

Faculty Continuity and Anatomy Practical Exams as the Real Behavioral Levers: What a Passing Threshold in a Curricular Thread Revealed About Student Engagement in an Integrated Curriculum

Daniela Armella Tangarife

`daniela.armella.tangarife.med@dartmouth.edu`

Geisel School of Medicine at Dartmouth

Ian R. Beals

Geisel School of Medicine at Dartmouth

Virginia T. Lyons

Geisel School of Medicine at Dartmouth

Research Article

Keywords: Integrated curriculum, longitudinal courses, curricular threads, anatomy, embryology, pass-fail curriculum, medical graduate school education

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11272301/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

PURPOSE

After integrating the Geisel pre-clerkship curriculum, USMLE Step 1 scores and clerkship performance revealed gaps in curricular-thread (CT) knowledge. Faculty hypothesized that fewer contact hours and the scarcity of CT items on assessments caused students to deprioritize these topics. This study explored whether instituting a passing threshold for a CT across a pre-clerkship integrated curriculum would increase student engagement and performance.

METHODS

Geisel implemented a pilot intervention requiring MD students to achieve $\geq 70\%$ in the Anatomy and Embryology (A&E) CT as a prerequisite for progressing to the clinical phase of training. Control data derived from the prior class who completed the same curriculum without this requirement.

Quantitative analyses compared performance, mean scores and passing rates on A&E assessment items. Cohort interviews were analyzed for students' engagement with longitudinal curricula, CT study prioritization, and barriers to mastery.

RESULTS

The experimental cohort had a higher failure rate and lower floor score than the control cohort. Analysis of interviews identified six themes: policy acceptance with low awareness of its novelty; equivalent study habits for threads and core content; exam-driven deprioritization of low-yield or difficult topics; confusion regarding what constituted a curricular thread; effective pedagogy as a catalyst of engagement and learning; and thread-specific exams motivated A&E study.

CONCLUSION

Results indicated that a passing threshold did not improve engagement or performance in the A&E CT. Instead, engagement was driven by standalone thread assessments and effective pedagogy. Understanding the factors that influence performance in longitudinally-distributed course materials is important for instructional design and assessment strategies.

INTRODUCTION

Over the past two decades, U.S. medical schools have increasingly adopted integrated pre-clerkship curricula to replace traditional, discipline-based courses with organ-system or case-based blocks in which foundational scientific content is woven together with clinical material [1, 2]. This integration is intended to promote horizontal and vertical connections across disciplines, support clinical reasoning earlier in training, and reduce redundancy [3]. However, a recognized trade-off of this model is that foundational science disciplines – such as anatomy, histology and pathology – lose their identity as free-

standing courses and exist as topics presented across multiple blocks, often in reduced form, over the pre-clerkship years [4, 5]. Several institutions that have undergone this transition have reported unintended consequences for these threaded curricula, including reduced contact hours, fragmented content delivery, and gaps in foundational knowledge detectable on licensing examinations and in clerkship performance [5–7].

Anatomical sciences and embryology appear particularly vulnerable to this effect. Serial national surveys of U.S. anatomical sciences course directors by Drake and colleagues have tracked these disciplines through the transition. In 2002, gross anatomy averaged 167 contact hours, most of them in cadaveric laboratory, whereas embryology averaged only 16 hours, was taught primarily by lecture, and was rarely a free-standing course [8]. By 2009, total gross anatomy hours and embryology laboratory hours had declined [9]; by 2013, a growing proportion of gross anatomy courses were embedded within integrated curricula [10]; and by 2017, few free-standing anatomy courses remained, with further reductions in gross anatomy and histology hours [4]. Gribbin and colleagues found that after their institution moved to a time-restricted integrated curriculum, histology-related academic outcomes declined and students reported adapting their study strategies specifically around the reduced weight of histology content on exams [6]. Similarly, Husmann and colleagues likewise found that a transition from a stand-alone to an integrated anatomy curriculum was associated with changes in students' study plans and self-reported study strategies [11]. Earlier work by McKeown and colleagues in the United Kingdom likewise documented declines in students' anatomical knowledge following curricular reform, findings that anticipated similar concerns now widely reported by U.S. schools transitioning to integration [7].

A common theme across this literature is the principle, well established in medical education, that assessment – not curricular intent – is often the primary driver of study behavior for students [12–14]. Wormald and colleagues found that as the weighting of anatomy within their institution's assessment scheme increased, students reported markedly greater motivation to learn the subject [12]. Similarly, Newble and Jaeger found that students redirected their study toward whichever domain an examination emphasized [13], and Cilliers and colleagues showed that students' learning behavior is shaped by how they appraise the consequences of an assessment and their own capacity to respond to it [14]. With this in mind, if curricular threads such as anatomy and embryology are assessed only in small proportion within larger, block-level examinations, this literature predicts that students will deprioritize thread content even when instructors emphasize its relevance for licensing exams or their future career.

