Deficiencies of Traditional Grading Systems and Recommendations for the Future

Jeff Cain, EdD, MS,a Melissa Medina, EdD,b,d Frank Romanelli,a,e Adam Persky, PhDc,d

a University of Kentucky, College of Pharmacy, Lexington, Kentucky
b University of Oklahoma, College of Pharmacy, Oklahoma City, Oklahoma
c University of North Carolina at Chapel Hill, Chapel Hill, North Carolina
d Associate Editor, American Journal of Pharmaceutical Education, Arlington, Virginia
e Executive Associate Editor, American Journal of Pharmaceutical Education, Arlington, Virginia

Submitted August 17, 2021; accepted November 15, 2021; published September 2022.

Objective. To review issues surrounding the use of grades in the educational process and provide evidence-based recommendations for how to redesign grading practices for optimal value.

Findings. Traditional tiered grading systems (ie, A, B, C, etc) have historically been a major component of the formal educational process. The way grades are used and interpreted are typically based on some commonly held assumptions, including that they are accurate measures of learning, that they motivate students to learn, and that they provide feedback to learners. However, much of the research regarding grades indicates that flaws exist in these assumptions. Grades may not always accurately measure learning, they can have adverse effects on student motivation, and they are not a good form of feedback.

Summary. The Academy should consider the evidence regarding the purpose, effects, and interpretation of grades in the educational process. Despite barriers and potential pushback, pharmacy educators should revise grading practices to be more accurate, interpretable, and beneficial to learner development.

Keywords: grades, assessment, motivation, feedback

INTRODUCTION

Grades are deeply embedded within educational systems, and they provide teachers, students, and others a simple way to quantify and interpret the complex and nuanced aspects of learning. We use them to measure learning, sort students, motivate learning behaviors, and provide feedback, yet some educators have begun to question the validity of and the assumptions around traditional practices, including how grades are determined and their effects on learner behavior.1 Alternative forms of grading, such as competency-based assessment,2 specifications grading,3 and “ungrading”4 have recently become more popular as educators have begun to realize the downsides of traditional grading systems. As discussions of grading reforms spread throughout health professions education, it is important to understand the associated issues and concerns with contemporary grading practices. The literature regarding grades and associated assessment practices is interwoven and expansive, making it difficult for educators to understand the depth and breadth of issues. This review provides an overview of the theories and practices surrounding grades, current limitations of traditional grading systems, and barriers to change. We conclude with recommendations for how colleges/schools of pharmacy and instructors can redesign grading practices to improve both student performance and accuracy of learning measurement.

Grades were first instituted in the 1700s in Europe to foster competition among students for prizes and rank order. In the late 1700s, Yale University introduced achievement-based rankings and classifications, and by 1800 it had paved the way for a 4.0 grade point average scale.5 By the mid- to late 1800s, Harvard University, University of Michigan, and Mount Holyoke College all had adopted a letter grading system with passing rates set arbitrarily as 26%, 50%, and 75%, respectively. Eventually, in this tiered grading system, an A came to mean that students earned at least 90% of the total possible points, a B at least 80%, a C at least 70%, and an F that they earned below 60% of the total possible points.5 Along with the quantifiers were qualifiers, with an A indicating an exceptional level of achievement, a B indicating a good but not outstanding level, and a C indicating a fair level. Over time, however, these meanings have eroded, and the uses and purposes of grades have changed, introducing challenges to the teaching and learning process.3
Because considerable global variations exist in how grades are structured and assigned, this review focuses on traditional grading schemes within the United States, although the same issues and concerns are relevant in other educational systems.

The most basic assumption concerning grades is that they accurately measure learning. However, numerous factors influence traditional grades’ validity in measuring learning and how those measurements are interpreted. First, the precision of how we measure and report learning (through points and grades) depends on the reliability of our assessments. Figure 1 illustrates how assessments with different Kuder-Richardson (KR-20) reliability scores and standard deviations can be interpreted differently when used to examine learning between a B student and a C student. If an assessment has a KR-20 of 0.9, which is excellent, then it is reasonable to assume that the student’s true score would reflect actual learning. If we assume a standard deviation of 10%, then there is no overlap between the distributions of the true score (X ± SD), and we may be confident we can differentiate between the two students. However, if the assessment’s KR-20 drops to 0.6 with the same 10% standard deviation, there now is overlap between the two true score distributions. The situation is further complicated as the standard deviation increases, preventing us from definitively stating that the B student has learned more than the C student.

