral discussions, one GSI brought up a student who asked if they could do it in person, since they had trouble understanding the questions (given in English) over Zoom. Another GSI admitted that they had given one oral discussion in person to work with a student who had accommodations. After some discussion, the instructors and course coordinator agreed that in order to keep things fair, they needed to use other methods if possible, such as the chat option on zoom. Though similar issues were not brought up during the final instructor meeting, one of the same GSIs reiterated the point on the final survey, indicating ESL students should be given more time since they thought "this [the oral discussion] isn't intended to be an English test." Though dealing with ESL students and students with other accommodations was challenging at first, the group worked together to create a consistent testing experience.

Another challenging interaction was interpreting student responses to test questions. Once both groups had given their first oral discussions, a few GSIs noted that students "meandered" more around certain questions. After the second oral discussion was given, the GSIs joked about certain questions where even with follow up, students would reword the answers in equally vague ways. "Don't prompt too much", one GSI mentioned in the final survey. "If student [sic] doesn't have a clear answer then move on" because "figuring out what students are saying is like squeezing water out of a stone". Though this one GSI felt strongly about the issue, no other GSI mentioned the same level of difficulty understanding students. It appears that most of the GSIs were able to develop strategies to work with deciphering difficult student responses.

A final challenging interaction connected to interpreting student responses was asking them follow up questions. During the first instructor meeting, the group talked about student answers around one question that seemed to cause some difficulty. One GSI prompted in a general way, and another prompted by giving explicit options for students to choose from. The course coordinator offered that the most important thing was consistency in how follow up questions were asked among students, even if they way they were asked was different among instructors as different sections could be assessed slightly differently (as with the homework

assignment grading). In the second instructor meeting, after both groups had given their oral discussions, there was a conversation on how better follow up prompts could be given. Basing them on what students might say, being consistent, and scaffolding responses were offered as solutions. Though asking follow up questions was a new thing GSIs were asked to do, through consulting with each other and the course coordinator they found ways to improve.

Self-Care

One final challenge separate from scheduling, technology, and student interactions was navigating self-care. During the first instructor meeting, after hearing about the oral discussion experience of other GSIs for the first time, one GSI asked, "How do you go to the bathroom"? Different strategies were offered, such as going before, positioning oneself near one, letting yourself run a little behind, and even asking students if you could go. Another important suggestion by the course coordinator was to build in guard time where a student couldn't enter the Zoom waiting room. With these initial suggestions, applicable to more than bathroom breaks, most of the GSIs did not bring up self-care again. One GSI still had trouble during the administration of the second oral discussion, noting in the final survey that even though the discussions took as long as grading HW, "you can't snack or use the restroom or walk around if you're getting sore from sitting". Even with initial suggestions for self-care, not every GSI was successful in implementing them.

Theme 2: The Oral Exam Itself

Bias and Fairness

On multiple occasions, GSIs wondered about potential biases in the oral discussion itself, as well as ways to make it more fair. One way they checked for fairness was by comparing scores across instructors. For example, during the first instructor meeting, different GSIs talked about one discrepancy in the discussion scripts they were using, with one person having used five questions and another person having used four. They compared their grading averages to make sure this mistake had not impacted their students. During the second instructor meeting, comparisons were made again in conjunction with questions about changing the grading scale. GSIs also brought up potentially hard questions and their impact on grades during the final instructor meeting. By comparing grades, they hoped to check for fairness across the oral discussion process.

Another way GSIs focused on issues of bias and fairness was through the grading scale. When asked in the first oral discussion survey "Do you think the grading of the exams was fair to the students?", most GSIs felt it was "mostly fair," though they wondered if more leniency and having different levels of grading might make improve it. This question spurred more discussion during the second instructor meeting, where one GSI wanted a more refined scale for each question in order to better distinguish between students who needed prompting and

eventually got the right answer and those that needed prompting but ultimately did not answer the question correctly. Another GSI and the course coordinator agreed with this idea. Overall, GSIs brought up changing the grading scale as a way to accommodate for a larger range of student outcomes and make things more fair.

