# Assigning a Valid and Reliable Grade in a Course

IDEA Paper #79 • May 2019

Thomas M. Haladyna • *Arizona State University*


## Abstract

The author discusses valid and reliable ways to assign grades in an academic course in any discipline. *Validity* means the accuracy of a grade's reflection of student learning and achievement. *Reliability* concerns the degree of random error that might be present and affect validity. First the author defines a grade as a measure of achievement, then identifies problems with existing grading practices and outlines principles that lead to valid grading. Grading criteria, which should be carefully and thoughtfully selected, are reviewed and categorized as "highly recommended," "controversial," and "inappropriate." Whichever grading method instructors choose should accurately measure their students' achievement, because a grade represents a contract between instructor and student that accurately reflects a particular level of learning. Finally, the author discusses grade inflation.


*Keywords:* College grading, grading criteria, validity, reliability

In this paper, I provide guidance for instructors in higher education and professional schools on valid and reliable grading for any course of study or discipline, based on a large body of scholarly essays, research, and my experience as a teacher educator and educational psychologist. I also draw on cognitive-learning and psychometric theories, as well as principles of effective teaching.

*Validity* and *reliability* are the central concepts in measuring student achievement. Briefly, *validity* addresses the accuracy of a test score, grade, or other measurement. In this context, a grade must accurately reflect a level of achievement. *Reliability* considers how much random error might invalidate a grade. In this context, an accurate grade of B might easily end up as a grade of A or C due to large random error. We must ensure that a grade is as accurate and error-free as possible.

## Defining a Grade

A course grade in any college, university, or graduate or professional school represents a student's achievement. Measurements of student achievement have typically been based on psychometric theory. Achievement is a quantitative variable that is customarily represented by a number that purportedly reflects the amount of learning that has occurred. The instructor calculates this number based on criteria that he or she has established for that particular course. It resembles a test score, which is often used as one of the many measures of achievement. The instructor then translates that number into a defined category in a hierarchy (A-B-C-D-F).

However, some information is lost in this translation. For example, within the B category, scores can range from a low B to a high B. Assigning a B to a student obscures the quantitative information represented by a low B or a high B. Thus, the grade simplistically reduces information regarding student achievement. Administratively, most colleges, universities, and professional schools nonetheless require a grade.

## The Importance of Grades

Student grades serve at least four important roles (Walvoord & Anderson (2009): (a) to evaluate the quality of a student's achievement; (b) to communicate to the student and also to graduate schools, employers, and other interested parties about the achievement; (c) to motivate the student; and (d) to demarcate the transition from the course to a future course or role in society. To advance in any program of study, a student usually requires a passing grade. Failing a course of study may eliminate a student from a program.

## Research on Grades

Brookhart and colleagues' comprehensive review of grading practices provides many insights into current-day grading practices (Brookhart et al., 2016), which, both historically and currently, are very diverse, perhaps as a result of academic freedom. However, a strong trend now exists toward making grading practices more uniform and attentive to validity. As courses increasingly become linked to state and national standards—some of which involve professional accreditation—uniform, valid, and reliable grading practices are now essential components of effective teaching and a good education. The following criticisms of poor grading practices are well documented in the Brookhart review.

First, grades often possess low reliability, especially when they are based on very little information collected during the course; an example is the traditional practice of using only a midterm exam and a final. A remedy is to collect more information to establish a more reliable basis for the grade. For example, an instructor can use more quiz scores, written assignments, activities, and other accomplishments that reflect learning.

Second, sometimes the instructor provides either scant grading criteria or none and subjectively determines a grade based on only his or her opinion; such subjective judgments often involve a rating scale, where the instructor estimates the degree of learning that has occurred. However, human judgment often undermines validity. An example is a student who performs a musical piece, with the instructor judging the performance using a five- or seven-point rating scale. In contrast, tests, quizzes, and performance that are scored without involving judgment result in objective information. Right versus wrong answers and timed performance, as would be observed with a test for typing speed, are examples of objective measures. The more we rely on objective data, the better.

Third, grades are sometimes based on student interest in and attitude toward the subject matter. Although these are important factors in teaching effectiveness, they should not influence grades.

Fourth, students may appear to be engaged in learning without actually being engaged. This pitfall is remedied by not attempting to superficially judge how much students are engaged.

Fifth, basing grades on student classroom behavior or misbehavior is another poor grading practice. Is how students behave in class a measure of how much they have learned?

A review by Yorke (2008) offers an additional critique that considers ambiguous grading criteria, low reliability, subjectively evaluated criteria, the mishandling of data to arrive at a grade, and variations in the influence of university departments on an instructor's grading. For instance, a department, college, or university might impose limits on grades to combat grade inflation, which I discuss later. One news report describes attempts at universities to impose standards in order to lower grades (Mansharamani, June 22, 2016).

## The Importance of National Standards

As stated in my definition of a grade, measuring student achievement validly and reliably is its

objective. Fortunately, we have considerable guidance, in the form of *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Although the *Standards* are intended for tests and testing programs, many of its guidelines are also useful for ensuring that a course grade truly represents a valid level of achievement for each student. The standards and discussions from the following chapters support the guidance offered in this paper.