Second, if grades comprise items other than assessment of knowledge and skills, then they may not accurately reflect learning. When points are awarded for subjective activities that aid the learning process (eg, participation or homework completion), results will vary from when we only consider objective measures of standards achievement (ie, quizzes and exams). Table 1 illustrates these variations through an example of three students with varying performance and different grading schemes.

Third, when assessing via exams, the number of questions that differentiate a grade is also important. For example, the difference between a student with an 80% versus a 70% on a 10-question assessment is only represented by one question (8/10 vs 7/10). The issue, then, is whether a single question is sufficient to discern an above-average student (80%) from an average student (70%). With a 100-question exam, the difference between average and above average is 10 questions. Whether those 10 questions relate to a single competency or four different competencies may also change an instructor’s interpretation of what the grade means.

Fourth, considerable variability can exist among instructors in their criteria for achieving high grades. Differences in grading rigor and leniency among instructors are well documented. For that reason, it is easy to comprehend how instructor variability complicates the interpretation of what grades really mean. Also, beginning in elementary school education, subjects such as social studies focus on performance and tests, while mathematics and English focus on enablers like effort, class participation, and homework. If elementary and high school educators formally trained in assessment exhibit variability in grading, it is reasonable to assume that pharmacy instructors will do so as well.

Fifth, differences in grading approaches complicate interpretations of grades. Many instructors use examinations to assess learning, but examinations tend to reflect performance and a student’s performance may reflect acute studying more so than learning. Sufficient evidence suggests that acute examination performances may not reflect longer-term retention. For example, student X scores 90% on an examination and student Y scores 85%. The interpretation is that student X learned more than student Y. However, the interpretation might change if a surprise repeat examination is administered a month later and student Y outperforms student X. Traditional approaches to grading usually do not consider growth over time nor

![Figure 1. Ability of an assessment to differentiate between two grades based on examination reliability and variability.](image-url)
the impact and importance of longitudinal growth. The complexity of grade interpretation is illustrated by comparing students with different trajectories of learning. For example, a student who starts the semester with low scores but progressively improves over time to achieve the best grade in the class will finish with the same grade as a student with the mirror opposite performance.

Taking all these assumptions together, while grades may reflect aspects of learning, they may not be a particularly accurate measure. Interpretations of grades within and across systems are clouded by numerous issues related to instructor variability, which brings into question the validity and usefulness of grades for sorting students by their learning outcomes.

A key issue regarding the validity of grades in the specific context of health professions education is whether grades align with clinical performance. In pharmacy education, correlations between didactic training and clinical performance have become increasingly critical as accreditation standards continue to emphasize students’ readiness to enter advanced pharmacy practice experiences (APPEs). The interface between grades and experiential coursework traverses two issues. First, investigators have raised the issue of whether performance within the didactic curriculum (as measured by grades) correlates to performance in clinical settings. Second, some have questioned whether, in clinical settings, pass-fail grading systems might be more logical than traditional A-F grading scales.

In this vein, Call and colleagues aimed to determine factors predictive of student failure or poor performance on APPEs by conducting a retrospective cohort study comparing students who failed any APPE to students who did not in 2012-2014. A total of 669 students at a single pharmacy school were included in the analysis, in which 28 students (4.2%) failed at least one APPE and 81 students (12.1%) were labeled as poor performers (grade of C or F). Results showed that failure of an APPE and poor APPE performance were both associated with poor academic performance and/or a professionalism issue that was identified during an introductory pharmacy practice experience (IPPE). In another study, Nyman and colleagues attempted to determine predictors of student readiness for APPEs by developing a predictive model that included 23 covariates. The analysis included 226 students from a single college between 2015 and 2018. The strongest predictor of performance was found to be grade point average within the pharmacy curriculum’s core courses. A modest yet consistent predictor of performance (as defined by APPE midpoint and final scores) was a student’s total score on the Pharmacy Curriculum Outcomes Assessment (PCOA). In a similar study of a cohort of PharmD students at a single college of pharmacy, McLaughlin and colleagues examined the relationship between admissions data, scores on objective structured clinical examinations (OSCEs), and APPE performance. The investigators reported that associations between all three variables and APPE scores were weak, with the strongest association being final OSCE scores. Low performers on the final OSCE performed worse than others on specific APPEs (acute care, ambulatory care, community). The researchers concluded that the complexities of assessing students and, in turn, predicting performance in the clinical arena are complex and likely a mixture of both cognitive and noncognitive factors.

