Medical Science Educator (2022) 32:1045–1054
https://doi.org/10.1007/s40670-022-01612-y

ORIGINAL RESEARCH

Assessment for Learning with Ungraded and Graded Assessments

Karly A. Pippitt[1,2] · Kathryn B. Moore[3] · Janet E. Lindsley[4,5] · Paloma F. Cariello[6,7] · Andrew G. Smith[8] · Tim Formosa[4] · Karen Moser[9] · David A. Morton[3] · Jorie M. Colbert-Getz[6,10] · Candace J. Chow[6,11]

Accepted: 25 August 2022 / Published online: 14 September 2022

Abstract

Introduction Assessment for learning has many benefits, but learners will still encounter high-stakes decisions about their performance throughout training. It is unknown if assessment for learning can be promoted with a combination model where scores from some assessments are factored into course grades and scores from other assessments are not used for course grading.

Methods At the University of Utah School of Medicine, year 1–2 medical students (MS) completed multiple-choice question quiz assessments and final examinations in six systems-based science courses. Quiz and final examination performance counted toward course grades for MS2017–MS2018. Starting with the MS2020 cohort, quizzes no longer counted toward course grades. Quiz, final examination, and Step 1 scores were compared between ungraded quiz and graded quiz cohorts with independent samples *t*-tests. Student and faculty feedback was collected.

Results Quiz performance was not different for the ungraded and graded cohorts ($p = 0.173$). Ungraded cohorts scored 4% higher on final examinations than graded cohorts ($p \leq 0.001$, $d = 0.88$). Ungraded cohorts scored above the national average and 11 points higher on Step 1 compared to graded cohorts, who had scored below the national average ($p \leq 0.001$, $d = 0.64$). During the study period, Step 1 scores increased by 2 points nationally. Student feedback was positive, and faculty felt it improved their relationship with students.

Discussion The change to ungraded quizzes did not negatively affect final examination or Step 1 performance, suggesting a combination of ungraded and graded assessments can effectively promote assessment for learning.

Keywords Assessment · Performance · Preclinical · Undergraduate medical education

✉ Karly A. Pippitt
karly.pippitt@hsc.utah.edu

[1] Department of Family and Preventive Medicine, University of Utah School of Medicine, 317 Chipeta Way, Suite A, Salt Lake City, UT 84108, USA

[2] Community Faculty, University of Utah School of Medicine, Salt Lake City, UT, USA

[3] Department of Neurobiology, University of Utah School of Medicine, Salt Lake City, UT, USA

[4] Department of Biochemistry, University of Utah School of Medicine, Salt Lake City, UT, USA

[5] Curriculum University of Utah School of Medicine, Salt Lake City, UT, USA

[6] Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA

[7] Health Equity, Diversity, and Inclusion, University of Utah School of Medicine, Salt Lake City, UT, USA

[8] Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, UT, USA

[9] Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT, USA

[10] Education Quality Improvement, University of Utah School of Medicine, Salt Lake City, UT, USA

[11] Education Research, University of Utah School of Medicine, Salt Lake City, UT, USA

Introduction

Assessment plays an important role in education; however, it may create an undue focus on points and scores instead of learning. Traditionally, assessment has been used as a checkpoint, where a score is used to quantify student learning for advancement or promotion. Focusing on grades deters students from addressing their weaknesses and can lead to less desirable behaviors like cramming and memorizing that are not as effective for long-term learning [1, 2]. There is increasing interest in using assessments that encourage learning and decrease student stress [3, 4]. As a result, the assessment paradigm in medical education is shifting from assessments designed to just measure a trainee's performance at a given timepoint (assessment *of* learning) to assessments designed to promote growth, or a "change in knowledge or behavior" (assessment *for* learning) [1, 5–7].

Assessment of learning focuses on judgment of performance while assessment for learning focuses on feedback for growth [8]. Importantly, assessment for learning re-centers the purpose of assessment on the learner's development [9] and shifts the onus of learning from the teacher to the learner [8]. One method of assessment of learning is ungraded assessments. Most of the literature supporting the use of ungraded assessments investigates their use within specific instructional modalities such as team-based learning [10–16] or in subject areas such as anatomy [17, 18]. The literature also suggests ungraded assessments may be beneficial in promoting learning because they are less stressful for students [14, 19]. This decrease in student stress may aid students in developing a mastery orientation to learning as opposed to continuing in a performance orientation, where students focus on achieving positive evaluations and avoiding negative judgements [5, 20, 21]. Additionally, when feedback accompanies ungraded assessments, students have opportunities to identify areas for growth [8, 9, 20, 22, 23]. Providing specific feedback on ungraded assessments is important because it informs students of learning gaps, helps them understand what next steps to take, and prepares them to transfer knowledge to new situations while familiarizing students with expectations of summative, graded assessments [8, 9, 22]. There is a growing consensus that "formative feedback" may be better for promoting self-improvement and learning among students than grades [5, 8, 9, 21, 24]. This may be because frequent assessments with feedback help reinforce the process and underlying goal of self-assessment as a part of learning [24–26].

Although assessments used primarily for learning may have their benefits, students will continue to encounter assessments where a score or pass/fail outcome to demonstrate content mastery impacts a high-stakes checkpoint decision (e.g., certifying exams, other criteria for progression). It has been proposed that undergraduate medical education should integrate formative and summative assessments [27], though literature on how to accomplish this is sparse. Our novel assessment strategy combined ungraded quizzes (to promote assessment for learning) with graded final exams (to determine mastery and criteria for progression) over multiple courses within 2 years of the curriculum. We investigated the impact on performance and perception of learning between students who had all assessments (quizzes and final exam) count toward a final course grade and students who had both ungraded (quizzes) and graded (final exam) assessments. Based on the assessment for learning paradigm, we hypothesized that ungraded assessments would encourage students to prioritize the mastery of course material over performance. Additionally, we hypothesized that students and faculty would view the change positively. If a combination of ungraded and graded assessments have similar catalytic and educational effects on learning as graded assessments alone, this has implications for how medical schools could structure such integrated assessment programs.