Chapter 1 addresses validity and emphasizes its importance. As noted previously, validity requires accuracy and truthfulness: a grade must accurately reflect student achievement. This chapter describes factors that may inflate or deflate measures of student performance and thus invalidate a grade. Chapter 2 discusses reliability as a precondition for validity. As noted previously, reliability involves reducing random error to improve validity. Chapter 3 presents threats to validity concerning fairness. Chapter 12 supplies some standards for student testing. Many factors affect the creation of accurate tests for assessing student achievement. One impediment to accuracy is imposing an unfair time limit. Most test takers need sufficient time to respond to test items. Imposing a too-strict time limit might unduly reduce performance. For instance, students whose primary language is not English tend to take longer to complete English-language tests.

## What Is Achievement?

Cognitive and educational psychologists have described achievement in various ways (Lohman, 1993; Messick, 1984; Sternberg, 1998). The phrase *knowledge, skills, and abilities* provides a useful basis for defining achievement: the subject matter in a course that its instructor expects students to master.

### *Knowledge*

*Knowledge* is a fundamental element of learning, in which facts, concepts, principles, and procedures are studied for recall or understanding. Recall involves rote memory—the lowest level of learning. Understanding requires comprehension, which is a slightly higher level of learning. For example, a student can memorize and repeat the definition of the word *metaphor*, which is recall. But when a student identifies an example of a metaphor not previously seen, that signals understanding. Elementary- and secondary-school teachers and undergraduate, graduate, and professional-school instructors all teach knowledge. The most efficient and effective way to measure knowledge is the multiple-choice test, as described in Haladyna (2018) and Haladyna & Rodriguez (2013); these references also discuss the cognitive complexity of knowledge in depth.

### *Skill*

A *skill* is a simple performed act. Examples include using correct grammar in writing a sentence, adding and subtracting, and accurate keyboarding. Skills are also demonstrated in music, art, and physical education. Knowledge is antecedent to skill. Knowing how to perform a skill helps students perform it. Many skills can be evaluated precisely with measuring instruments (e.g., a yardstick, a timer, a scale) or visually, as in running, jumping, and throwing something. Occasionally, a skill must be judged by an expert using a rating scale, such as in musical performance or artistic creation, which is a subjective way to measure a skill.

### *Ability*

*Ability* is a very complex mental construction. Lohman (1993) refers to a *fluid ability*, Messick (1984) to a *developing ability*, and Sternberg (1999) to *developing expertise*. The common thread in these definitions is, simply, that knowledge and skills are combined in a complex way to perform a task: examples include problem solving, critical thinking, creative thinking or production, evaluation, or analysis. Most professions and professional schools, for instance, are familiar with a domain of tasks that are typically performed in their field. Each domain is developed in consultation with practitioners in that profession (Raymond, 2016).

The education and licensing of a dentist is a good example. The technical report for the annual examination provides a comprehensive description of the processes necessary to validate a testing program

(American Dental Association, 2017). In dentistry, a domain of tasks that dentists perform is identified. In dental school, the student acquires knowledge and develops skills and the ability to practice dentistry. To earn a license in a state, a candidate must pass two important knowledge tests and one performance test that samples from the domain of tasks typically performed by dentists in practice. In school, the dental student practices tasks from that important performance domain. The same process holds true for virtually all professions. In undergraduate, graduate, and professional programs, courses of study should emphasize the knowledge, skills, and abilities to be learned and select grading criteria that are consistent with validity theory (Kane, 2006; 2016).

If a direct measure of validity is desired, grading criteria should focus on performing complex tasks. Each of us has lifelong abilities, examples of which are listed in the first column of Table 1. The second column contains developing abilities. For example, if one is training to be a criminologist, he or she must master domains of knowledge, skills, and complex tasks before earning a credential or license. Most lifelong and developing abilities are interrelated. All are slow-growing. Each takes a lifetime to improve. A course of study is an opportunity to grow one or more abilities. Every instructor should help students further develop lifelong abilities but also move students along the continuum of one or more developing abilities. Developing abilities are those introduced later in life, such as those we cultivate when we choose an occupation or profession. Developing abilities rely on lifelong abilities but mainly include unique knowledge, skills, and abilities.

Table 1

*Abilities That Are Developed in a Lifetime*

| Lifelong abilities | Examples of developing abilities |
|---|---|
| Reading, writing, speaking, listening<br>Mathematical problem solving<br>Scientific problem solving<br>Critical thinking<br>Creativity<br>Analytical thinking | Accounting, architecture, athletics, engineering, criminology, dietetics, electrical trade, finance, fire prevention, law, medicine, nursing, policing, plumbing, poetry, pharmacology, psychiatry, physical therapy, sculpting, social work, teaching |

## Validity

*Validity* refers to the truthfulness and accuracy of the evaluation of a student's achievement. A grade should validly reflect what the student has learned in the course. The content of any course can be reified as domains. Each domain is typically very large. Thus, the best any instructor can do is draw from a representative sample of course material when establishing grading criteria. Validity rests on not only defining the domains but on showing students that the criteria you select to measure their learning is a fair, unbiased sample. A comprehensive discussion of the use of domains and the reasoning underlying validation can be found in Kane (2006; 2016).