Finally, GSIs also commented on the quality of the questions themselves, and the potential problems they could cause for students. During the first instructor meeting, one person brought up a question related to R objects, wondering if the question was too vague based on student responses. All of the GSIs teaching R agreed that another question was likely too hard. When asked in the first survey "Do you think the grading of the exams was fair to the students?", one GSI mentioned the broadness of some questions and that "maybe [it's] not clear what answer is expected". However, three of the GSIs explicitly said they thought the grading was 'fair' or 'mostly fair' with a fourth mentioning their request for a finer scale to make it more fair. Through discussion, GSIs and the course coordinator were able to pinpoint questions that needed modification.

Utility

One GSI mentioned that it coincided with what they had been seeing in class and in the homework already. In the second survey a similar sentiment was shared by another GSI, who pointed out that "One student had no clue what the `%>%` [a common operator used in the course] was...so we are catching what we want to". One GSI also noted that the oral discussions were a "very fast and effective way to see who knows what's going on in the class" and who does not. Overall, the GSIs found the oral discussions useful, specifically with assessing knowledge and doing it in a timely matter in relation to a single student.

When asked if they would do oral exams in their own classes in the last instructor meeting, a few GSIs compared the oral discussions to another kind of homework or quiz, but one that was harder to grade. Their responses indicated that it would depend on how much time they had and the number of students.

Theme 3: Reactions to the Oral Exams

Beyond challenges in administration and the oral exam itself, the third theme centered around GSI and student reactions to the oral exam experience as a whole. During the first instructor meetings, all the GSIs agreed that it "went fine". After both groups had gone, the GSIs in the first survey rated the smoothness of administration of the oral discussion at an average of 8 out of 10. When asked to elaborate on their scores, the GSIs expressed that it "went well", "[it was] very smooth" and "I had fun! Not sure if my students had fun". Only one GSI expressed a negative sentiment, stating, "It made me very sad". However, feelings were more negative after the second round of oral discussions were given. Though these feelings were not expressed in the final instructor meeting, in the final survey one GSI said it caused them "undue suffering and misery." Another expressed frustration with student preparation,

and another used the word "grueling" to describe the experience (although they also said in another response that they 'thought it was a very fast and effective way to see who knows what's going on in the class and who isn't with us at all'). A final GSI wondered if they could be an instructor if they were required to do this in the future due to the time commitment. GSI reactions initially started out positive, but as their own workload increased they became more negative.

Student reactions, unlike the GSI reactions, tended to become more positive. In the first and second instructor meetings, several GSIs reported students being annoyed with oral discussion times running behind. They also discussed potential confusion students may have with getting immediate feedback after completing their discussion. However, by the time the second oral discussion happened, one GSI reported students seeming much more relaxed with the experience, and even performing better (another GSI had the opposite experience with students performing better, but saw similar levels of relaxation.) Further, one GSI reported that a few students spontaneously told them that they "like the oral discussions". After being exposed to the new assessment format, it seems students became more comfortable with it.

Theme 4: Preparation for Future Oral Exams

The final theme we saw centered on GSIs looking to the future and wondering how to prepare their students for an oral discussion experience. This came from the final survey and instructor meeting as GSIs reflected on the experience as a whole and how they might do it differently. One idea was to incorporate an oral component into in class activities, maybe by having students explain things to each other in groups. Another GSI wondered about having opportunities for students to present in class and build slides using code. Overall they agreed that making the first one weighted less than the second would be a good idea. On the survey, another GSI offered another idea: "I would encourage future TAs to get their students talking about R or SAS first thing in the semester and enforce the vocabulary". This would also help them answer questions orally correctly. Thus, there were many avenues, ranging from changing activities to focusing on vocabulary, that were mentioned to help students prepare better for the oral discussions.