For myriad reasons, some health professions programs have moved to pass-fail grading for some didactic as well as experiential coursework. The advent and increased use of OSCEs as components of assessment has, in some instances, helped bridge the gap between pass-fail and traditional grading systems. Medical schools began investigating the prospects of pass-fail grading as early as the late 1960s, when concerns arose that students were achieving for the sole purposes of earning certain grades rather than for the sake of learning. Pass-fail grading was hypothesized to be able to better identify students who would be motivated to learn and achieve even in the absence of grades. Movement in this direction was also supported by research results that did not demonstrate correlations between traditional grades and

<table>
<thead>
<tr>
<th>Student</th>
<th>Participation</th>
<th>Homework</th>
<th>Midterms</th>
<th>Final</th>
<th>Traditional grading</th>
<th>Participation: 20%</th>
<th>Homework: 25%</th>
<th>Midterm: 25%</th>
<th>Final: 30%</th>
<th>Standards-based grading</th>
<th>Participation: 0%</th>
<th>Homework: 0%</th>
<th>Midterm: 50%</th>
<th>Final: 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100%</td>
<td>100%</td>
<td>76%</td>
<td>88%</td>
<td></td>
<td>89% (B)</td>
<td></td>
<td></td>
<td></td>
<td>91% (A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>40%</td>
<td>40%</td>
<td>84%</td>
<td>96%</td>
<td></td>
<td>73% (C)</td>
<td></td>
<td></td>
<td></td>
<td>91% (A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>90%</td>
<td>98%</td>
<td>83%</td>
<td>90%</td>
<td></td>
<td>90% (A)</td>
<td></td>
<td></td>
<td></td>
<td>87% (B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
clinical performance but did report decreases in competition among students.18,20

The motivational aspects of grades is another assumption that should be examined. A common sentiment among instructors is that grades motivate students to learn and that students might ignore any activity that does not have a direct link to their grade. In some cases, points may be used to incentivize behavior more so than measure learning. While grades do influence which tasks students complete, the motivational nature of grades is more complicated and nuanced. Motivation is a complex construct that can be examined from a variety of perspectives and theoretical models. However, at the most basic level, motivation can be described as either intrinsic or extrinsic. Students with intrinsic motivation will complete learning activities out of a natural curiosity and desire to obtain knowledge and skills; those with extrinsic motivation will engage in activities for some external reward (eg, getting a good grade or pleasing the instructor) rather than an internal desire to learn.21 Extrinsic motivation is on display when students plead for partial credit points, whereas intrinsic motivation drives student behavior to learn for the sake of learning, not because there is some reward for doing so. With grades as an external motivator, students may still learn and devote appropriate attention if they self-endorse the value of the activities beyond the grade; however, decades of research have shown that using external rewards to motivate learning is much less powerful than intrinsic motivation.21

The most notable problem with grades as an external motivator is not that they fail to motivate students but that they exert problematic psychological effects. Grades can motivate students, but in the wrong way for the wrong reasons. In many cases, grades can have a negative effect on the more desirable intrinsic motivation. Results of Deci and colleagues’ meta-analysis of 128 studies showed that tangible rewards (eg, grades, points, etc) undermined intrinsic motivation across a variety of activities.22 When grades are introduced, the internal drive to engage in learning may be replaced with the desire to achieve the external reward.23 In economics theory, this is called the “crowding out” effect, which means that the introduction of external rewards pushes away intrinsic motivation.24 It is a subtle but notable switch. The result of using grades to incentivize student behavior is that it may lead to gamesmanship, whereby students focus more on obtaining points and desired grades instead of maximizing learning.25