Simply stated, a grade is validly assigned and reflective if the following conditions are met:

1. Domains, the subject matter, have been identified and shared with students.
2. Instruction provides opportunities to acquire knowledge and skills from these domains and apply each in complex ways to tasks.
3. The grading criteria represent a reasonable and fair sample of each domain.
4. Grading principles are identified and shared with students.
5. Grading standards are fair and publicly revealed on or before the first day of class.

## Threats to Validity

Many threats can undermine the accuracy of a test score or the collection of data leading to a grade; major ones include *systematic error* and *domain misrepresentation*. Each has posed a major concern for those involved in measuring student achievement.

*Systematic error* is bias, which distorts measures of student achievement by either increasing or decreasing the measure. Some examples of systematic error include poor or no instruction, student cheating, poor health or illness of the student, test anxiety, and a student's lack of motivation. Chapter 12 of the *Standards* (AERA et al., 2014) covers sources of systematic error associated with administering the quizzes, tests, and other assignments on which a grade is based.

Scoring is another source of systematic error. This is especially true for subjectively scored performance. One of the most underrated and underestimated challenges is student English-language proficiency. Many foreign students' command of English is not as good as that of their native tongue. They may have mastered course content but not fully mastered the English language. The remedy is to use appropriate vocabulary and less-complex sentence structure. A huge and growing body of research addresses this significant threat to validity (Abedi, 2016). Abedi advocates creating testing practices in which the language complexity is suitable for the students. Among his suggestions is the use of simpler vocabulary and less-complex sentences, while retaining terminology and jargon that are necessary to the content. This accommodation helps students who struggle with the English language.

*Misrepresentation* involves a poorly drawn sample of knowledge, skills, or the complex tasks of an ability, which may bias a domain. For example, some instructors tend to avoid performance tasks as a basis for grading and resort to multiple choice to measure knowledge and skills. This may misrepresent the important learning outcomes that they desire. For instance, would a multiple-choice test suffice for driver licensing? Critics of multiple-choice testing can rightly point out that measuring knowledge well does not make up for not measuring the more important tasks involved in an important ability (Haladyna, 2018). Chapter 12 of the *Standards* (AERA et al., 2014) is devoted to the importance of test design in achieving a high degree of validity where instruction is being offered and test results have important consequences. The chapter offers the following guidelines for course design.

1. Any test or quiz should fairly represent the domain of knowledge and skills being taught. A table of specifications (also known as test specifications or a two-way grid) is typically recommended for achieving fair domain representation. The table of specifications shows the percentages of test items that an instructor allots for each topic in the course. It is useful for designing quizzes, tests, and other grading criteria. Typically, the multiple-choice test format is best suited for a table of specifications (Haladyna, 2018).
2. Any assignment or project should represent the domain of tasks that most directly demonstrate the ability being developed. In most instances, this domain involves tasks that require performances.
3. The weighting from knowledge, skills, and abilities should be public (i.e., in the syllabus) and well known to students.

## Reliability

*Reliability* summarizes how much random error might exist in any test score. Random error can be large or small, positive or negative. We never know. However, we can try to ensure that the number representing student achievement is as reliable as possible, with a margin of random error so small that a grade accurately reflects what a student has learned. In a course, the use of several measures of student achievement results in a number representing the amount of learning. One might think of it as points earned versus points possible (750 points earned out of 1,000 possible points is 75%). If an instructor tested all items or tasks in a domain—so that all knowledge, skills, and abilities were tested—a true score would result. A true score is error-free. Unfortunately, as previously discussed, it is impossible

or impractical to sample an entire domain. So, the points earned are what a student achieves on a fair sample of all items or tasks, plus the existence of random error.

Psychometric theory and the *Standards* (AERA et al., 2014) both maintain that more information leads to a more reliable result and a lower degree of random error. To maximize reliability, it is therefore useful to have as many sources of information about student achievement in the course as possible. A good example is the use of many grading criteria, such as quizzes, tests, projects, and class-participation activities. Moreover, the multiple indicators cover the content more comprehensively, which improves validity. In other words, more grading criteria not only increases reliability but increases validity.

In contrast, low reliability casts doubt on the validity of a grade. For example, if a final examination is the only basis for a grade, the instructor has only one source of data. If the test is not sufficiently reliable, the grade will be compromised by a large amount of random error. A student might thus receive a course grade of A, B, or C merely due to random error and not actual achievement.

## Principles of Grading

Next, let's consider a set of recommended principles for grading (Haladyna, 1999). These principles are not axiomatic but were derived from various sources, including a large number of textbooks that focus on measuring student achievement and grading. Individual instructors will likely choose principles that reflect their ideas of good instruction and valid grading.

