Oral-Exam-Paper

Oral Exam Paper

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 a 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 a 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 as the instructors of record for the course 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.

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
- 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 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 a thorough 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 to 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 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.

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 nearly 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, it was branded as an 'oral discussion' to hopefully lessen the anxiety students felt.

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 appropriate responses at the 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 was slated to be taken in a five minute window, two of each type of question was 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. They were instructed to be prepared to discuss code only. 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 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.

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 used in the oral discussion 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 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 of students (roughly 30 students).

Methods

Monitoring Instructor Sentiment and Issues

In order to understand problems that arose, and sentiment around, administering the oral discussions, the course coordinator requested feedback at various points of the process formally and informally.

Change language in image to 'Oral discussions' and 'Graduate student instructor and course coordinator meeting'.

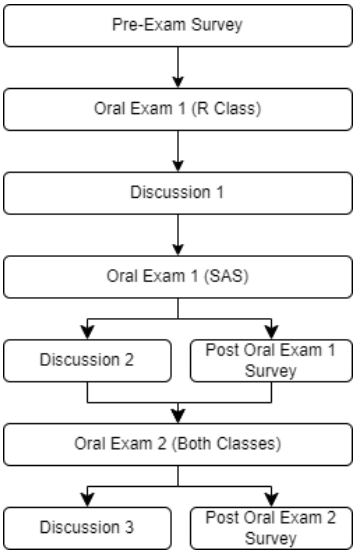


Figure 1: Flow Chart of when exams, discussions, and surveys took place.

Initial thoughts about giving, and experience with, oral discussions were requested on a Google form prior to administering the 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 was 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.

Qualitative Data Analysis: Analyzing Instructor Sentiment and Issues

Most of the data collected was qualitative as it involved open ended questions on forms and informal discussions during meetings. For this type of data, the most common approach used to analyze it involves iteratively going through the data, creating codes, and then grouping pieces of coded data into groups (usually called "themes"). This was the approach taken to analyze our survey and meeting data, formally called thematic analysis (Zhang and Wildemuth 2009).

More specifically, in the first iteration, data was coded based on the use of particular words (such as "time" or "technology") and also on whether or not it came from a survey or a discussion. Pieces of text were then grouped together manually to create five starting themes. In the second and third iterations, these themes were refined after recognizing that they may be better represented by including more sub-themes; for example, instead of discussion of time being a theme, breaking this down into discussions of the quantity and quality of time. Finally, as we passed through the data, we looked over which themes had sufficient support to be included. Themes were included only if they had 5 or more pieces of evidence. **We could either just change this or we could say there is a caveat for one theme.**