Discussion

To our knowledge, this is the first study done of a large scale oral exam in the statistics and data science education literature. Further, due to the pivotal role graduate student instructors play in the success of the oral exams, this study focused on their attitudes and experiences as they administered them. As we look at the themes from the conversations and open-ended survey questions, it should be noted that many of the discussions were prompted by questions about the negatives. For instance, "What went wrong?" and "What can we do better for next time?". These questions naturally lead to themes being captured that focus on the perceived issues and not the perceived benefits. Based on the data we received and analyzed, there were a

range of attitudes, mainly focusing on the time commitment of the exams, the potential biases and equity issues of the exams, concerns about their efficacy, and positive student reactions. This left us with two questions: what do our findings mean for designing large scale oral examinations, and what future research directions could be pursued?

The Importance of GSI Feedback

When conducting exams on a larger scale, GSI feedback is pivotal in order to assess whether the oral exam is working the way it is intended. In several instances, GSIs were the ones who brought up issues of fairness and bias, such as expanding the grading scale or whether questions were too broad. These recommendations came through direct experience with their students, experience that, at large scales, we can't obtain as course coordinators. GSIs also detailed what the experience was like on the administrative end, with conversations centering around time, follow up questions, and issues with technology. This allowed for help to be given in the right places and assess problem areas that we may not have thought of previously. Feedback is immensely helpful at crafting the oral exam experience you want.

Asking for feedback through regular meetings and surveys is effective. Meetings allow for the exchange of ideas between GSIs, where, for example, issues can be brought up in a supportive environment with others participating in the same experience. They also allow for multiple solutions to be offered, and are great way way to hear instructor thoughts without having to wait or read through comments. However, surveys are more anonymous. We saw that strong and negative feelings were much more likely to be expressed in a survey rather than in a discussion. Though they took more time to read and understand, they gave us a better overall picture of how the GSIs actually felt. A combination of multiple feedback methods, in our case meetings and surveys, would be most helpful to really understanding our GSIs and how the process is going.

On reflecting on the surveys and meetings, these feedback opportunities could have been elevated by asking a few more prepared, specific questions. Most of the questions asked, at first, were along the lines of, "How did it go?", "Any issues come up?", which are important. However, the responses we got became much more nuanced and informative when we asked more specific questions, such as "Compared to a written exam, what is the utility of an oral exam? Is it any better?". Other ones relating to time, the student experience, follow up questions, and bias and fairness would have elevated the feedback we got and allowed us to implement even more impactful changes. A mix of open, general questions with specific ones prepared more ahead of time is a great way to get feedback.

Though instructors will have experiences of their own, having all of their GSIs report in helps gain a much better picture of different things that do or can go wrong, and ways that the administration of oral exams at a large scale can be made more efficient and bearable.

GSI Training and Managing GSI Load

Scheduling

The most offered advice the GSIs gave to future oral exam administrators centered around scheduling. Scheduling was the main way GSIs coped with the load of giving hundreds of oral discussions in a week. Taking their advice to heart, those who want to administer large scale oral exams need to make sure that there is adequate space for breaks for GSIs to walk, eat, and use the restroom. There needs to be built in time for those who may have to reschedule their exams. The systems used to schedule times need to have mechanisms to allow for these things, and we strongly encourage setting up guardrails ahead of time for at least the first oral exam.

Beyond student scheduling, making sure the oral exams aren't given during a time that is also busy for the GSIs would have alleviated their load in our study. It was when they had to administer exams during their own projects and finals that the most intense negative feelings were reported. We recognize that, in most situations, this is unrealistic. One recommendation is to allow for 2 weeks instead of one for students to take the oral exams, though different solutions will change based on different circumstances. Focusing on scheduling the exams in a way that allows for a balance with other responsibilities (such as grading homework) might be a better method to accommodate instructor schedules. Thus, giving GSIs strategies to effectively schedule students and effectively schedule the exams themselves on weeks when school work is not as busy for the GSIs will alleviate their load.