1. Grades should be based on what each student has learned. Nothing else matters. We infer a level of learning, and thus assign a grade, based on the data collected on each student during the course.
2. Instructors want their students to earn high grades. High grades ideally reflect high achievement and effective learning and teaching. Of course, low standards can also yield high grades, but that situation should be avoided.
3. As noted previously, grades are important. Of course, some very successful people have had low grades. Nevertheless, if grades reflect a level of learning in a course of study, high achievement, and therefore earning a high grade, is important for every student. In addition, grades provide an important criterion when a student wishes to continue their education at a higher level. Moreover, they foster student morale and improve motivation.
4. Grading is confidential. The Family Educational Rights and Privacy Act, a federal law with which all higher education institutions must comply, protects the confidentiality of a student's grade.
5. The assignment of a grade should be based on valid and reliable information.
6. Grades should be objectively determined, because subjective judgment has many shortcomings, as reported in many essays and books (Meehl, 1954; Egisdottir et al., 2006). However, sometimes, subjective judgment is the best an instructor can offer. In those instances, students should have the opportunity or the right to appeal. Granted, objective determination of a grade ultimately depends on a subjectively determined set of grading standards, but such subjectivity is unavoidable. The argument defending the subjective establishment of grading standards is that an instructor's expertise and experience weighs in favor of setting fair and defensible standards.
7. Your grading policy and procedure should be written and clearly presented in your course syllabus on the first day of class. As noted previously, this is a hallmark of highly effective instruction. It alerts students about what to learn and how to plan their studying.
8. The choice of the grading method should incorporate your definition of student learning, the learning activities you choose, and the body of evidence (i.e., grading criteria) that you will use to assign the grade.

## Grading Criteria

Traditionally, quizzes and tests provide criteria for assigning a grade. Table 2 lists grading criteria that are deemed either highly recommended, controversial,

or inappropriate. Instructors may want to consult with colleagues to see their grading criteria, but the responsibility for choosing criteria is strictly up to the individual instructor. Sometimes a college, university, or professional school might have a standardized grading policy for multiple sections of the same course. In that case, the grading policy is developed by a team of instructors.

Table 2
*Grading Criteria*

| Highly recommended | Controversial | Inappropriate |
|---|---|---|
| Creative work<br>Critique<br>Demonstration<br>Essay<br>Exhibit<br>Experiment<br>Group project<br>Homework<br>In-class activity<br>Individual project<br>Paper<br>Performance<br>Portfolio<br>Quiz<br>Review<br>Test | Attendance<br>Class participation<br>Extra credit<br>Improvement over time<br>Subjective judgment<br>Violation of a deadline | Appearance<br>Attitude<br>Behavioral problems<br>College- or department-imposed grading standard<br>Disability<br>Effort<br>Enthusiasm<br>Emotional need<br>Gender<br>Hygiene<br>Intelligence<br>Interpersonal skills<br>Neatness<br>Personality<br>Race/ethnicity<br>Religion<br>Reputation<br>Verbal ability |

The categorization of criteria in Table 2 warrants some discussion. The criteria in the first column are examples of appropriate measures, assuming that standards of validity and reliability are followed. For greater depth on the design of these highly recommended criteria, consult Haladyna and Rodriguez (2013), which provides comprehensive treatment of the design and construction of test items and tasks.

### *Controversial Grading Criteria*

1. Many instructors think that attendance is important and include it in their grading criteria. Would you argue that a student who attends class learns more than a student who does not? Or could a student skip class and learn the subject matter just as effectively in an alternative way?
2. Do students who contribute to class discussion and ask questions learn more than those who simply sit and listen (or who daydream or play with their phone, tablet, or laptop)? In some classes, such as one devoted to learning a language, participation is very important.
3. Should students be given opportunities to improve their grade by doing extra work? Some types of extra credit repeat what was to be learned, whereas other types might serve to extend what was learned. In addition, extra credit might be assigned to remediate some knowledge, skills, and abilities that have not yet been learned. The key point is that extra credit extends time to learn and is a good thing if it is linked to course objectives and fills a gap or fulfills a need.

4. Some students start at a very low level but show considerable improvement. Should they be given a higher grade than a student who showed little improvement? Keep in mind that a floor and ceiling effect may exist. A student who starts near the ceiling cannot show much improvement. In contrast, a student with very low performance at the beginning of the course has more potential for gain. If the grade is based on improvement, the higher performing student may, ironically, receive a lower grade.
5. Sometimes an instructor must evaluate performance subjectively. In the performing and creative arts, instructors must make expert judgments about the quality of a performance or a product—such as a play, sculpture, or dance. Invariably, a rating scale is used. As noted previously, human judgment can be flawed in many ways. Bias can occur when the grade is based on a criterion other than achievement. Severity (underrating) and leniency (overrating) are examples of bias. Other biases in rating a performance or the quality of a product include *central tendency*, *compression*, *inconsistency*, *idiosyncrasy*, and *halo*.

   *Central tendency* can occur when ratings fall mainly in the center of the rating scale. *Compression* results when all students receive the same rating. *Inconsistency* involves rating the same level of performance highly sometimes and poorly other times, which breeds unreliability. *Idiosyncrasy* means simply an irrational rating. For example, one reviewer of a student essay might find the subject matter chosen by the student to be repulsive. An undeserved low rating follows, not because of the writing quality but because of the subject matter, which may or may not be relevant (Haladyna & Rodriguez, 2013). *Halo* results when several traits are being judged on a rating scale, and the first impression carries over to all traits without discrimination.

