Midterm oral exams add value as a predictor of final written exam performance in engineering classes: A multiple regression analysis

Abstract

What is gained when midterm oral exams are implemented in the undergraduate engineering classroom? This research paper examines whether midterm oral exam scores add value above and beyond midterm written exam scores in predicting students' final written exam scores. The purpose of this study is to evaluate the potential utility of oral exams as formative assessments: if oral exam scores provide additional information beyond written exam scores, they may add meaningful value for students and instructors. The current study investigates this question using data from 10 undergraduate engineering classes (N = 925), representing 6 different courses and 5 different instructors. Though course and exam context differed, all classes implemented a low-stakes midterm oral exam, a midterm written exam, and a midterm final exam. We compared two multiple regression models: a smaller model with only the midterm written exam score as a predictor for the final written exam score, and a full model with both the midterm written exam score and the midterm oral exam score as predictors for the final written exam score. We found that the fuller model with oral exam score was a better fit for our data, indicating that including oral exams explains more variance in students' final exam performance than midterm written exams alone. Further analyses tentatively indicate that the granularity of the rubric used to score oral exams matters, with finer-grained rubrics more consistently providing predictive value. This study has implications for developing a theory of oral exams, as it leaves room for the possibility that oral exams tap deeper learning processes than written exams. These results show that oral exams provide actionable information instructors can use to make interventions to foster students' meaningful learning before the end of the term. The quantitative analysis also provides instructors with a simple statistical measure to assess the role of oral exams in student's learning.

Keywords: oral exams, instructional design, assessment, multiple regression, quantitative analysis

Introduction

Exams taken by students in college classes serve the goals of both formative assessment (i.e., midterm feedback that students and instructors can use to shape their ongoing learning and teaching strategies, respectively) and summative assessment (i.e., end-of-term evaluations of students' achievement of course learning goals). While written exams are standard in engineering classes, oral exams have the potential to provide additional information about a students' current level of learning. The current paper uses multiple regression to investigate the diagnostic value of

adding low-stakes midterm oral exams to traditional written exams across multiple undergraduate engineering courses. Put simply, we will investigate whether low-stakes midterm oral exam scores predict final written exam scores above and beyond midterm written exam scores.

Assessment quality and learning

The evaluation of student learning can serve functions of making judgments about students' achievements (i.e., summative assessment) and improving students' learning (i.e., formative assessment [1]). Formative assessments provide useful information that instructors and students can use to regulate their teaching and learning strategies. For example, a college student can use the feedback received from a midterm exam to make judgments about the effectiveness of their learning strategies and help identify content areas for improvement. Two important variables in the value of a formative assessment are the extent to which the assessment corresponds to the course learning objectives and the utility of the information the assessment provides for the regulation of teaching and learning.

The match between the type and content of the questions and the course learning goals is paramount for the utility of an assessment, whether its application is formative or summative. Learning that leads to rote memorization and learning that leads to understanding and comprehension involve distinct cognitive processes [2], with conceptual understanding requiring mentally representing the abstract relationships between concepts [3]. The type of test that students prepare for and take, such as a detail-based or comprehension-based test, affects their learning outcomes [4], [5]. Measurement theories also suggest that additional measures of the same construct help validate the scientific relationship between the construct and the measures. [6]. The nature of the engineering major, with the building of knowledge across a structured curriculum and the disciplinary focus on problem solving, means that focusing on processes that will lead to flexible and long-lasting learning outcomes is especially critical. Formative assessments must provide information that students can effectively use to regulate their learning. This information can come from the exam experience itself and from information and feedback provided during and after the exam. For example, much of the foundational research on the testing effect shows that retrieval practice alone can enhance memory [7], and this effect has been extended to problem-solving transfer tests [8]. Though such studies show testing can be a useful learning event even in the absence of feedback, the effectiveness of testing for learning is enhanced when feedback is given. Minimal feedback has learning benefits compared to no feedback [9], but care for the quality of feedback can help optimize its benefits [10], [11].

Therefore, the quality of a formative assessment is key if it is to provide accurate and actionable information for instructors and learners. The following section will discuss the relationship between that quality and exam modality.

Affordances and effects of different exam modalities

Like instructional modality, different exam modalities have different affordances and drawbacks: the ease of grading multiple-choice exams makes them infinitely scalable, but can be limiting when the goal is to probe deep knowledge; written exams give space for students to demonstrate their thinking process, but make delivering prompt feedback challenging; computer-based exams can adapt to student performance, but with limited data this adaptation can be miscalibrated. Oral exams add social context, allowing the examiner to probe shallow answers or precisely titrate

guidance to the level needed by the examinee. An ostensible drawback of oral exams is scalability, as they require individual attention and scheduling, however with appropriate resources this problem can be surmountable (e.g., the current analysis includes classes implementing oral exams with enrollment from 24 students to over 180[1]).