At the Geisel School of Medicine at Dartmouth, the pre-clerkship curriculum (“Phase 1”) became fully integrated in fall of 2019. In the years that followed, USMLE Step 1 performance data and anecdotal reports from clerkship directors – particularly from the surgery clerkship – identified gaps in students' foundational knowledge of the Anatomy and Embryology (A&E) curricular thread. Medical education faculty proposed two complementary causal hypotheses: (1) that the integration-induced reduction of dedicated contact hours for A&E content had a negative impact on student learning [7], and (2) that A&E questions constitute only a small percentage of individual course examinations, creating an incentive

structure in which students deprioritize this content, consistent with the assessment-driven behavior described above [6, 12–14].

To investigate the second hypothesis, Geisel piloted a policy in July 2024 requiring students in the Class of 2028 to achieve a minimum score of 70% in the longitudinal A&E curricular thread, aggregated across the pre-clerkship phase, as a prerequisite for progression into the clinical phase. This policy was designed to function as a behavioral lever: by attaching a concrete, individually consequential threshold to thread-level performance, the intervention was intended to increase student engagement with A&E content and overall study time devoted to it, with the downstream goal of improving knowledge and preparedness for clerkships.

Minimum passing thresholds have already been widely applied in medical education to certify readiness for clinical training and licensure, ranging from NBME clinical subject ('shelf') examinations to the USMLE Step 1, Step 2 CK, and Step 3 examinations. However, most literature about exams with passing thresholds concerns broad, integrative competency examinations rather than thresholds applied to a single longitudinal content thread embedded within an otherwise integrated grading structure. Additionally, little empirical work has examined whether such a targeted policy actually changes how students study as opposed to simply raising the bar for an outcome that would have been reached regardless. Understanding whether a thread-specific passing threshold changes engagement, and if not, what does drive engagement with threaded content, has direct relevance for the many U.S. medical schools that have adopted or are considering integrated curricula and are grappling with similar concerns about foundational science attrition [1, 2, 4].

Consequently, the purpose of this study was to evaluate whether implementing a minimum passing threshold in the A&E curricular thread at Geisel improved student engagement and performance within its integrated pre-clerkship curriculum. A mixed-methods design was utilized, comparing quantitative A&E performance data between a pre-intervention control cohort (Class of 2027) and the first intervention cohort (Class of 2028), and conducting semi-structured interviews with students from both cohorts to characterize how – and whether – the policy shaped study behavior, and what students identified as the actual drivers of their engagement with curricular thread content.

METHODS

Study Design and Setting

A single-institution, mixed-methods study was conducted at Geisel School of Medicine at Dartmouth using a combined quantitative between-cohort comparison of A&E curricular thread performance with a qualitative thematic analysis of semi-structured student interviews between cohorts.

Geisel's 18-month pre-clerkship curriculum was fully integrated in Fall of 2019, and A&E content has since been delivered longitudinally across multiple organ-system blocks. Although the entire curriculum undergoes continuous quality improvement, the A&E curriculum, pedagogy and assessment questions

remained relatively stable during the study period (fall 2023 – spring 2026). Content has been consistently delivered in two primary formats – 1) dissection laboratories with anatomical donors and 2) case-based learning classroom sessions. Written examination questions were selected from a pool of previously vetted items, with approximately 85% of the bank reused across the most recent sequential years. For practical examinations, the list of structures students were expected to dissect, identify, and study was identical each year; only the specific structures tagged for examination varied, based on the quality and availability of structures in each year's dissections.

In July 2024, Geisel's Medical Education Committee implemented a pilot policy requiring students in the MD Class of 2028 to achieve a minimum aggregate score of $\geq 70\%$ in the A&E curricular thread across the pre-clerkship phase as a prerequisite for progressing into clinical clerkships. Students who did not reach the threshold would not receive any type of notation on their transcript; instead, they would be required to participate in a remedial activity during the dedicated Step 1 study period tailored to their knowledge gaps. The MD Class of 2027, which completed the same integrated curriculum *without* this passing requirement, served as the historical control cohort.

Ethical Considerations

Our study was reviewed and approved as exempt human subjects research by the Dartmouth College Committee for the Protection of Human Subjects (CPHS).

Quantitative performance data from both cohorts were obtained as de-identified, aggregate output from Geisel's Office of Assessment, Evaluation and Accreditation as part of routine, already-collected educational assessment data; individual consent was not required for this component.

