REVIEW   **Open Access**

Alternative grading practices in undergraduate STEM education: a scoping review

Emily L. Hackerson[1*] , Tara Slominski[1], Nekeisha Johnson[2], John B. Buncher[2], Safana Ismael[2], Lauren Singelmann[3], Alexey Leontyev[4], Alexander G. Knopps[5], Ariana McDarby[4], Jonathan J. Nguyen[6], Danielle L. J. Condry[6], James M. Nyachwaya[7], Kathryn T. Wissman[5], William Falkner[1], Krystal Grieger[4], Lisa Montplaisir[1], Angela Hodgson[1] and Jennifer L. Momsen[1]

Abstract

Alternative grading strategies are increasingly popular in higher education, but research into the outcomes of these strategies is limited. This scoping review aims to provide an overview of the relevant research regarding alternative grading strategies in undergraduate STEM and identify gaps in the literature to inform future research. This scoping review was done using the stages as described by Arksey and O'Malley (Int J Soc Res Methodol 8(1):19–32, 2005). The results of this review indicate there is a lack of consensus on the theoretical foundation for the benefits of alternative grading and, therefore, limited validated tools being used to capture these benefits. Additionally, we find that research into alternative grading methods tends to occur in both disciplinary and practice-based silos.

Keywords Alternative grading, Undergraduate education, STEM, Scoping review, Grading strategies, Standards-based grading, Specifications grading, Ungrading

*Correspondence:
Emily L. Hackerson
Emily.hackerson@ndsu.edu
[1]Department of Biological Sciences, North Dakota State University, Fargo, ND 58108, USA
[2]Department of Physics, North Dakota State University, Fargo, ND 58108, USA
[3]Minnesota State University-Mankato, Iron Range Engineering, Hibbing, MN, USA
[4]Department of Chemistry and Biochemistry, North Dakota State University, Fargo, ND 58108, USA
[5]Department of Psychology, North Dakota State University, Fargo, ND 58108, USA
[6]Department of Microbiological Sciences, North Dakota State University, Fargo, ND 58108, USA
[7]Departments of Chemistry and Biochemistry and School of Education, North Dakota State University, Fargo, ND 58108, USA

Introduction

Examining the development of grading in the United States reveals a progression from its informal origins in 1646 to the establishment of the standardized A-F system that characterizes academic accomplishment today (Bowen & Cooper, 2022). Records from the late 1700s indicate early grades were assigned and influenced by social factors such as student socio-economic status (Bowen & Cooper, 2022). With records dating back to 1898, the A-F grading scale was first used to objectively reflect classroom achievement but did not become common until the mid-1900s (Bowen & Cooper, 2022; Schinske & Tanner, 2014). The rising popularity of intelligence testing in the mid-1900s prompted educators to shift away from the previous norm of assigning grades influenced by social factors and instead assign grades

based on their perception of student merit and achievement, often curving to fit a bell curve (Bowen & Cooper, 2022; Feldman, 2018; Schinske & Tanner, 2014). Aside from lowering the threshold for failure from 75 to 60% in the mid-20th century, and adjusting the letter-grade bins accordingly, the traditional A-F grading scheme remains in effectively identical form today with grades of A=90–100%, B=80–89%, C=70–79%, D=60–69%, and F=0–59%. The practice of assigning grades based on this scale has come to be a hallmark of traditional grading practices across higher education.

Traditional grading practices are increasingly critiqued as perpetuating systemic inequities (Feldman, 2018) by conflating the outcome of learning with behaviors in the process of learning (Lipnevich et al., 2020, 2021). Alternative grading practices which include, but are not limited to, standards-based grading (SBG; Lewis, 2022a), specifications grading (Nilson & Stanny, 2023), and ungrading (Kohn & Blum, 2020), all share a goal of more accurately communicating what a student knows and can do (Clark & Talbert, 2023; Nilson & Stanny, 2023; Schinske & Tanner, 2014; Townsley & Schmid, 2020). The popularity of alternative grading practices in undergraduate Science, Technology, Engineering, and Mathematics (STEM) courses, evidenced through dedicated conferences (e.g., *The Grading Conference*), substack newsletters (e.g., *Grading for Growth*), and a multitude of conversations on social media, highlights several needs for faculty in higher education. However, there is confusion over what elements constitute specific alternative grading practices. Alternative grading practices have been given many names with very little universality in definitions or implementation. Additionally, interested instructors are often overwhelmed by the many options for alternative grading, and struggle to systematically implement classroom changes. Finally, there is skepticism that alternative grading practices positively impact student learning. It is this last need—the need for evidence of efficacy—that we address in this scoping review.

An interdisciplinary scoping review

We used an interdisciplinary approach to explore alternative grading practices across undergraduate STEM disciplines for several reasons. First, STEM disciplines are "not a monolith" (Reinholz et al., 2019); disciplinary differences can have profound impacts on classroom instructional practices, including the uptake of new practices. For example, research finds that adoption of evidence-based teaching strategies is not uniform across STEM fields (Lund & Stains, 2015; Shadle et al., 2017; Stains et al., 2018). Second, grading practices also vary extensively across STEM disciplines (Lipnevich et al., 2020). As a result, we might expect to see differences in how STEM disciplines adopt and adapt alternative

grading practices. Such variation has repercussions for students, who must navigate a curriculum that includes a suite of introductory STEM courses, all while making sense of their distinct grading systems. Finally, there is a tendency for discipline-based education research (DBER) to occur in silos, with limited cross-talk across disciplines (Slominski et al., 2023, 2020; Trujillo & Long, 2018). If we are to make *systemic changes* to our grading practices, it is essential to use interdisciplinary approaches, so that we can build a broad consensus about how grading practices impact student learning, and ultimately, whether they result in more inclusive classrooms.

As an interdisciplinary group, the DBER community at NDSU is uniquely positioned to tackle a scoping review exploring alternative grading practices across undergraduate STEM. We are a collaborative community with faculty, post-doctoral researchers, and graduate students from Biology, Chemistry, Engineering, Physics, and Psychology. We are also active practitioners, teaching courses in which we have implemented alternative grading practices. Represented in our community are faculty who teach nearly every introductory science course, which are typically large-enrollment and often gatekeeping, prerequisite courses. We first began exploring alternative grading practices through a book club in 2020 that read *Grading for Equity* (Feldman, 2018). Out of this small group grew an interest in alternative grading, both in practice (i.e., how do we implement this in a large enrollment, first-year course) and research (i.e., what impact do these practices have on student outcomes, both cognitive and affective), across the NDSU DBER community.

Scoping review—a type of literature review

The growing interest in alternative grading approaches in undergraduate STEM education, particularly the calls for evidence of their efficacy, warrants an exploration of existing literature. However, the use of alternative grading practices in undergraduate STEM education is relatively recent; as a result, the literature corpus is limited and disparate, and not conducive to more traditional literature reviews like a systematic review or meta-analysis.

Unlike a systematic review, which is narrowly focused and driven by a well-defined research question, a scoping review is suited to rapidly map or describe the current state of an emerging research field (Arksey & O'Malley, 2005; Khalil et al., 2016). Commonly used in healthcare research, scoping reviews follow a systematic approach to determine the extent of research on a particular topic (Arksey & O'Malley, 2005; Munn et al., 2018). In the present study, we adopt the five stage framework of Arksey and O'Malley (2005), which includes (1) identifying the research question, (2) identifying relevant studies, (3) study selection, (4) analysis, and (5) reporting results.

Scoping reviews report the extent and nature of current research on a particular topic and can be used to clarify key concepts, examine research methods, and identify knowledge gaps in the literature (Munn et al., 2018); however, scoping reviews do not describe the quality of existing research (Arksey & O'Malley, 2005). In our current study, we were particularly interested in gaps in the existing literature on alternative grading. Miles (2017) describes seven categories of research gaps which includes evidence gaps, knowledge gaps, practical knowledge gaps, methodological gaps, empirical gaps, theoretical gaps, and population gaps. An *evidence gap* is indicated when results in the body of literature are contradictory to each other. A *knowledge gap* is when research is not found; evidence does not exist or has not been published. *Practical knowledge gaps* are a gap between what is considered empirically as the most supported action, but this does not translate into practice. A *methodological gap* occurs when a variety of methods are needed to generate new results. An *empirical gap* occurs when findings or predictions have yet to be empirically verified. When the research area lacks an overarching theoretical backing, a *theoretical gap* is present. Finally, a *population gap* is indicated when the current literature lacks representation of a population (e.g. race, gender).

We conducted a scoping review to describe the extent and nature of recent research on alternative grading and the impacts on undergraduate student outcomes (e.g., grades, motivation, etc.) across STEM disciplines, and to identify the types of gaps present in the collective body of literature. This study describes the alternative grading research landscape through three mechanisms: (1) descriptive statistics summarizing the context of existing publications, (2) analysis describing the study characteristics with a focus on the measurements and metrics used, use of any validated instruments, and results, and (3) direct citation and co-citation analyses to understand how publications in this body of work are citing the broader literature.

Methods
Context: developing our collaboration
For the last 15 years, our DBER community at NDSU has hosted a vibrant Journal Club. Though we primarily serve the DBER community, we routinely host faculty from the broader NDSU community who have a growing interest in teaching and learning. All attendees - faculty, postdocs, graduate students, and undergraduate students - contribute to the Journal Club as leaders, facilitators, and contributors. Each Friday during the fall and spring semesters, we gather to discuss contemporary research (whether our own or others), which often focuses on evidence-based pedagogical practices. Over the past several years, we held multiple Journal Club sessions that centered on alternative grading approaches. Given our group's interest in this topic, we decided to allocate a subset of our weekly discussions to initiate this scoping review. While the emphasis of the current work is on the findings from the scoping review, our interdisciplinary approach on this project exemplifies the strengths and advantages that come from this type of collaboration (see Henderson et al., 2017), a point we return to and elaborate on in the Discussion.

During the Fall 2022 semester, we devoted several Journal Club meetings to the scoping review process (Table 1). Our first meeting was faculty-led (AL, JM), with the goal of scaffolding the beginning of the process by exploring different types of literature reviews to confirm that a scoping review was an appropriate tool given our research interests. Subsequent sessions were co-led by faculty and students, who expressed particular interest in later topics associated with the scoping review process. At the end of the Fall 2022 semester, a smaller group of faculty and students formed from those interested

Table 1 Timeline of our collaborative approach to the scoping review

Date	Topic	Description
Meeting 1 (Fa '22)	Choosing our approach	Explore different types of reviews; discussion of why scoping review is most appropriate; brainstorm research questions for our project Resources: Arksey and O'Malley (2005) and Graulich et al. (2021)
Meeting 2 (Fa '22)	Developing our methods	Initial discussion of general criteria (e.g., which disciplines to include); discipline-specific discussion of which databases/journals we will search; brainstorm ideas for where we will store our data
Meeting 3 (Fa '22)	Getting organized	Defining our search terms; how to use Zotero; creating discipline-specific project notebooks
Meeting 4 (Fa '22)	Thematic analysis	Developing familiarity with thematic analysis; practice using thematic analysis Resources: Braun and Clarke (2006)
Meeting 5 (Fa '22)	Taking stock	Reflect on progress; develop plan for finalizing literature searches; discuss status of Zotero library
Meetings 6–11 (Sp '23)	Working groups	Finalize literature search; develop authorship agreement; establish smaller working groups based on individual interest; discuss progress and define goals within working groups on a bi-weekly basis
Meetings 12–15 (Su '23)	Working groups	Discuss progress and define goals within working groups on a bi-weekly basis
Meetings 16–22 (Fa '23)	Putting it all together	Data analysis; figure generation; manuscript preparation

in a deeper and ongoing involvement with the scoping review; participation in this group was voluntary. During the Spring 2023 semester, our group met on a bi-weekly basis outside of Journal Club to make progress on the scoping review. At each meeting, smaller working groups discussed their progress over the previous two weeks and outlined goals for the upcoming two weeks. These meetings also provided the opportunity for working groups to bring questions and seek input from the larger group. During the Summer of 2023, we continued to make progress in our smaller working groups (focused primarily on thematic analysis, direct citation/co-citation, and manuscript preparation) and met on a biweekly basis. During the Fall 2023 semester, we met biweekly and our focus shifted to data analysis, figure generation, and writing the final manuscript.

Our group also created and signed an Authorship Agreement (*Supplementary Materials*), which was intended to help us establish and maintain clear expectations regarding authorship. Authorship was based on criteria described in and derived from the International Committee of Medical Journal Editors (2023).