Another issue with grades is the type of feedback they represent. In traditional education settings, students typically receive one of two different types of feedback on their work or assessments: evaluative or descriptive.26 Evaluative feedback includes grades (ie, letter grades, points, percentage marks, or even statements of praise or concern such as “try harder next time,” “your skills have improved,” or “good work”27), and this type of feedback tells learners how they compare to others and provides a judgment summarizing the quality of learning or the characteristics of the learner.27 In contrast, descriptive feedback focuses on depicting the characteristics of the work product in order to help students understand how to improve academic knowledge and performance for future assessments.27 One example of descriptive feedback on a student’s communication skills might be, “You made some mistakes with the use of open-ended questions; in the future make sure you ask, ‘What questions do you have about your medication?’” versus “Do you have any questions?”

As a form of evaluative feedback, grades can interfere with learning. Lipnevich and Smith demonstrated that including a grade with descriptive feedback depressed future learning performance versus providing descriptive feedback alone.26 Koenka and colleagues’ meta-analysis of the impact of grades on learning showed that compared to written feedback and no feedback at all, students receiving grades had poorer achievement and less-optimal motivation.28 Therefore, simply telling students their test grade or indicating which questions they missed does not help improve their knowledge or skills as much as providing feedback about why a certain answer is correct or incorrect.

Even descriptive feedback is meaningless, though, if students only passively review it or do not read it at all. Research has shown that although students want feedback, they may not engage with it (by not reading it or forgetting it), reflect on it, interpret it, or apply it.29,31 Further, this problem may be made worse by grades: As Keupper-Tetzel and colleagues showed, the presence of a grade can interfere with students’ engagement with feedback, because students may prioritize and focus on the grade versus processing and attending to written descriptive feedback. Engagement with feedback and future academic performance were improved when feedback was given before the grade was released compared to groups who received their grade first and then got feedback.32 Therefore, for feedback to be used effectively to improve performance, it must encourage students to be “mindful” when responding and it must move the learner to actively process the information.34 Grades do not appear to stimulate either of these situations.

**DISCUSSION**

The evidence surrounding the use of traditional tiered grading systems suggests that we need to move away from them and toward practices that are more valid and reliable. However, before recommending widespread grading
reform, it is important to recognize the difficulties associated with such a change.

Barriers to Revising Grading Practices

Barriers to reforming grading practices are philosophical, cultural, and logistical in nature and exist at an Academy level as well as with individual instructors. One of the long-standing concerns about changing grading systems at a programmatic level is the potential effect on how residency and/or other postgraduate programs evaluate candidates.35 As postgraduate experiences such as residencies and fellowships becoming increasingly popular and competitive, the need to discriminate between candidates has become increasingly relevant. In a 2015 study of 4056 graduating pharmacy students, Caballero and colleagues showed that those with an alphanumeric grading scale were 6% more likely to match with a residency than students from a pass-fail grading system.36 Critics of non-traditional grading systems have argued that in the absence of a traditional grading scale, discrimination between passing students becomes arduous if not impossible.19 However, sentiments about using grades as a primary determinant of a student’s suitability for residency may be changing. Arguably, sorting and discriminating candidates is not the purpose of academic programs but rather of employers. Several pharmacy educators, citing the flaws of grading and grade point averages that were outlined above, now advocate that grade point averages are a poor tool for screening residency candidates.37,38 Pincus and colleagues examined the effects of letter grade versus pass-fail grading on pharmacy residency match rates between 2013 and 2015.39 Their analysis, which included 100 schools, found higher match rates in the pass-fail cohort (versus in the traditional grading cohort) in 2013 but no subsequent differences in 2014 and 2015. The authors hypothesized that in some instances, letter grades may actually harm applicants by calling attention to lower performance.