Similar to the effect of test type (e.g., detail-based vs. concept-based) described in the previous section, the student experience of exam modality can also vary along metacognitive lines. For instance, lab-based research shows that students learn more deeply when studying in preparation to teach content to another student than they do when preparing for a written exam [13], [14], and that explaining on video can be a more effective review technique than writing explanations or restudying [15]. Other research shows that studying in preparation for a high-stakes video-based exam can cause students to strategize their learning in ways that overcome poor instructional design, compared to a low-stakes lab-based assessment [16].

Oral exams and traditional assessments in the classroom

The affordances of oral exams create an avenue for assessing student learning in a meaningful, dynamic way. Implementing oral exams also responds to calls to introduce multiple modes of measurement [17]. While researchers have increasingly investigated the potential role of oral exams in the classroom, this research base remains relatively limited – perhaps due to the rarity of implementing such exams at all [18]. Of the existing research, much investigates students' experience of oral exams through surveys; relatively less investigates the relationship between oral exam and traditional exam scores.

A research approach relevant to the current investigation is to quantitatively analyze the relationship between scores on oral exams and more traditional assessments. For example, some authors [19] report similar distributions of scores across oral and written exams, while others [20], [18] report significantly better performance on oral than written exams. Ramella [17] likewise describes teacher reports of higher scores on oral exam performance, crediting the opportunity to self-correct and think out loud as possible explanations. Other researchers have taken an approach more similar to the current investigation by using correlation coefficients to demonstrate that oral exam scores correlate, but not perfectly so, with written exam scores [21], [22]. While such an analysis is informative, we believe that using further statistical modeling will help give a clearer answer about the value of oral exams. Specifically, we investigate whether adding midterm oral exam scores explains additional variance in final written exam scores above and beyond what is explained by midterm written exam scores. To ask such a question requires the statistical analysis of multiple regression, which will be explained in the following section.

Brief introduction to multiple regression

What follows is a brief introduction to the multiple regression technique used in this paper. This is intended to be a conceptual introduction for readers without a formal background in statistics.

A familiar technique for investigating whether two variables tend to vary together is correlation. A correlation tells you both the direction and strength of a linear relationship between two

[1]The oral exam method will be briefly described in the Method section of this paper, but see [12] for more information on how these exams have been implemented.

variables. Regression is a related technique that goes beyond merely summarizing the relationship between variables to evaluate the strength of the predictive value of a variable or combination of variables over an outcome. A benefit of the regression approach is that we can model how more than one predictor variable predicts an outcome variable. This is called multiple regression.

The results of a multiple regression analysis show how much of the variance in an outcome variable is attributable to the variance in one or more predictor variables. To give an example, imagine you work in college admissions and are trying to decide whether to require students to provide SAT scores on their application. You already have information about students' potential for college success from their high school GPA and their application essay, which is scored on a rubric. You want to know whether SAT scores provide any additional value in predicting students' college success above and beyond what's already provided by students' high school GPA and application essay. In other words, it is possible that the information provided by SATs is essentially redundant with the information already provided by these other measurements? You can test this question using multiple regression in three steps: First, build a model with high school GPA and essay rubric scores as predictor variables and college GPA as an outcome variable. The outcome of this analysis will tell you how well these predictors explain the outcome variable. Second, build a larger model, now including SAT scores alongside high school GPA and essay scores as predictors of college GPA. Third, statistically compare these two models. If the larger model does do a better job of predicting college GPA than the first model, that means that SAT score is explaining additional variance in college GPA, *above and beyond* the other predictors. If the larger model does not do a better job of predicting college GPA, then SAT scores are not adding any information.

The example above demonstrates a specific application of multiple regression called *nested model comparison*. The approach is *model comparison* because we are comparing the strength of two models in how well they explain the outcome variable, and it is *nested* in that the smaller model consists of a subset of the predictors in the larger model. If model A and model B look at the same outcome variable, but model A has one predictor variable and model B has two predictor variable where one overlaps with the one from Model A, by conducting nested model comparisons of models A and B, we can answer whether adding the extra predictor variable in model B better fits the dataset than model A.

