SPECIAL ISSUE
Historical and Contemporary Perspectives on Educational Evaluation

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 26 Number 49 April 16, 2018 ISSN 1068-2341

A Critique of Grading: Policies, Practices, and Technical Matters

Lorin W. Anderson
University of South Carolina (Emeritus)
United States

Citation: Anderson, L. W. (2018). A critique of grading: Policies, practices, and technical matters. *Education Policy Analysis Archives, 26*(49). http://dx.doi.org/10.14507/epaa.26.3814 This article is part of the Special Issue, *Historical and Contemporary Perspectives on Educational Evaluation: Dialogues with the International Academy of Education,* guest edited by Lorin W. Anderson, Maria de Ibarrola, and D. C. Phillips.

Abstract: In recent years there was been a raft of criticisms of the way that grades (or marks) are assigned to students. The purpose of this paper is to examine the strengths and weaknesses of grading systems and grading practices, drawing upon both historical and contemporary research and writing. Five questions are used to frame the review and organize the paper. They are: (1) Why do we grade students? (2) What do grades mean? (3) How reliable are students' grades? (4) How valid are students' grades? and (5) What are the consequences of grading students? The results suggest that (1) The are several purposes for grading students; the way that grades are assigned and reported should be consistent with the specified purpose. (2) Grades mean different things to different people (including the teachers who assign them). (3) Grades on a single task (e.g., a test or project, a

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 1/15/2018
Revisions received: 2/18/2018
Accepted: 2/18/2018

homework assignment) are quite unreliable, whereas cumulative grades (that is, those based on several data sources) are reasonably reliable. (4) The validity of grades on a single task is virtually impossible to determine; however, the evidence suggests that cumulative grades are reasonably valid. (5) Grades influence a variety of student affective characteristics (e.g., self-esteem). However, their influence is no greater, nor less than, a host of other school-related factors.
Keywords: Grading; standards; fairness; education policy

Una crítica a las calificaciones: Políticas, prácticas y asuntos técnicos
Resumen: En años recientes ha habido una racha de críticas a la manera como las calificaciones o grados se asignan a los estudiantes. El propósito de este artículo es examinar las fortalezas y las debilidades de los sistemas y las prácticas de calificación, con base en texto e investigaciones históricas y contemporáneas. Se aprovechan cinco preguntas para marcar los límites y organizar el texto. 1) Por qué calificamos a los estudiantes. 2) Qué significan las calificaciones. 3) Qué tan confiables son las calificaciones de los estudiantes. 4) Qué tan válidas son. y 5) Cuáles son las consecuencias de calificar a los estudiantes. Los resultados sugieren que: 1) Hay diferentes propósitos para calificar a los estudiantes; la manera como se asignan y se reportan debiera ser consistente con el propósito especifico. 2) Las calificaciones tienen diferentes significados para diferentes personas, (incluyendo a los profesores que las asignan). 3) Las calificaciones asignadas a una sola tarea (esto es una prueba, un proyecto, una tareas) son muy poco confiables, mientras que las calificaciones acumuladas (esto es las que se basan en diferentes fuentes) son razonablemente confiables. 4) La validez de las calificaciones asignadas a una sola tarea es casi imposible de determinar, en cambio la evidencia sugiere que las calificaciones acumuladas son razonablemente válidas. 5) Las calificaciones influyen sobre las características afectivas de los estudiantes (por ejemplo, la autoestima), sin embargo su influencia no es mayor ni menor que un montón de otros factores escolares relacionados.
Palabras-clave: calificar; estándares; justicia; política educativa