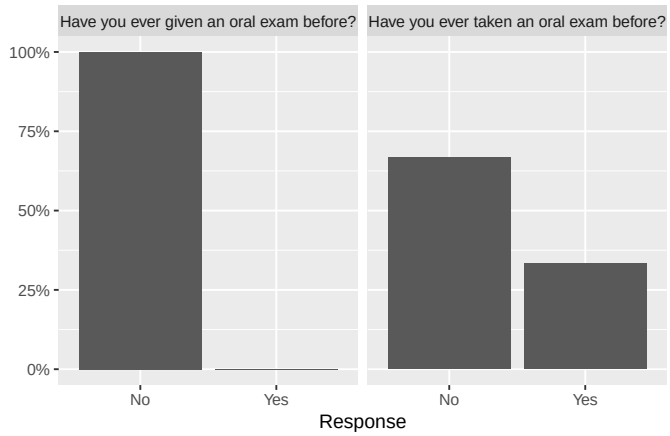
Quantitative Data Analysis

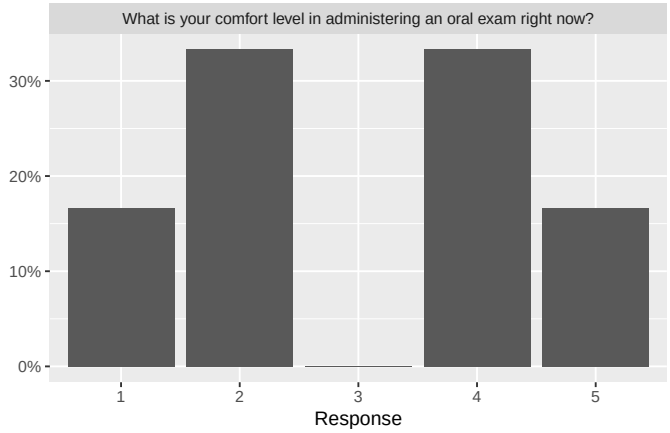
Along with the themes generated from the open ended responses on surveys and whole instructor discussions, closed ended responses from the surveys were also tabulated.

Results

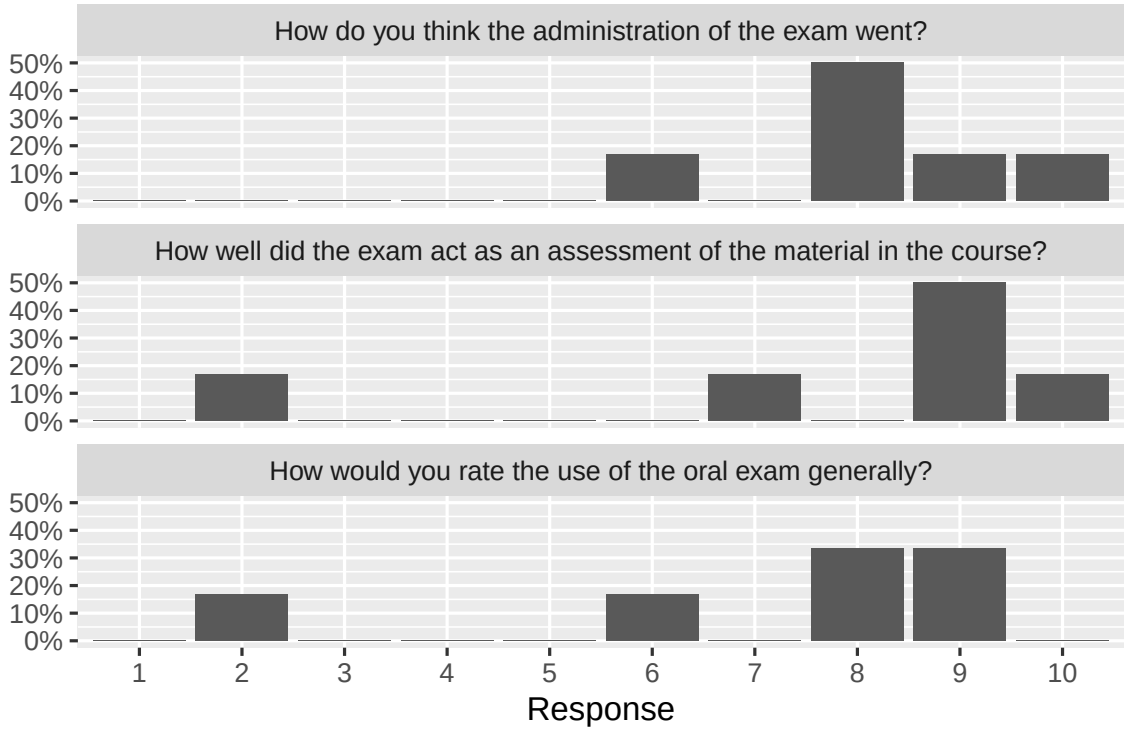
Closed Ended Survey Responses

The closed ended questions asked prior to administering the oral discussions centered around experience with oral exams. None of the GSIs had experience giving an oral exam before and four of the six had never taken an oral exam.