For the qualitative component, verbal informed consent was obtained from all interview participants using a Dartmouth CPHS-approved verbal consent script prior to conducting the interviews and transcribing them via Zoom (Zoom Video Communications, Inc., San Jose, CA). Participation was voluntary and participants could decline to answer any questions or withdraw at any time without academic consequence. Interview recordings and transcripts were de-identified prior to analysis, and participants were not referred to by name or other identifying information in any part of the study. Participants who completed an interview were eligible for a \$10 gift card, limited to the first ten respondents from each class to enroll due to limited compensation resource availability.

Quantitative Component

The quantitative sample included all MD students in the Class of 2027 (control cohort, $n = 90$) and Class of 2028 (intervention cohort, $n = 95$) who completed the standard pre-clerkship A&E curriculum at Geisel. Students who transitioned to a decelerated academic schedule during the study period ($n = 4$, Class of 2027; $n = 1$, Class of 2028) were excluded from quantitative analysis because their assessment timelines

did not align with their nominal cohort. De-identified performance data were provided by Geisel's Office of Assessment, Evaluation and Accreditation.

The primary outcome was each student's composite A&E curricular thread score, defined as the average performance across all A&E-related assessment items completed over the pre-clerkship phase, including both in-house course examination questions tagged to the A&E thread and anatomy laboratory practical examination items. Cohort-level pass/fail status on the A&E thread was defined using the $\geq 70\%$ threshold that applied prospectively to the Class of 2028 and was applied retrospectively to Class of 2027 data for comparison purposes only (this threshold was not a graduation requirement for the Class of 2027).

Mean A&E composite scores, score ranges, and thread pass/fail rates were compared between the Class of 2027 and Class of 2028 cohorts. All calculations were performed in Microsoft Excel (Microsoft Corporation, Redmond, WA), with statistical significance set at $\alpha = 0.05$.

Qualitative Component

Students from the Class of 2027 and Class of 2028 were recruited via an advertisement distributed through Geisel's weekly *Medical Student Research Newsletter*. Participation was voluntary and open to all students in both classes; no other inclusion or exclusion criteria were applied.

Ten semi-structured interviews were conducted with students from each cohort (total $n = 20$). Interviews lasted approximately 15 minutes and used a standardized, predetermined interview guide developed by the study team to explore policy awareness, study prioritization, barriers to A&E mastery, and general engagement with curricular thread content. All interviews were conducted via Zoom by two medical student researchers (D.A.T., Class of 2028, and I.R.B., Class of 2029) who were not involved in grading or assessment of participants. Interviews were transcribed and de-identified transcripts were saved for qualitative analysis.

Interview transcripts were analyzed using thematic analysis following a six-phase framework proposed by Braun and Clarke [15]: 1) familiarization with the data, 2) generation of parent and child codes, 3) searching for candidate themes, 4) reviewing themes, 5) defining and naming themes, and 6) producing the final analytic narrative.

The two interviewers (D.A.T. and I.R.B.) first independently reviewed all transcripts to develop a preliminary parent- and child-code structure, which was then reviewed together and finalized as a single, official study code. Both coders then independently applied this codebook to every transcript in the dataset; the team subsequently met to compare coding, resolve discrepancies by consensus, and finalize a single, unified code for each transcript. After coding finalization, a meeting was held to review the transcripts and derive overarching themes as well as select illustrative quotations from the texts for each theme. Coding was managed using ATLAS.ti. Sample size was guided by the expectation of thematic saturation within approximately 9 to 17 interviews per relatively homogeneous population,

consistent with empirical work on saturation in semi-structured interview studies [16, 17]; saturation was confirmed qualitatively, as no substantively new codes emerged in the final interviews of either cohort. All illustrative quotations included in our final results were de-identified and attributed by cohort, with instructor names within quotations replaced with role descriptors in brackets.

RESULTS

Quantitative Analysis

Baseline characteristics of the two cohorts are summarized in Table 1. The control (Class of 2027) and intervention (Class of 2028) cohorts were comparable on available measures of prior academic preparation, with near-identical MCAT total scores (515 and 514, respectively) and similar mean undergraduate GPAs (3.75 and 3.81). The cohorts were likewise similar in mean age, sex composition and college major (science vs. non-science). This baseline comparability supports the use of the Class of 2027 as a historical control.

Table 1

Baseline characteristics of the control and intervention cohorts. Cohort-level performance is presented in Table 2. Composite Anatomy and Embryology (A&E) curricular-thread scores were available for 90 students in the control cohort (Class of 2027) and 95 students in the intervention cohort (Class of 2028); five students on a decelerated schedule (n = 4, Class of 2027; n = 1, Class of 2028) were excluded as described in the Methods.