Understanding the Purpose of an Oral Exam

Along with optimal scheduling, helping GSIs understand the purpose of an oral exam is an important part of training. Though our GSIs complied with the discussions, engaged in meetings, and offered feedback, some that had less interest in teaching still had fuzzy ideas as to why they were being asked to administer an exam in this way. They brought up how an open response quiz may do the same thing (and would certainly take less time). Though they admitted that being able to test knowledge on the fly, allowing partial credit, and avoiding cheating were good aspects of an oral exam, half missed the main point of Theobold (2021), which was that oral exams allowed a deeper look into what students knew, and that being able to probe allows instructors to judge that to a much higher degree than with a normal testing format.

Further, the GSIs were unaware of the benefits from the literature, including authenticity, students preparing more for oral exams, and practice for the future. Understanding and teaching these things will allow GSIs, even if they don't like them, to understand why they are giving oral exams in the first place and help them feel more motivated.

Preparing for the Administration of Exams

Beyond the more practical aspects of scheduling and building an understanding of oral exams, GSIs may require more specific guidance. Many of the GSIs faced unexpected issues

such as technology failing, students answering questions in odd ways, and having to navigate accommodations. Further, none of them had ever given an oral exam before.

Preparing GSIs for the experience and for the potentially unexpected by roleplaying or other approximations of practice (grossman citation here) could help them to feel more comfortable when these situations arise. For example, in our course, we had our GSIs do mock oral exams with each other. What may work better is asking other undergraduates or other peers to participate in roleplaying. Further, providing them with guides on what to say and potentially how to say it helped provide a necessary scaffold before GSIs felt more comfortable doing it on their own. This looked like providing follow up questions for our GSIs to use the first oral exam, but then allowing them more autonomy as they became more comfortable. During our roleplaying of oral exams, we simply discussed them in person since we were all present. It would have been advantageous to have the roleplaying done using the technology of the oral exams. This would have given them a more authentic example of administering the exams prior to the real thing.

Overall, training in the form of practicing interacting in the exam setting, practice with the technology, imagining difficulties that may come, and providing inital scaffolding will help GSIs be more successful in navigating oral exams.

Technology & Preparing Students in a Large-Scale Setting

In order to accommodate the amount of students taking an oral exam in this class, technology was a necessity. For the class, we used the software Zoom. Zoom allowed for not having to schedule out a physical space for hundreds of students, and also to record the meetings in case any issues of unfairness were brought up by the student. We recommend that when trying to do large scale oral exams, video conferencing software should be used in order to save time, space, and easily record exams.

Along with video conferencing software, we also used an LMS scheduler to schedule exams for students. This worked extremely well and students had no issues signing up for time slots. The LMS also provided the functionality for a digital rubric that GSIs could fill out while administering the exam. This saved time as the entire exam for one student was completed in the five minute time period. Whatever is used, whether it's the LMS or Google sheets, proper choice of technology should be considered to help schedule students and ease the grading of the exams.

As mentioned above, we found that roleplaying was extremely useful. However, we highly recommend roleplaying with exactly the same technology as will be used in the oral exams. While our GSIs did practice using the scoring rubric via our LMS, we did not practice with our video conferencing software. Issues with the waiting room behavior may have been avoided and a smoother implementation may have been had if this were practiced.

Preparing Students for the Oral Exam

Prior to each oral exam week, the students were given the specific instructions for the exam, an example program to help them narrow down the topics to study, and the grading rubric that we would use to assess their learning. We believe this was vital in easing student anxiety and discomfort with the exam format.

One of the final pieces of GSI feedback we received revolved around better preparing students for the non-standard experience of taking an oral exam. The other assessments in class did not require students to verbalize their understanding of content. Several GSIs offered suggestions such as asking students to practice oral explanations to peers in class or even giving small presentations of topics using slides. These would be easy additions to the usual in-class activities used in the course.

In addition to these, presenting a video of a mock exam would also help give students a better idea of what they would be asked to do, at least prior to the first oral exam. This would be along with other strategies that were already implemented in order to help students prepare and understand how to be successful in this oral exam setting.

Our recommendation is to find ways to help students practice oral skills and to have the oral exam process laid out in detail for them.