Methods

Participants

Participants were University of Utah School of Medicine (UUSOM) medical students (MS) who graduated in 2017, 2018, 2020, and 2021 (MS2017, MS2018, MS2020, MS2021). The MS2019 class experienced a mix of conditions due to the timing of the change to ungraded quizzes and thus was excluded from the study. Student demographics and pre-matriculation performance for the graded (MS2017, MS2018) and ungraded (MS2020, MS2021) cohorts are provided in Table 1. Due to changes in the MCAT exam, we did not compare scores between the cohorts statistically. Undergraduate GPAs were not significantly different for graded and ungraded cohorts ($p = 0.579$). UUSOM has a 4-year program and all courses in the first 2 years are graded using a pass/fail scale. Student perception data were gathered from the most recently studied cohort (MS2023), though their quantitative data are excluded as they had not finished the preclinical curriculum at the time of this study.

Foundational Science Quizzes and Final Examinations

The foundational science curriculum includes a first semester fall introductory course (Foundations), where all assessments were graded for all cohorts. There are six subsequent systems-based courses between spring semester

Table 1 Demographics and incoming MCAT scores and GPA for the four University of Utah School of Medicine cohorts studied. *M* mean, *SD* standard deviation

	Graded cohorts	Ungraded cohorts
Gender		
Male (%, *N*)	50% (97)	57% (136)
Female (%, *N*)	50% (99)	43% (104)
Ethnicity/race group		
Black, Indigenous, person of color (%, *N*)	22% (43)	15% (36)
White (%, *N*)	76% (148)	81% (195)
Unknown (%, *N*)	3% (5)	4% (9)
Pre-matriculation performance		
MCAT—old version	$M=30$ (SD = 3) Percentile rank: 79	$M=31$ (SD = 4), $N=103$ Percentile rank: 83
MCAT—new version	n/a	$M=511$ (SD = 5), $N=137$ Percentile Rank: 85
Undergraduate GPA	3.72 (SD = 0.2)	3.73 (SD = 0.2)

of year 1 and spring semester of year 2, where quiz grading varied by cohort. The assessment program in each course includes multiple-choice question (MCQ) quizzes (typically 35–60 questions given every other week), assignments which make up a variable percentage of the overall course grade, and a final, graded MCQ exam that accounts for 60–85% of the overall grade. With the change to ungraded quizzes, the final exam weight toward the overall grade increased by about 35% in the systems-based courses. Each assessment was administered using ExamSoft (ExamSoft, Dallas, TX, 2020). Students take the USMLE Step 1 exam after the second year.

Terminology

At our institution, we intentionally use the terms "graded" and "ungraded" as opposed to "summative" and "formative" to refer to the changes in how quizzes are used in courses. We continued to score each quiz (the number of correct answers out of the total) and provide feedback on questions, regardless of cohort. Our intent with continuing to provide a score for the ungraded quizzes was to allow students to use these data on their performance as feedback to guide their learning. The quiz score and subsequent review allowed students to identify the concepts they did not understand and/or how they had interpreted the question stem incorrectly. The cumulative course final exams remained unchanged; those exam scores still contributed to the final course grade.

Changes to Systems-Based Courses Quizzes

Starting in Fall 2016, we adopted ungraded quizzes for all the systems-based courses for first- and second-year medical students. The MS2019 cohort was a transitional cohort experiencing graded quizzes in two systems-based courses and ungraded quizzes in four systems-based courses quizzes, and was excluded from analyses. The MS2020 class was the first cohort to experience ungraded quizzes in all six systems-based courses (Fig. 1). Graded cohorts consist of MS2017–MS2018 and the ungraded cohorts include MS2020–MS2021 students.

Importantly, many elements of the quizzes did not change over the course of this study. They occurred on Mondays, approximately every other week, in a lecture hall, during specific testing hours. All students, regardless of graded or ungraded format, were held to the same standards: unable to use notes, use the internet, or consult with classmates; all students signed an honor code to uphold these standards. Quizzes consisted of high-quality, National Board of Medical Examiners (NBME)-format questions reviewed by course directors, by content experts, and by a PhD scientist with experience in NBME item writing. Quizzes were created using a question bank. All questions were revised by course directors for quality assurance on a regular basis; items were omitted based on data on item discrimination, item difficulty, and overall reliability statistics. Quiz changes were made annually as described above, but were minimal: there was < 5% change to questions between cohorts. The percentage of questions that were recall-based was not different between cohorts (11% graded, 10% ungraded, $p=0.513$), suggesting no difference in Bloom's level of cognition. Each question had a detailed explanation and students were provided time after each quiz (both graded and ungraded cohorts) to review this information. Notes taken by the graded cohorts were collected after the review. Both cohorts had access to the same feedback reports detailing strengths and weaknesses by question topic and Bloom's level of cognition.