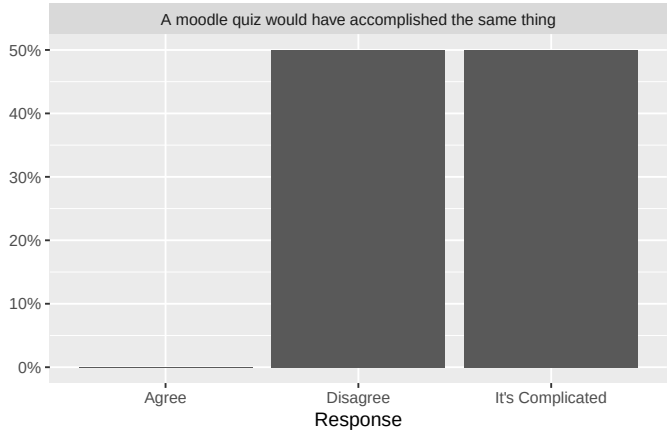
When asked about their comfort administering an oral exam on a scale from 1 (least comfortable) to 5 (most comfortable), the responses varied. Half of the GSIs were on the more comfortable side and half were on the less comfortable side.



Between the first and second oral discussions, the GSIs were asked three questions. The first was how they thought the administration of the oral discussions went on a scale of 1 (very poorly) to 10 (very smoothly). This question was meant to look at how well the scheduling or exams, use of zoom for administering, and other set up related things were viewed. Most of the GSIs viewed the administration favorably. Based on this, it seems the logistics of the oral discussions were well thought out and went pretty smoothly.

The second question asked them how well the discussions acted as an assessment of the material on a scale of 1 (very little utility) to 10 (very high utility). This question meant to investigate the GSIs' value placed on this type of assessment. Here one GSIs rated the exams poorly (two) while one rated a seven and four rated a nine or a ten. This indicates that most felt the assessment was useful compared to their view of other types of assessments.

The third question asked them to rate the oral discussions overall on a scale of 1 (very poor) to 10 (very useful). Again, we see one lower score, one mildly positive score and four higher scores (two eights and two nines). This indicates an overall positive view of the oral discussions by the GSIs after the administration of one of the discussions.



After the second oral exam, the GSIs were asked one closed ended question. Would a quiz accomplish the same thing as the oral discussion? The possible responses were 'Agree', 'Disagree', or 'It's Complicated'. As quizzes were the other timed, in-person, assessment used in the course, this question was meant to see if they thought the oral discussion added value above and beyond the traditional in-class quiz. None of the students agreed that a standard quiz would accomplish the same thing. Half disagreed and half said it's complicated. Open ended questions allowed the GSIs to elaborate on this response.

Open Ended Survey Responses and Instructor Discussions

Theme 1: The graduate student instructors focused on time

Sub-theme 1: They focused on the quantity of time

On the pre-exam survey, half of the GSIs expressed concerns about the "time" or "timeframe" needed to administer the oral discussions not only due to the number of students but also because it had to happen in a week. This concern was confirmed **confirmed... well they have 20 hour a week appointments, most weeks they do much less than 20** after the first oral exam during the first discussion, where some of the GSIs gave estimates of 12 to 18 hours to complete their oral discussions. During the second discussion after the SAS GSIs had also administered their own oral discussions, similar thoughts were voiced with one instructor who joked,"I was losing my sanity being on Zoom for so long", and another instructor on the second survey feeling that it "took way too long".

Unfortunately, after the second oral exam was given, these concerns only intensified. Though only one GSI commented that they would not "do this to themselves" again with a large class, but perhaps a smaller one, during the final discussion, almost every GSI had something negative to say about the time commitment in the final survey. "Overall, very time consuming" one said. Words such as "grueling", that administering the discussions was something they

had to "endure", that it was "draining" and "extremely time inefficient", were also used to express their feelings towards the time commitment.

Two reasons for focus on the time commitment were given by two different GSIs. One commented that the time commitment to give oral exams was bleeding into other classes and exams they had towards the end of the semester. The other remarked that they felt they were promised less work as a first-year graduate student, but that this was not the case in this position."If the oral exam format goes unchanged", they said, "I am unsure if I am fit to be an instructor next semester."