Scoping review
Our scoping review followed the five stages described in Arksey and O'Malley (2005) and meets the items from the PRISMA checklist for scoping reviews (Peters et al., 2020; Tricco et al., 2018). We briefly describe each stage as it pertains to our study.

Stage 1: identifying the research question
Alternative grading is both an emerging practice in STEM education and a developing research area in DBER. Early discussions in our group (Table 1) centered on determining our research question. Through iterative discussion, we developed two initial guiding research questions: (1) what is currently known about the impacts of alternative grading practices on student outcomes across STEM disciplines, and (2) what gaps currently exist in the literature. As our scoping review progressed, we identified a third research interest, namely if the research on alternative grading in STEM was occurring in discipline- or methods-based silos. We recognize 'alternative grading' is a broad term, one we chose to specifically encompass the diversity of grading practices faculty are currently adopting (e.g., specifications grading, ungrading, standards-based grading, etc.). We also note the uptake of alternative grading practices is uneven across disciplines, hence our need to use a broad and encompassing term.

Stage 2: identifying relevant studies
As with systematic reviews and meta-analyses, we searched a variety of sources to answer our guiding research questions. In Fall 2022, databases and journals were searched by the disciplines of the graduate students and faculty in our group: Biology, Chemistry, Engineering, Physics, and Psychology (Supplementry Materials Table 1). Some of these sources also indexed published conference proceedings, while others did not. After discussion, we decided to include peer-reviewed conference proceedings as they are reflective of the current state of the research for several disciplines.

As a community, and after a discussion with our Science Librarian, we developed a set of search terms related to alternative grading (Supplementry Materials Table 2). Each keyword was combined with our focal disciplines using Boolean operators (e.g., "standards-based grading AND biol*"). This initial search yielded 467 records (Fig. 1). Duplicate records were removed, leaving 332 distinct records. These 332 records underwent blind inclusion sorting (Ouzzani et al., 2016) using Rayyan, a freely available, web-based software designed to support collaborative literature reviews. If the necessary details were not present or clear in the record abstract while applying the inclusion criteria at this stage, records underwent a full-text review.

Stage 3: study selection
After an initial review of our 332 records, we recognized many studies were not relevant to the goal of our scoping review, which was to characterize the landscape of the empirical research on alternative grading. We thus limited ourselves to data-driven peer reviewed studies, and omitted essays and opinion pieces. We also excluded entire books, theses, and dissertations because we wanted research that was widely available and, importantly, peer reviewed. We also chose to limit our scoping review to studies in formal learning environments from the undergraduate level because our community at NDSU is focused, and has expertise specifically on DBER at that level. Additionally, we limited our studies to those completed at schools in the United States because grading practices differ substantially in the U.S. from those in other countries. Finally, based on an early exploration of the literature, we did not limit the publication dates of our search, resulting in a corpus including studies as early as the 1970s, with most being published after 2014. While the norms of teaching and learning in higher education may have evolved since the 1970s, what we now consider to be traditional grading practices have not changed since they were first introduced in the 1890s. Therefore, alternative grading practices have been "alternative" since well before the 1970s. Further, the alternative practices presented in early studies are similar in motivation and execution to those presented in the studies published later, so we opted to include all studies (regardless of publication date) that met our inclusion criteria to avoid biasing our sample while also maximizing the number of
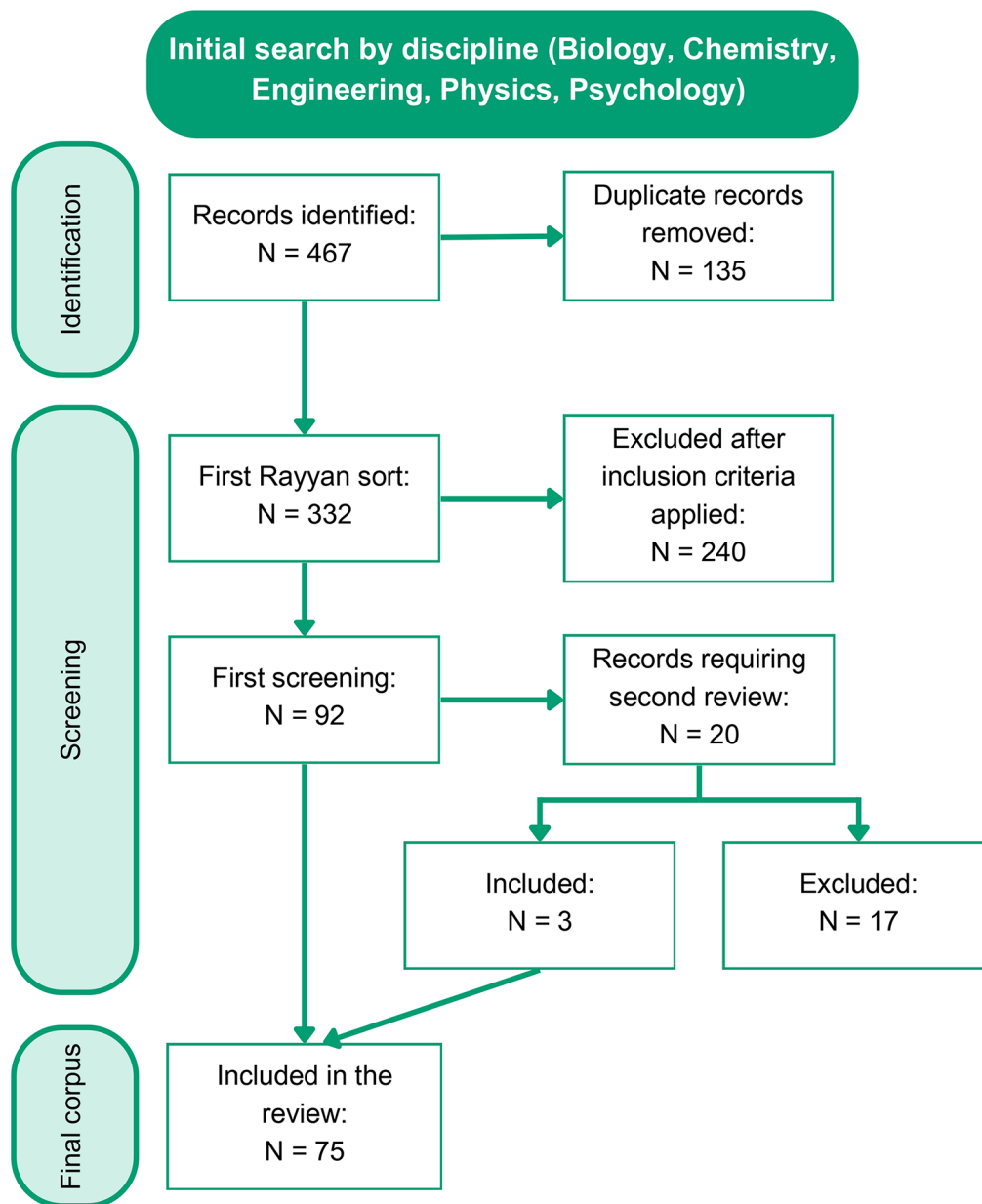
Fig. 1 Flow diagram of the study selection process

included studies. Indeed, both earlier and later studies published on this topic are relevant to our goal of characterizing the landscape of the research into alternative grading practices. After applying the inclusion criteria (see Table 2) there were 92 records.

Upon a deeper reading of the records during analysis, we identified 20 records that required further review due to potentially not fully meeting our inclusion criteria. Some records included reflective manuscripts with no data collected and it was unclear if they had undergone peer review. After multiple rounds of discussion with both the full group and coding teams (see Stage 4), we

Table 2 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Empirical papers	Essays and opinion pieces, Grey literature
Undergraduate STEM focal population	K-12, graduate, and professional school focal population
Formal learning environment	Informal learning environments
U.S. institutions	Institutions outside the U.S.

excluded 17 records. We also note our search returned two studies in pre-print. One of those studies (Lengyel et al., 2023) was published during our analysis in August 2023 and was therefore included. The other pre-print

has not been published at the time of writing in a peer-reviewed journal and was one of the 17 not included. The final number of studies included in our data corpus for this scoping review was 75 (Supplementry Materials Table 3).

Stage 4: analysis of chosen studies—charting the data

We identified two types of coding, study context and study characteristics. We also intended to conduct thematic analysis to characterize the motivations, theoretical framework, and implementation of alternative grading in each study. Unfortunately, we found few studies included sufficient details to enable such an analysis. As a result, we were limited in our ability to make any meaningful thematic conclusions about the corpus as a whole, and thus abandoned this aspect of our analysis.

To characterize the context of the studies included in our data corpus, we coded each study for course delivery, type, audience, discipline, enrollment, name of the alternative grading practice used, and Carnegie classification of the institution where the study took place (Table 3). At this stage of the scoping review process, our coding returned additional disciplines beyond our search terms from Stage 2, including geology, computer science, and mathematics.

Each study was initially coded for context by two independent coders (either NJ, AK, JJN, JMN, WF, KG). Following this first round of coding, a subset of coders (NJ, AK, JJN) met to compare codes and flag disagreements. These disagreements were identified, reviewed, and assigned to a third independent coder for further review; all disagreements were discussed asynchronously via Slack until consensus was reached.

A separate coding team (ELH, TS, AM) coded the characteristics of each study included in our data corpus. Specifically, we focused on describing the variables reported in each study, identifying seven categories or types of variables (Table 4) and the tools used in the research. Performance variables included measurements of perceived student performance, performance in subsequent classes or throughout a program, student grades (e.g., GPA, course grade, exam grades), and course-level grade measures (i.e. D/Failure/Withdraw or DFW rate and grade distribution). Some variables with theoretical foundations included affective construct variables such as anxiety, mindset, motivation, self-efficacy, etc. Attitude variables assessed less specific student beliefs about alternative grading (e.g. whether students liked the alternative grading practice) and included course evaluations. The Learning code was assigned to studies that compared end of course performance to initial performance. Retention was also reported in some studies, and these measures were coded as Retention.

Table 3 Study contexts, context codes, and number of studies in each code

Context	Context codes	n
Course delivery	In person	58
	Other/unknown	11
	Hybrid	5
	Online	1
Course type	Lecture	50
	Lab	9
	Other/unknown	16
Course audience	STEM majors	55
	General education	14
	Other/unknown	6
Course level	Introductory	60
	Upper division	9
	Other/unknown	6
Course enrollment	Large (60+ per section)	34
	Medium (20–60 per section)	26
	Small (0–20 per section)	7
	Other/unknown	8
Institution type (current classification)	Doctoral universities—very high research activity	29
	Doctoral universities—high research activity	10
	Master's colleges and universities—larger programs	10
	Other/unknown	10
	Baccalaureate colleges: arts & sciences focus	8
	Master's colleges and universities—smaller programs	4
	Baccalaureate colleges: diverse fields	2
	Doctoral/professional universities	1
	Master's colleges and universities—medium programs	1
Discipline	Engineering	30
	Chemistry	21
	Mathematics	9
	Physics	7
	Biology	3
	Geology	2
	Psychology	2
	Computer Science	1
Name of alternative grading practice[a] (as used in the study)	Standards based grading (SBG)	17
	Mastery	16
	Specifications	14
	Other	7
	Contract	5
	Keller	5
	Competency	4
	Pass/fail	5
	4.0 scale	2

[a]Within the "Name of Alternative Grading Practice" context, practices found in only one study were collapsed into "Other" (Cafeteria, Criterion, Do It Right (DIR), Multiplier, Outcome-Based, Portfolio, and Ungrading)

There were also a subset of studies that reported information about the instructor experience when

Table 4 Study characteristics

Variable	Variable type	n
Performance (n = 74)	Student performance (GPA, exam scores, and final course grades)	46
	Course level grade distribution	17
	Student self-reported performance	6
	Course-level DFW rate	3
	Long-term performance (performance in subsequent courses—GPA)	2
Affective constructs (n = 20)	Anxiety	5
	Motivation	5
	Self-efficacy	4
	Mindset	2
	Communication apprehension	1
	Goal orientation	1
	Interest in science	1
	Perceptions of control over grade	1
Attitudes (n = 51)	Student attitudes	49
	Student satisfaction with grading system	2
Learning (n = 7)	Content knowledge	5
	Skill development	2
Retention (n = 6)	Retention (in a program)	6
Other (n = 10)	Frequency of retests, Grade inflation, Student behaviors, Student rate of progress, Student use of instructional components, Student-faculty interaction, Students' self-assigned grade, Time prepping for assessment	10
Instructor measures (n = 11)	Faculty perceptions	8
	Graduate TA perceptions	3

implementing alternative grading practices. The Instructor Measures code captures faculty or graduate TA perceptions of their time investment and other experiences with alternative grading.