Current study and predictions

The current study analyzes student performance across 925 undergraduate students in six engineering courses. In traditional exam structures, instructors benefit from looking at midterm written exam scores to guide the latter half of their course, with the goal of seeing a higher performance in the final written exam. In this study, we aim to test the diagnostic value of adding a low-stakes midterm oral exam to the traditional exam structure. That is, we examine whether adding a midterm oral exam to a class with the traditional exam structure (a midterm written exam and a final written exam) help the instructors better diagnose their students' performance halfway through their course. This is done as one way of assessing oral exams' value as a formative assessment: If midterm oral exam scores provide additional diagnostic information in predicting final exam performance above and beyond midterm written scores, then it may be worthwhile for instructors to expend the effort to implement oral exams in their classes. We analyze this by comparing two models: a smaller model with just the traditional written exam

scores (where the midterm exam score predicts the final exam score), and a bigger model where the midterm oral exam score serves as an additional predictor for the final exam score.

As the second goal of this paper, we break down our data at the course level to explore what format of oral exam scoring shows the highest predictive value toward final written exam scores. Our dataset includes multiple instructors implementing several methods of conducting and scoring oral exams. We explore the granularity of the oral exam rubric (e.g., 3pt scale, 5pt scale, scores summed across single vs. multiple questions) and examine whether it affects the reliability of oral exam scores as an additional predictor for final written exam scores. Our observations will help generate preliminary guidelines for developing grading rubrics for oral exams.

Methods

Participants

Participants were 925 students enrolled across 10 total offerings of 6 different courses in two engineering departments (Mechanical and Aerospace Engineering, Electrical Engineering) of a large research university. The dataset was gathered as part of an initiative among engineering faculty to conduct oral exams, and includes all courses that implemented oral exams as part of this initiative from Winter Quarter 2021 to Fall Quarter 2021 (see Criteria for Our Dataset). Table 1 lists these classes, and demonstrates the heterogeneity in course content and exam structure across courses. Three exam scores collected from participants were used for our analysis: the midterm written exam score, the midterm oral exam score, and the final written exam score. An additional 28 participants who missed at least one of these target exams were dropped from this dataset.

Design

The outcome variable of interest is the final written exam score of students in the courses. There are two predictor variables of interest: the midterm written exam score and the midterm oral exam score. The midterm written exam preceded the midterm oral exam in all included courses. All students in the final dataset had taken all three exams. Random variables that were taken into account were the instructor and the session, where the session was defined as a combination of both the type of the course (e.g. ECE, lower-level) and the term of the course (e.g. spring quarter, summer session).

Criteria for our dataset

Continuous scale for oral exams: As part of an initiative on campus, instructors implemented low-stakes midterm oral exams as an assessment in their respective courses. Each instructor had the flexibility to choose the format of the midterm oral exam based on their course needs and pedagogical goals. Therefore, courses differed in whether the midterm oral exams were graded and, if they were graded, the nature of the scales (e.g. on a binary or continuous scale, or as a participation credit). For our analysis, we only include courses with midterm oral exams that were graded on a continuous scale with at least 3 reference points to capture the variance in student performance in the oral exams.

Context of oral exam: All midterm oral exams took place remotely on the Zoom video

conferencing interface, with one assessor and one student attending the session for a 10 to 25 minute time window. Students were tested on questions that they were either previously tested on in a midterm written exam (3 courses) or a take-home exam (1 course), were asked to answer some questions on the MATLAB codes that they had submitted as a homework assignment (1 course), or were asked to solve novel questions that they have not seen before in written assessments (1 course).

Selection of exam scores: Instructors implemented varying exam schedules, including different numbers of midterm written and oral exams throughout the term (see Table 1). Additionally, the timing of midterm oral exams sometimes differed within a class. To address this variability and account for the novelty of students' experiences with oral exams, we only analyzed students' performance on the very first oral exam score of the course. For the midterm written exam scores, we also selected the very first written exam of the course that preceded the midterm oral exam. This was done to reduce confounding factors that affected the relationship between midterm written exam and the final written exam, so that any additional variance explained outside the traditional written exam structure could be attributed to the midterm oral exam. Therefore, the three exam scores used as the key variables of this dataset were the very first written exam in this course, the very first midterm oral exam that happened after the midterm written exam, and the final written exam that happened at the end of the term.

Data Analysis

Standardizing the dataset: To account for differences in in raw exam scores and sample size across multiple courses in our dataset, standardized z-scores were used for analysis.