This theme needs to be counter balanced I think. Of course they are going to say it took too long and other things but nothing was outside their scope of time commitment - just not what they want to be doing.

Sub-theme: They focused on what they were doing during that time

Though instructors had an inkling during the pre-exam survey that what they would be doing would be time consuming, the mentions of what they were doing qualitatively during the oral exams began in the first discussion. For example, several GSIs joked about not always being able to use the bathroom, since you have to be "100% focused". However, part of the issue was not following previous recommendations to leave extra time in their schedules, and later on in the discussion they discussed how to do that. This appeared to help, since no mention of being focused or other issues related to what they were actually doing during that time did not appear in the second discussion or first post exam survey.

However, during the final discussion, one conversation centered around the qualitative difference between grading HW and administering the exams. One GSI commented that, for them, the exams took longer, and, unlike with homework grading, for the exams you had to "block out time". Another disagreed, saying the time commitment felt similar, but later, acknowledging that it could take similar time.**This doesn't make sense... they said it was similar already** One final GSI commented that grading homework was more "flexible" for them. **So I think this may be the theme for both of these first ones. Flexibility of time to do the work.**

Though five of the six GSIs did not discuss any concerns with how the time was being used or differences between the oral exams and how time is used for other things, one GSI did express that even with well spaced time blocks, they were still unable to eat a snack, use the restroom, or walk around. **Is this worth mentioning?**

Sub-theme: They focused on when things went wrong with timing

During the pre-exam survey only one GSI commented that they worried about "the student's scheduled slots being too spread out". Issues about student timing really began to come out during the first discussion. One talked about the fragility of the online set up **but looking at the numeric responses, they said it went smoothly generally - perhaps we are letting one TA push a larger narrative on some of these themes? Either way, we should be countering it with the note on the numerical data**, where if even one student

comes late, it can "ruin the timing for all of the students". Another mentioned how students would immediately sign in when there was an open spot without letting the instructors know. However, the issue of student scheduling did not appear again in the second discussion, with only one question in the first follow up survey about "how to make everyone register and show up on time".

After the last oral exam, though, much of the final discussion time was taken up by the issue again. Some GSIs talked about having to reschedule with students who missed, either to Saturday or to the next week, and how to avoid it. One instructor mentioned doing this and "regretting" it. Even more telling, on the final survey taken at the beginning of the discussion, an overwhelming amount of the advice the GSIs gave to future instructors in their shoes was about optimal time management due to student scheduling issues. For example, graduate student instructors offered comments such as "If possible, avoid scheduling any time on Friday" so as to plan for students who miss their time slots, "try to make everyone schedule and show up on time", and "build in more than a 5 minute break every 30 minutes" or else, due to students not arriving to the meeting on time, "you'll need to use [it]".

Theme 2: The graduate student instructors focused on bias and equity

Even before administering exams, GSIs worried worried about their role not just as administrators but as graders. One worried that the grading might be "too subjective". Another worried about their "poor memory" and bias that could come from familiarity, and a final GSI worried about the flexibility of being the grader may be too much.

In the first discussion, there was some clarification on how to ask and grade one question, but it wasn't until after both groups had administered exams that they began to share more fully their thoughts about equity and bias. Part of their focus was on the exam itself. Several GSIs wanted a larger grading scale in order to more fairly grade students. Another GSI wondered if the questions were too broad, which prompted a conversation about anticipating student answers. Later in the discussion, it was found that different instructors had been phrasing certain questions differently, which created some different student responses.

Another part of their focus was on accommodating and being fair to all students. For example, the course coordinator and one GSI discussed appropriate ways to help ESL students. Another GSI had given an oral exam in person to accommodate a student who had gotten sick. A worry expressed in the first follow up survey was the confounding of the oral exam with being nervous or with speaking ability.

Some of these same concerns and questions came up during the final discussion and second follow up survey, including making the test less biased and being fair to all students. One GSI brought up a question they thought was unfair. Another wondered about the length of the oral discussions (which were 5 minutes each) and whether that was enough time to rate everyone well. In the follow up survey, one GSI offered a recommendation to eat snacks to

avoid biased grading from hunger. Another GSI suggested more time for ESL students, since "this is not an English test".