We (ELH, TS, AM) also coded each study for the instrument or tool used to measure the variables identified in Table 4. Surveys developed by the authors of a study were coded as researcher-generated surveys (RGS) and data collected through a focus group or interview was coded as FG/I. When the survey or instrument used in a study was from previously cited work it was coded as a Validated Tool. Tools for reporting grades included final grades, GPA, and exams. A code of 'Gradebook' was assigned when course-level DFW rates and grade distributions were measured. When no instrument was reported in a study, we coded it as 'None'.

Finally, we coded the findings reported by each study. Given that each study could have multiple variable types, the results of each variable type were coded separately. Regardless of the variables measured or tools used, we were interested in whether the research outcomes supported the efficacy of alternative grading practices on student learning. A result was coded as Positive when there was statistical significance indicating a positive impact of

alternative grading on a given variable of interest; a Negative code was used when the statistical analysis indicated a statistically significant negative impact. Trending Positive or Trending Negative codes were used when studies reported a trend in favor of or against alternative grading, but that trend was not statistically significant. Studies that present the results of a variable being both positive and negative (e.g. some students liked the alternative method and some did not) were coded as Mixed. Studies that reported no trend in either direction were coded as Neutral.

Each study was coded for these characteristics by two independent coders. Following this coding, ELH, TS, and AM met to compare codes. Disagreements were identified and discussed until consensus was reached.

Citation network analysis

As mentioned previously, STEM disciplines are "not a monolith'" and disciplinary differences inevitably manifest themselves in instructional practices. This can manifest not only in what grading practices a discipline values, but also how they communicate with each other and how they may approach trying to "solve" the problem of traditional grading in the classroom. With alternative grading practices being relatively new, we were interested in whether all of the STEM disciplines we sampled were citing a common body of literature on alternative grading practices and whether they were building on work across disciplines. To answer this question, we conducted both direct citation and co-citation analyses of the records in our data corpus. This allowed us to gauge whether disciplines are building on the same foundational knowledge of these methods and whether there are shared practices across disciplines.

The data set for the network analysis started with PDFs from the full data corpus identified in stage 3 of the scoping review. These PDFs were scanned for references using Scholarcy (Gooch, 2021) to create a database of reference papers cited by the corpus papers. Each entry was manually checked and corrected for accuracy and completeness (NJ, JB, LS, JJN, DLJC, LM). A matrix was created where each column referred to a paper in the data corpus, and each row corresponded to a paper cited by a paper in the corpus. A "1" was entered into the corresponding cell if the reference was cited by the paper in a given column, and a "0" was entered otherwise.

This matrix was then converted into a direct citation network file and a co-citation network file using Python. In the network files, each row and column corresponds to a paper in the corpus. In the direct citation network file, a "1" was put into a cell if the paper in the row paper cited the column paper. For example, if cell [8, 40] has a "1" in it, that means paper 8 cited paper 40. In the co-citation network file, the value in each cell corresponds to the

number of commonly cited papers between the row and column papers. For example, if cell [3, 12] has a "5" in it, that means papers 3 and 12 shared 5 citations. These network files were then put into Gephi (Bastian et al., 2009) for visualization. Colors and shapes were used to visualize both disciplines and alternative grading practices.

To determine if disciplines or alternative grading practices were citing only within their same discipline or grading practice, the statistical significance of each discipline and grading practice community was calculated using the methods presented in He et al. (2021). This method considers the weights between items within a discipline or grading-type community and compares it to the weights between the items within the community and items outside the community. It then uses a significance testing approach to determine if the community is statistically significant ($p < 0.01$).

Results
Stage 5: reporting results
Descriptive overview
Our scoping review identified 75 studies (Fig. 1), including 44 peer-reviewed journal publications and 31 published conference papers (Table 2). The studies range in publication date from 1970 to 2023, with a majority of studies having publication dates of 2016 and later (Fig. 2).

Study context
Studies from Chemistry (n=21) and Engineering (n=30) made up 68% of our data corpus. Biology (n=3), Computer Science (n=1), Geology (n=2), Mathematics (n=9), Physics (n=7), and Psychology (n=2) comprised the remaining 32% (Fig. 3). The 21 publications in Chemistry were all published in peer-reviewed journals, while 28 of the 30 publications in Engineering were published in conference papers (Table 5). Engineering comprised almost the entirety of the 31 studies we found in published conference papers, with the remaining 3 coming from Physics (2 conference papers) and Computer Science (1 conference paper).

Standards-based grading was the most commonly identified alternative grading practice (n=18) followed closely by Mastery grading (n=16) and Specification grading (n=14). There were seven grading strategies that appeared only once in our corpus: Cafeteria, DIR, Multiplier, Outcome based, Portfolio, and Ungrading. The Keller Method (n=5) and the 4.0 scale (n=2) are the only grading methods to only occur in a single discipline, Chemistry and Physics, respectively, and appear in more than one study. Engineering had the largest number of alternative grading practices (n=6), followed by Chemistry (n=5) (Fig. 3).
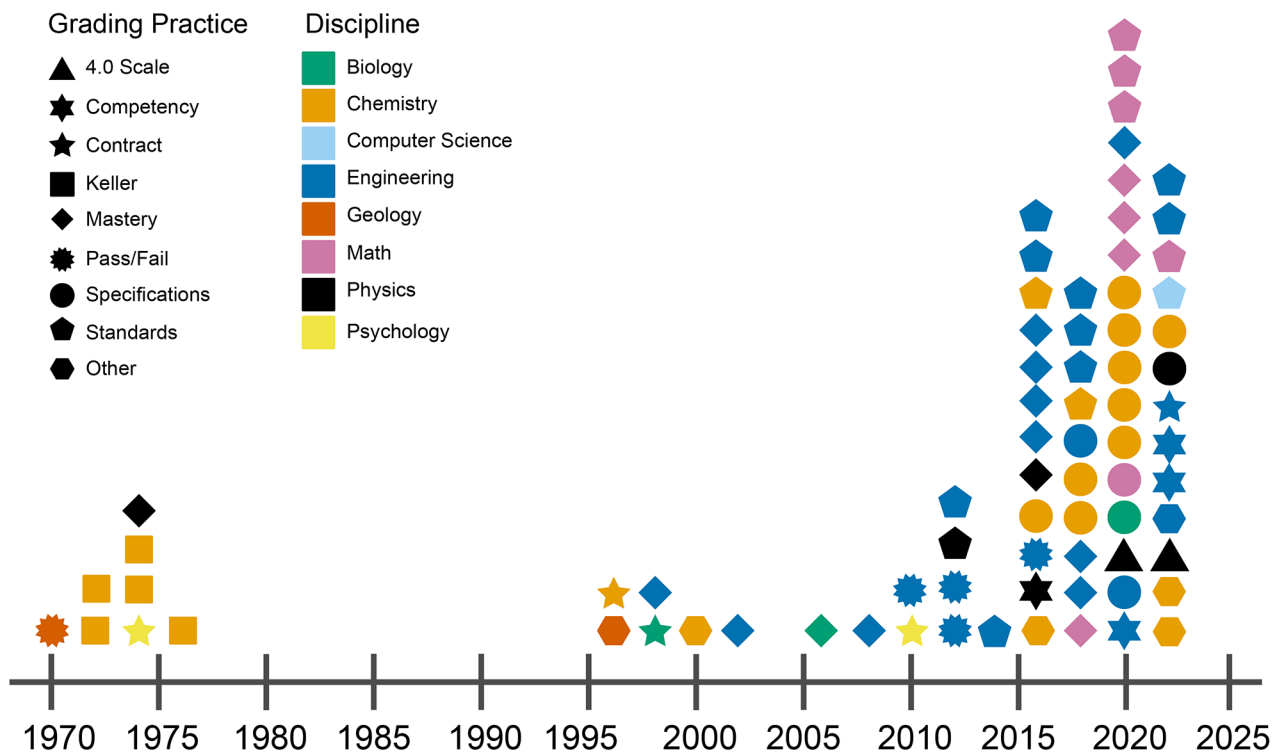

Fig. 2 The publication timeline of studies included in the review. Color indicates the discipline represented in the study. Studies are assigned a shape based on their alternative grading practice, with "Other" being any grading practice with only one study (Cafeteria, Criterion, DIR, Multiplier, Outcome-Based, Portfolio, and Ungrading)
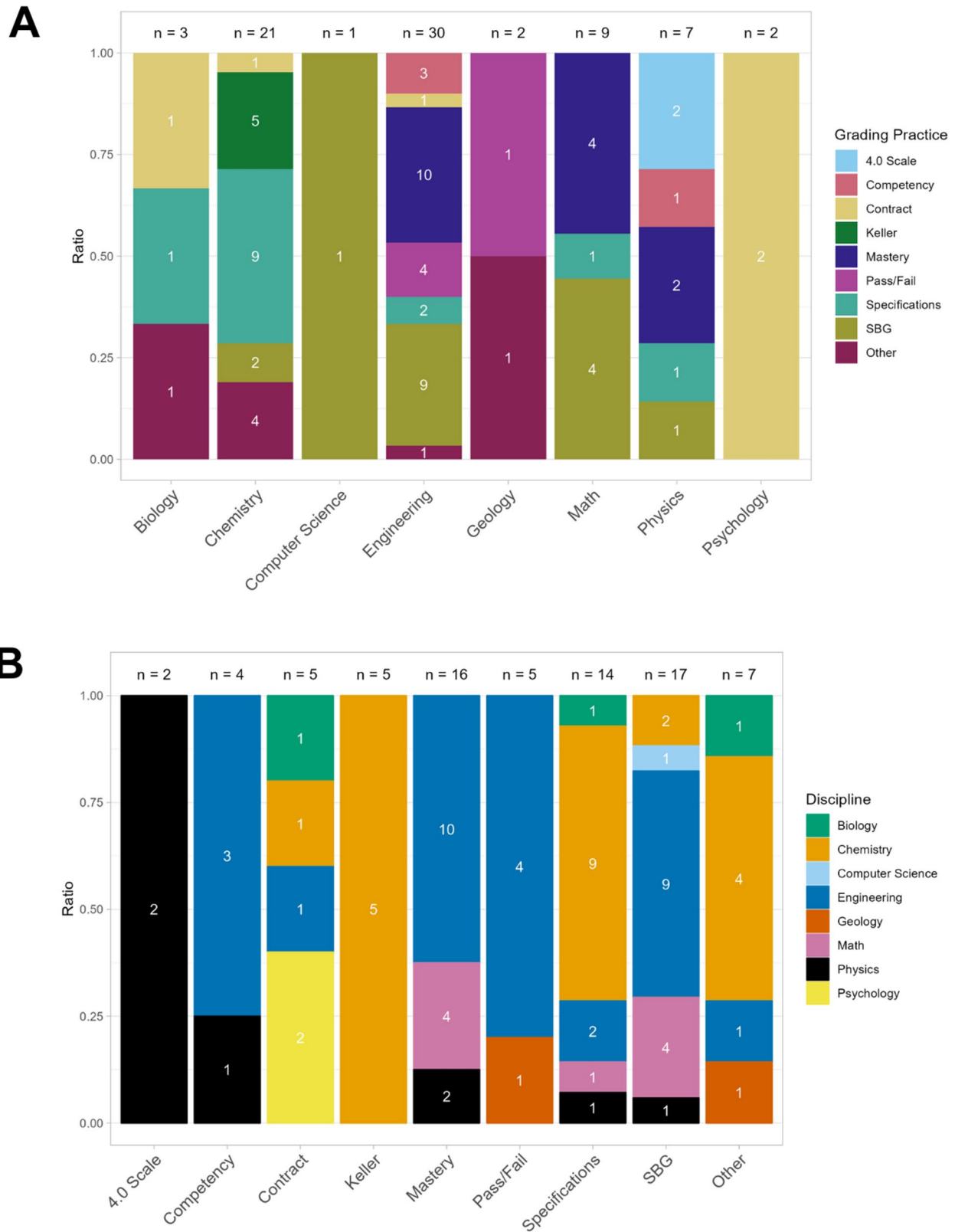
Fig. 3 (**A**) Ratios of alternative grading strategies identified in the studies in our data corpus, broken down by discipline. Studies are assigned a color based on their alternative grading practice with "Other" being any grading practice that had only one record. (**B**) Ratios of disciplines identified in the studies in our data corpus, broken down by alternative grading strategy. Studies are assigned a color based on their discipline

Table 5 Publication type broken down by discipline

	Journal article	Conference paper	Total
Biology	3	0	3
Chemistry	21	0	21
Computer Science	0	1	1
Engineering	2	28	30
Geology	2	0	2
Mathematics	9	0	9
Physics	5	2	7
Psychology	2	0	2
Total	44	31	75

Course enrollment was reported in 67 out of the 75 studies (Fig. 4), which were binned as Small Enrollment (<20 students, n=7), Medium Enrollment (20–60 students, n=26), and Large Enrollment (>60 students, n=34) (Fig. 4). Within each size category there were at least 5 different alternative grading practices used, and no single alternative grading practice emerged as dominant. The most common alternative grading practice reported in large enrollment courses was mastery grading (n=8 of 36); standards-based grading was most common in medium (n=6 of 23) and small (n=3 of 8) enrollment courses.