   Overall, subjectivity undermines validity and should be avoided if possible. If subjectivity must be used, it should be employed carefully, and additional judges should be brought in to confirm a judgment.
6. There are many valid reasons that a student might fail to meet a deadline. Should violating a deadline affect a student's achievement? This is a hard call for some faculty members. On the one hand, in the real world outside the classroom, deadlines are often fixed, and missing one has real consequences. However, courses have real time limits, because a semester ends and a grade is expected at that point. On the other hand, some deadlines can be extended. If a student is given more time, his or her work should be better as a result. Some instructors may thus choose to allow students to file for incompletes and finish the work later.

### *Inappropriate Grading Criteria*

The list in the third column of Table 2 may seem unthinkable, but occasionally some items on this list are actually used to grade students. It would be very difficult to defend using any of these inappropriate grading criteria. In fact, some are illegal.

## Grading Methods

Although many grading methods exist, several are recommended for those who want to provide students with some advantages that a more traditional grading method may not provide. The source for these recommendations is an extensive review of past and current grading practices and advice provided in standard textbooks on measurement and evaluation in the classroom (Haladyna, 1999). One stands out as traditional and highly recommended: absolute standards. Each of these grading methods, however, is subject to the instructor's consideration and evaluation. What works for some instructors may not work for others.

### *The Normal (Bell-Curve) Method*

The normal distribution (bell curve) is well known to statisticians. Shown in Figure 1, this grading method assumes that student achievement distributes itself normally in a class. Assigning grades should thus follow the bell curve. However, this method is not recommended, because students in higher education are above average on most cognitive measures, and therefore this assumption is seldom true.

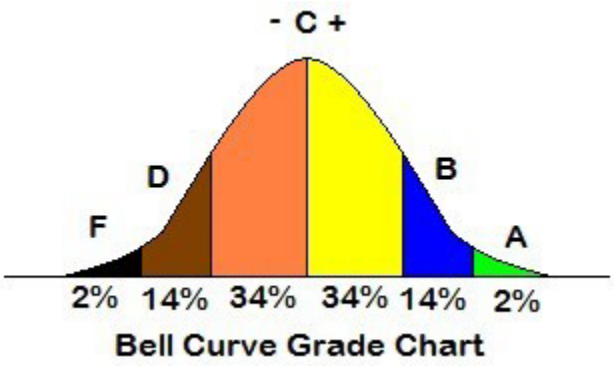


*Figure 1.* Bell-curve grade chart.

Proponents argue that normal (bell-curve) grading prevents grade inflation because the class grade-point average is 3.0, or C. However, consider the fact that grades for university and professional-school students are distributed as in Figure 2—a negative skew. Also, the goal for any instructor is to skew the distribution more negatively, because that is what teaching and learning is all about.

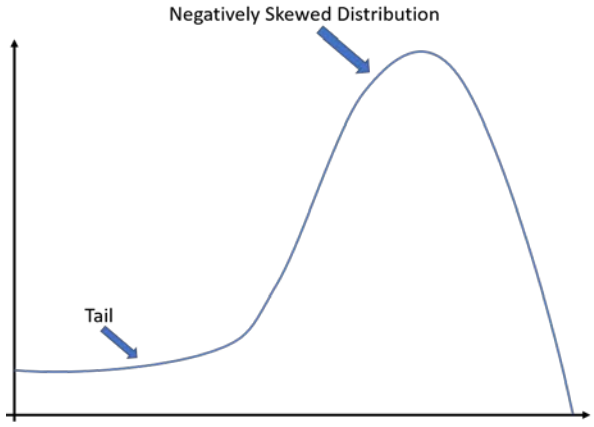


*Figure 2.* Negatively skewed distribution.

### *Negatively Skewed Distribution*

The most extensive critique of the normal-curve method can be found in the book *The Myth of the Normal Curve* (Dudley-Marling & Gurn, 2010). Another good source is the review of research and discussion by Brookhart et al. (2016). In truth, normal-curve grading has an ugly history, because it was used to cull the "feebleminded" students from school (Wallin, 1912). Nonetheless, some may argue that normal-curve grading is objective and mathematical, even though it is hardly fair to students. Because the method promotes unhealthy competition, students seldom want to help their peers or work cooperatively. Moreover, even if all students learn and do well, those scoring at a lower level will receive a lower grade. To summarize, normal-curve grading has no value in modern education.

### *Pass/Fail*

In the pass/fail (PF) system, students earning A, B, C, or D get a "pass." The rest get a "fail." This method essentially combines a variety of achievement levels into a single summary (pass). Melrose (2017) provided an excellent review of this grading method, including a set of pros and cons. The most obvious shortcoming is that a D student's pass has the same value as an A student's pass. Although they actually achieved at different levels, the pass makes it look as if they achieved to the same degree.

If a D-level amount of learning is sufficient for a pass, why bother to achieve more than a D? In fact, one study done long ago found that most students tended to do work in the C and D range in a PF class (Stallings & Smock, 1971). Another knock against the PF grading method is that a letter grade has greater value when a student applies to continue advanced learning, such as in a graduate or professional school.