Reliability coefficients were computed for course exam scores by cohort year using the Kuder-Richardson 20 formula (KR-20). A coefficient ≥ 0.70 suggests good internal consistency [28]. Reliability coefficients (KR-20) for final exam score by course and by cohort year are provided in Table 2. Reliability coefficients for final exam

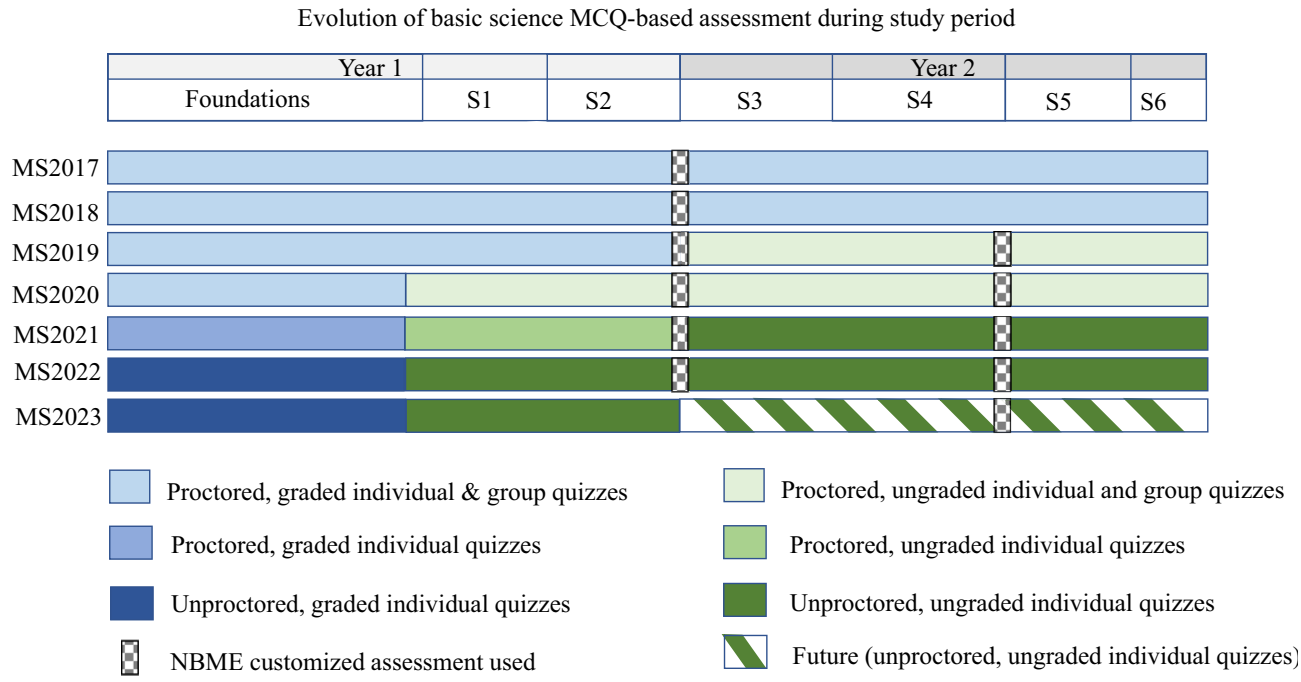

Fig. 1 Evolution of Foundational Science MCQ-based assessment during study period at University of Utah School of Medicine. The courses are listed at the top and include a first semester fall introductory course (Foundations), where all assessments were graded for all cohorts. There were six subsequent systems-based courses (S1–S6) between spring semester of year 1 and spring semester of year 2 where grading varied based on the cohort

scores were ≥ 0.70 for every cohort year for three courses (S1, S3, S5). For two of the course final exams (S2, S4), the coefficient was 0.67 and 0.68 for one of the ungraded years (MS2020, MS2021, respectively) and ≥ 0.70 for all other cohorts, resulting in an average coefficient ≥ 0.70 for ungraded and graded cohorts. For the last final exam (S6), coefficients were ≥ 0.70 for one graded year and one ungraded year and 0.62 for one graded year and 0.64 for one ungraded year, resulting in an average coefficient of 0.67 for both graded and ungraded cohorts. Thus, exam scores had good internal consistency and/or similar reliability coefficients regardless of being graded or ungraded.

Table 2 Reliability coefficients (KR-20) for final examination scores by cohort year and course for University of Utah School of Medicine students

Final exam	Graded years		Ungraded years	
	MS2017	**MS2018**	**MS2020**	**MS2021**
S1	0.81	0.79	0.83	0.73
S2	0.77	0.72	0.67	0.72
S3	0.76	0.81	0.71	0.80
S4	0.80	0.71	0.73	0.68
S5	0.84	0.91	0.71	0.74
S6	0.72	0.62	0.64	0.70

To summarize, in the graded cohorts, the purpose of all quiz and final exam assessments was more in line with assessment of learning; for the ungraded cohorts, quizzes served as assessments for learning, while final exams functioned as the assessments of learning. To further promote assessment for learning for ungraded cohorts, students were allowed to take and keep notes during review of the ungraded quizzes so they could use the feedback for subsequent study. In addition, while our institution has always had an honor code, we gradually relaxed the testing environment to one that was time-flexible and without proctors; these changes treat students as professionals as a signal to them that we trust them.

Data Analysis

Student Performance To determine if the cohorts started the first systems-based course with similar performance on multiple-choice questions, we compared final examination performance for the Foundations course between graded and ungraded cohorts with an independent samples *t*-test. Average quiz performance and final examination performance for systems-based courses were calculated for the graded and ungraded cohorts and compared with independent samples *t*-tests.