Theme 3: The graduate student instructors wondered about the efficacy of the oral exam

Though no views about the efficacy of the oral exam were expressed prior to the first one or during the first or second discussions, anonymously during the first follow up survey a few GSIs expressed doubt over the potential benefits of oral exams. "Not sure what this accomplishes over HW/moodle quiz" one said. Another voiced the same feelings, writing, "[a] paper quiz could serve the same goal". Noticing this, the GSIs were asked about their feelings explicitly in the second follow up survey.

In the survey, they clarified that there were some benefits to the oral exam. Words like "integrity" and the concept of open vs closed book were used, alluding to how hard it is to plagiarize or cheat with an oral assessment. Another pointed out that it tested whether someone could produce "on the fly knowledge". This mechanism, however, was noted as being similar to, and essentially accomplished by, a well crafted free response quiz. Though half of the GSIs shared this view, two did acknowledge the advantage of being able to ask follow up questions, and it leading to better grading potentially being more lenient than written exams, since "you can understand what students are trying to say better".

The last survey question asked pretty explicitly if they thought a quiz would do the same thing and none said yes.

I don't know if we should keep this theme 4, or incorporate it else where. I didn't want everything to be all doom and gloom, but there really wasn't much in the way of positive things....it was more like they enjoyed time with their students, but they barely brought it up, just three times, which doesn't meet our criteria of 5 pieces of evidence. What do you think? #### Theme 4: The graduate student instructors had few positive things to say

When asked directly about positives, half of the graduate student instructors responded. All of the comments were centered around having increased student interactions where, in such a large class, they usually didn't. One graduate student instructor commented that, unprovoked, a few students talked about liking the format.

I didn't go through the verbal recordings but did look over the post oral exam form responses. These were the positive items of note:

- post oral exam 2
  - integrity improved over moodle quiz
  - better than quiz at evaluating how well students can produce knowledge "on command"
  - ability to keep closed book (integrity again)

  - More follow-up things, easier to know how deep students understand questions
  - Students who clearly knew the material were able to showcase this understanding, while highlighting areas in which others need to improve.
  - Ability to catch students that haven't been doing their work
  - Some students independently told me they really like the oral exam format.
  - Noticeably more confident, comfortable on the second oral exam. (noted twice)
  - Enjoyed talking, feel more comfortable with students
  - Students getting a little more comfortable speaking in programming lingo.
  - Very fast and effective way to see who knows what's going on in the class and who isn't with us at all.

- post oral exam 1

  - I think it made clear who had a good idea of what they've gotten well
  - I had fun **Seems to be that there are more positives to talk and bring into these themes.**

Discussion

To our knowledge, this is the first large scale oral exam study done in the statistics and data science education literature and one of the largest oral exam studies done in the oral exam literature as a whole. Further, due to the pivital role graduate student instructors play in the success of the oral exams, this study focused on their attitudes and experiences as they administered them.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, wonderings about their efficacy, and positive student reactions.This left us with three lingering questions: why did our GSIs react the way they did, what do our findings mean for designing large scale oral examinations, and what future research directions could we pursue?

Why did the graduate student instructors react the way they did?

Time

GSIs probably had such an intense focus on time because they were tyring to surivive the process. This reminded us of Maslow's Hierarchy of needs (Maslow something), a framework detailing the need of people to have satisifed basic, foundational needs (like food and shelter) before being able to satisfy higher level needs (sense of belonging, becoming a better person). In the same way, we hypothesize our GSIs did not talk much about the experience of administering the exam, the wider pros and cons, and how students felt about them because they themselves were focused on fulfilling the basic need of surviving the process (doing 12-18 hours of exams) while also having other coursework and responsibilities on top of that. If it had been a less stressful or time consuming experince, they would have had the cognitive energy and capacity

to reflect on and improve on their skills, focus on the positives, bias, reliabilty, validity, grading better?…maybe not and focus on the student experience versus just trying to figure out how to administer them and work with the logistical issues.