Of studies that reported course delivery mode (n=64), the majority (n=58) represented in-person courses, with a few (n=5) being hybrid delivery; only one was online. Two-thirds of the studies (n=50) reported the courses as lectures, while only a handful (n=9) were reported as labs. A majority of studies (n=55) also reported serving mostly students in a STEM major. Studies were overwhelmingly from the introductory level (n=60). Additionally, a large proportion of these studies were from doctoral universities with very high (n=29) or high (n=10) research activity and larger master's colleges/universities (n=10).

Study characteristics

We identified 179 variables measured across the 75 studies in our data corpus, indicating most studies measured multiple variables. The most common variable reported in our corpus was Performance (n=74 of 179), which was reported via final course grades (n=25), an exam grade (n=15), GPA (n=3), student self-reported performance (n=4), performance in subsequent courses (n=2), a
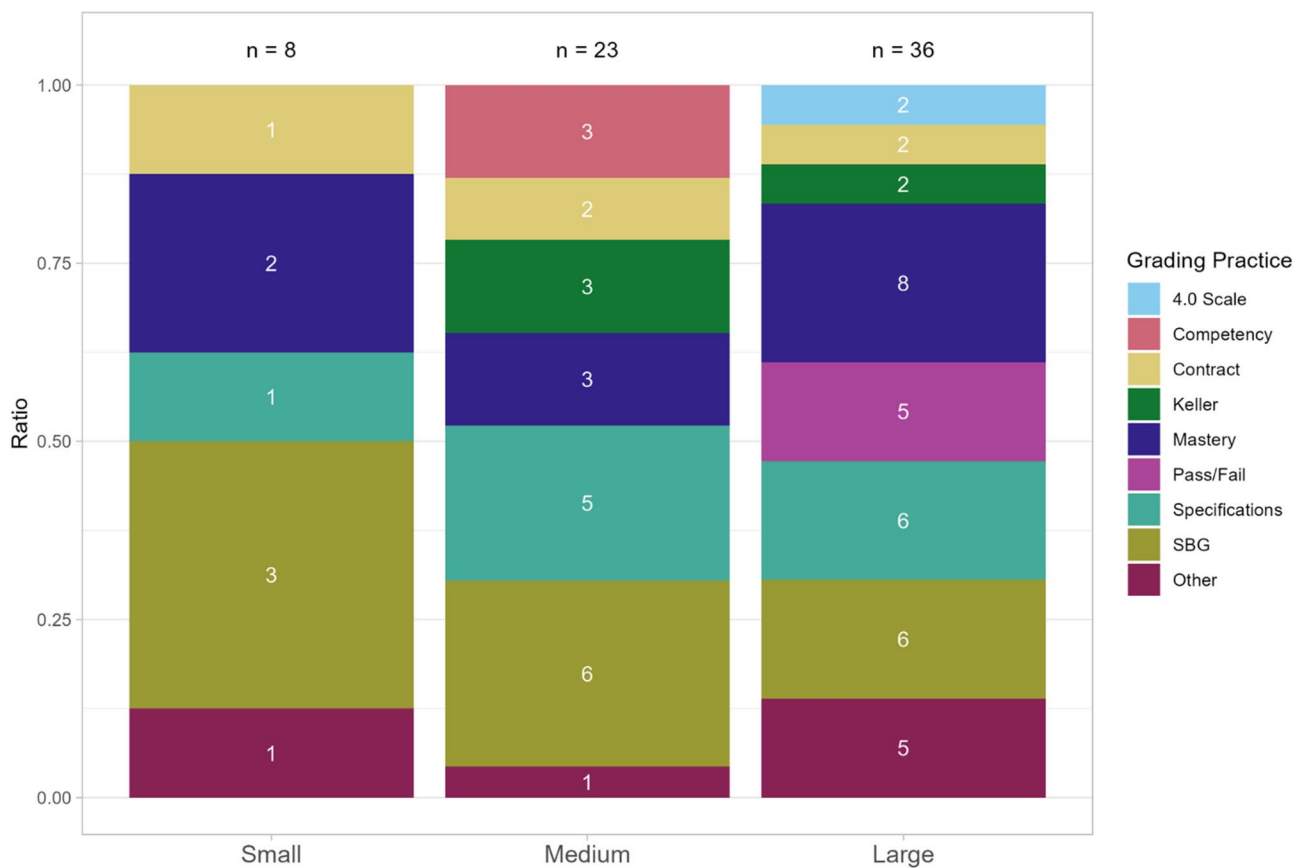

Fig. 4 Ratios of alternative grading strategies identified in the studies in our data corpus, broken down by enrollment size. Records with less than 20 students per section were classified as "Small", between 20 and 60 students were classified as "Medium", and greater than 60 students classified as "Large". Studies are assigned a color based on their alternative grading practice with "Other" being any grading practice that had only 1 record

validated tool (n=2), or another classroom artifact measure (e.g. learning objectives; n=2). The remaining 21 instances of the performance variable represent course-level DFW rates and grade distributions (Fig. 5).

Another commonly reported variable was general student Attitudes towards the alternative grading strategy (n=51 of 179). Student attitudes were largely captured using researcher-generated surveys (RGS; n=36), but were also captured through Course Evaluations (n=8), focus groups or interviews (FG/I; n=3), and in one instance, a Validated Tool (Fig. 5). The remaining studies reporting on attitudes did not specify an instrument or tool.

We were interested in the use of validated tools because these studies could support future research looking to compare the impacts of alternative grading practices across contexts. Validated tools were used to capture 17 variables across 10 studies (Supplementry Materials Table 4). Concept inventories were used to characterize

learning gains in two studies (Fig. 6). Those concept inventories were the Force Concept Inventory (FCI) and the Strength of Materials Concept Inventory (SMCI) from physics and engineering, respectively. Standardized tests were also used to capture student performance by two studies and include the Dunning-Abeles Physics Test and the American Chemical Society (ACS) exam, each used by a single study. Surveys were typically used to characterize affective constructs, but only Dweck's Implicit Theory of Intelligence Scale (Dweck 3; Dweck, 2006) was used in more than one study, and even then, it was used in only two studies. An additional nine validated surveys were each used only in one study.

The studies included in our data corpus largely reported findings that found positive impacts of alternative grading practices on student learning and attitudes. Across the 75 studies and 179 variables, we coded 106 outcomes as Positive and 38 as Trending Positive (Fig. 5). Variables measured through RGS were the single largest
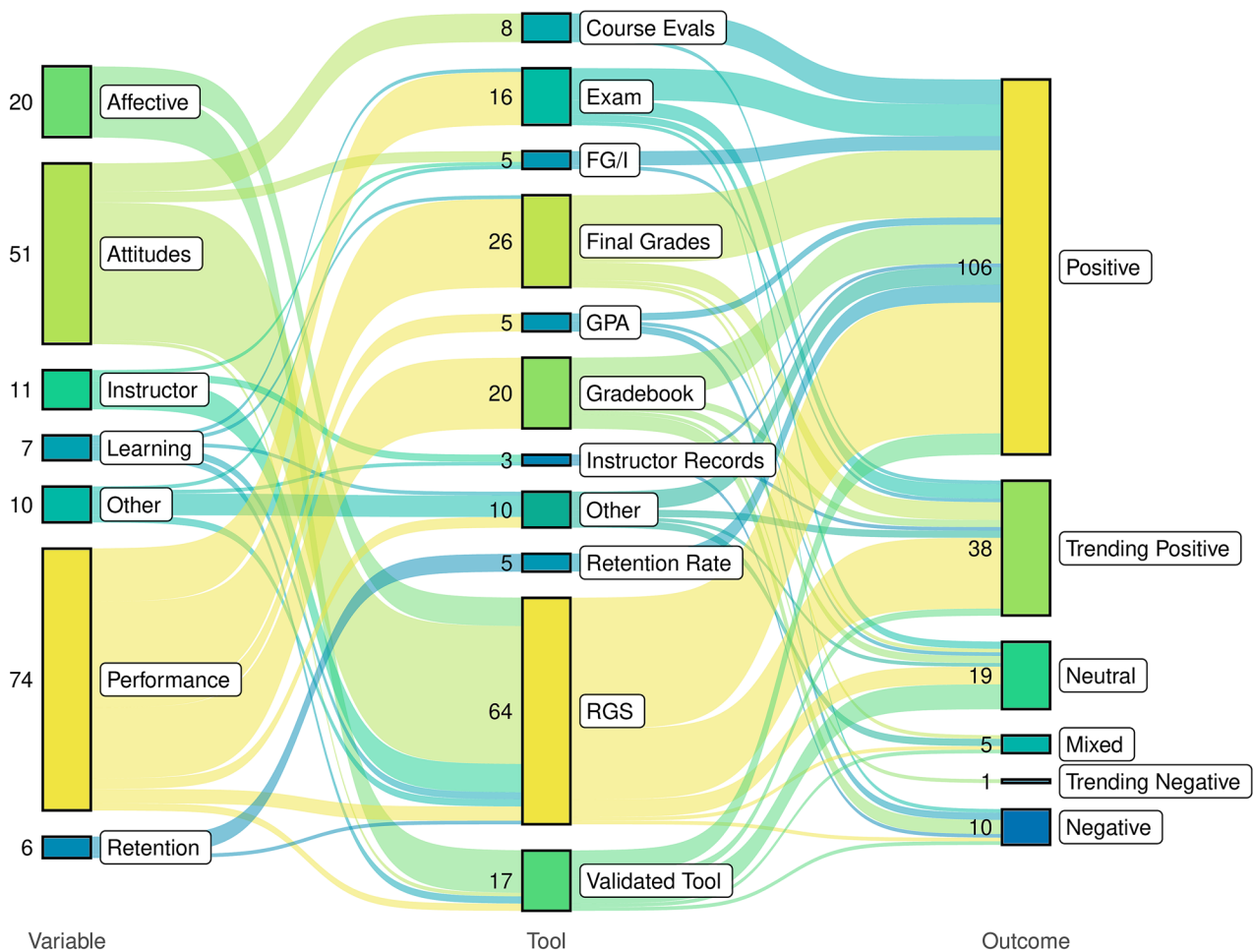