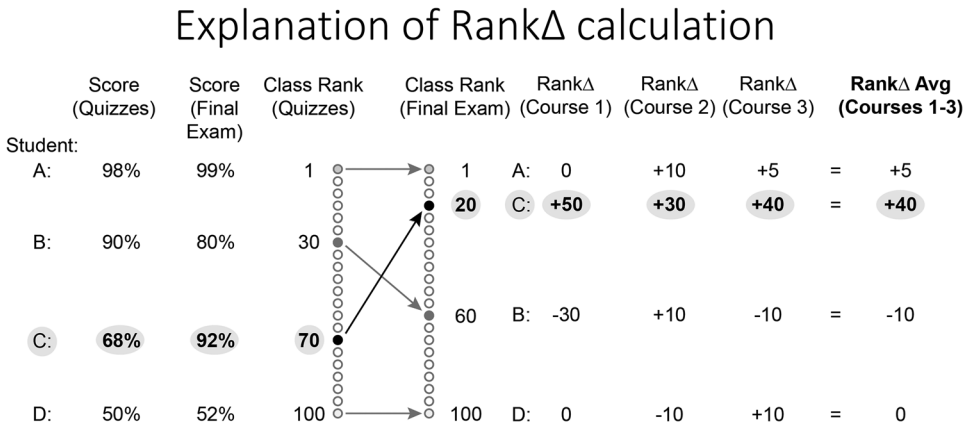

Fig. 2 Student performance examples to demonstrate the logic behind idea of the rankD metric utilized in this study

To address the concern that ungraded quizzes might not provide adequate motivation to drive student effort, we looked for lower performance by individual students on ungraded quizzes relative to performance on the final exam as an indication that quizzes were taken less seriously. Each student was ranked, relative to their classmates, by their mean quiz score and again by their final exam score for each systems course (see Fig. 2 for an example). Our baseline assumption, supported by the demographic information, MCAT scores, and pre-matriculation GPA described in Table 1, is that all students entered our medical school with similar capabilities and took all assessments seriously; our hypothesis is that a student's class rank would therefore remain unchanged on each assessment. Any results that do not support this hypothesis could be ascribed to a lack of effort on quizzes (assuming all students exert maximal effort on graded exams), resulting in lower ungraded quiz scores compared to the summative final exam within a course. We therefore ranked the students by their averaged ungraded quiz scores within a course and compared this to their rank based on the final exam score in that course.

The change in rank (rankΔ) from quiz to final was calculated for each student for each course; then, the average rankΔ was computed for all six courses. This metric is an attempt to detect behavior in a novel way, and has not been used previously. A positive rankΔ value indicates a lower ranking on the quizzes than on the final exam, suggesting that ungraded quizzes provided less motivation or garnered less effort to learn the material being tested for that individual student. Examination of the graded cohorts showed that few students had a rankΔ over 20, so this was taken as an arbitrary cutoff for comparison with the ungraded cohorts. Calculations based on other values produced different absolute results but supported the same conclusions.

In addition to the rankΔ, correlation coefficients were calculated to determine the strength of the relationships between quiz performance and Step 1 performance for graded and ungraded cohorts. To determine the relationship between quiz and exam performance, Pearson's *r* was calculated [29].

Student Perception To better understand the immediate perception of the ungraded relative to graded quizzes, we surveyed one cohort of year 1 students (MS2023) after they had completed the Foundations course (graded quizzes) and the first systems-based course (ungraded quizzes). These students were asked to anonymously indicate the extent to which graded quizzes and ungraded quizzes allowed them to meet practice-based learning objectives on the required end-of-course survey.

Course Director Perspective We gathered feedback from all five faculty who served as course directors both prior to and after the change to ungraded quizzes to understand their perceptions. Although these faculty were open to the change to an ungraded quiz format, the initial idea came from the assistant curriculum dean for foundational sciences and the final decision to implement the change came from the UUSOM's curriculum committee. Four of the faculty participated in a 45-min focus group and one faculty member who was unable to attend the focus group answered the questions (see Appendix) in writing. The focus group was conducted by a member of the team (CJC) who has extensive qualitative research experience and who was the lone team member who neither directed/taught the courses nor analyzed course evaluation data. The focus group was transcribed using Descript version 3.6.1 (San Francisco, CA). The transcript and the written answers were thematically coded [30] using Dedoose Version 8.3.21 (SocioCultural Research Consultants, LLC, Los Angeles, CA, 2020). The data were open coded in order to create a codebook. Codes were further refined through focused coding, which involved comparing, collapsing, and expanding the initial codes. These codes were subsequently organized into three themes. To ensure trustworthiness, the themes and associated quotes were sent to the five faculty for member checking.

The UUSOM Institutional Review Board deemed this study exempt.

Results

Student Performance

Performance data is limited to students who completed all six systems-based courses with their cohort and did not opt-out of sharing their academic data per the UUSOM umbrella IRB consent process (graded cohorts, $N=196$; ungraded cohorts, $N=240$). At least 97% of each MS cohort was represented in the sample. The higher N for ungraded cohorts was due to a planned class size increase unrelated to the quiz grading intervention. There was no difference in the Foundations final examination performance between the graded (mean = 85%, SD = 8%) and ungraded (mean = 84%, SD = 6%) cohorts ($p=0.421$). At the end of year 2, ungraded cohort students scored higher on Step 1 (mean = 237, SD = 14, z-score = −0.15) than graded cohort students (mean = 226, SD = 20, z-score = 0.34) ($p \le 0.001$, $d=0.64$). During the study period, national Step 1 scores increased by 2 points while at our institution, they increased by 11 points for ungraded compared to graded cohorts.

Average quiz and final examination performance in systems-based courses is provided in Fig. 3. Quiz performance was not significantly different for the ungraded (mean = 85%, SD = 5%) and graded (mean = 86%, SD = 5%) cohorts ($p=0.173$). Final exam performance was significantly higher for the ungraded cohorts (mean = 87%, SD = 4%) compared to the graded cohorts (mean = 83%, SD = 5%) ($p \le 0.001$, $d=0.88$).

There were strong positive relationships (> 0.50 Pearson's r) between quiz and exam performance for both graded ($r=0.90$) and ungraded ($r=0.81$) cohorts. The strength of the relationships was also strong between quiz and Step 1 performance for graded ($r=0.79$) and ungraded ($r=0.59$) cohorts.