Uma crítica das qualificações: Políticas, práticas e assuntos técnicos
Resumo: Nos últimos anos, tem havido uma série de críticas sobre como as notas ou notas são atribuídas aos alunos. O objetivo deste artigo é examinar os pontos fortes e fracos dos sistemas e práticas de qualificação, baseados em textos e pesquisas históricas e contemporâneas. Cinco perguntas são usadas para marcar os limites e organizar o texto. 1) Por que classificamos os alunos? 2) O que as qualificações significam 3) Quão confiáveis são as notas do aluno. 4) Qual a sua validade e 5) Quais são as consequências de qualificar os alunos? Os resultados sugerem que: 1) Existem diferentes propósitos para qualificar os alunos; a maneira como são designados e relatados deve ser consistente com o propósito específico. 2) As notas têm significados diferentes para pessoas diferentes (incluindo os professores que as designam). 3) As classificações atribuídas a uma única tarefa (isto é, um teste, um projeto, uma tarefa) são pouco confiáveis, enquanto as qualificações acumuladas (isto é, aquelas baseadas em fontes diferentes) são razoavelmente confiáveis. 4) A validade das avaliações atribuídas a uma única tarefa é quase impossível de determinar, enquanto as evidências sugerem que as qualificações acumuladas são razoavelmente válidas. 5) As notas influenciam as características afetivas dos alunos (por exemplo, autoestima), porém sua influência não é maior ou menor do que muitos outros fatores relacionados à escola.
Palavras-chave: classificação; normas; justiça; política educacional

A Critique of Grading: Policies, Practices, and Technical Matters

> When we consider the practically universal use in all educational institutions of a system of marks, we can but be astonished at the blind faith that has been felt in the reliability of the marking systems. (Finkelstein, 1913)

> Isn't it hypocritical to preach about the importance of innovation in education while simultaneously clinging to a system of grading which is almost as archaic as it is useless. (Ferriter, 2015)

These two quotations, written a century apart, illustrate the negativity associated with the ways in which grades (or marks) are assigned to students in schools. Even a cursory search of Google Scholar or JSTOR will yield scores of articles with similar points of view. Several educators, notably Alfie Kohn (1999, 2011) and Thomas Guskey (2002, 2011), have published extensive criticisms of grading policies and practices.

Not only are the criticisms timeless, they are widespread. In addition to academicians, teachers and educational consultants have railed against grading. Writing more than a half century ago, teacher Dorothy de Zouche called the giving of grades one of her "10 educational stupidities." More recently, consultant Mark Barnes (2014) gave a TED talk in which he addressed the apparently rhetorical question, "Isn't it time to eliminate grades in education?"

Despite a century of fairly constant criticism, however, the practice of grading students remains a cornerstone of our educational system. Why is this so? Could it be that grades, despite the problems inherent in grading policies and practices, have some value? My purpose in this chapter is to offer a critique of grades, grading policies, and grading practices. By critique I mean a "careful judgment in which [one gives an] opinion about the good and bad parts of something" (*Merriam Webster Learner's Dictionary*). To facilitate my critique I will focus on five basic questions.

1. Why do we grade students?
2. What do grades mean?
3. How reliable are students' grades?
4. How valid are students' grades?
5. What are the consequences of grading students?

Let me begin with some definitions. "Grade" can be either a noun or a verb. When applied to education and used as a noun, a grade is a position on a continuum of quality, proficiency, intensity, or value. The continuum can be expressed numerically (e.g., 1 to 100), by letters (e.g., A, B, C, D, F), or using a set of verbal descriptors (e.g., exemplary, proficient, basic, below basic). When applied to education and used as a verb, "to grade" means to place a student on the aforementioned continuum based on impressions, evidence, or, more than likely, some combination of the two.

Before continuing it should be noted that early writers in the field (e.g., Rugg, 1918) as well as some British higher education institutions today (e.g., University of Liverpool, 2015) use the term "marks" rather than "grades" and "marking systems" rather than "grading systems." However, most dictionaries (e.g., *the Oxford English Dictionary*) use the terms synonymously as will I.

Why Do We Grade Students?