Fig. 5 Sankey diagram illustrating outcome measurements (N = 179), tool, and whether the results are in favor of (Positive) or against (Negative) alternative grading. Measurements and tools with 2 or fewer occurrences were grouped as "Other"
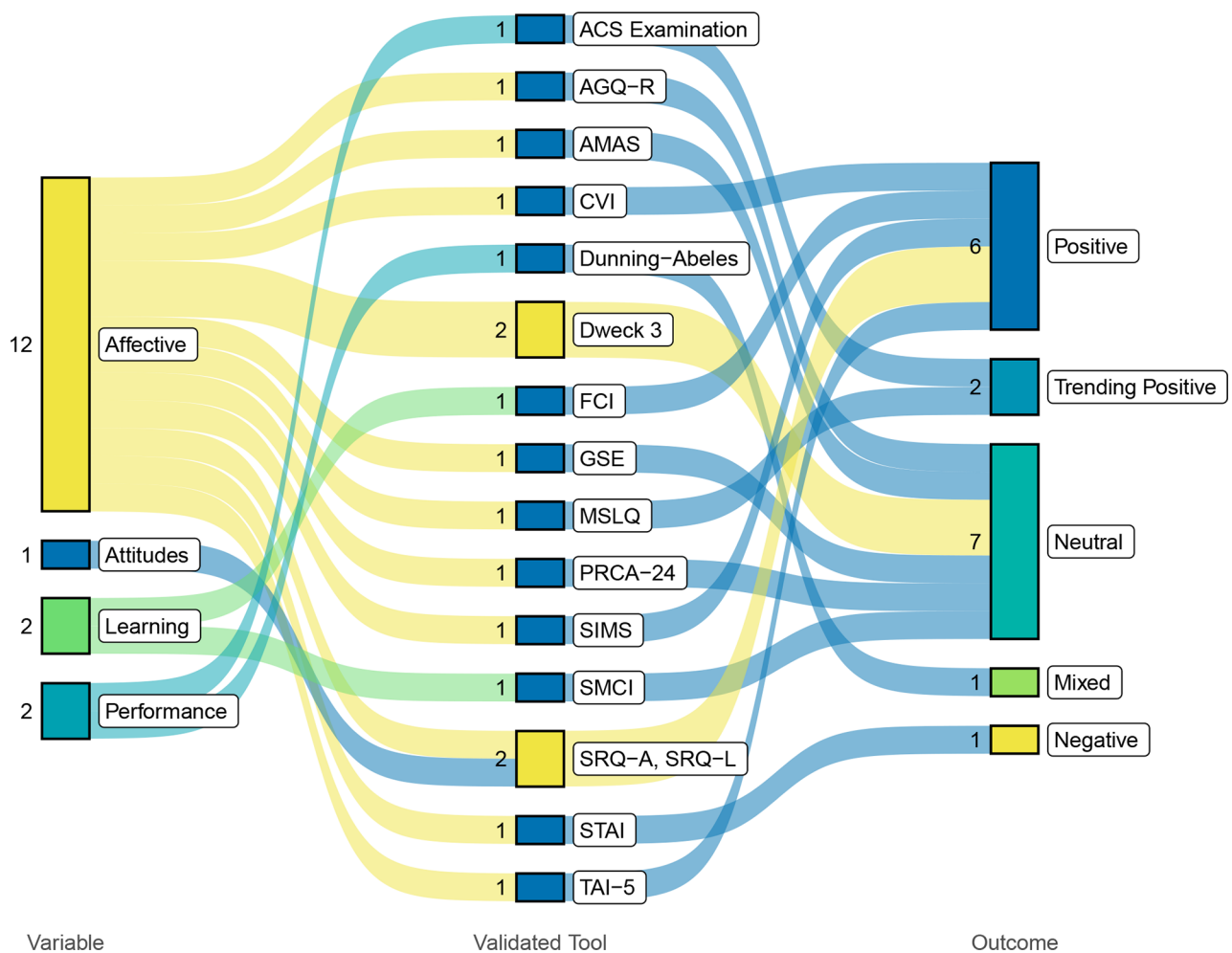
Fig. 6 Sankey diagram illustrating the outcome measurement (n = 11), specific validated tool, and whether the results are in favor of (Positive) or against (Negative) alternative grading

contributor to Positive results (n=35). Results from validated tools were mostly Positive (n=6) or Neutral (n=7) (Fig. 6).

There were ten outcomes that were coded as Negative, where alternative grading practices had a statistically significant negative association with learning outcomes and attitudes. These negative results predominantly came from performance measures, specifically course-level DFW rates and grade distributions (n=4), impacts on GPA (n=2), and exam scores (n=1). Negative results also came from measures of skill development through an RGS (n=1) and student anxiety measured with a validated survey (n=1).

Citation analysis

To determine whether and how the studies in our data corpus were citing one another, we created a direct citation network of the data corpus (Fig. 7). Each study was given a shape corresponding to the alternative grading method used and a color corresponding to discipline. The network representation shows that many citations occur within a discipline and/or grading practice. Of the 90 direct citations that occurred within the dataset, 48 citations had the same discipline and alternative grading practice, 16 had the same discipline with different grading practices, 17 had the same grading practice with different disciplines, and 13 had neither the same discipline nor grading practice – suggesting that both alternative grading method and discipline play a large role in who is citing who in the alternative grading literature.

To further extend this citation analysis to explore if disciplines and alternative grading practices cite a similar body of literature as a whole, co-citation networks of the studies were created. Significance testing of the alternative grading practice communities shows that SBG, mastery, specifications, pass/fail, Keller, and contract grading are statistically significant independent communities that each cite consistent but different bodies of
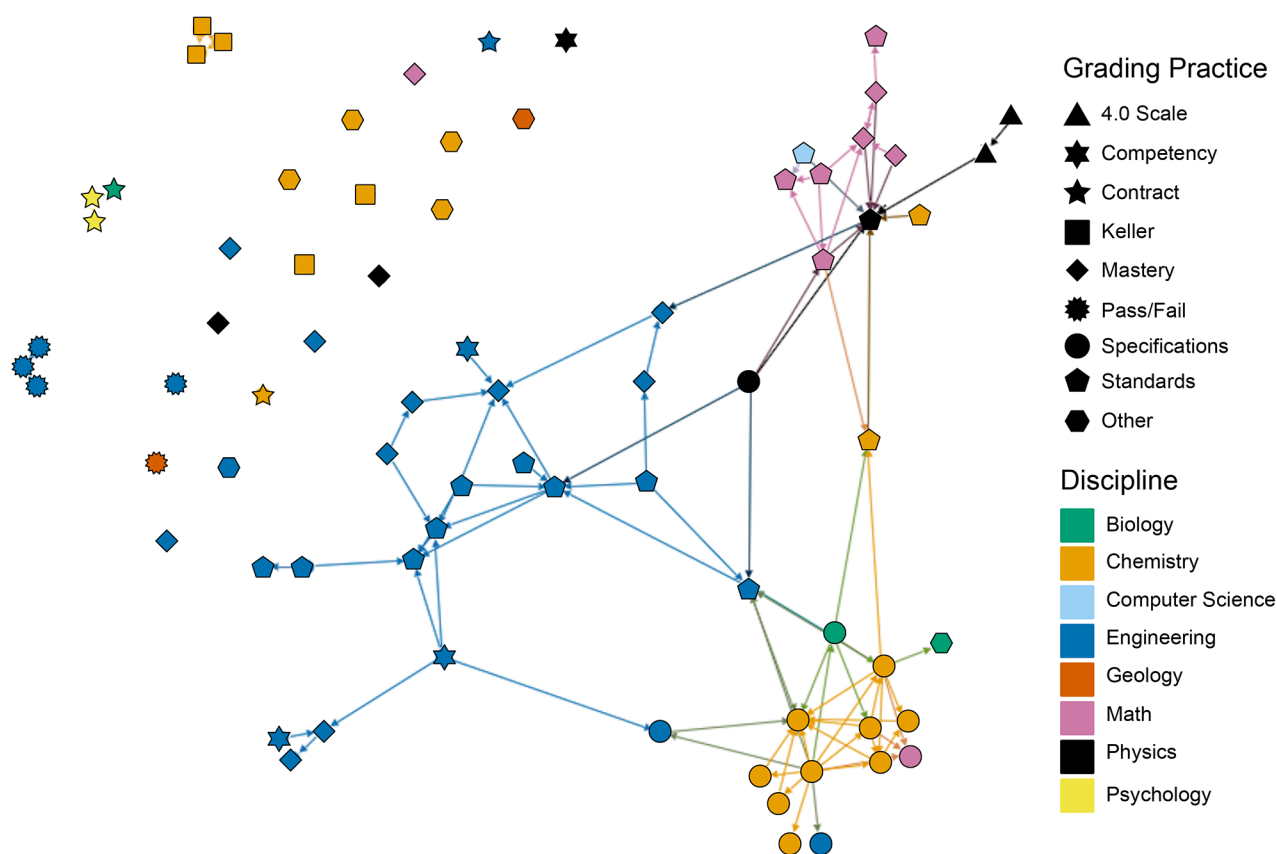
Fig. 7 Direct citation network illustrating what records cited one another. Each node in the network represents a study in the data corpus, and the directional edges represent which records have cited each other. Studies are given a shape based on the alternative grading practice and a color based on their discipline

literature ($p<0.01$; Fig. 8). Similarly, engineering, chemistry, mathematics, and physics are statistically significant independent communities that each cite consistent but different bodies of literature ($p<0.01$; Fig. 8). Other communities have less than five members, which may explain why those communities were not detected as statistically independent, either due to limited statistical power to detect differences in smaller communities, or the small community size forces its members to cite across communities.

Not surprisingly, studies that use the same alternative grading practice are more likely to cite one another. However, both the direct citation and co-citation analysis show that discipline also plays an equally important role in describing who is citing who within the alternative grading literature.

Discussion

Alternative grading practices are increasingly popular in STEM classrooms, yet as our scoping review documents, empirical evidence supporting their efficacy on learning outcomes is currently limited. We further find a fragmented landscape with inconsistent terminology, a dizzying array of variables studied, and limited theoretical underpinnings.

Describing the landscape of research on alternative grading practices

Our initial research question sought to describe what we currently know about the impacts of alternative grading practices on student outcomes across STEM disciplines. Unfortunately, we struggled to answer this question beyond superficial findings, namely that most studies do find a positive effect of alternative grading practices on student learning and attitudes. Two factors impeded our analysis. First, the studies in our data corpus had limited connections to theoretical frameworks (discussed in more detail below) and second, these studies used a wide array of tools to measure learning and attitudes. Together, these factors made it impossible to thematically code the studies and limited our ability to more fully generalize the current state of the research on alternative grading practices in STEM.
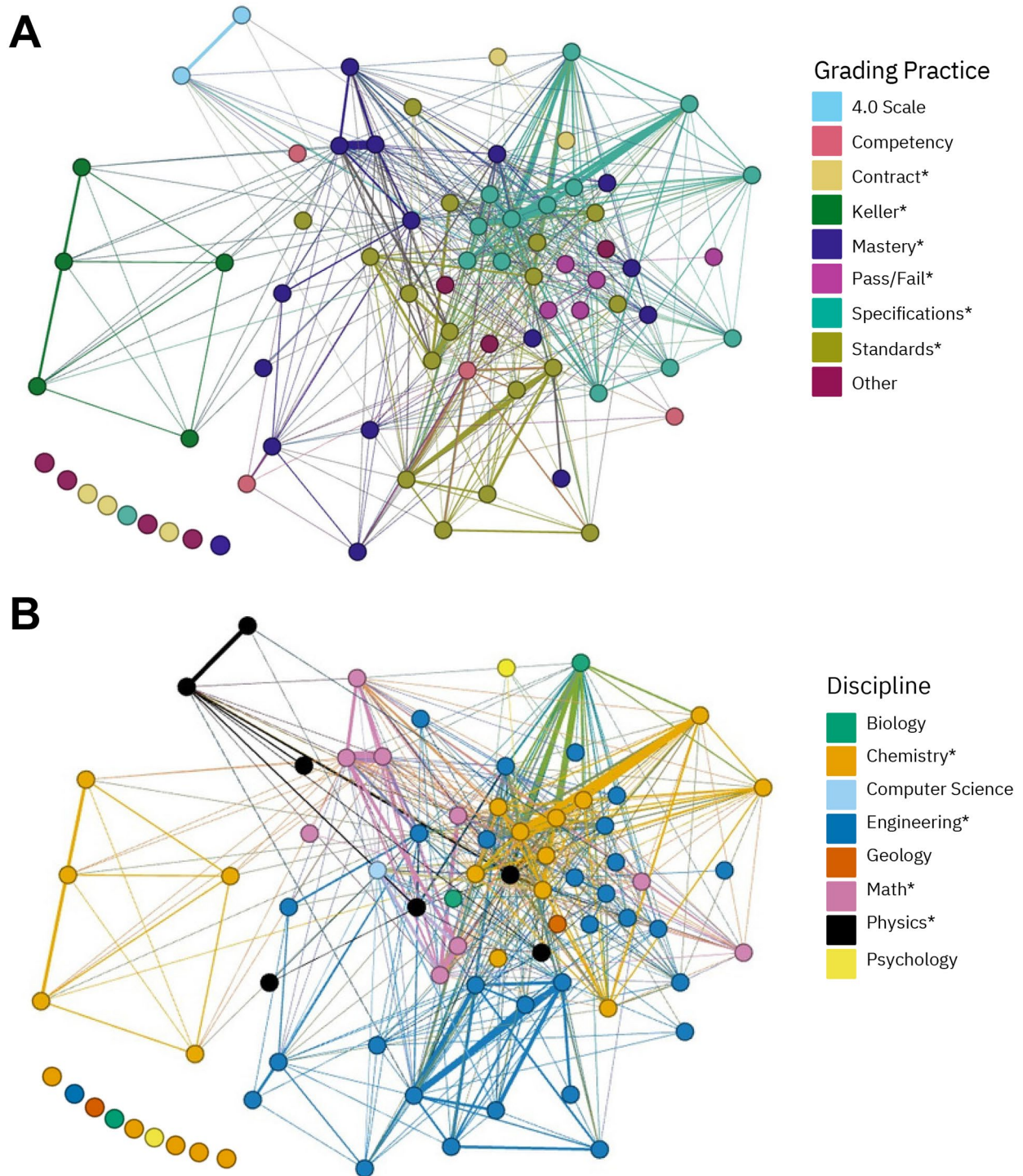
Fig. 8 Co-citation networks illustrating the citations shared between studies in the data corpus and the presence of communities based on alternative grading practice or discipline. Each node in the network represents a study in the data corpus, and the thickness of the edges represents the number of shared citations between two nodes. Studies are colored by (**A**) alternative grading practice and (**B**) discipline. Groups with an asterisk (*) represent statistically significant communities ($p < 0.01$)

Identifying gaps in the research

An important outcome of our scoping review was the identification of gaps in the literature. Guided by the types of gaps described by Miles (2017), the authorship team discussed and identified three types of gaps found in the alternative grading practices research.