Analysis plan: We analyzed the diagnostic value of midterm oral exam scores on top of traditional written exam scores (midterm, final) by examining two multiple regression models: (1) a small model that reflects traditional grading practices where the midterm written exam score predicts the final written exam score

$$FinalWrittenExam \sim MidtermWrittenExam + (1|Instructor) + (1|Session) \quad (1)$$

and (2) a bigger model where a midterm oral exam is added as another predictor along with the midterm written exam to predict students' performance at the final written exam

$$FinalWrittenExam \sim MidtermWrittenExam + MidtermOralExam + (1|Instructor) + (1|Session) \quad (2)$$

If the bigger model is a better fit for student performance data compared to the smaller model, the nested model comparison will support the idea that midterm oral exam scores explain more variance in students' improvement in the scores from the midterm written exams to final written exams.

Results

We constructed two linear regression models predicting students' performance on the final written exam, including a smaller model that reflects the traditional class structure with just written exam

Instructor; Dept; Course; Level	**Course Exam context (All oral exam sessions on Zoom)**	**Sample size**	**Rubric Format**
Instructor A: ECE; Components and Circuits Laboratory; Lower-division	1 midterm written exam, 2 take-home (TH) tests, 1 oral exam, and 1 final written exam. For the oral exam, half the students were tested on materials from written TH test 1 (and did a peer-review sessions on TH test 2); the other half did the reverse (peer-review for TH test 1; oral exam for TH test 2).	69	3pt scale based on judgment of students' understanding; oral exam was used mostly as an extra credit activity
Instructor B: ECE; Introduction to Analog Design; Lower-division	4 written quizzes (first written quiz counted as midterm written exam in this dataset), 1 midterm oral exam, 1 final written exam. Students answered TA's prompts on the written quiz materials during the oral exam. Half the students were tested on Quiz 1 material, and the other half were tested on Quiz 2 materials for the oral exams.	233 total (FA21A: 122; FA21B: 111)	3pt quality-based scale; oral exam was used mostly as an extra-credit activity. Students were granted participation credit, but the instructor and TAs graded their performance for internal purpose.
Instructor C: MAE; MATLAB Programming for Engineering Analysis; Lower-division	There were 2 midterm written exams, 1 midterm oral exam, and 1 final written exam. During the oral exam, students were asked about the MATLAB codes they had submitted as homework.	302 total (SP21: 122; WI21: 180)	5pt scale based on judgment of students' understanding; intended to be a low-stakes conversation to promote interaction and learning.
Instructor D: ECE; LabVIEW Programming: Design and Applications; Upper-division	There were 2 midterm written and 2 midterm oral exams, with 1 final written exam. The oral exam had a new question (not related to written exams), and students were asked to write a program that satisfies several design requirements. Hints were provided only if needed.	55 total (FA21: 24; SP21: 31)	Subtractive rubric starting from 10 full points; content-based; subtracted when TAs had to provide hints.
Instructor E: Course E1: MAE; Statics and Introduction to Dynamics; Lower-division Course E2: MAE; Solid Mechanics; Upper-division	During a regular academic quarter: 2 midterm written and 2 midterm oral exams, with 1 final written exam and 1 final oral exam. During a summer session: 1 midterm written exam, 1 midterm oral exam, 1 final written exam and 1 final oral exam. The oral exams were follow-ups on written exam questions. Students were tested on one of the 3 questions from the written exam, and answered about 3 prompts on that question.	E1: 125 total (WI21: 95; SS21: 30) E2: 141 total (SS21: 20; SP21: 121)	A summed score across three prompts for one exam question, where each prompt was graded on a 6pt scale with 3 checkpoints (0,1,3,5 pt), with intermediate scores awarded; based on both content and delivery.

Table 1: List of courses with oral exam structure and grading rubric. Courses are arranged by the granularity of the grading rubric (coarse to fine-grained). Department codes are ECE: Electrical Engineering; MAE: Mechanical and Aerospace Engineering. Term codes are FA21: Fall 2021 (10-week); SP21: Spring 2021 (10-week), WI21: Winter 2021 (10-week), SS21: Summer 2021 (5-week). A and B after the quarter label refers to the sections within a course.

scores (e.g., one midterm written exam and one final written exam), and a bigger model that adds one midterm oral exam on top of the two written exams. We hypothesized that the midterm oral exam scores explain additional variance in students' final written exam scores above and beyond the midterm written exam scores, in which case we will observe our bigger model with the midterm oral exam to be a better fit for our data (vs. the smaller model without the midterm oral exam). For all models, we have included the grouping variables as random effects, such as the instructor for the course (Instructor A,B,C,D,E) and the course session (a combination of the type of the course and the term of the course).