Another big challenge to grading reform is the time, expertise, and resources required to develop and implement practices that overcome the shortcomings of grades mentioned throughout this paper. Most faculty have not been adequately trained in the complexity and nuances of how to measure learning, how to provide evaluative feedback, and how to motivate students to learn without using points or grades. For example, assessing learning via competency demonstrations (eg, OSCEs) requires skill in case development and can be logistically time-consuming and expensive to implement on a wide scale. Even with more traditional forms of testing, creating assessments that are true measures of learning requires expert knowledge in item construction, sampling, and test development. Furthermore, when attempting to measure medical competence, applying numerical scores gets increasingly difficult, as these assessments incorporate not just knowledge but also personality traits, skills, attitudes, and problem-solving abilities.40

A hurdle to eliminating the use of points and grades for motivation is that they have been a proven method to compel students to complete activities. Designing instructional practices that incentivize students to try, fail, study, and learn, not just to put forth token effort, is challenging. It may be difficult for instructors to imagine alternative ways to motivate, and this may seem like too much of a risk to take, especially if an instructor is alone in their attempts to eschew grades as motivators. At least initially, some students may choose to continue “gaming” the educational system by focusing more on classes that do use grades to incentivize behavior. Finally, like the challenge of designing valid learner assessments mentioned above, it can take considerable time and effort to change a familiar grading system to one that accurately reflects competency as opposed to learner behaviors.

With regard to providing learners feedback through means other than points or grades, time and expertise are constraints once again. Providing descriptive feedback on assessments takes longer, and extensive effort and support may be needed to overhaul a programmatic system that has relied on evaluative feedback. Finally, separating the feedback component from grades themselves, so that students focus only on the feedback, may also be a cultural shift for both faculty and students.

Recommendations for Revising Grading

Despite the challenges mentioned above, the Academy should examine, purposely redesign, and implement grading practices in ways that are more evidence based and less reliant on historical tradition. To improve upon the shortcomings of how we often use grades, the Academy should consider the following recommendations (Table 2).

First, colleges and schools should consider what elements of their curricula, from didactic to clinical, could be transitioned from tiered grading systems to pass-fail scales. Traditional tiered grading systems may fail to accurately discriminate authentic learning and, in turn, produce flawed rankings of students. Foregoing tiered grading for competency-based or pass-fail grading systems has added benefits of improving intrinsic motivation and overcoming students’ focus on “chasing points.”41,42 In these types of systems, students must simply prove they meet predefined learning objectives as opposed to striving for the external reward of a particular grade. Common concerns that
pass-fail and similar grading systems set a “low bar” can easily be overcome by defining stringent criteria and setting the threshold as high as necessary to ensure that students who pass have achieved the necessary skills and knowledge outcomes. This means that a pass standard need not necessarily be linked to an equivalent score of D (60%). There is sufficient evidence within health professions and pharmacy education to suggest that removing tiered grading generally does not reduce motivation or performance.19,42-45 Students who learn in environments focused on mastery versus competing for grades exhibit higher levels of motivation and achieve greater learning outcomes.46 This type of grading system also has the added benefit of increasing various forms of personal well-being19,47 because it reduces competition, encourages collaboration, and promotes self-regulation.42

Second, assessment of clinical skills should be based upon entrustable professional activities (EPAs) rather than points-based grades. Quantitative scores in clinical settings are limited in that they are often anchored and weighted on unknown or ambiguous factors. EPAs are a more holistic method of assessment that likely reflect clinical skills more precisely and authentically than do quantitative scores. Assessing students against EPA attainment levels may also provide more information to postgraduate programs (eg, residency programs) and potential employers.

Third, when it is appropriate to use exams, they should be created through blueprinting, sufficient sampling, and careful item writing. Test blueprinting is a systematic process to create valid exams by matching and linking test items with content delivered during instruction.48 Best practices for test blueprinting suggest that the relative weighting of items should depend on importance, such as frequency, urgency, or impact of the topic. For example, hyperkalemia, which may be common and life-threatening, may garner more weighting than alkalosis, which is less common and serious but not life-threatening, which, in turn, may be weighted more heavily than polyuria, which is rarely seen and nonurgent.49 To the point of not burdening students with overassessment, the more sampling that can be accomplished, the more valid our interpretations of achievement become. Attention should be given to individual questions to ensure validity and to reflect the intended competency. Recognizing resource limitations, the increased use of OSCEs as a means of assessment should be encouraged, as they may provide a more accurate assessment of a student’s ability to perform the functions of a pharmacist.