Knowledge gap

A knowledge gap is indicated when there is a lack of research or a lack of research with desired measures (Miles, 2017). Given that grades and grading are omnipresent in higher education and have immense influence on students' undergraduate and professional careers (Feldman, 2018), we argue the limited number of studies identified in this scoping review is evidence of a knowledge gap: we lack necessary empirical research investigating the impacts of alternative forms of grading on learning in undergraduate STEM courses.

While the research available on alternative grading is still limited, we are encouraged by recent momentum (Fig. 2), largely led by publications from individual STEM disciplines (namely chemistry and engineering). The skewed disciplinary representation in the aforementioned knowledge gap raises important considerations. First, disciplines are citing more within themselves than across the broader literature (Fig. 8). While there is momentum, the siloed nature of the research makes it more challenging for these findings to be extended into other disciplines from both a research and practice perspective. Second, the absence of research from disciplines like biology or physics may not necessarily reflect inactivity or disinterest, but rather a difference in the venues through which research is shared. In our corpus, we see research in engineering largely coming from published conference papers (Table 2). Not all DBER communities have formal venues that routinely publish conference papers (as is the case with American Society for Engineering Education (ASEE)). As such, there may be more research on alternative grading practices happening in other STEM disciplines, but publication practices limit their indexing by databases and search engines. As an example, the American Association for Physics Teachers (AAPT) is one of the largest gatherings of physics educators and physics education researchers; however, their conference abstracts are not indexed nor are papers published as part of conference proceedings. Research on alternative grading practices presented at these conferences that is not subsequently published in a peer-reviewed journal is unlikely to contribute to a broader or interdisciplinary conversation on alternative grading practices in STEM.

Our scoping review identified eight STEM disciplines engaged in research on alternative grading practices; however, the disciplines seem to be largely unaware of each other. The co-citation analysis finds that Chemistry, Engineering, Mathematics, and Physics were each statistically significant independent communities; in other words, studies within each of those disciplines build on a body of work that is different and independent from other disciplines. For example, Engineering papers cite the same body of work as other Engineering papers, but cite different papers from Chemistry, Mathematics, and Physics. The direct citation analysis supports this siloing of disciplines as almost no studies in our corpus cite other corpus studies outside of their respective discipline. Additionally, many studies in our corpus (n=22) do not cite any other study in the corpus (Fig. 7). This finding may be expected in a relatively young field (half of the studies in our corpus were published in 2016 or beyond), but it also presents an opportunity for alternative grading researchers and practitioners to become familiar with relevant work outside of their disciplinary expertise.

Methodological gap

A methodological gap is present when there is little variation in methodological approaches (Miles, 2017). Our scoping review revealed two major types of methodological gaps, one related to research practices and the second related to implementation of alternative grading practices.

We found evidence to suggest a methodological gap with respect to research practices as most studies in our corpus used tools that have not undergone validation efforts (Fig. 5). While nearly 90% of the outcomes captured by studies that used RGS found positive impacts of alternative grading practices (i.e., Positive or Trending Positive), these studies often provided little description or rationale for the survey design and none included validation efforts. While these surveys may be accurately capturing the researchers' variables of interest (e.g., attitudes, perceptions, learning gains, affective constructs, etc.), the findings obtained through these measures are limited in their generalizability and not necessarily replicable, which in turn makes it challenging for researchers and practitioners to make comparisons across populations or time.

Just over 13% of the 75 studies used validated tools to measure affective constructs (e.g., Dweck's implicit theory of Intelligence Scale) and conceptual learning (e.g., Force Concept Inventory; Fig. 6). In these studies, far fewer (47%) found a positive impact of alternative grading. By using validated measures, these studies enable comparisons across populations and time, support generalizability, and build a more robust understanding of how alternative grading strategies impact undergraduate students in STEM courses.

Further, there were few studies measuring content or conceptual learning. Most studies seemed to focus on

affective constructs or final grades. There are several issues with this: first, we need empirical support that alternative grading positively impacts learning of content and skills because at the end of the day, this is what most college instructors really care about. Second, the dizzying array of affective measures dilutes the findings and we simply cannot synthesize across studies. This dilution highlights the need for theory-driven methodology - because these studies are not grounded in theories about learning and teaching, they cannot help us understand the mechanisms by which grading practices may impact the learner.

The second methodological gap we found stemmed from a lack of universality in the definitions used by the studies in our data corpus of different alternative grading practices. We had initially intended to characterize the grading systems used in each study using the definitions provided by Clark (2023); however, we found many records did not include a thorough enough description of their grading practices for us to independently characterize the method. Thus, we relied on the name of the method used by the authors in each study. So, while standards-based grading was the most common grading system in our corpus (Fig. 3), there is likely variation in implementation. This lack of detail also made it challenging to identify themes across grading strategies in a meaningful way, and we were not able to discern whether seemingly discipline-specific grading methods (e.g., the Keller Method) shared attributes with more broadly seen methods (e.g., SBG; Fig. 7). In addition, discipline-specific grading methods may contribute to the similarity in citation patterns seen between grading methods (Fig. 8) furthering the idea that defining and implementing alternative grading practices is currently a discipline-specific endeavor. This ultimately precluded our ability to make conclusions about the outcomes or efficacy of any specific alternative grading practice.

Theoretical gap

A theoretical gap is indicated when there is a lack of theory underlying the research in a given area (Miles, 2017). As noted in our methods, we initially intended to capture the underlying theory motivating the research described in the studies included in our data corpus; however, we found these were often not sufficiently named or described and impeded any thematic analysis.

Theoretical frameworks influence all aspects of research - from the questions asked and the methods used to collect data, to the interpretation and discussion of the results (Luft et al., 2022). By naming the theoretical framework used in a given study, researchers (both the original authors and other scholars) can better situate the findings and claims presented within the existing body of research. Articulating the theoretical framework gives greater meaning to the choices made by researchers and reveals the lens researchers applied in their attempt to understand a given phenomenon. Currently, only a fraction of the research exploring the impacts of alternative grading on student outcomes explicitly draws on theories of learning and other relevant theoretical frameworks. The effects of this scant, disjointed theoretical footing can be seen in the many disparate variables and tools observed in the current body of research (Fig. 5). With no theoretical foundations in place, researchers working to understand the impact of grading practices on student outcomes are left to place stock in variables and tools that may ultimately be ill-suited for their intended research aims, making it all the more challenging to develop a robust understanding of this complex phenomenon.

Limitations

Our search for empirical research into alternative grading was limited by the disciplines represented by our interdisciplinary team. We have many STEM experts, but not across all STEM disciplines. As a result, we opted to limit our search to those disciplines where we had one or more members with expertise. Through our search process, we did identify and ultimately include several studies that fell outside of our collective expertise (i.e., computer science, geology, and mathematics).

Additionally, while our list of alternative grading practices was exhaustive to the best of our knowledge, given the lack of consensus on names and definitions of the many alternative grading practices, it is likely our search missed studies that did not explicitly name their alternative grading practice or that used a practice not on our list.

Our search of databases and subsequent study analysis revealed that the alternative grading conversation is not restricted to journal articles. The presence of peer-reviewed and published articles from conference proceedings in engineering leads us to believe there are other venues (i.e., conferences) that include work about alternative grading that were not identified through our database search and thus not represented in our data corpus. Therefore, we believe the research into alternative grading practices is broader than we can currently characterize and report.

In addition, the communication channels about alternative grading are not limited to peer-reviewed journals and conference proceedings but includes blog posts, books, and social media conversations. While these dissemination platforms were not included in our search criteria, they contribute substantially to the broader conversation around alternative grading in higher education. These venues typically advocate for the adoption of alternative grading practices and are often based on anecdotal or limited empirical evidence. While these less

formal dissemination pathways may not contribute to the empirical findings of alternative grading, their role of rapid communication is an important consideration for the landscape as a whole and warrant further exploration.

Implications

Alternative grading is rapidly increasing in popularity in STEM classrooms, resulting in calls for empirical evidence of its efficacy. Our scoping review provides an initial map of the research landscape (Fig. 2) and identifies areas of research needs. First, research in alternative grading needs to be grounded in theoretical frameworks, enabling us to develop informed hypotheses about how and when alternative grading practices should impact learning and other affective constructs. Such grounding will subsequently impact the variables we measure, allowing us to develop a more robust and unifying understanding of how alternative grading practices impact student learning.

Second, it is critical that authors fully describe their implementation of alternative grading practices, including defining their terms using common language. While term definition may seem a trivial task, this lack of consensus is currently hindering our ability to uncover patterns across courses, disciplines, or institutions. Robust descriptions of implementation practices will enable us to develop clearer definitions of alternative grading practices, resulting in better research.

Third, using validated tools rather than researcher-generated instruments will support richer comparisons within and across contexts such as grading systems, disciplines, class sizes, etc. Validated tools also enable us to ask deeper questions, such as whether a particular grading system is better at developing learners' self-regulation skills or if alternative grading practices create more equitable learning environments. Researcher-generated surveys and course evaluations help us gain insights into an individual course but when placed in the overall landscape of research, do not fully enable meaningful comparisons.

Fourth, the STEM disciplines represented in our corpus are citing different literature (Figs. 6 and 7), which may contribute to the lack of a unifying theory or universal definitions of alternative grading strategies. This disciplinary siloing may also lead to many instances of "reinventing the wheel", where each discipline does not avail itself of the lessons learned by other disciplines. By extension, students, who are often enrolled in courses across STEM disciplines, may face confusion and shifting expectations. Interdisciplinary efforts are critical to capturing the entire landscape of a research area in STEM education and will be important in building a more in-depth understanding of the costs and benefits of alternative grading practices moving forward.

Conclusions

Our scoping review does not allow us to make comparisons across the studies in our corpus. However, the high proportion of positive results is promising and warrants further investigation. First, future research should explicitly connect to theoretical frameworks to explore how alternative grading practices impact students' learning of skills and content. Indeed, a primary goal of education is to support students' learning, so evaluating the extent to which alternative grading practices produce tangible and positive effects on memory and comprehension of material is critical. Research investigating the efficacy of these practices should aim to involve a variety of experimental techniques and draw from various cognitive science frameworks (Creswell & Plano Clark, 2018; Spivey, 2023) to increase generalizability and applicability across disciplines. Second, alternative grading practices might have an impact on students' development of other skills such as self-regulated learning skills. For example, many alternative grading approaches involve components that provide students with autonomy in their learning experiences and numerous opportunities to demonstrate competence–two key factors in self-determination theory. Self-determination theory (see Deci & Ryan, 1985; also see 2008) is a macro-theory of motivation that takes into account an individual's psychological needs and factors that impact individual's growth and development, with research showing that incorporating activities that support autonomy and competence has a positive effect on motivation and learning in education environments (for a review, see Niemiec & Ryan, 2009). Content learning and the development of self-regulated learning skills are both areas in which validated tools can be employed, allowing the broader community to draw comparisons between traditional and alternative grading practices. Finally, though preliminary, outcomes from this review indicate an increased usage of alternative grading across STEM disciplines, which suggests a need to support effective implementation of these practices. Thus, another fruitful avenue for future research could be to scaffold faculty's transformation of their grading practices by guiding them through the five stages of Roger's Diffusion of Innovation Framework: knowledge, persuasion, decision, implementation, and confirmation (see Reinholz et al., 2021; also see Rogers, 1995, 2004).