Figure 4 provides the rankΔ between an individual student's quiz and exam performance. Students were sorted by their score on each exam and their rank relative to peers was assigned as their position in this list. For the graded cohorts, only 1% of students (2) had a rank$\Delta > 20$ (suggesting less motivation for quizzes relative to finals) and 3.1% (6) had a rank$\Delta < 20$ (suggesting less motivation for finals relative to quizzes). For the ungraded cohorts, 9.6% of students (23) had a rank$\Delta > 20$, and 6.7% (16) had a rank$\Delta < 20$.

Student Perception

Table 3 shows the percentage and frequency of students surveyed (response rate 100%, 125/125) who felt graded or ungraded quizzes helped them better meet practice-based learning program objectives. This cohort of students (MS2023) experienced graded quizzes in their first Foundations course and then had ungraded quizzes in their next semester's systems-based course.

The majority of students reported ungraded quizzes better helped them meet all four objectives: identify strengths (64%, 80), identify deficiencies (62%, 77), set learning goals (59%, 74), and address gaps in knowledge (62%, 78) relative to graded quizzes. Fifteen percent or fewer students reported graded quizzes relative to the ungraded quizzes better helped them identify strengths (13%, 16) and deficiencies (12%, 15), set learning goals (15%, 19), or address gaps in knowledge

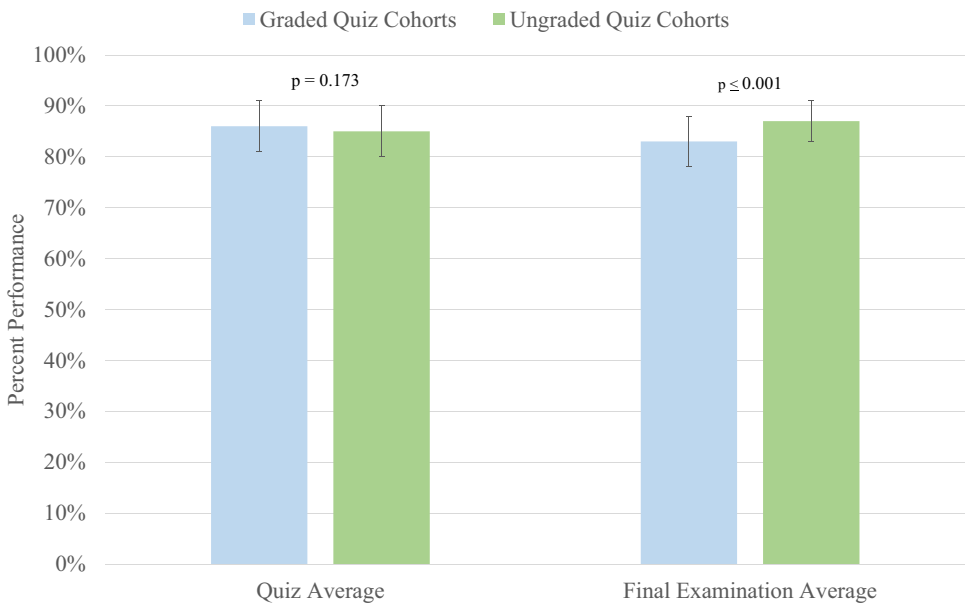

Fig. 3 Average quiz and final exam performance for graded and ungraded cohorts at the University of Utah School of Medicine. Error bars represent ± 2 standard deviation from the mean

Table 3 First year students' perception of which quiz format allowed them to better meet practice-based learning objectives at the University of Utah School of Medicine MS2023

Practice-based learning objectives	Graded quizzes		Both equally	Ungraded quizzes	
	Much more	Slightly more		Slightly more	Much more
Identify strengths in my knowledge	4% (5)	9% (11)	23% (29)	28% (35)	36% (45)
Identify deficiencies in my knowledge	5% (6)	7% (9)	26% (33)	22% (27)	40% (50)
Set appropriate learning and improvement goals	5% (6)	10% (13)	26% (32)	16% (20)	43% (54)
Address gaps in my knowledge in time for the final knowledge exam	3% (4)	9% (11)	26% (32)	14% (18)	48% (60)

(12%, 15), and the remaining students found no difference between the quiz types.

Course Director Perception

Course Director feedback was organized into three themes: constants, changes, and challenges.

Constants Course Directors stated that their approaches to the quizzes generally remained the same. One said, "we probably spent about the same amount of time preparing for quizzes and …making sure the questions were up to date." They also did not see a big difference in quiz performance or quiz taking attitudes from students, stating, "I don't recall seeing any major differences in … performance data on our quizzes" and "most of the students do take this seriously."

Changes Course Directors explained that students no longer fought over quiz questions and answers: "It was … the standard [when] everything was a graded evaluation… the students would nitpick everything." Course Directors said students were more honest about why they scored badly on a quiz if they did, "… when we check in with students who are below passing on each assessment in the ungraded realm… we're getting a lot more responses from students of, 'Oh, I know I blew that off or whatever.'" Course Directors said students' general approach to quizzes was also different as they became more interested in learning for learning's sake: "I feel now that the students are probably more interested in content for clarity as opposed to for getting test questions" and "…the ungraded quizzes have allowed for more questions about concepts and… 'I didn't understand this.'" Finally, Course Directors noted that students were more stressed at the end of the course rather than throughout: "… the antagonism and the entitlement that … we used to see from the students in the entire course is now really only in the last two weeks."

Course Directors said their own behaviors changed as well. They are now willing to write harder questions: "we're a little bit more willing to test out some questions." Course Directors also noted that they were more anxious for their students late in the course: "my anxiety has gone up for the final."