Characteristic	Class of 2027 (Control, n = 90)	Class of 2028 (Intervention, n = 95)
MCAT total score average	515	514
Undergraduate GPA average	3.75	3.81
Sex, n	46 women, 45 men	56 women, 39 men, 1 genderqueer
Age, mean (years)	25	24
Students with non-science major, %	22.2%	22.1%

Table 2
Anatomy and embryology curricular-thread performance by cohort.

Characteristic	Class of 2027 (Control, n = 90)	Class of 2028 (Intervention, n = 95)
Composite A&E thread score, mean $\pm$ SD	82.0 $\pm$ 7.0	81.1 $\pm$ 7.8
Composite A&E thread score median (IQR)	82.3 (78.2–86.6)	82.4 (76.8–86.9)
Composite A&E thread score, range	61.7–94.3	52.7–95.9
Students scoring below 70%, n (%)	3 (3.3)	10 (10.5)
Range of sub-threshold scores	61.7–67.1	52.7–69.3

The great majority of students in both cohorts met the $\geq 70\%$ threshold. In the control cohort, 3 of 90 students (3.3%) scored below 70%, compared with 10 of 95 students (10.5%) in the intervention cohort. This numerically higher proportion of sub-threshold students in the intervention cohort did not reach statistical significance, per Fisher's exact test, two-sided $p = 0.083$. Additionally, odds of scoring below threshold was approximately 3.4 times higher in the intervention cohort, though the confidence interval was wide given the small number of events. Median scores were nearly identical between cohorts (82.4 vs. 82.3), whereas mean scores were 81.1 (SD 7.8) in the intervention cohort and 82.0 (SD 7.0) in the control cohort. The distribution of composite A&E scores was also wider in the intervention cohort. Scores ranged from 61.7 to 94.3 in the Class of 2027 and from 52.7 to 95.9 in the Class of 2028; as well the intervention distribution was more left-skewed (skew = 1.01 vs. -0.39).

The numbers of students that did not meet the grading threshold in each pre-clerkship course are shown in Table 3; values were similar in all courses throughout the phase so it is unlikely that the content or assessments in any particular course caused students to miss the threshold. Although the A&E average across the entire phase determined passing or failing status in the thread, students were only required to remediate the subjects that they were deficient in.

Table 3

Numbers of A&E assessment questions in Phase 1 courses and failures for the intervention cohort.

Phase 1 Course	Number of A&E questions (Class of 2028)	Number of students below threshold (Class of 2028)
Foundations of Medical Science and Practice	16	8
Cardiovascular Medicine	30	6
Respiratory Medicine	30	7
Endocrinology	4	5
GI, Metabolism, Nutrition	31	6
Reproductive Medicine	29	7
Dermatology, Rheumatology and Orthopedics	98	7
Renal Medicine	8	7
Neuroscience and Neurology	67	6

Qualitative Analysis

Ten interviews were conducted and analyzed per cohort; saturation was reached within this population and analysis revealed six primary themes. Each theme is described below with selected quotes from interview participants.

Theme 1. Attitudes toward the policy

An important distinction emerged between awareness that a passing standard existed and awareness that it was new. Most intervention-cohort students knew that a minimum grade had to be achieved in the A&E thread, but the majority did not realize that this threshold was a new requirement specific to their class rather than a long-standing feature of the curriculum. Where students were aware of the policy, they generally regarded it as a reasonable, fair guardrail that set a floor of anatomical competence without adding meaningful stress.

“To be completely honest, I didn’t even know they had changed that for us.” (Class of 2028)

“It makes sense to have it- a good motivator without being too hard to actually achieve.” (Class of 2028)

When control-cohort students were asked to consider the same policy hypothetically, responses were mixed: some anticipated added anxiety, particularly for low-volume, high-difficulty content such as embryology, while others saw value in a motivational floor.

"I would have been more stressed for an exam- specifically embryology. If there were three questions on a particular exam, it probably would have made me freak out." (Class of 2027)

"I feel like it's probably not a bad thing to have a minimum requirement. I would have been more motivated to study." (Class of 2027)

Theme 2. No change in study habits (Class of 2028)

Awareness of the threshold did not translate into changed study behavior. Most intervention-cohort students reported that the policy did not prompt any targeted A&E study strategy; instead, they described a holistic approach oriented toward strong performance across all block material. A recurring explanation for this from the interviewees was that the pervasiveness of anatomy across the longitudinal curriculum provided sufficient opportunities to engage with and master the material; thus, meeting the threshold did not require any redirection of effort.