Analysis of the full dataset

To ask whether adding a midterm oral exam to a traditional exam structure provides additional explanatory value of the final written exam above and beyond the midterm written exam alone, we performed a nested model comparison on the full dataset by comparing two models: a bigger model with the oral exam as a predictor along with the midterm written exam (Model 1); versus a smaller model without the midterm oral exam as a predictor (Model 2, where only the midterm written exam is a predictor).[2] With the full data (N = 925 students), nested model comparisons of linear regressions showed that Model 1 with the midterm oral exam included as an additional predictor was a better fit for our data than Model 2 without the midterm oral exam as a predictor ($\chi^2(1) = 13.18, p < 0.001$). This nested model comparison result indicates that adding the midterm oral exam to the traditional written exam structure added explanatory value in predicting the final written exam performance of students.

Analysis by course: Rubric granularity

As a follow-up analysis, we conducted nested model comparisons at the course level to explore whether the diagnostic value of midterm oral exams differed by the granularity of their grading rubric. We grouped the courses across different terms that were taught by the same instructor and therefore used the same rubric for oral exams. We observed five different levels of granularity among the instructors: 3pt scale used by instructor A that was primarily used to give full credit to students (one session); 3pt scale used by instructor B on students' general performance during the oral exam (one session with two within-course sections); 5pt scale used by instructor C on students' general performance (two sessions); 10pt scale based on subtracting points from full point used by instructor D (two sessions); and an 18pt scale summed across three prompts that were graded on 6pt scale each, used by Instructor E (two courses, two sessions each). These rubrics visibly differed in the distribution of student scores (see Figure 1).

Overall, we confirmed that in 4 out of 6 of our courses, midterm oral exam additionally explained

[2] A note on the interaction term: We first tested a model that covers all possible predictors, including the midterm written exam score, the midterm oral exam score, and the interaction between the two predictors (midterm written exam x midterm oral exam). In this model, the main effects of both predictors on the final written exam were significant (midterm written exam score, $t(921) = 17.95, p < 0.0001$; midterm oral exam score, $t(921) = 4.19, p < 0.0001$). The interaction of the two predictors in predicting the final written exam score was also significant ($t(921) = 3.60, p < 0.001$). This model with the interaction term was a better fit for the data compared to the model without the interaction term (Model 1 above), $\chi^2(1) = 12.92, p < 0.001$. However, since we did not have specific predictions on interpreting the interaction term and were mainly interested in the main effect of midterm oral exams, we will be using Model 1 without the interaction term as the full model for nested model comparisons, to be compared with the model without the midterm oral exam as a predictor (Model 2).

the variance in students' final written exam performance above and beyond the midterm written exam (see Table 2): in Instructor B's lower-division ECE course that used a 3pt scale ($N = 233, \chi^2(1) = 9.25, p = 0.0023$); in Instructor D's upper-division ECE course that used a subtractive rubric from a 10pt question ($N = 55, \chi^2(1) = 7.54, p = 0.0060$); and the two MAE courses taught by Instructor E that used 6pt scale for each of the three prompts, in both the lower-division course ($N = 125, \chi^2(1) = 4.14, p = 0.0419$) and the upper-division course ($N = 141, \chi^2(1) = 5.69, p = 0.0171$) (see Figure 2). Therefore, the model including midterm oral exams was consistently a better fit for the data among the finer-grained rubrics in our sample (i.e., the 10pt and 18pt scales). Additionally, Instructor B's 3pt scale used for internal purposes significantly improved model fit for that class, despite the lower granularity of the rubric.

Inst.	**Granularity of rubric**	**Is the bigger model a better fit for our data than the smaller model?**	**Sample size**
A	3pt scale (mainly EC, 1 course)	No ($F(1) < 1, p = 0.95$), no random effect	69
B	3pt scale (1+ Q, 1 course)	Yes ($\chi^2(1) = 9.25, p = 0.0023$)	233
C	5pt scale (1+ Q, 2 sessions)	No ($\chi^2(1) = 0.20, p = 0.6582$)	302
D	10pt (subtractive, 1Q, 2 sessions)	Yes ($\chi^2(1) = 7.54, p = 0.0060$)	55
E1	6pt scale x 3 prompts (2 sessions)	Yes ($\chi^2(1) = 4.14, p = 0.0419$)	125
E2	6pt scale x 3 prompts (2 sessions)	Yes ($\chi^2(1) = 5.69, p = 0.0171$)	141

Table 2: Results of nested model comparison by course.

Discussion

Using data from six engineering courses (10 total offerings) that used varied implementations of midterm oral exams, we analyzed whether midterm oral exam scores help predict final exam scores above and beyond midterm written exam scores. We found that a model that included midterm oral exam score and midterm written exam score as predictors of final written exam score was a significantly better fit for the data than a model that included only midterm written exam score as a predictor. This result supports the idea that midterm oral exam scores provide additional predictive value above and beyond midterm written exam scores.