More generally, this review underscores the need for further interdisciplinary research efforts in STEM, echoing calls like the ones from Henderson and colleagues (Henderson et al., 2017). The structure and composition of the NDSU Journal Club facilitates our ability to conduct cross-disciplinary research on teaching and learning practices in STEM, with findings from the current review highlighting the value, strength, and richness that can come from such collaborations. The interdisciplinary

approach employed in this scoping review illustrates how future investigations into alternative grading practices in STEM would be strengthened by increased interdisciplinary communication and collaboration.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43031-024-00106-8.

> Supplementary Material 1

Declarations

Competing interests
The authors declare that they have no competing interests.

References
An * indicates an entry included in our corpus

*Ahlberg, L. (2021). Organic chemistry core competencies: Helping students engage using specifications. In *ACS symposium series* (Vol. 1378). https://doi.org/10.1021/bk-2021-1378.ch003

*Ankeny, C. J., Adkins, A., & O'neill, D. P. (2022). Impact of two reflective practices in an engineering laboratory course using standards-based grading. *ASEE annual conference and exposition, conference proceedings*. https://doi.org/10.18260/1-2%9640716

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

*Armacost, R. L., & Pet-Armacost, J. (2003, November). Using mastery-based grading to facilitate learning. In *33rd annual frontiers in education, 2003. FIE 2003*. (Vol. 1, pp. T3A–20). IEEE. https://doi.org/10.1109/FIE.2003.1263320

*Astwood, P. M., & Slater, T. F. (1997). Effectiveness and management of portfolio assessment in high-enrollment courses. *Journal of Geoscience Education*, *45*(3), 238–242. https://doi.org/10.5408/1089-9995-45.3.238

*Atwood, S. A., Siniawski, M. T., & Carberry, A. R. (2014, June). Using standards-based grading to effectively assess project-based design courses. *ASEE annual conference & exposition* (pp. 24–1345). https://doi.org/10.18260/1-2%9623278

*Baisley, A., & Hjelmstad, K. D. (2021). What do students know after statics? Using mastery-based grading to create a student portfolio. *ASEE annual conference and exposition, conference proceedings*. https://doi.org/10.18260/1-2%9638041

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *International AAAI conference on weblogs and social media*.

*Boesdorfer, S. B., Baldwin, E., & Lieberum, K. A. (2018). Emphasizing learning: Using standards-based grading in a large nonmajors' general chemistry survey course. *Journal of Chemical Education*, *95*(8), 1291–1300. https://doi.org/10.1021/acs.jchemed.8b00251

Bowen, R. S., & Cooper, M. M. (2022). Grading on a curve as a systemic issue of equity in chemistry education. *Journal of Chemical Education*, *99*(1), 185–194. https://doi.org/10.1021/acs.jchemed.1c00369

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

*Carberry, A. R., Siniawski, M., Atwood, S. A., & Diefes-Dux, H. A. (2016, June). Best practices for using standards-based grading in engineering courses. *ASEE annual conference & exposition*. https://doi.org/10.18260/p.26379

*Carlisle, S. (2020). Simple specifications grading. *PRIMUS*, *30*(8–10), 926–951. https://doi.org/10.1080/10511970.2019.1695238

*Chen, L., Grochow, J. A., Layer, R., & Levet, M. (2022). Experience report. *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education*, *1*, 221–227. https://doi.org/10.1145/3502718.3524750

Clark, D. (2023, January 30). An alternative grading glossary. Grading for Growth. https://gradingforgrowth.com/p/an-alternative-grading-glossary

Clark, D., & Talbert, R. (2023). *Grading for growth*. Routledge. https://doi.org/10.4324/9781003445043

*Collins, J. B., Harsy, A., Hart, J., Haymaker, K. A., Hoofnagle, A. M., Janssen, M. K., Kelly, J. S., Mohr, A. T., & O'Shaughnessy, J. (2019). Mastery-based testing in undergraduate mathematics courses. *PRIMUS*, *29*(5), 441–460. https://doi.org/10.1080/10511970.2018.1488317

*Cooper, A. A. (2020). Techniques grading: Mastery grading for proofs courses. *PRIMUS*, *30*(8–10), 1071–1086. https://doi.org/10.1080/10511970.2020.1733151

*Craugh, L. (2017). Adapted mastery grading for statics. *2017 ASEE annual conference & exposition proceedings*. https://doi.org/10.18260/1-2%9627536

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, *19*(2), 109–134.

Deci, E. L., & Ryan, R. M. (2008). Self-determination theory: A macrotheory of human motivation, development, and health. *Canadian Psychology/Psychologie Canadienne*, *49*(3), 182.

Defining the Role of Authors and Contributors. (2023). International committee of medical journal editors. https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html#two

*DeGoede, K. (2019). Board 16: Mechanical engineering division: Competency based assessment in dynamics. *2019 ASEE annual conference & exposition proceedings*. https://doi.org/10.18260/1-2%9632278

*Diefes-Dux, H., & Ebrahiminejad, H. (2018). Standards-based grading derived data to monitor grading and student learning. *2018 ASEE annual conference & exposition proceedings*. https://doi.org/10.18260/1-2%9630981

*Dosmar, E., & Williams, J. (2022, August). Student reflections on learning as the basis for course grades. *2022 ASEE annual conference & exposition*.

*Dougherty, R. C. (1997). Grade/performance contracts, enhanced communication, cooperative learning and student performance in undergraduate organic chemistry. *Journal of Chemical Education*, *74*(6), 722. https://doi.org/10.1021/ed074p722

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random house.

*Evensen, H. (2022, August). Specifications grading in general physics and engineering physics courses. *2022 ASEE annual conference & exposition*.

Feldman, J. (2018). *Grading for equity: What it is, why it matters, and how it can transform schools and classrooms*. Corwin Press.

Gooch, P. (2021). *How scholarcy contributes to and makes use of open citation*. Scholarcy.

*Goodwin, J. A., & Gilbert, B. D. (2001). Cafeteria-style grading in general chemistry. *Journal of Chemical Education*, *78*(4), 490. https://doi.org/10.1021/ed078p490

*Grau, H. J. (1999). "Streamlined" contract grading–a preformance-measuring alternative to traditional evaluation methods. *Journal of College Science Teaching*, *28*(4), 254.

Graulich, N., Lewis, S. E., Kahveci, A., Nyachwaya, J. M., & Lawrie, G. A. (2021). Writing a review article: What to do with my literature review. *Chemistry Education Research and Practice*, *22*(3), 561–564. https://doi.org/10.1039/D1RP90006D

*Hamilton, N. B., Remington, J. M., Schneebeli, S. T., & Li, J. (2022). Outcome-based redesign of physical chemistry laboratories during the COVID-19 pandemic. *Journal of Chemical Education*, *99*(2), 639–645. https://doi.org/10.1021/acs.jchemed.1c00691

*Harsy, A., Carlson, C., & Klamerus, L. (2021). An analysis of the impact of mastery-based testing in mathematics courses. *PRIMUS*, *31*(10), 1071–1088. https://doi.org/10.1080/10511970.2020.1809041

*Harsy, A., & Hoofnagle, A. (2020). Comparing mastery-based testing with traditional testing in calculus II. *International Journal for the Scholarship of Teaching & Learning*, *14*(2), 1–13. https://doi.org/10.20429/ijsotl.2020.140210

He, Z., Chen, W., Wei, X., & Liu, Y. (2021). On the statistical significance of communities from weighted graphs. *Scientific Reports*, *11*(1), 20304. https://doi.org/10.1038/s41598-021-99175-2

*Helmke, B. P. (2019, June). Specifications grading in an upper-level BME elective course. *2019 ASEE annual conference & exposition*.

Henderson, C., Connolly, M., Dolan, E. L., Finkelstein, N., Franklin, S., Malcom, S., Rasmussen, C., Redd, K., & St. John, K. (2017). Towards the STEM DBER alliance: Why we need a discipline-based STEM education research community. *International Journal of Research in Undergraduate Mathematics Education*, *3*(2), 247–254. https://doi.org/10.1007/s40753-017-0056-3

*Hicks, N., & Diefes-Dux, H. (2017). Grader consistency in using standards-based rubrics. *2017 ASEE annual conference & exposition proceedings*. https://doi.org/10.18260/1-2%9628416

*Hollinsed, W. C. (2018). *Applying innovations in teaching to general chemistry* (pp. 145–152). https://doi.org/10.1021/bk-2018-1301.ch009

*Houseknecht, J. B., & Bates, L. K. (2020). Transition to remote instruction using hybrid just-in-time teaching, collaborative learning, and specifications grading for organic chemistry 2. *Journal of Chemical Education*, *97*(9), 3230–3234. https://doi.org/10.1021/acs.jchemed.0c00749

*Howitz, W. J., McKnelly, K. J., & Link, R. D. (2021). Developing and implementing a specifications grading system in an organic chemistry laboratory course. *Journal of Chemical Education*, *98*(2), 385–394. https://doi.org/10.1021/acs.jchemed.0c00450

*Hylton, J., & Funke, L. (2022, August). Journey towards competency-based grading for mechanical engineering computer applications. *2022 ASEE annual conference & exposition*.

*Katzman, S. D., Hurst-Kennedy, J., Barrera, A., Talley, J., Javazon, E., Diaz, M., & Anzovino, M. E. (2021). The effect of specifications grading on students' learning and attitudes in an undergraduate-level cell biology course. *Journal of Microbiology & Biology Education*, *22*(3). https://doi.org/10.1128/jmbe.00200-21

Khalil, H., Peters, M., Godfrey, C. M., McInerney, P., Soares, C. B., & Parker, D. (2016). An evidence-based approach to scoping reviews. *Worldviews on Evidence-Based Nursing*, *13*(2), 118–123. https://doi.org/10.1111/wvn.12144

*Kirschenbaum, D. S., & Riechmann, S. W. (1975). Learning with gusto in introductory psychology. *Teaching of Psychology*, *2*(2), 72–76.

*Kitchen, E., King, S. H., Robison, D. F., Sudweeks, R. R., Bradshaw, W. S., & Bell, J. D. (2006). Rethinking exams and letter grades: How much can teachers delegate to students? *CBE—Life Sciences Education*, *5*(3), 270–280. https://doi.org/10.1187/cbe.05-11-0123

Kohn, A., & Blum, S. D. (2020). *Ungrading: Why rating students undermines learning (and what to do instead)*. West Virginia University Press.

*Ladd, G. T. (1971). Factors affecting student choice of pass-fail in an introductory science course. *Journal of College Science Teaching*, *1*(2), 17–19.