Putting aside the arguments against PF, a clear advantage is the positive impact on students' emotional well-being. For example, in one study of medical school students, no achievement differences were found between conventional grading and the PF method. However, students in the PF-graded class showed more improvement in emotional well-being (Bloodgood et al., 2009). Some colleges, universities, and professional schools consequently use the PF method or offer it as an option. Instructors may then merely follow the guidelines provided by the institution.

*Innovative Grading Methods*

Instructors interested in the merits of some innovative and nontraditional grading methods might consider one of the following. All are recommended, but only in certain circumstances that only the instructor can determine.

***Mastery.*** In a mastery approach, the student is given as much time and as many chances to succeed as possible. The theory posits that, with effective teaching and additional time to learn, students with low achievement will improve (Bloom, 1974). Research supports the theory (Guskey, 2010). However, mastery is very difficult, expensive, and time-consuming to design and administer. When mastery learning is implemented, grading takes on a different meaning. Students who persevere will doggedly complete units of instruction and attain a high level, justifying a grade of A—which looks like grade inflation. However, because students spent more time learning, they learned more, so a higher grade is justified. In a professional program, mastery makes much sense. Who wants their daughter's brain tumor removed by a C student in surgery? Mastery may thus be an ideal method for teaching, learning, and grading, but its primary obstacle is feasibility.

***Individual program.*** The individual education program is a very familiar device in special education. Each student gets an individual plan based on his or her disabilities and other factors. Individual programs also can be developed and used in undergraduate and graduate education. However, in professional schools, individual programs do not make any sense, because the domain of tasks to be learned is immutable. A readings-and-conference course is the best example of the individual program. Another example is when a graduate student requests a specialized, individual course that supports an individualized program of study leading to an advanced degree. Some instructors can accommodate this need, although it is time-consuming and has limitations. Beyond a single individual course, an entire individual program can be customized for a special purpose or objective. As science becomes more diffuse with specialties and subspecialties, individual graduate programs make sense.

***Contract grading.*** The instructor negotiates with a student or a group of students that they will accomplish certain assignments, projects, or performances in exchange for a particular grade. For theatrical, musical, or other types of group performances, group projects, or specialized education programs, a contract may make sense. The instructor has to set the guidelines, rules, criteria, and consequences for the grade. This is not easy, but it is very student friendly, because the student has the opportunity to negotiate.

***Blanket grading.*** The blanket grade is an administrative grading method in which a group of students performs as a team and each student receives an identical grade. Occasionally, blanket grading makes sense, such as in collegiate competition, where awards are based on a team result. There is no discrimination between team members, no matter which one contributed the most or least.

*Absolute Standards (Highly Recommended)*

Setting absolute standards is a widely used method, for many good reasons. The instructor weights criteria before determining grades. For example, a portfolio might be weighted 50% of a grade. The instructor then adds up all weighted performance points to obtain that number, representing a level of achievement. Following this, the instructor applies the points earned by each student to the subjectively determined standard, as shown in Table 3. However, the values displayed in Table 3 are arbitrary. Therein lies a criticism of absolute standards: Individual professional judgment is the criterion for establishing these values.

An assumption underlying absolute standards is that the instructor has experience and high standards and will thus ensure that an A is earned and fair. If, however, the instructor's standards are harsh, few students will earn high grades. If, on the other hand, standards are lenient, most students will earn As and Bs. However, instructors are ordinarily subject to annual reviews of their teaching. During this peer-review process, harsh or lenient tendencies are observable. Thus, faculty members can be encouraged to adjust their standards to be fair to students and improve the validity of their grades.

Table 3

*Absolute Standards*

| A | B | C | D | F |
|---|---|---|---|---|
| 100–90 | 89–80 | 79–70 | 69–60 | Below 60 |

Because a moderate correlation exists between expected grades and student ratings of instruction (SRI), some argue that SRI has led to grade inflation; that is, in order to achieve high ratings, some instructors might apply lenient absolute standards (Brookhart et al., 2016). However, these researchers also report that the relationship is more complicated than simply attributing ratings to the leniency or severity of standards. In a review of studies on SRI, Benton and Cashin (2010) reported low-to-zero correlations between grades assigned and student ratings, based on several sources. However, this low correlation may be due to three eventualities. First, students who achieve more earn higher grades and give higher ratings, which is called the *validity hypothesis*. Second, some lenient instructors may still assign higher grades than students deserve in order to earn higher ratings than the instructor deserves (*leniency hypothesis*), even if that is not the outcome. Third, student variables, such as interest or motivation, may lead to greater learning and higher grades, which results in higher ratings.