> What could have prompted the first teacher to start a marking system? Was it a desire to stimulate the pupils through emulation to stronger effort? Or could it have been

> through a desire to record individual shortcomings and so enable the teacher to modify his instruction accordingly? (Campbell, 1921, p. 510)

Note that Campbell mentioned two possible reasons for grading students: (1) to motivate them to put forth greater effort and (2) to provide information that teachers can use to improve their instruction. More recently, a third reason for grading has been proffered, namely, to communicate information about student learning to a variety of audiences who want and/or need information about how well students are learning or progressing in order to make decisions about the students (Bailey & McTighe, 1996).

Motivating Students

> Anyone who doubts [that] grades are not a spur needs only to recall which was uppermost in his thought during his schooldays at the end of the report periods—What is my grade? (Rorem, 1919, p. 671)

> [Although] our marking systems are fraught with innumerable weaknesses and inconsistencies … they do serve as a spur to the laggard, even their most outspoken opponents must admit. (Campbell, 1921, p. 511)

As these two quotations indicate, the belief that grades are inherently motivating is longstanding. Furthermore, because most educators at that time believed that motivation was enhanced when students competed among themselves, many, if not most, early grading systems were based on rankings among students, rather than ratings of the quality of individual student's work or learning (Cureton, 1971).

Even when critics of grading accept that grades have some motivational value (Bull, 2013), they maintain that grades foster the "wrong" kind of motivation. They point out that working harder to achieve better grades is not the same as working harder to learn more. In fact, the results of several studies suggest that these two "orientations" are inversely related (Kohn, 1999). Furthermore, students who are motivated by grades rather than learning are less likely to be interested in what they are learning (Kohn, 2011), more likely to avoid challenging tasks (Schinske & Tanner, 2014), and more likely to engage in "gamesmanship" that allows them to achieve the highest grades (or, in some cases, "acceptable" grades) with the least amount of effort (Schwartz & Sharpe, 2011).

Schinske & Tanner (2014) have provided a concise summary of what is currently known of the relationship between grading and motivation. "At best, grading motivates high-achieving students to continue getting high grades—regardless of whether that goal also happens to overlap with learning. At worst, grading lowers interest in learning and enhances anxiety and extrinsic motivation, especially among those students who are struggling" (p. 161).

Providing feedback to teachers

> In practice, the ordinary marking system simply registers relative standing with respect to other pupils in the class; … it certainly does not furnish a prescription for the teacher to follow. It is here that our marking systems break down; they do not provide for treatment. (Campbell, 1921, p. 510)

This statement is as valid today as it was almost a century ago. Grades typically do not provide information that can be used by teachers to improve their instruction (or by students to improve their learning). To be useful for improvement purposes, grades must provide information about

what students, individually and/or collectively, have and have not learned … know and do not know … can and cannot do. Advocates of "standards-based grading systems" (Scriffiny, 2008) argue that their systems provide the necessary level of detail.

In standards-based grading students are evaluated on the basis of their mastery of a clearly articulated set of course objectives (widely known as academic standards or, simply, standards) (Tomlinson & McTighe, 2006). Students receive a separate grade for each standard; they may also receive an overall grade for the curriculum unit in which the standards are embedded. Table 1 contains a sample of a standards-based grade report for a single student in chemistry.

Table 1
A Portion of a Standards-Based Report

Student' Name: Olivia George	GRADE
Uses laboratory equipment properly and safely	4
Calculates density correctly	4
Applies the concept of density to relevant problems	2
Recalls the formulas for gas laws (e.g., Boyle, Gay-Lussac)	4
Selects appropriate gas laws to solve given problems	1

The report begins by identifying five standards associated with a unit entitled "Density and Gas Laws." For each standard a grade of 4 (excellent), 3 (proficient), 2 (approaching proficiency), or 1 (well below proficiency) is given. As shown in Table 1, the student (Olivia George) is "proficient" or "excellent" in three of the five standards. She "approaches proficiency" in her ability to apply the concept of density and is "well below proficiency" in her ability to select the appropriate gas law to solve a problem. Such information can at the very least help teachers understand <u>where</u> they need to spend additional time and effort. However, the information does not inform teachers as to <u>how</u> they should change their instruction in order for the student to improve their learning relative to these two standards (Campbell's "treatment").