Table 2. Summary of Issues with Traditional Grading Systems and Recommendations for Reform

<table>
<thead>
<tr>
<th>Issue with traditional grading</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grades often contain information other than achievement of learning objectives (eg, attendance), which diminishes accuracy</td>
<td>Use only exams and skills assessment to assess intended learning outcomes</td>
</tr>
<tr>
<td>Grades are an imprecise measure of learning</td>
<td>Consider using a mastery learning approach</td>
</tr>
<tr>
<td>Tiered grading can create an adversarial relationship between faculty and students</td>
<td>Use a competency-based grading system with set criteria for assessing competency</td>
</tr>
<tr>
<td>Traditional grading may not accurately predict clinical performance</td>
<td>Use a different grading scale, such as entrustable professional activities, that assesses the whole student (their knowledge, skills, attitudes, etc), especially for clinical skills and attitudes</td>
</tr>
<tr>
<td>Grades are an external motivator that hinder intrinsic motivation to learn</td>
<td>Use instructional practices that promote intrinsic motivation, including those that enhance curiosity and relevance to practice</td>
</tr>
<tr>
<td>Grades, as an evaluative form of feedback, are less effective at improving learning than descriptive feedback</td>
<td>Separate grades from feedback</td>
</tr>
</tbody>
</table>

Fourth, grades should be used solely for the purposes of assessing learning versus incentivizing students to attend class, participate in class activities, or engage in other competency-developing behaviors. As stated earlier, when grades are used as motivators, a student’s focus turns away from learning and toward obtaining the points. Instead of using grades as an external reward to motivate students, we should focus more on cultivating intrinsic motivation. Instructors should find ways to provoke curiosity, provide autonomy, model enthusiasm, and emphasize that the goal is content mastery to optimize the health of patients.30

Fifth, grades should be separated from feedback, and that feedback should be descriptive rather than evaluative. When grades and feedback are paired, students may pay less attention to the feedback because of the focus on grades. Even in programs where quantitative or tiered grading systems remain the only option, faculty can still frame courses so that students are provided constructive feedback and where credit is earned through continual improvement and editing of assignments and test material. This emphasis on continued professional development can help move the focus from numeric grades to learning products and tangible outcomes. In situations where grades must be provided to students, providing feedback first while temporarily withholding grades generally results in better outcomes than providing grades and feedback simultaneously. Although the research literature does not provide best practices on the timing of sequencing, the important factor is that students are forced to engage with the feedback before receiving the grade.32

Finally, inherent challenges to widespread change exist, as various programs across the Academy use different assessment and grading methods. These differences are not only a barrier to reform but also hinder employers’ abilities to interpret grades. The Academy should consider hosting conversations and workshops combining pharmacy faculty and employer stakeholders to bring greater awareness to grading systems, their effects on learning, and what employers desire and use when interpreting learner competence. Workshops should include training and guidance on how to effectively create and use evidence-based grading practices that are superior to traditional points-based tiered grading for measuring learning, motivating students, and providing feedback.

CONCLUSION

Sufficient evidence exists to show that traditional grading systems do a relatively poor job of accurately measuring knowledge and clinical skills, motivating students, and providing feedback. Despite barriers and potential pushback from some faculty, students, and employers, programs need to consider the evidence and revise grading systems to meet several criteria. While there is no single ideal way to approach grades, a good grading system should uphold high academic standards, accurately reflect student learning outcomes/competencies, motivate students to learn, reduce undue student stress, make students feel responsible for their learning, minimize conflict between faculty and students, and give students feedback they will use.

ACKNOWLEDGEMENTS

We thank Dr Tina Brock for her constructive feedback on the initial draft of this manuscript.

REFERENCES


14. Call WB, Grice GR, Tellor KB, Armbruster AL, Spurlock AM, Berry TM. Predictors of student failure or poor performance on
37. Soric MM, Robinson JD, Ulbrich TR. Grade point average is an inappropriate measure of a residency candidate’s knowledge and skills. *J Am Coll Clin Pharm.* 2019;2(2):177-180. doi:10.1002/jac.5.1076