We further broke down our data by rubric type, in order to investigate whether rubric granularity affects the predictive value of oral exams. This analysis was conducted due to the heterogeneity of rubric type in our dataset, ranging from 3-point rubrics based on overall performance to 10-point-per-question rubrics. We predicted that more fine-grained rubrics would provide additional information that increases the predictive value of oral exam scores. We found tentative support for this prediction, with the most fine-grained rubrics consistently improving the predictive value of the model, and less-consistent improvement among the less fine-grained rubrics. However, one of the three low-granularity rubrics did significantly improve model fit. Additional work is needed to investigate the oral exam grading rubrics qualities that optimize the usefulness of these exams.

This work has implications for building a theory of oral exams. One argument for the inclusion of oral exams is that they have the potential to tap deeper levels of learning than written exams. Though this analysis does not directly access the quality of students' learning across exam type,

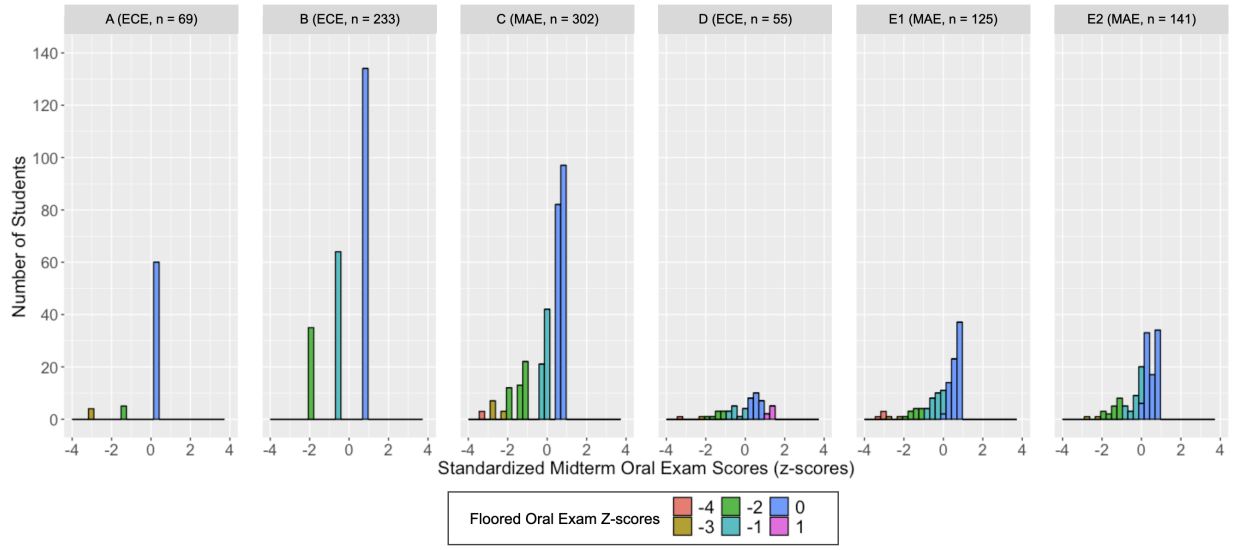

Figure 1: Histogram of midterm oral exam scores by each course (merged across sessions)