Table 2 illustrates how standards-based systems can provide information about the learning strengths and weaknesses of group of students. We see once again that student achievement relative to the third and fifth standards is quite weak. Such information should be useful to teachers who are interested in improving the learning of an entire class of students.

Table 2
A Portion of a Standards-Based Report for a Group of Students

Density and Gas Laws	4	3	2	1
Uses laboratory equipment properly and safely	80%	20%	0%	0%
Calculates density correctly	40%	40%	10%	10%
Applies concept of density to relevant problems	20%	30%	25%	25%
Recalls formulas for gas laws	60%	40%	0%	0%
Selects appropriate gas law to solve problems	5%	10%	20%	65%

Note: The numbers in the cells represent the percentage of students receiving each grade on each standard.

Finally, although rarely discussed by advocates of standards-based grading, the grades assigned to students (that is, individual <u>ratings</u>) can easily be converted to comparisons between and among students (that is, a student's <u>ranking</u> within a group or class). In Table 1, Olivia George has a grading pattern of 4-4-2-4-1 across the five standards. Her achievement would be greater than a student with a pattern of 3-3-1-3-1, but less than a student with a pattern of 4-4-4-4-3. Therefore, she would rank somewhere between these two students.

Communicating with a Variety of Audiences

> The primary purpose of grades is to communicate student achievement to students, parents, school administrators, postsecondary institutions, and employers. (Bailey & McTighe, 1996, p. 120)

The statement above, either copied verbatim or slightly paraphrased, has found its way into grading policy statements in numerous school districts throughout the United States. Upon first reading, this statement is quite straightforward. The primary purpose of grading is communication; furthermore, there is a need to communicate with many different audiences. Upon further reading, however, we become aware that (1) there is an exclusive focus on student achievement, and (2) the list of audiences is incomplete.

Because I will deal with the exclusive focus on achievement later, for now I will focus on three "missing" audiences. The first audience is teachers: not those who assigned the grades, but those who would likely benefit from having information about those students upon entry to their classrooms in subsequent terms or years. "Olivia received a grade of B in Chemistry I. Does this mean that she is ready to meet the demands of Chemistry II?" The second audience is policy makers. Recently in the state of South Carolina, the State Board of Education replaced a 7-point grading scale (that is, A = 93 to 100; B = 85 to 92) with a 10-point grading scale (A = 90 to 100, B = 80 to 89) The State Superintendent of Education stated that the change would "level the playing field" and "benefit those students who transfer into the state." Whether it accomplishes these two goals is debatable. What is not debatable is the fact over a four-year period approximately 6,000 additional students will receive state-supported scholarships to post-secondary institutions, costing the state an additional $50 million. The third audience is members of the media. A recent headline in the *Washington Post* read, "Is it becoming too hard to fail? Schools are shifting toward no-zero grading policies" (Balingit & St. George, 2016). "No-zero grading policies" are those that discourage teachers from assigning percentage grades lower than 50 if a student makes a "reasonable attempt to complete the work." Is there evidence that the policy has reduced or will likely reduce the number of failing students? And, why is this a concern? Do we, as a society, desire more failing students?

Schneider and Hutt (2013) have argued that there is a "seemingly inescapable tension in modern schooling between what promotes learning and what enables a massive system to function" (p. 203). This "inescapable tension" can be seen in the information needs of the various audiences mentioned above. Teachers are (or should be) primarily concerned with promoting learning. Students and parents are likely to join teachers in this concern. Replacing letter or number grades with standards-based reports, written narratives (Kohn, 1999), and/or conferences (Pitler, 2016) is likely to serve these audiences well. At the same time, however, the detail provided by such grading systems in combination with the qualitative nature of much of the data make it difficult to aggregate the data in a way that is useful for other audiences (e.g., administrators at higher education institutions, policy makers, and members of the media). Nowhere is this "inescapable tension" more apparent than in selective universities with admissions officers who have begun to place a greater value on interviews, essays, and written reports in admission decisions (Hoover, 2012), while, at the same time, their offices of communications and public relations continue to release to the media the number of valedictorians in, or the mean SAT scores of, their incoming freshman classes.