## Developing a Grading Method for a Course

Whatever principles the instructor adopts, the next step is to describe the body of evidence that will be used to determine the grade. The course syllabus is the vehicle for presenting and describing the chosen grading method. The following actions are recommended:

- Reiterate that the grade is based on achievement in that course, not on other factors.
- Make clear what students are expected to learn. Expected outcomes, class activities, and studies and activities outside of class should be connected directly with those written expectations.
- State the criteria used to calculate the grade. This body of evidence will likely include quizzes, tests, completion of activities, and a portfolio.
- Use points with absolute standards.
- Provide the weights for each criterion for the grade. As shown in Table 3, assign points for each item in the body of evidence.
- Keep detailed records of student point counts.
- Develop a policy and procedure for borderline situations. Some students will invariably fall one point above or below a cutoff. Is there any sympathy for those falling below that point? Are there mitigating circumstances that might help the student? For instance, student health, family emergencies, and financial exigency are some circumstances where flexibility might be exercised. Another approach is to offer remedial work or extra credit to help the student move across a cutoff from a lower grade to a higher grade.

## An Abbreviated Example of Grading Criteria for Students Training to Be Teachers

Because I aspire to a high level of achievement for every student, I incorporate a mastery component in the class Measurement in Education. These students are going to be teachers; therefore, all need to perform at a high level in this class. If some students fail to meet this standard, they have additional chances to improve by increasing study time and redoing assignments. They create a portfolio, which is a collaborative project, with me as their resource. I encourage them to develop original material that will be useful in their own teaching, but the material has to meet high standards. Therefore, sometimes I return their portfolios for further improvement. In extreme circumstances, a student may need to apply for an incomplete grade or repeat the course. Under this mastery condition, some students spend additional learning time to achieve at a high level. Once they attain this level of achievement, they earn the appropriate grade—usually an A. This might appear to be grade inflation, but following Bloom's mastery theory, it is defensible (Bloom, 1974). All criteria are linked to state requirements for teacher licensing.

Table 4

*Example of Absolute Standards and Weightings for Grading Criteria*

| A | B | C | D | F |
|---|---|---|---|---|
| 1000–900 | 899–825 | 824–750 | 749–700 | Below 700 |

*Note.* Eight quizzes, each worth 50 points—400 total points; oral presentation on a current issue—100 points; cooperative project-—100 points; portfolio— 400 points.

## Grade Inflation

According to one source, approximately 15% of all students in higher education in 1940 earned a grade of A. In 2012, that rate increased to 45%. If student learning increased during these 72 years, the 45% figure is accurate. However, the following competing hypotheses could explain this result.

1. Previously, grading standards were too harsh. This may be true if the normal-curve method was used. The normal-curve method underreports achievement for a typical high-achieving university, graduate, or professional-school student.
2. Grading standards currently are too lenient. Rampell (July 14, 2011) reported that 43% of letter grades in higher education were As, whereas in 1960, 28% were As. What accounts for this difference and what can be done about it? If grading leniency is the cause, it represents a threat to validity, which involves accuracy and truth. The solution is the prudent use of absolute standards coupled with a valid and reliable body of evidence regarding student achievement, which should result in greater discrimination among the grades that students receive.
3. If an instructor uses mastery learning strategies to improve learning, higher achievement is the outcome, because a mastery approach forces many students to spend more time learning—resulting in more learning. Thus, students achieve higher grades because of their persistence. A dilemma is whether a student who takes more time to learn should earn a lower grade?

Which of these three hypotheses is true? It is hard to determine. One interesting perspective, from Kohn (2002), is that grade inflation is a myth. An examination of college transcripts sampled across the nation suggests that grades have not risen as sharply as students have reported in surveys.

However, national standardized tests do not bear out that today's students are more able and have learned more. In fact, the National Assessment of Educational Progress reports that student achievement has remained relatively level for many years. National Center for Educational Statistics (2013). This finding might support the hypothesis that grading standards currently are too lenient.

What is clear is that there is considerable variety in grading practices across 35 university websites in different countries (Brookhart et al., 2016). Some universities use fixed standards, such as assigning a passing grade to any student who earns 50% of all possible points in a course. Other universities use essays extensively, which are subjectively scored and very problematic regarding validity and reliability. Another point made in the Brookhart et al. review is that, historically, grading criteria have varied considerably and changed over time. The authors' central premise is that a grade should represent student achievement and nothing else. However, there is a strong tendency to incorporate student progress, behavior, participation, attitude, effort, and attendance into grades. This practice seems deficient, because the meaning of a grade differs from one instructor to another.

A final influence on grade inflation is normal-curve grading. The science of psychometrics often relies on ranking student performance. However, if student

performance is clustered within a tight range of scores, this ranking procedure does not hold up well. Brookhart et al.'s review (2016) reveals the breadth of research on grading and grade inflation, but the reasons for apparently higher grades are still a mystery.

One possibility is that, historically, colleges and universities adopted the normal-curve grading method, which guaranteed an average grade of C. As this method was rightfully abandoned, grading practices in higher education became more diverse. Grades improved perhaps due to a fairer system of grading, leniency, greater motivation as the competitive normal-curve method was abandoned, or mastery approaches to learning that were student friendly and more effective.

If you as an instructor base student grades on your criteria and standards, and students achieve highly, they deserve the grade they earned. It is part of a contract between you and them. If you allow low-achieving students more time to learn, they will likely learn more and both earn and deserve a higher grade. This may seem like grade inflation, but the goal of teaching is to increase learning. Allowing more time to learn thus improves achievement for low-achieving students.