Finally, Course Directors noted that overall, their relationships with students were better as a product of the change. One said, "I think our relationship with students is better… I feel like we're more of a coaching environment … we're working as a team to help you understand this." Another noted, "instead of their feelings all being hurt that they didn't get it, they see now that … we're trying to help them."

Challenges Course Directors described new challenges that accompanied the change. One was that it was difficult to know which students truly needed help when some students did not take the quizzes as seriously: "to really identify those at-risk students can be a little more challenging." Another challenge was that everything rests on the final exam: "I... feel for them... [it's] such a high stakes exam." Finally, Course Directors stated that the challenge of Step 1 remained a constant: "[they are] overwhelmed by this life-defining score that they're going to get on this test that I don't feel like they can see beyond anything."

Discussion

Results of this study support the concurrent use of ungraded and graded assessments in medical education. Specifically, students did not perform any differently when quizzes were ungraded compared to graded (educational effect) and almost all continued to give their full effort on ungraded quizzes (catalytic effect). Additionally, ungraded assessments did not negatively impact performance on the graded final course exam or Step 1. Although there could be a number of contributors to the increased final exam and Step 1 scores in the ungraded cohorts, similar to prior studies, our results suggest that students still take ungraded assessments seriously and these assessments can drive future learning on graded assessments [5, 14, 19].

As shown in Fig. 4, the standard deviation for the rankΔ metric increased and the number of students above the arbitrary cutoff also increased to about 10% with the ungraded cohorts. We interpret this to mean that some students indeed took formative quizzes less seriously, but this behavior was limited to a small number of students. Perhaps this finding could also reflect an increased willingness for students to experiment with new study plans or test-taking strategies in the setting of an ungraded assessment. The increased variance in the rankΔ measurement suggests that students took a more variable, individualized approach to ungraded quizzes, but without a dramatic overall decrease in the perceived importance of their performance on quizzes. This metric provides some insight into the behavior being examined that does not rely on self-reporting.

The majority of students felt the ungraded quizzes promoted course material mastery by helping them change their learning behaviors to better identify strengths and deficiencies, set learning goals, and address gaps in knowledge as compared to graded quizzes, all elements of assessment for learning. Perhaps not surprisingly, students liked ungraded quizzes, even with the increased emphasis on the final exam toward their overall course score. We were clear about the purpose of these ungraded quizzes, that though they were not graded, students could use their scores and accompanying feedback as indicators of how well they were doing in the course. As such, students still perceived these ungraded assessments as worthwhile, perhaps because we were explicit about the purpose of such assessments and our expectations that all students take the quizzes and approach them seriously [8, 31–33]. Additionally, we found that quiz performance, regardless of grading, predicted final exam performance. This study provides a possible answer to questions [3] about how faculty can combine the use of ungraded and graded assessments.

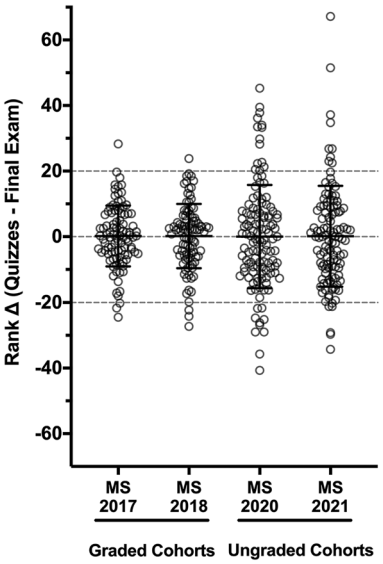

Fig. 4 The mean and standard deviation for the change in rank (rank Δ= average class ranking for quizzes minus the class ranking on the final exam)for each of the four cohorts in the study period at University of Utah Schoolof Medicine. This is based on the assumption that all students would give theirfull effort on all assessments. Positivechange values indicate improved performance relative to peers on the final examcompared to the ungraded quizzes, possibly suggesting decreased effort on the quizzes.The spread of the distribution of this metric increased in the ungraded quizcohorts as measured by the standard deviation, indicating that ungraded quizzesdid affect behavior. However, the absolute number of students outside thearbitrarily chosen gates of plus or minus 20 points was small

Moreover, the faculty focus group data showed that student and course director post-quiz discussions centered around knowledge acquisition and understanding, not student grades. This is an important shift in attitude because students need to understand that learning is not just a collection of facts that one regurgitates when needed, but is about applying those facts to real-life situations—which is what students will do when they practice medicine [9]. These findings model how combining ungraded assessments with meaningful feedback can foster assessment for learning, which is supported by literature that shows that ungraded quizzes can promote long-term educational success for students [3, 5, 8, 9, 14, 19, 23, 24].

Assessment for learning turns the focus of assessment back on the learner by identifying strengths and weaknesses, so students can better learn the information instead of performing for a single assessment. Our study utilized multiple ungraded assessments throughout six courses in the preclinical years, and without changing the final exam significantly, performance on these summative course final exams improved. Each of these ungraded assessments included detailed feedback for students to use for their learning and further development, and indeed, students felt they learned the material better and were more supported in this endeavor.

Finally, our data on faculty perspectives of ungraded quizzes helps provide a more complete picture about how this change has affected multiple stakeholders. Like students, our course directors found ungraded quizzes to be a positive change. Faculty were more willing to experiment with questions that focused on concepts, instead of worrying about writing "perfect" questions to avoid arguments with students, changing the dynamic with students. They felt that they had freedom to create more difficult, multi-step questions. Asking more complex questions often requires students to apply their knowledge in novel ways. This supports Boulet's research suggesting that assessments which involve active production of knowledge result in better learning [34]. Perhaps this lower-stakes environment encouraged faculty to see whether students were indeed learning rather than just performing [5].