What Do Grades Mean?

> What merit is required for an A grade? Is there anything about grade merit that can be standardized? Until a standard is established, every whim of a teacher will be the

> grading-plan. "I like to have my pupils think?" said one teacher. … "Pupils must be able to remember what they study?" said another. (Rorem, 1919, p. 670-671)
> I leaned over the student's shoulder … and asked him if he could show me his teacher's feedback on his work and his current marks. He opened his electronic folder of social studies on his laptop and there was a list of assignments. … Besides one of the assignments, it said 100%. I asked him what that meant—"well I handed that in on time," he said. (Tinney, 2014, p. 1)

When it comes to the meaning of grades, there is general agreement that high grades are "good" and low grades are "bad." Parents, particularly, want their children to achieve "good grades." However, there is a lack of agreement as to what constitutes a "good" grade. As Rorem (1919) suggested almost a century ago, a student may receive a "good" grade in one teacher's class if he or she memorizes what was taught, while in another teacher's class he or must demonstrate an ability to critically analyze what was taught. A student may receive a "good" grade if work was handed in on time in one teacher's class (Tinney, 2014), but must submit work that means a teacher's quality standards in another class. Table 3 summarizes four ways in which a grade can be represented and interpreted.

Table 3
A Summary of Differences in What Grades Represent

A GRADE MAY REPRESENT		
performance on a single task	OR	performance on multiple tasks
achievement at one point in time	OR	changes in achievement over time
achievement only	OR	achievement, effort, attendance, participation
achievement of intended learning outcomes (that is, ratings)	OR	achievement in comparison with peers (that is, rankings)

First, a grade can represent a student's performance on a single task (e.g., a quiz or test, an essay, a research report). These are "single task grades." Alternatively, a grade can represent a student's performance on multiple tasks over time (e.g., a semester or course grade) and, even, across subject matters and teachers (e.g., grade point average). These are "cumulative grades." Cumulative grades require some form of data aggregation, be it a simple arithmetic average of the single task grades, a simple arithmetic average after the highest and lowest grades have been eliminated, a weighted average (as when a unit test counts twice as much as homework assignments), or some other method.

Second, a grade can represent a student's achievement <u>at a particular point in time</u> or how much a student has learned <u>over time</u> (that is, how much a student's achievement has improved from Time A to Time B). The majority of grading systems focus on achievement at one point in time (e.g., a unit test, a course project). Grading on improvement, in fact, has been criticized because (1) it is a difficult thing to measure, and (2) it is unfair to initially high achieving students who have little if any room to improve (Davis, 1993; McKeachie, 1999). Other educators, however, suggest that "grading on improvement" is preferable because it does not penalize students who enter a course with less knowledge than their peers (Esty & Teppo, 1992, p. 616). In the words of one music educator "some students that start out 'woefully behind' can, with hard work, emerge as outstanding musicians; yet if they are judged against some arbitrary standards in their early careers they might wrongly infer (or even be told) that they don't 'measure up'" (Everett, 2013).

Third, a grade may represent academic achievement only (as recommended by Bailey & McTighe, 1996) or some combination of academic achievement and one or more other factors (e.g., effort, attendance, class participation, and/or conduct). Interpreting a grade representing academic achievement only is a far easier task. If grades are based on some combination of "scores from major exams, compositions, quizzes, projects, and reports, along with evidence from homework, punctuality in turning in assignments, class participation, work habits, and effort, the result in a 'hodgepodge grade' that is just as confounded and impossible to interpret as a 'physical condition' grade that combined height, weight, diet, and exercise would be" (Guskey, 2011, p. 18). Nonetheless, there is evidence that teachers tend to avoid grading on achievement only and consider factors in addition to achievement when they assign grades (Andersson, 1998).