"I didn't really place emphasis on studying the specific threads- I just studied the course material." (Class of 2028)

"I shoot for an 85 on everything, because then you're comfortably above the passing threshold and you're also not overworking yourself." (Class of 2028)

"There's so much anatomy over the course of Phase 1 that, as long as I have a decent average, I know there's some good buffer room." (Class of 2028)

Theme 3. Exam-driven de-prioritization

Study allocation was often governed by anticipated question volume and perceived exam yield. Content expected to contribute only one or two examination items, or content perceived as conceptually difficult, was routinely de-prioritized when time was limited. Embryology was the most frequent casualty of this reasoning; students in both cohorts independently expressed retrospective regret at having de-prioritized it once they encountered it in USMLE Step 1 preparation.

"If I was running out of time... I would drop things that I knew were going to be one or two questions on the exam." (Class of 2027)

"I'm probably not going to understand [embryology] because it's so difficult...it was easy to push it to the side since it's only worth two questions... But then, coming back and studying for STEP, I thought 'Wow, that actually is so important.'" (Class of 2028)

Interestingly, if thread content was integrated well or aided understanding of other material, some students were less likely to deprioritize it during exam preparation since they perceived it as part of the "core material" of a course.

Theme 4. Low curricular-thread identity

Particularly in the control cohort, students demonstrated variable understanding of which content constituted a longitudinal thread and how threads differed from core block material, with some describing thread sessions as discrete, one-off classes rather than a continuous strand. Labeling content as a separate thread appeared to encourage students to mentally separate it from the surrounding block material, which could lead to de-prioritization.

"[Curricular thread sessions] feel like one-off classes in the block- if you skipped it, you just missed that info." (Class of 2027)

Anatomy largely escaped this dynamic in the intervention cohort: delivered pervasively across blocks and anchored by consistent faculty and practical examinations, it was experienced as foundational course content rather than as a separable thread.

"I don't really think of anatomy as a separate course... it's the foundation we build on for every block." (Class of 2028)

Theme 5. Faculty consistency as an engagement driver

Students in both cohorts identified continuity of teaching faculty that utilized an effective pedagogy as among the strongest determinants of whether a thread felt coherent, predictable, and worth investing in, a pattern that held independent of the grading policy. The more generalizable form of this finding concerned its converse: where a thread lacked a consistent instructor, particularly where sessions were delivered by rotating or one-off "guest" lecturers, students described the thread as disjointed, harder to prioritize, and less consequential to study. Continuity built familiarity with an instructor's pedagogy, expectations and study resources that fostered trust; its absence left a gap in engagement likelihood, regardless of the underlying content.

"Knowing how they teach and what worked before [for study strategies] - that was enormously helpful." (Class of 2027)

"[Lack of continuity] lent an air of guest-lecture-y-ness that made it feel even less consequential to really study." (Class of 2027)

Theme 6. Practical examinations as the real behavioral lever (Class of 2028)

When intervention-cohort students made a concerted effort to study A&E thread content, they attributed it to upcoming anatomy laboratory practical examinations rather than the $\geq 70\%$ threshold. Questions on laboratory practical exams contributed to the longitudinal A&E thread grade, and these exams were administered separately from block exams. Therefore, more time could be allocated for exam preparation, which provided students with the opportunity to strive towards enhancing their overall longitudinal curriculum score.

"If you have a practical coming up, it's like: okay, start working on that." (Class of 2028)

"Practicals force me to learn a lot of anatomy, but outside of studying for practicals, I spend very little time studying anatomy." (Class of 2028)

DISCUSSION

This mixed-methods pilot evaluated whether a thread-specific minimum passing threshold increased student engagement and performance in the Anatomy & Embryology curricular thread within a fully integrated 18-month pre-clerkship curriculum. Neither method of analysis supports the hypothesis that the threshold achieved its intended effect. Quantitatively, the intervention cohort did not outperform the control cohort; a larger proportion of intervention students fell below the threshold (10.5% vs. 3.3%), a difference that was not statistically significant. Central performance, however, was essentially unchanged: the near-identical medians alongside a lower, more variable mean indicate that the difference reflected a heavier low-scoring distribution tail rather than a general decline in performance. Qualitatively, most intervention students felt that having a passing standard for foundational content was appropriate, and few reported changing their study behavior in response to it. Instead, students identified two other factors as the real drivers of their engagement with A&E content: (1) anatomy laboratory practical exams and (2) a consistent, predictable teaching approach, in which students knew what each session would involve, how to prepare for it, and how to best learn from it. We interpret the central finding of this study as a policy-behavior gap: a grading threshold was insufficient on its own to change study behavior. It is also worth noting that the motivations behind study behavior and course engagement are diverse and the interpretations are based on a subset of students, not the entire cohort.