As many medical educators have experienced, when high-stakes testing dominates the picture (i.e., prior to Step 1), the classroom dynamics change, and it complicates using assessment for learning [32]. Akin to using ungraded assessments to create an assessment for learning culture, using a pass/fail grading system during the preclinical years can contribute to assessment for learning, as compared with a tiered grading system (e.g., honors/pass/fail, A/B/C/D/F), which is known to increase burnout and stress in students [35]. A cross-sectional study of preclinical grading and United States Medical Licensing Exam (USMLE) scores across 96 allopathic medical schools showed that pass/fail grading does not negatively impact USMLE scores [4]. This further supports the tenets of assessment for learning in that grades need not be the driver for learning, in spite of high-stakes exams. The recent move of USMLE Step 1 to pass/fail and away from using numerical scores to rank students [36] may also decrease student stress around this high-stakes exam.

Several limitations of our study should be considered. With any curriculum, iterative changes occurred in the curriculum, including changes in course directors, adding Team-Based Learning modules to provide some of the content, and changes in quiz questions. Some faculty included more difficult questions on ungraded quizzes, though this was a very small number. It is possible, at least in part, that these changes impacted our study findings. It is also unclear if the decreasing strength of association between quiz performance and final exam performance is due to students thinking that the scores they attain on ungraded quizzes are less important than the learning, the change in the weight of the final exam toward the course score, or some other unknown factor. We do not yet know how the association between quiz and final exam performance may be impacted by Step 1 changing to pass/fail. Our data indicate that only a small portion of students put less effort into ungraded quizzes, but this will require further monitoring. There may be other confounding factors we have not considered.

Multiple reports show changing the assessment environment to support a learning culture is imperative to students' professional identity as physicians and continued curiosity as lifelong learners [25, 32]. Additionally, Heeneman posited that learner perception of an assessment's purpose is key to supporting learning [31]. We believe that shifting to a time-flexible and non-proctored testing environment and allowing students to keep notes they took after reviewing ungraded quizzes created a non-threatening testing environment. This type of climate fosters the use of formative assessment feedback for growth [22]. We continued to use a quality assurance process to ensure quiz questions promote content mastery. Thus, we recommend that if other schools move to ungraded assessments, they ensure that assessments have validity and appropriate item difficulty and discrimination [37, 38].

Conclusions

Medicine is a field of lifelong learning, and developing self-directed learning and reflective skills early in training is critical to success [39]. Our study showed a combination of ungraded and graded assessments can promote self-directed student learning, while still ensuring content mastery. By moving to ungraded quizzes early in medical school, we intended to send a clear message to students that our institution's educational focus is on effective learning. We involved students as part of this conversation, which can help ensure that students understand the benefits of formative assessment for their learning [8, 40]. Given that student and faculty satisfaction is strong and that ungraded quizzes do not compromise overall learning, we plan to continue with ungraded quizzes in the preclinical years of our curriculum. Further, we have now expanded the use of ungraded quizzes to the very first course of our curriculum (Foundations) and will continue to monitor student performance.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Pugh D, Regehr G. Taking the sting out of assessment: is there a role for progress testing? Med Educ. 2016;50(7):721–9. https://doi.org/10.1111/medu.12985.
2. van der Vleuten CPM. A programmatic approach to assessment. Med Sci Educ. 2016;26:9–10. https://doi.org/10.1007/s40670-016-0343-7.
3. Norcini J. What's next? Developing systems of assessment for educational settings. Acad Med. 2019;94(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 58th Annual Research in Medical Education Sessions):S7-S8. https://doi.org/10.1097/ACM.0000000000002908.
4. Kim S, George P. The relationship between preclinical grading and USMLE scores in US allopathic medical schools. Fam Med. 2018;50(2):128–31 https://doi.org/10.22454/FamMed.2018.145163.
5. Scott IM. Beyond 'driving': the relationship between assessment, performance and learning. Med Educ. 2020;54(1):54–9. https://doi.org/10.1111/medu.13935.
6. Harrison C, Wass V. The challenge of changing to an assessment for learning culture. Med Educ. 2016;50(7):704–6. https://doi.org/10.1111/medu.13058.
7. Schuwirth LW, Van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478–85. https://doi.org/10.3109/0142159X.2011.565828.

8. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ. 2019;53(1):76–85. https://doi.org/10.1111/medu.13645.
9. Kulasegaram K, Rangachari PK. Beyond "formative": assessments to enrich student learning. Adv Physiol Educ. 2018;42(1):5–14. https://doi.org/10.1152/advan.00122.2017.
10. Carrasco GA, Behling KC, Lopez OJ. Implementation of team-based learning: a tale of two new medical schools. Med Sci Educ. 2019;29(4):1201–10. https://doi.org/10.1007/s40670-019-00815-0.
11. Carrasco GA, Behling KC, Lopez OJ. A novel grading strategy for team-based learning exercises in a hands-on course in molecular biology for senior undergraduate underrepresented students in medicine resulted in stronger student performance. Biochem Mol Biol Educ. 2019;47(2):115–23. https://doi.org/10.1002/bmb.21200.
12. Behling KC, Gentile M, Lopez OJ. The effect of graded assessment on medical student performance in TBL exercises. Med Sci Educ. 2017;27:451–5.
13. Koh YYJ, Rotgans JI, Rajalingam P, Gagnon P, Low-Beer N, Schmidt HG. Effects of graded versus ungraded individual readiness assurance scores in team-based learning: a quasi-experimental study. Adv Health Sci Educ Theory Pract. 2019;24(3):477–88. https://doi.org/10.1007/s10459-019-09878-5.
14. Deardorff AS, Moore JA, McCormick C, Koles PG, Borges NJ. Incentive structure in team-based learning: graded versus ungraded Group Application exercises. J Educ Eval Health Prof. 2014;11:6. https://doi.org/10.3352/jeehp.2014.11.6.
15. Thompson BM, Haidet P, Borges NJ, et al. Team cohesiveness, team size and team performance in team-based learning teams. Med Educ. 2015;49(4):379–85. https://doi.org/10.1111/medu.12636.
16. Carrasco GA, Behling KC, Lopez OJ. Evaluation of the role of incentive structure on student participation and performance in active learning strategies: a comparison of case-based and team-based learning. Med Teach. 2018;40(4):379–86. https://doi.org/10.1080/0142159X.2017.1408899.
17. Azzi AJ, Ramnanan CJ, Smith J, Dionne E, Jalali A. To quiz or not to quiz: formative tests help detect students at risk of failing the clinical anatomy course. Anat Sci Educ. 2015;8(5):413–20. https://doi.org/10.1002/ase.1488.
18. Mogali SR, Rotgans JI, Rosby L, Ferenczi MA, Low BN. Summative and formative style anatomy practical examinations: do they have impact on students' performance and drive for learning? Anat Sci Educ. 2020;13(5):581–90. https://doi.org/10.1002/ase.1931.
19. Rudland JR, Golding C, Wilkinson TJ. The stress paradox: how stress can be good for learning. Med Educ. 2020;54(1):40–5. https://doi.org/10.1111/medu.13830.
20. Wiliam D. What is assessment for learning? Stud Educ Eval. 2011;37(1):3–14.
21. Seligman L, Abdullahi A, Teherani A, Hauer KE. From grading to assessment for learning: a qualitative study of student perceptions surrounding elimination of core clerkship grades and enhanced formative feedback. Teach Learn Med. 2020:1–19. https://doi.org/10.1080/10401334.2020.1847654.
22. Evans DJ, Zeun P, Stanier RA. Motivating student learning using a formative assessment journey. J Anat. 2014;224(3):296–303. https://doi.org/10.1111/joa.12117.
23. Butler AC, Roediger HL 3rd. Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Mem Cognit. 2008;36(3):604–16. https://doi.org/10.3758/mc.36.3.604.
24. Larsen DP, Butler AC, Roediger HL 3rd. Test-enhanced learning in medical education. Med Educ. 2008;42(10):959–66. https://doi.org/10.1111/j.1365-2923.2008.03124.
25. Swan Sein A, Rashid H, Meka J, Amiel J, Pluta W. Twelve tips for embedding assessment for and as learning practices in a programmatic assessment system. Med Teach. 2020:1–7. https://doi.org/10.1080/0142159X.2020.1789081.
26. van der Vleuten CP, Dannefer EF. Towards a systems approach to assessment. Med Teach. 2012;34(3):185–6. https://doi.org/10.3109/0142159X.2012.652240.
27. Konopasek L, Norcini J, Krupat E. Focusing on the formative: building an assessment system aimed at student growth and development. Acad Med. 2016;91(11):1492–7. https://doi.org/10.1097/ACM.0000000000001171.
28. Cortina JM. What is coefficient alpha? An examination of theory and applications. J Appl Psychol. 1993;78(1):98.
29. Cohen J. A power primer. Psychol Bull. 1992;112(1):155–9. https://doi.org/10.1037//0033-2909.112.1.155.
30. Kiger ME, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. Med Teach. 2020;42(8):846–54. https://doi.org/10.1080/0142159X.2020.1755030.
31. Heeneman S, Oudkerk Pool A, Schuwirth LW, van der Vleuten CP, Driessen EW. The impact of programmatic assessment on student learning: theory versus practice. Med Educ. 2015;49(5):487–98. https://doi.org/10.1111/medu.12645.
32. Harlen W. Teachers' summative practices and assessment for learning - tensions and synergies. The Curriculum Journal. 2005;16:207–23.
33. Eva KW, Bordage G, Campbell C, et al. Towards a program of assessment for health professionals: from training into practice. Adv Health Sci Educ Theory Pract. 2016;21(4):897–913. https://doi.org/10.1007/s10459-015-9653-6.
34. Boulet J. Teaching to test or testing to teach? Med Educ. 2008;42(10):952–3. https://doi.org/10.1111/j.1365-2923.2008.03165.x.
35. Reed DA, Shanafelt TD, Satele DW, et al. Relationship of pass/fail grading and curriculum structure with well-being among preclinical medical students: a multi-institutional study. Acad Med. 2011;86(11):1367–73. https://doi.org/10.1097/ACM.0b013e3182305d81.
36. United States Medical Licensing Examination. Change to pass/fail score reporting for Step 1. Secondary Change to pass/fail score reporting for Step 1. 2020. https://www.usmle.org/incus/#decision. Accessed 1 Oct 2021.
37. Colbert-Getz JM, Ryan M, Hennessey E, et al. Measuring assessment quality with an assessment utility rubric for medical education. MedEdPORTAL. 2017;13:10588. https://doi.org/10.15766/mep_2374-8265.10588.
38. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206–14. https://doi.org/10.3109/0142159X.2011.551559.
39. Boud D, Falchikov N. Aligning assessment with long-term learning. Assessment and Evaluation in Higher Education. 2006;31:339–413 https://doi.org/10.1080/02602930600679050.
40. Harrison CJ, Konings KD, Schuwirth LWT, Wass V, van der Vleuten CPM. Changing the culture of assessment: the dominance of the summative assessment paradigm. BMC Med Educ. 2017;17(1):73. https://doi.org/10.1186/s12909-017-0912-5.