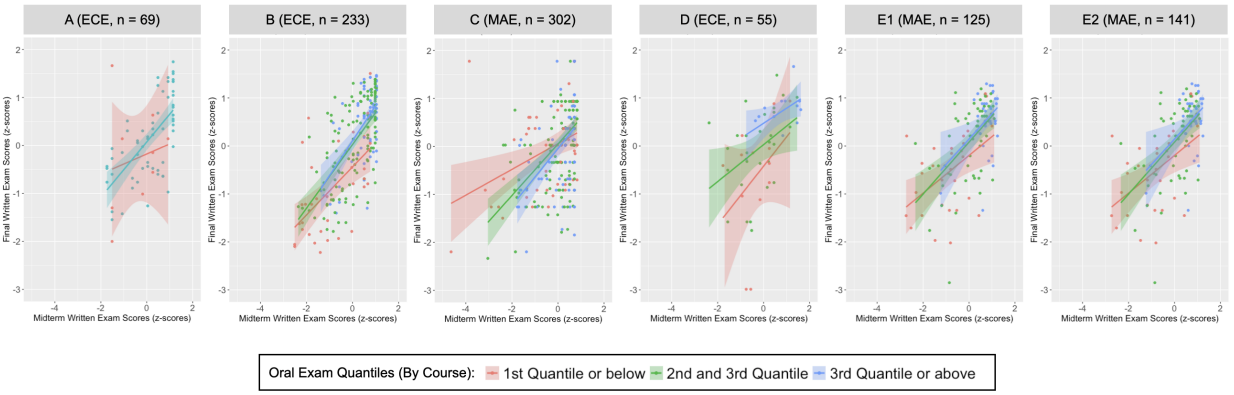

Figure 2: Linear regressions on how the predictive value of midterm written exam on the final written exam varies by the midterm oral exam scores by courses. Three linear regressions are being plotted(red:1st Quantile or below; green: 2nd and 3rd Quantile; blue: 3rd Quantile or above). The gray areas indicate the 2SD confidence intervals. Each graph shows the courses.

our results indicate that oral exams are capturing something relevant to final exam performance above and beyond the midterm written exam. These results create a justification for further exploration of this effect. While deeper learning processes are one potential explanation for this additional predictive power, further empirical work is required to make a clear process-based account of the value of oral exams.

This study helps develop practical guidance for teachers regarding oral exams. For a formative assessment to be worth implementing in the classroom, it must provide useful, diagnostic information of a students' learning. Due to limitations of time and resources, it is also important that separate formative assessments are not completely redundant with one another in the information that they provide. These results show that midterm oral exams provide additional diagnostic information predicting final written exam scores above and beyond the information provided by midterm written assessments. This study therefore provides support for the value of adding midterm oral exams as a formative assessment alongside midterm written exams in engineering classes. Our analyses also investigate the effect of rubric granularity on the predictive value of oral exams. Here, we found more consistent predictive value among the more fine-grained rubrics than the less fine-grained rubrics, though the relationship is not perfectly linear.

The classes included in this analysis were heterogeneous in their content, course structure, exam format, exam timing, and exam scoring. Some classes had multiple midterm written and oral exams, which may have also affected students' performance in the final written exam. Such multicollinearity may have not been captured in our analysis. Further, there may be an effect of content level, as oral exam scores more consistently improved model fit in the upper-division courses in our analysis than the lower division courses, though further investigation is needed.

Future work should investigate how to optimize the predictive value of oral exams in terms of content and assessment strategy. We attempted an initial analysis of assessment strategy by investigating the effect of rubric, but more theory-driven and experimental approaches would help develop useful guidelines. Along with rubrics, the quality and modality of exam preparation may also affect students' performance on oral exams. For example, matching the modality of exam preparation with oral examination may impact students' performance on oral exams (e.g., see [23] for the use of video assignments for students' exam preparation). Further, the current study investigates only oral exam scores, which is only one respect in which a formative assessment can provide useful information about student performance. Future work should include further contextual factors including feedback during and after the exam.

Additional work is needed to investigate factors of students' experience, performance, and assessment that may differ between oral and written exams. The student perception of oral exams (in comparison to written exams) may affect students' performance in oral exams (e.g., see [24] for an exploratory analysis and [12] for an overview of students' experience in oral exams). For example, personal background and identity factors may affect students' performance (e.g., stereotype threat, familiarity of the topic, first-generation status, transfer status), or the grader's perception of students' performance (e.g., race/ethnicity, language proficiency, gender) may introduce factors that specifically affect the utility of oral exams. These factors should also be taken into account when assessing change in students' learning experience after oral exams. For

example, participating in oral exams may affect students' motivation to learn, and this relationship may not be separable from students' personal background and identity factors (e.g., see [25] for discussion on academic support, oral exam experience, and student motivation).

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2044472. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank the following colleagues for the helpful discussion: Carolyn Sandoval, He Liu, Josephine Relaford-Doyle, Leah Klement, Maziar Ghazinejad, Mia Minnes, and Nathan Delson. We thank all our undergraduate and graduate instructional assistants who helped our instructors with administering the oral exams. We would also like to thank the project advisory committee members - Adriana Kezar, Christine Alvarado, and Sheri Shepherd for their feedback and suggestions to our project.

References

[1] B. S. Bloom, T. Hasting, and G. Madaus, *Handbook of formative and summative evaluation of student learning*. McGraw-Hill, 1971.

[2] R. E. Mayer, "Rote versus meaningful learning," *Theory into practice*, vol. 41, no. 4, pp. 226–232, 2002.

[3] W. Kintsch, "Learning from text," *Cognition and instruction*, vol. 3, no. 2, pp. 87–108, 1986.