Impact of the policy on student behavior

The qualitative data suggests that the threshold's ineffectiveness stemmed from several factors. First, students indicated that the A&E curricular thread is particularly well-integrated in the pre-clerkship curriculum and thus many students perceived it as "core course content". Other students described anatomy as a foundation that was essential to understanding the subsequent organ-system content. Due to these perceptions, additional motivation was likely not needed for students to study the A&E material; students described a holistic approach to assessment preparation where A&E content was studied in a manner equivalent to other course content. A grading threshold might be more effective for increasing engagement in threads that students perceive as insignificant or as isolated topics in the curriculum. Second, since the grading threshold was longitudinal in nature with consequences only at the end of the pre-clerkship phase, most students did not feel pressured to devote significant study time to A&E content assessed within course examinations; poor thread performance on one assessment could be buffered by superior performance on a subsequent assessment. In contrast, anatomy practical exams were viewed as isolated assessments that tangibly impacted the A&E curricular thread score, motivating students to devote considerable time to exam preparation. These observations are consistent with the long-standing principle that assessment, rather than curricular intent, is the primary driver of study behavior [12], and with evidence that students' learning responses depend on how they appraise the consequences of a given assessment [13, 14]. A passing threshold changes the stated stakes of an

outcome, but it does not, by itself, alter the approximate assessment structure that students actually respond to. Where behavior did shift toward A&E content, it did so in response to practical examinations - not to the threshold. Finally, the fact that no official notation was recorded on the transcripts of students not meeting the threshold may have diminished the impact of the intervention. Students that scored $\leq 70\%$ were obligated to attend remedial sessions to review anatomy and embryology content prior to taking Step 1. Students typically appreciate review sessions, thus the consequence may have been viewed as a benefit rather than a penalty. In fact, when word got around about the review sessions, students who were not in the remedial group attempted to "opt-in".

Instructional consistency and thread identity as deeper structural findings

Beyond the specific policy question, the strongest and most consistent cross-cohort signal in the qualitative data was the role of instructional consistency. Threads taught by stable, recurring instructors using an effective predictable format were experienced as coherent and worth engaging with, because students knew what each session would involve, how to prepare for it, and how to learn from it. In contrast, threads taught by a rotating group of faculty, each using a different teaching approach, were experienced as fragmented and less useful, and were deprioritized. This finding reframes the problem: for a longitudinal thread, the consistency with which it is taught may matter more for engagement than the grading rules attached to it. For schools working to improve the integration and visibility of their threads, this points to instructional continuity – a successful consistent teaching approach, typically sustained by the same faculty or teaching team recurring across the curriculum – as a concrete, low-cost lever, and identifies inconsistent or rotating staffing as a specific, addressable cause of disengagement.

The interviews suggest several possible, non-mutually exclusive explanations for this pattern. When students could not anticipate what a session would involve, they appeared less willing to invest their time, whereas a predictable format lowered the preparatory cognitive load and made the value of attending clear. Most students attributed their engagement primarily to the quality of the content and teaching itself, and attendance appeared to be further reinforced when curricular thread assessments closely reflected what was taught in person, and when no adequate alternative resource existed to replace a session. Some students also described how a personal connection with a recurring instructor fostered a comfortable learning environment, which in turn made them more likely to attend, though this appeared to be a secondary factor. Because our sample was small and students' perspectives varied, we cannot determine the relative contribution of each factor; they are best understood as hypotheses for future work.

A related and somewhat paradoxical finding concerns thread labeling. Students reported that explicitly designating content as a separate thread encouraged them to cognitively separate it from the surrounding block, and that this partition became a basis for strategic deprioritization. Sessions perceived as one-offs were more readily dropped from students' priority lists.

Interpreting the higher sub-threshold rate with caution

The numerically higher proportion of intervention-cohort students scoring below 70% should not be interpreted as evidence that the policy harmed learning. Several considerations argue against overinterpretation. First, the comparison is underpowered: with only 3 and 10 events across two cohorts, the difference did not reach significance ($p = 0.083$). Second, the two cohorts were comparable on available measures of incoming preparation, with near-identical MCAT scores (515 vs. 514) and similar undergraduate GPAs (3.75 vs. 3.81), which argues against the higher rate simply reflecting a less-prepared intervention class. However, the design relies on a historical control, so unmeasured temporal differences between cohorts, such as concurrent curricular changes, testing conditions, or shifts in competing demands on students' time, cannot be excluded. Additionally, more students in the control cohort ($n = 4$) vs. the intervention cohort ($n = 1$) were placed on a decelerated schedule due to academically difficulty and excluded from the data; it is likely that at least some of these students would have been part of the remedial group, potentially leading to a smaller difference in group size between the two cohorts. Third, the threshold may function less as a behavior-change lever than as a detection mechanism: by formally measuring thread-level pass/fail status for the first time, the policy may have surfaced pre-existing gaps that, in the control cohort, simply went unflagged.