## Conclusion

Grading is a contract between instructors and students. Validity and reliability are important principles to consider when developing a grading policy. Being clear about what students are to learn and how they are to learn it is a critical foundation of accurate grading. Instructors should explain their grading principles and the criteria that they will use. The syllabus is the best way to communicate this information. If the grading criteria are correctly applied, as I describe in this paper, students will recognize the contract as fair.

## Author Biography

Thomas Haladyna, PhD, has served in many roles in his long career in education. He began as an elementary school teacher and assistant principal and was a university professor for 24 years, a test director at the American College Testing (ACT) Program, and a research professor in the Oregon State System of Higher Education. He was also a visiting scholar at the Navy Personnel Research and Development Center and a National Assessment of Educational Progress Visiting Scholar at the Educational Testing Service. His university work mainly involved classroom assessment in the teacher-education program. In his position at ACT, he managed many national testing programs in medicine and health-related fields. His position in Oregon involved a variety of funded and other sponsored research and development in testing and faculty development.

Aside from these formal duties, Dr. Haladyna has consulted for a wide variety of clients in the development and validation of testing programs and item development. He has also evaluated testing programs in various fields. His scholarly achievements include 14 books, several other monographs, 60 refereed publications, and several hundred conference presentations, white papers, opinions, invited presentations, and book chapters. He is a regular reviewer for many journals and continues to conduct research, consult, and participate in professional activities in his retirement. He is one of the most cited authors in the *Handbook of Test Development*.

## References


Abedi, J. (2016). Language issues in item development. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development*, 2nd ed., (pp. 355–373). New York: Routledge.

American Dental Association (2017). Technical report: National Board Dental Examinations. Chicago: American Dental Association.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Benton, S. L., & Cashin, W. E. (2010). *Student ratings of teaching: A summary of research and literature* (50). Manhattan, KS: The IDEA Center

Bloodgood, R. A., Short, J. G., Jackson, J., & Martindale, J. R. (2009). A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. *Academic Medicine*, *84*(5), 655–662.

Bloom, B. S. (1974). Time and learning. *American Psychologist, 29*(9), 682–688.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., et al. (2016). *Review of Educational Research, 86*(4), 803–848.

Dudley-Marling, C., & Gurn, A. (Eds.) (2010). *The myth of the normal curve*. New York: Peter Lang.

Egisdottir, S., White, M. J., Spengler, P. M., Maugerman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus actuarial prediction. *The Counseling Psychologist, 34*(3), 341–382.

Guskey, T. R. (2010). Lessons of mastery learning, *Educational Leadership, 68(2)*, 52-57.

Haladyna, T. M. (1999). A *complete guide to student grading*. Boston: Allyn & Bacon.

Haladyna, T. M. (2018). *Developing test items for course examinations* (70). Manhattan, KS: The IDEA Center.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.

Kane, M. T. (2006). Content-related validity evidence. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–154). Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. T. (2016). Validation strategies: Delineating and validating proposed interpretations. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development,* 2nd ed. (pp. 64–80). New York: Routledge.



T: 785.320.2400
T: 800.255.2757

301 South Fourth St., Ste. 200
Manhattan, KS 66502

Email: info@IDEAedu.org
IDEAedu.org

Kohn, A. (2002). The dangerous myth of grade inflation. *Chronicle of Higher Education*, *49*(11), B7–B9.

Lohman, D. F. (1993). Teaching and testing to develop fluid abilities. *Educational Researcher, 22*, 12–23.

Mansharamani, V. (2016). How an epidemic of grade inflation made A's average. Public Broadcasting Sysem ((https://www.pbs.org/newshour/economy/column-how-an-epidemic-of-grade-inflation-made-as-average. Retrieved April 22, 2019)

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Melrose, S. (2017). Pass/fail and discretionary grading: A snapshot of their influences on learning. *Open Journal of Nursing*, *7*, 185–192.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, *21*, 215–237.

National Center for Education Statistics (2013). The Nation's Report Card: *Trends in Academic Progress*. Washington, DC: US Department of Education.

Rampell, C. (July 14, 2011). A history of grade inflation. Retrieved February 25, 2019 from https://economix.blogs.nytimes.com/2011/07/14/the-history-of-college-grade-inflation/

Raymond, M. R. (2016). Job analysis, practice analysis, and the content of credentialing examinations. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development,* 2nd ed. (pp. 144–164). New York: Routledge.

Stallings, W. M., & Smock, H. R. (1971). The pass-fail grading option at a state university: A five semester evaluation. *Journal of Educational Measurement*, *8*(3), 153–160.

Sternberg, R. J. (1998). Abilities are forms of developing expertise. *Educational Researcher, 27*(3), 11–20.

Wallin, J. E. W. (1912). Experimental studies of mental deficiency: A critique of the Binet-Simon Tests and a contribution to the psychology of epilepsy. *Journal of Psycho-Asthenics, XVII*(2), 76–80.

Walvoord, B. E., & Anderson, V. J. (2009). *Effective grading: A tool for learning and assessment in college*, 2nd ed. San Francisco: Jossey Bass.

Yorke, M. (2008). *Grading student achievement in higher education*. London and New York: Routledge.