[4] N. Sagerman and R. E. Mayer, "Forward transfer of different reading strategies evoked by adjunct questions in science text." *Journal of Educational Psychology*, vol. 79, no. 2, p. 189, 1987.

[5] A. K. Thomas and M. A. Mcdaniel, "Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions," *Psychonomic Bulletin & Review*, vol. 14, no. 2, pp. 212–218, 2007.

[6] K. R. Murphy and C. O. Davidshofer, "Psychological testing," *Principles, and Applications, Englewood Cliffs*, vol. 18, 1988.

[7] J. D. Karpicke and H. L. Roediger, "The critical importance of retrieval for learning," *Science*, vol. 319, no. 5865, pp. 966–968, 2008.

[8] C. I. Johnson and R. E. Mayer, "A testing effect with multimedia learning." *Journal of Educational Psychology*, vol. 101, no. 3, p. 621, 2009.

[9] E. L. Thorndike, *Human learning*. Century, 1931.

[10] V. J. Shute, "Focus on formative feedback," *Review of Educational Research*, vol. 78, no. 1, pp. 153–189, 2008.

[11] M. H. Trowbridge and H. Cason, "An experimental study of Thorndike's theory of learning," *The Journal of General Psychology*, vol. 7, no. 2, pp. 245–260, 1932.

[12] H. Qi, M. Lubarda, C. Schurgers, C. L. Sandoval, L. Klement, M. Ghazinejad, J. Relaford-Doyle, M. Kim, M. Minnes, S. Baghdadchi, A. M. Phan, H. Liu, X. E. Gedney, C. Pilegard, and N. Delson, "Enhancing

students' learning through scalable oral assessment in undergraduate engineering classes: insights from Year 1," Accepted for publication in *2022 ASEE Annual Conference and Exposition*.

[13] L. Fiorella and R. E. Mayer, "The relative benefits of learning by teaching and teaching expectancy," *Contemporary Educational Psychology*, vol. 38, no. 4, pp. 281–288, 2013.

[14] K. Kobayashi, "Learning by preparing-to-teach and teaching: A meta-analysis," *Japanese Psychological Research*, vol. 61, no. 3, pp. 192–203, 2019.

[15] V. Hoogerheide, L. Deijkers, S. M. Loyens, A. Heijltjes, and T. van Gog, "Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them," *Contemporary Educational Psychology*, vol. 44, pp. 95–106, 2016.

[16] L. Fries, M. S. DeCaro, and G. Ramirez, "The lure of seductive details during lecture learning." *Journal of Educational Psychology*, vol. 111, no. 4, p. 736, 2019.

[17] D. Ramella, "Oral exams: A deeply neglected tool for formative assessment in chemistry," in *Active Learning in General Chemistry: Specific Interventions*. ACS Publications, 2019, pp. 79–89.

[18] M. Huxham, F. Campbell, and J. Westwood, "Oral versus written assessments: A test of student performance and attitudes," *Assessment & Evaluation in Higher Education*, vol. 37, no. 1, pp. 125–136, 2012.

[19] A. S. Theobold, "Oral exams: A more meaningful assessment of students' understanding," *Journal of Statistics and Data Science Education*, pp. 1–4, 2021.

[20] L. K. Davids, "A study on the effectiveness of team-based oral examinations in an undergraduate engineering course," in *2012 ASEE Annual Conference & Exposition*, 2012, pp. 25–108.

[21] A. Ahmed, A. Pollitt, and L. Rose, "Assessing thinking and understanding: can oral assessment provide a clearer perspective," in *8th International Conference on Thinking, Edmonton, Canada*, 1999.

[22] R. Boedigheimer, M. Ghrist, D. Peterson, and B. Kallemyn, "Individual oral exams in mathematics courses: 10 years of experience at the air force academy," *Primus*, vol. 25, no. 2, pp. 99–120, 2015.

[23] A. M. Phan and H. Qi, "Matching preparation with examination: Effectiveness of video assignments on oral examination outcomes," Accepted for publication in *2022 ASEE Annual Conference & Exposition*.

[24] S. Baghdadchi, H. Qi, M. Lubarda, A. M. Phan, and N. Delson, "An exploratory study of student perceptions of oral exams in undergraduate engineering courses," Accepted for publication in *2022 ASEE Annual Conference & Exposition*.

[25] N. Delson, S. Baghdadchi, M. Ghazinejad, M. Lubarda, M. Minnes, A. M. Phan, C. Schurgers, and H. Qi, "Can oral exams increase student performance and motivation?" Accepted for publication in *2022 ASEE Annual Conference & Exposition*.