Implications for integrated curricula

Taken together, these findings carry several practical implications for schools using integrated curricula. When addressing the concern of decreasing engagement and fractured delivery in an integrated curriculum the findings from this pilot study could be considered when making policy change. Grading thresholds should not be expected, on their own, to change study behavior; if a threshold is adopted, its likeliest useful role is as a safety net that identifies struggling students, which argues for pairing it with active progress monitoring and, potentially, a requirement that students acknowledge their standing rather than assuming they will be contacted. Second, stand-alone assessments that can heavily impact final grades (e.g., anatomy practicals) are powerful, low-cost behavioral levers; embedding thread content into such assessments is likely to do more for engagement than adjusting pass criteria. We propose another interventional middle ground to this finding. Rather than adjusting pass criteria, schools may do more by attending to how threads are represented within integrated assessment: content students expect to be worth only one or two items is readily deprioritized, so ensuring each thread is assessed in proportion to its intended importance, rather than strict contact hours, could be a considered intervention. This need not mean returning thread content to separate examinations, which would be in direct contrast to the goal of curricular integration, but rather designing balanced integrated assessments so that longitudinal threads are not perceived as inconsequential. Finally, faculty continuity emerges as a high-yield, relatively inexpensive intervention. Threads with inconsistent or rotating instructors were described by students as disjointed and dismissible, thus protecting instructional continuity is a concrete step schools can take to improve the coherence and visibility of their longitudinal threads. Importantly, however, what students valued was not necessarily specific faculty, but content that was delivered reliably with predictable expectations, a consistently effective presentation style, and

high-quality study materials. Stable faculty is thus interpreted as a vehicle for pedagogical consistency. Schools should protect this consistency when implementing curricular threads to improve coherence and visibility of the thread. Finally, the labeling paradox suggests that structurally de-identifying longitudinal content as threads may reduce the cognitive compartmentalization that drives deprioritization.

Limitations

This study has several limitations. It is a single-institution pilot comparing two sequential cohorts. As a result, the quantitative comparison is correspondingly underpowered; the non-significant difference in pass rates cannot support definitive conclusions. The use of a historical (Class of 2027) control introduces the possibility of temporal confounding; although the cohorts were similar on available baseline measures (MCAT and undergraduate GPA), unmeasured differences and concurrent curricular changes cannot be excluded, and the $\geq 70\%$ threshold was applied to control-cohort data only retrospectively, for comparison. The primary quantitative outcome was a composite in-course thread score that does not predict clinically meaningful outcomes that are the intention of the intervention. Five students on decelerated schedules were excluded ($n = 4$ vs. $n = 1$). For the qualitative component, the volunteer interview sample is subject to self-selection and social-desirability bias; control-cohort reflections on the policy were hypothetical and thus vulnerable to recall and speculation effects. Interviews were conducted by peer student-researchers, bringing in another source of bias. The qualitative findings are intended to characterize behavioral mechanisms and student reasoning rather than to be statistically generalizable.

Future directions

Several future directions stem from these findings. First, adding the Class of 2029 upon completion of the pre-clerkship phase would increase statistical power for the pass-rate comparison and allow assessment of whether the observed pattern is stable. Disaggregating A&E assessment by exam type would clarify whether performance is driven by practical examinations or by course multiple-choice items. Thread subject-level analysis could determine if anatomy and embryology content contributed equally to failure rates or if students struggled more with one subject than the other. Also, taking an in-depth dive into the question type within the multiple choice tests and practicals (e.g., recall vs. application questions, osteology, soft-tissue identification) could reveal patterns of performance gaps. The threshold approach could also be examined in other longitudinal threads, such as histology, radiology, pharmacology, and pathology. This would allow for deeper quantitative analysis of the impact of longitudinal faculty.

Conclusion

In this pilot study, a thread-specific minimum passing threshold did not increase student reported engagement or performance in the Anatomy and Embryology curricular thread; although most students were aware that a passing standard existed, few recognized it as a new requirement, and fewer still

altered their studying in response. The factors that actually governed engagement were the proximity and concreteness of assessment, embodied by the anatomy laboratory practical, and the continuity of teaching faculty. For integrated curricula confronting performance and engagement in foundational-sciences, these results suggest that the more promising levers are assessment design and faculty continuity rather than a distal grading threshold imposed on an otherwise unchanged curricular structure.

Declarations

Statements and Declarations

The authors have nothing to declare.