Expectations between what they thought they were going to do and what they actually did, what was fair to them verus what was actually spelled out in the contract.

The efficacy of the oral exam

Questioning whether oral exams are worth it is not a new phenomena. In the literature (*find citation*), this instructor also noticed that though their students saw oral exams as good, based on previous socialization in the school system they saw closed book written exams as the gold standard. This may have lead some GSIs to wonder whether….

Wonderings were also caused by the time commitment, since a well designed moodle quiz would certainly take less time, maybe more flexible grading time, no need to be there giving the exam with the students.

We also hypothesize that part the feelings towards the efficacy of the oral exams may come from the novice instructors discomfort with assessing students generally, and especially assessing them on the fly in real time. This is based on the experience of the course coordinator, who found that, especially in the second round of exams when students were more comfortable with the format, the short meetings were an excellent and refreshing way to assess students knowledge. For GSIs, the lower pressure grading of quizzes and homework assignments give more time to be sure of answers given and deem what is an appropriate score should be.

Test bias and equity issues

I'm not sure about this one, maybe you have some ideas Justin? Why would they be so focused on fairness and accommodations, as well as trying to create a more fair test and reduce their own bias as they gave the exams?

Trying to explain why they reacted the way they did will help us give recommendations for how to design future large scale oral exams ### What do our findings mean for designing large scale oral examinations?

Instructors are valuable sources of feedback. They are there in the trenches, they had wonderful comments, though at times biting, highlighting structural issues with the oral exam set up and format as it stood. They were willing to offer advice to others and among each other, and work with the coordinator as well. They also had keen insights and cared about fairness for students.

You can see how they become more critical over time, and that also the last discussion and the final comments are all about time and timing and scheduling, while the second discussion is

a lot more about getting the test better, and grading, and making sure things are fair, things we'd want the GSIs to focus on.

Very negative feelings towards the time commitment, not the format itself, event though they were contracted to work 20 hours. The type of work was particularly hard for them, maybe not used to it, expecting it, etc. Regardless, it would be impossible to continue oral exams in this way. Think about how peer to peer knowledge of this would spread, those who administer may grow to hate it, maybe even biasing things, or missing why they are given. Maybe it doesn't matter what the benefit is, the cost of time is too much.

Thinking of this like a GSI hierarchy of needs, we'd have to make sure that logistics are in place well, that the time commitment is either not too large or that there are appropriate expectations in place, and that effective time management is implemented and monitored.

We also will probably need some training on how to do these things well, how to give exams well, and perhaps some modeling of what it can look like.

On a positive note, these graduate student instructors were aware of and sensitive to potential biases and equity issues, and had an excellent discussion of ways to attempt to mitigate these problems.

Overall, we would suggest that under realistic circumstances, administering oral exams at larger scales may greatly weaken what makes oral exams important/interesting/powerful in the first place. With the necessary shortening of their administration comes a loss in reliability ((Memon, Joughin, and Memon 2010)). Possible to find ways around this?

What future research directions could we pursue?

Future research questions include I think these center on looking at students, but also more so how we can get the same quality that small scale exams have at large scale.

Overall, we would suggest that under realistic circumstances, administering oral exams at larger scales may greatly weaken what makes oral exams important/interesting/powerful in the first place. With the necessary shortening of their administration comes a loss in reliability ((Memon, Joughin, and Memon 2010)). Possible to find ways around this?

- AI enhanced oral assessment?
- Student perceptions positive in literature, but how does doing it for 5 minutes in zoom change things? Having to compare and contrast experiences
- **Others?**

Conclusion

Matthew and Justin will work on this part after discussion is complete.

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.

Zhang, Yan, and Barbara M Wildemuth. 2009. "Qualitative Analysis of Content." *Applications of Social Research Methods to Questions in Information and Library Science* 308 (319): 1–12.