*Lee, E., Carberry, A. R., Diefes-Dux, H. A., Atwood, S. A., & Siniawski, M. T. (2018). Faculty perception before, during and after implementation of standards-based grading. *Australasian Journal of Engineering Education*, *23*(2), 53–61. https://doi.org/10.1080/22054952.2018.1544685

*Lengyel, G. A., Boron, T. T., Loe, A. M., & Zirpoli, S. (2023). Implementation of the multiplier, an alternative grading system for formative assessments. *Journal of Chemical Education*, *100*(1), 186–191. https://doi.org/10.1021/acs.jchemed.2c00628

*Leonard, W. J., Hollot, C. V., & Gerace, W. J. (2008). Mastering circuit analysis: An innovative approach to a foundational sequence. *2008 38th annual frontiers in education conference*, F2H-3-F2H-8. https://doi.org/10.1109/FIE.2008.4720568

*Lewis, D. (2020a). Gender effects on re-assessment attempts in a standards-based grading implementation. *PRIMUS*, *30*(5), 539–551. https://doi.org/10.1080/10511970.2019.1616636

*Lewis, D. (2020b). Student anxiety in standards-based grading in mathematics courses. *Innovative Higher Education*, *45*(2), 153–164. https://doi.org/10.1007/s10755-019-09489-3

Lewis, D. (2022a). Impacts of standards-based grading on students' mindset and test anxiety. *Journal of the Scholarship of Teaching and Learning*, *22*(2). https://doi.org/10.14434/josotl.v22i2.31308

*Lewis, D. (2022b). Impacts of standards-based grading on students' mindset and test anxiety. *Journal of the Scholarship of Teaching and Learning*, *22*(2). https://doi.org/10.14434/josotl.v22i2.31308

*Lewis, D. K., & Wolf, W. A. (1973). Implementation of self-paced learning (Keller method) in a first-year course. *Journal of Chemical Education*, *50*(1), 51. https://doi.org/10.1021/ed050p51

*Lewis, D. K., & Wolf, W. A. (1974). Keller plan introductory chemistry. Students' performance during and after the Keller experience. *Journal of Chemical Education*, *51*(10), 665. https://doi.org/10.1021/ed051p665

*Lindemann, D. F., & Harbke, C. R. (2011). Use of contract grading to improve grades among college freshmen in introductory psychology. *SAGE Open*, *1*(3), 215824401143410. https://doi.org/10.1177/2158244011434103

Lipnevich, A. A., Guskey, T. R., Murano, D. M., & Smith, J. K. (2020). What do grades mean? Variation in grading criteria in American college and university courses. *Assessment in Education: Principles, Policy & Practice*, *27*(5), 480–500. https://doi.org/10.1080/0969594X.2020.1799190

Lipnevich, A. A., Murano, D., Krannich, M., & Goetz, T. (2021). Should I grade or should I comment: Links among feedback, emotions, and performance. *Learning and Individual Differences*, *89*, 102020. https://doi.org/10.1016/j.lindif.2021.102020

Luft, J. A., Jeong, S., Idsardi, R., & Gardner, G. (2022). Literature reviews, theoretical frameworks, and conceptual frameworks: An introduction for new biology education researchers. *CBE—Life Sciences Education*, *21*(3). https://doi.org/10.1187/cbe.21-05-0134

Lund, T. J., & Stains, M. (2015). The importance of context: An exploration of factors influencing the adoption of student-centered teaching among chemistry, biology, and physics faculty. *International Journal of STEM Education*, *2*(1), 13. https://doi.org/10.1186/s40594-015-0026-8

*Marti, E. J. (2022, July). WIP: Contract grading as an alternative grading structure and assessment approach for a process-oriented, first-year course. 2022 first-year engineering experience.

*Martin, L. J. (2019). *Introducing components of specifications grading to a general chemistry I course* (pp. 105–119). https://doi.org/10.1021/bk-2019-1330.ch007

*McKnelly, K. J., Morris, M. A., & Mang, S. A. (2021). Redesigning a "writing for chemists" course using specifications grading. *Journal of Chemical Education*, *98*(4), 1201–1207. https://doi.org/10.1021/acs.jchemed.0c00859

*McKnelly, K. J., Howitz, W. J., Thane, T. A., & Link, R. D. (2022). Specifications grading at scale: Improved letter grades and grading-related interactions in a course with over 1,000 students [Preprint]. *Chemistry*. https://doi.org/10.26434/chemrxiv-2022-77wr7

*Mendez, J. (2018). Standards-based specifications grading in a hybrid course. *2018 ASEE annual conference & exposition proceedings*. https://doi.org/10.18260/1-2%9630982

*Mikula, B. D., & Heckler, A. F. (2017). Framework and implementation for improving physics essential skills via computer-based practice: Vector math. *Physical Review Physics Education Research*, *13*(1), 010122. https://doi.org/10.1103/PhysRevPhysEducRes.13.010122

Miles, D. A. (2017, August). A taxonomy of research gaps: Identifying and defining the seven research gaps. In *Doctoral student workshop: finding research gaps-research methods and strategies, Dallas, Texas* (pp. 1–15).

*Moore, J. (2016). Mastery grading of engineering homework assignments. *2016 IEEE frontiers in education conference (FIE)*, 1–9. https://doi.org/10.1109/FIE.2016.7757584

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*(1), 143. https://doi.org/10.1186/s12874-018-0611-x

*Naegele, C. J., & Novak, J. D. (1975, March). An evaluation of student attitudes, achievement, and learning efficiency in various modes of an individualized, self-paced learning program in introductory college physics. *Annual meeting of the national association for research in science teaching*. https://eric.ed.gov/?id=ED129556

Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education*, *7*(2), 133–144.

Nilson, L. B., & Stanny, C. J. (2023). *Specifications Grading*. Routledge. https://doi.org/10.4324/9781003447061

*Novak, H., Paguyo, C., & Siller, T. (2016). Examining the impact of the engineering successful/unsuccessful grading (SUG) program on student retention. *Journal of College Student Retention: Research, Theory & Practice*, *18*(1), 83–108. https://doi.org/10.1177/1521025115579674

*Oerther, D. (2017). Reducing costs while maintaining learning outcomes using blended, flipped, and mastery pedagogy to teach introduction to environmental engineering. *2017 ASEE annual conference & exposition proceedings*. https://doi.org/10.18260/1-2%9628786

*Okamoto, N. (2020). Implementing competency-based assessment in an undergraduate thermodynamics course. *2020 ASEE virtual annual conference content access proceedings*. https://doi.org/10.18260/1-2%9634780

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

*Pascal, J., Vogel, T., & Wagstrom, K. (2020). Grading by competency and specifications: Giving better feedback and saving time. *2020 ASEE virtual annual conference content access proceedings*. https://doi.org/10.18260/1-2%9634712

*Paul, C. A., & Webb, D. J. (2022). Percent grade scale amplifies racial or ethnic inequities in introductory physics. *Physical Review Physics Education Research*, *18*(2), 020103.

Peters, M. D. J., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis*, *18*(10), 2119–2126. https://doi.org/10.11124/JBIES-20-00167

*Ponce, M. L. S., & Moorhead, G. B. G. (2020). Developing scientific writing skills in upper level biochemistry students through extensive practice and feedback. *The FASEB Journal*, *34*(S1), 1–1. https://doi.org/10.1096/fasebj.2020.34.s1.00661

*Rajapaksha, A., & Hirsch, A. S. (2017). Competency based teaching of college physics: The philosophy and the practice. *Physical Review Physics Education Research*, *13*(2), 020130. https://doi.org/10.1103/PhysRevPhysEducRes.13.020130

*Rawlins, M., & Junsangsri, P. (2022, April). Refining competency-based grading in undergraduate programming courses. In *ASEE-NE 2022*.

Reinholz, D. L., Matz, R. L., Cole, R., & Apkarian, N. (2019). STEM is not a monolith: A preliminary analysis of variations in STEM disciplinary cultures and implications for change. *CBE—Life Sciences Education*, *18*(4), mr4. https://doi.org/10.1187/cbe.19-02-0038

Reinholz, D. L., White, I., & Andrews, T. (2021). Change theory in STEM higher education: A systematic review. *International Journal of STEM Education*, *8*(1), 37.

*Ring, J. (2017). ConfChem conference on select 2016 BCCE presentations: Specifications grading in the flipped organic classroom. *Journal of Chemical Education*, *94*(12), 2005–2006. https://doi.org/10.1021/acs.jchemed.6b01000

Rogers, E. M. (1995). *Diffusion of innovations*. Free Press.

Rogers, E. M. (2004). A prospective and retrospective look at the diffusion model. *Journal of Health Communication*, *9*(S1), 13–19.

*Salzman, N., Cantley, K., & Hunt, G. (2019). Board 64: Work in progress: Mastery-based grading in an introduction to circuits class. *2019 ASEE annual conference & exposition proceedings*. https://doi.org/10.18260/1-2%9632397

Schinske, J., & Tanner, K. (2014). Teaching more by grading less (or differently). *CBE—Life Sciences Education*, *13*(2), 159–166. https://doi.org/10.1187/cbe.cbe-14-03-0054

*Schlemer, L., & Vanasupa, L. (2016). Grading for enhanced motivation and learning. *2016 ASEE annual conference & exposition proceedings*. https://doi.org/10.18260/p.27305

Shadle, S. E., Marker, A., & Earl, B. (2017). Faculty drivers and barriers: Laying the groundwork for undergraduate STEM education reform in academic departments. *International Journal of STEM Education*, *4*(1), 8. https://doi.org/10.1186/s40594-017-0062-7

*Siller, T., & Paguyo, C. (2012). Use of pass/fail grading to increase first year retention. *SAT*, *1200*, 1230.

*Siniawski, M. T., Carberry, A., & Dionisio, J. D. N. (2012). Standards-based grading: An alternative to score-based assessment. *Proceedings of the 2012 ASEE PSW section conference*.

*Slade, D. J. (2017). Do it right! Requiring multiple submissions of math and NMR analysis assignments in the laboratory. *Journal of Chemical Education*, *94*(10), 1464–1470. https://doi.org/10.1021/acs.jchemed.7b00136

Slominski, T., Fugleberg, A., Christensen, W. M., Buncher, J. B., & Momsen, J. L. (2020). Using framing as a lens to understand context effects on expert reasoning. *CBE—Life Sciences Education*, *19*(3), ar48. https://doi.org/10.1187/cbe.19-11-0230

Slominski, T., Christensen, W. M., Buncher, J. B., & Momsen, J. (2023). The impact of context on students' framing and reasoning about fluid dynamics. *CBE—Life Sciences Education*, *22*(2). https://doi.org/10.1187/cbe.21-11-0312

*Smith, H. A. (1976). The evolution of a self-paced organic chemistry course. *Journal of Chemical Education*, *53*(8), 510. https://doi.org/10.1021/ed053p510

Spivey, M. J. (2023). Cognitive science progresses toward interactive frameworks. *Topics in Cognitive Science*, *15*(2), 219–254.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., Rodela, T. M., Sanders, E. R., Schimpf, N. G., Schulte, P. M., Smith, M. K., Stetzer, M., Van Valkenburgh, B., Vinson, E., Weir, L. K., Wendel, P. J., Wheeler, L. B., & Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468–1470. https://doi.org/10.1126/science.aap8892

*Stange, K. E. (2018). Standards-based grading in an introduction to abstract mathematics course. *PRIMUS*, *28*(9), 797–820. https://doi.org/10.1080/10511970.2017.1408044

*Stanton, K., & Siller, T. (2011). A pass/fail option for first-semester engineering students: A critical evaluation. *2011 frontiers in education conference (FIE)*, T2D-1-T2D-6. https://doi.org/10.1109/FIE.2011.6143057

*Stanton, K. C., & Siller, T. (2012). A first look at student motivation resulting from a pass/fail program for first-semester engineering students. *2012 frontiers in education conference proceedings*, 1–6. https://doi.org/10.1109/FIE.2012.6462266

*Toledo, S., & Dubas, J. M. (2017). A learner-centered grading method focused on reaching proficiency with course learning outcomes. *Journal of Chemical Education*, *94*(8), 1043–1050. https://doi.org/10.1021/acs.jchemed.6b00651

Townsley, M., & Schmid, D. (2020). Alternative grading practices: An entry point for faculty in competency-based education. *The Journal of Competency-Based Education*, *5*(3). https://doi.org/10.1002/cbe2.1219

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., Lewin, S., Godfrey, C. M., Macdonald, M. T., Langlois, E. V., Soares-Weiser, K., Moriarty, J., Clifford, T., Tunçalp, Ö., & Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine.*, *169*(7), 467–473. https://doi.org/10.7326/M18-0850

Trujillo, C. M., & Long, T. M. (2018). Document co-citation analysis to enhance transdisciplinary research. *Science Advances*, *4*(1). https://doi.org/10.1126/sciadv.1701130

*Vournakis, J. N. (1974). A noncompetitive introductory organic chemistry course for premedical students. *Journal of Chemical Education*, *51*(11), 742. https://doi.org/10.1021/ed051p742

*Webb, D. J., Paul, C. A., & Chessey, M. K. (2020). Relative impacts of different grade scales on student success in introductory physics. *Physical Review Physics Education Research*, *16*(2), 020114. https://doi.org/10.1103/PhysRevPhysEducRes.16.020114

*White, J. M., Close, J. S., & McAllister, J. W. (1972). Freshman chemistry without lectures. A modified self-paced approach. *Journal of Chemical Education*, *49*(11), 772. https://doi.org/10.1021/ed049p772

*Wierer, J. (2022, August). WIP: Standards-based grading for electric circuits. 2022 ASEE annual conference & exposition.

*Williams, S. M., & Newberry, B. P. (1998). First year experiences implementing minimum self paced mastery in a freshman engineering problem solving course. *1998 annual conference proceedings*, 3.287.1–3.287.12. https://doi.org/10.18260/1-2%967131

*Zimmerman, J. K. (2020). Implementing standards-based grading in large courses across multiple sections. *PRIMUS*, *30*(8–10), 1040–1053. https://doi.org/10.1080/10511970.2020.1733149

Publisher's Note