Author Contribution

D.A.T., I.R.B., and V.T.L. conceived and designed the study. D.A.T. and I.R.B. conducted the interviews and performed the thematic analysis. D.A.T., I.R.B., and V.T.L. wrote the main manuscript text and contributed to creation of tables. All authors reviewed, edited, and approved the final manuscript.

Acknowledgement

Dr. Abigail Konopasky, PhD, for her guidance in designing the mixed-methods methodology for this study as well for helping the team select a journal of choice for manuscript submission.

Data Availability

The datasets generated and analyzed for this study are not publicly available because they consist of student educational performance records, provided by the Geisel School of Medicine Office of Assessment, Evaluation and Accreditation, and interview transcripts collected under conditions of participant confidentiality. De-identified data are available from the corresponding author upon reasonable request and with the permission of the Geisel School of Medicine and the Dartmouth College Committee for the Protection of Human Subjects.

References

1. Brauer DG, Ferguson KJ. The integrated curriculum in medical education: AMEE Guide 96. *Med Teach*. 2015;37(4):312–22. <https://doi.org/10.3109/0142159X.2014.970998>.
2. Blood AD, Farnan JM, Fitz-William W. Curriculum Changes and Trends 2010–2020: A Focused National Review Using the AAMC Curriculum Inventory and the LCME Annual Medical School

- Questionnaire Part II. *Acad Med.* 2020;95(Supplement 1):S5–14.
<https://doi.org/10.1097/ACM.0000000000003484>.
3. Schmidt H. Integrating the Teaching of Basic Sciences, Clinical Sciences and Biopsychosocial Issues. *Acad Med.* 1998;73 Supplement (September): S24-S31.
 4. McBride JM, Drake RL. National Survey on Anatomical Sciences in Medical Education. *Anat Sci Educ.* 2018;11:7–14. <https://doi.org/10.1002/ase.1760>.
 5. Hortsch M, Mangrulkar RS. When Students Struggle with Gross Anatomy and Histology: A Strategy for Monitoring, Reviewing, and Promoting Student Academic Success in an Integrated Preclinical Medical Curriculum. *Anat Sci Educ.* 2015;8:478–83. <https://doi.org/10.1002/ase.1519>.
 6. Gribbin W, Wilson EA, McTaggart S, Hortsch M. Histology education in an integrated, time-restricted medical curriculum: Academic outcomes and students' study adaptations. *Anat Sci Educ.* 2022;15(4):671–84. <https://doi.org/10.1002/ase.2127>.
 7. McKeown PP, Heylings DJ, Stevenson M, McKelvey KJ, Nixon JR, McCluskey DR. The impact of curricular change on medical students' knowledge of anatomy. *Med Educ.* 2003;37(11):954–61. <https://doi.org/10.1046/j.1365-2923.2003.01670.x>.
 8. Drake RL, Lowrie DJ Jr, Prewitt CM. Survey of gross anatomy, microscopic anatomy, neuroscience, and embryology courses in medical school curricula in the United States. *Anat Rec.* 2002;269(2):118–22. <https://doi.org/10.1002/ar.10079>.
 9. Drake RL, McBride JM, Lachman N, Pawlina W. Medical education in the anatomical sciences: The winds of change continue to blow. *Anat Sci Educ.* 2009;2(6):253–9. <https://doi.org/10.1002/ase.117>.
 10. Drake RL, McBride JM, Pawlina W. An update on the status of anatomical sciences education in United States medical schools. *Anat Sci Educ.* 2014;7(4):321–5. <https://doi.org/10.1002/ase.1468>.
 11. Husmann PR, Gibson DP, Davis EM. Changing study strategies with revised anatomy curricula: A move for better or worse? *Med Sci Educ.* 2020;30:1231–1243. doi10.1007/s40670-020-00998-x.
 12. Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? *Anat Sci Educ.* 2009;2(5):199–204. <https://doi.org/10.1002/ase.102>.
 13. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ.* 1983;17(3):165–71. <https://doi.org/10.1111/j.1365-2923.1983.tb00657.x>.
 14. Cilliers FJ, Schuwirth LW, Adendorff HJ, Herman N, van der Vleuten CP. The mechanism of impact of summative assessment on medical students' learning. *Adv Health Sci Educ Theory Pract.* 2010;15(5):695–715. <https://doi.org/10.1007/s10459-010-9232-9>.
 15. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol.* 2006;3(2):77–101. <https://doi.org/10.1191/1478088706qp063oa>.
 16. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. *Field Methods.* 2006;18(1):59–82. <https://doi.org/10.1177/1525822X05279903>.

17. Hennink MM, Kaiser BN, Marconi VC. Code saturation versus meaning saturation: how many interviews are enough? *Qual Health Res.* 2017;27(4):591–608.
<https://doi:10.1177/1049732316665344>.