Future Research Directions

When administering oral exams at scale, it is essentially required that the time commitment per student is shortened as compared to most oral exams studied in the literature. An important question to consider is if shortening the exams, say to a five minute time period, still provides the benefits of a standard oral exam. A previous study on shorter oral exam periods showed a loss in reliability ((Memon, Joughin, and Memon 2010)). Can we find possible ways to mitigate this issue and keep the quality?

Understanding the student perceptions of shortened oral exams is also an open question. Students generally have been shown to have positive perceptions of oral exams (**does this need a citation?**) but does that change through the use of technology or due to the shorter interaction time with their instructor?

Another question is if the recent advancements in generative AI can be leveraged for a similar purpose. For instance, could asking our students to have a probing conversation with AI where they must turn in their 'conversation' be a valuable assessments. In this case, evaluators could assess the transcript to consider the quality of questions asked and the dialogue. As it is now easy to oral converse with generative AI, could a question and answer session in that format serve a similar purpose?

References

Appendix

Student instructions for the oral discussions.

- We want to ask you a few questions about a SAS program (or an R Markdown document) we'll share with you.
- I'll share my screen, ask you to consider particular pieces of code and describe to me what that code does or why we might run it.
- I may ask clarification questions or follow-up questions if you don't fully answer the question.
- If you don't know the answer, that's ok. Just let us know and we'll move to the next item.
- We do have firm time limits on answers to questions. We may have to cut you off so we can get all of the questions in a timely manner.
- Any questions?

Asklund, Ulf, and Lars Bendix. 2003. "Oral Vs. Written Evaluation of Students." *Pedagogisk Inspirationskonferens, Lunds Tekniska Högskola, Sid*, 45–46.

Beccaria, Gavin. 2013. "The Viva Voce as an Authentic Assessment for Clinical Psychology Students." *Australian Journal of Career Development* 22: 139–42.

Carver, Robert, Michelle Everson, John Gabrosek, Nicholas Horton, Robin Lock, Megan Mocko, Allan Rossman, et al. 2016. "Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016."

Huxham, Mark, Fiona Campbell, and Jenny Westwood. 2012. "Oral Versus Written Assessments: A Test of Student Performance and Attitudes." *Assessment & Evaluation in Higher Education* 37: 125–36.

Iannone, Paola, and A Simpson. 2012. "Oral Assessment in Mathematics: Implementation and Outcomes." *Teaching Mathematics and Its Applications: International Journal of the IMA* 31: 179–90.

Joughin, Gordon. 2010. *A Short Guide to Oral Assessment*. Leeds Met Press in association with University of Wollongong.

Kang, Dredge, Sara Goico, Sheena Ghanbari, Kathleen Bennallack, Taciana Pontes, Dylan O'Brien, and Jace Hargis. 2022. "Providing an Oral Examination as an Authentic Assessment in a Large Section, Undergraduate Diversity Class." *International Journal for the Scholarship of Teaching and Learning* 13 (2).

Memon, Muhammed Ashraf, Gordon Rowland Joughin, and Breda Memon. 2010. "Oral Assessment and Postgraduate Medical Examinations: Establishing Conditions for Validity, Reliability and Fairness." *Advances in Health Sciences Education* 15: 277–89.

Newell, Samantha J. 2023. "Employing the Interactive Oral to Mitigate Threats to Academic Integrity from ChatGPT." *Scholarship of Teaching and Learning in Psychology*.

Ohmann, Peter. 2019. "An Assessment of Oral Exams in Introductory Cs." In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 613–19.

Ramella, Daniele. 2019. "Oral Exams: A Deeply Neglected Tool for Formative Assessment in Chemistry." In *Active Learning in General Chemistry: Specific Interventions*, 79–89. ACS Publications.

"Structured Interviews." 2024. https://www.opm.gov/policy-data-oversight/assessment-and-selection/structured-interviews/.

Theobold, Allison S. 2021. "Oral Exams: A More Meaningful Assessment of Students' Understanding." *Journal of Statistics and Data Science Education* 29: 156–59.

Wiggins, Grant. 2019. "The Case for Authentic Assessment." *Practical Assessment, Research, and Evaluation* 2.