Fourth, a grade may represent achievement relative to intended learning outcomes (that is, criterion-referenced) or achievement relative to the achievement of his or her peers (that is, norm-referenced). Virtually all grading systems in the early 20th century were norm-referenced. In 1963, Robert Glaser argued that educators should move away from "norm-referenced" measurement to what he termed "criterion-referenced" measurement. In terms of grading, then, students should be rated in terms of their learning relative to pre-determined curricular standards or learning expectations, rather than ranked in terms of their peers.

With this variety of representations and interpretations, it should not be surprising that the standardization sought by Rorem, Rugg, and others almost 100 years ago has not come to fruition and, quite likely, never will. Rather, the meaning of any grade is context- or situational-specific.

One is reminded of the conversation between Humpty Dumpty and Alice in Lewis Carroll's *Through the Looking Glass*. "'When I use a word,' Humpty Dumpty said, in rather a scornful tone, 'it means just what I choose it to mean—neither more nor less.' 'The question is,'" said Alice, 'whether you can make words mean so many different things.'" When it comes to grades it appears that the answer to Alice's question is, "Yes, indeed!"

So, what should be done? Rather than work toward a standardization of grades, a more reasonable strategy would be to embrace the contextual- or situational-specific nature of grades. Each teacher (or group of teachers) would be responsible for communicating clearly the meaning of each of the grades they are likely to assign. Table 4 illustrates one attempt to do so (adapted from Frisbie & Waltman, 1992). Note that it is possible (and, in some cases, may be desirable) to provide

Table 4
Criterion- and Norm-Referenced Descriptors of Letter Grades

Grade	Criterion-Referenced	Norm-Referenced
A	Firm command of knowledge domain, high level of skill development, exceptional preparation for later learning	Far above class average
B	Command of knowledge beyond the minimum, advanced development of most skills, has prerequisites for later learning	Above class average
C	Command of only the basic concepts and principles, demonstrated ability to use basic skills, lacks a few prerequisites for later learning	At the class average
D	Lacks knowledge of some fundamental concepts and principles, some important skills not attained, deficient in many of the prerequisites for later learning	Below class average
F	Most of the basic concepts and principles not learned, most essential skills not demonstrated, lacks most of the prerequisites needed for later learning	Far below class average

both criterion-referenced and norm-referenced interpretations. For example, a student may possess a "command of knowledge beyond the minimum, advanced development of most skills, and the prerequisites for later learning" (that is, a criterion-referenced grade of "B"), while at the same time being "at the class average" (that is, a norm-referenced grade of "C").

One grading system, contract grading, requires teachers to clearly communicate their expectations for different letter grades at the beginning of a semester or course. Teachers describe the achievement and/or performance levels that are needed to earn each letter grade (see Table 5). Based on this information, each student can decide on the letter grade that he or she intends to pursue and then sign a contract in which the teacher is committed to award the agreed upon grade if the student meets or exceeds those levels (Taylor, 1980).

Because Table 4 is more generic than Table 5, the information contained in that table can be used with multiple audiences (e.g., students, parents, potential employers). Table 5, by contrast, is only appropriate for the students enrolled in a specific course. Although neither is perfect, both can be considered "good faith efforts" to solve the problem of the ambiguity inherent in the meaning of grades. Without such attempts, the interpretation of a grade rests solely with the recipient of the grade, typically, the student (and his or her parents). When this happens, we are left with an entire classroom, school, or educational system composed of Humpty Dumptys.

Table 5
A Sample Contract System

To Receive an A	To Receive a B	To Receive a C
Submit 90% of in-class writing assignments	Submit 80% of in-class writing assignments	Submit 70% of in-class writing assignments
Complete 100% of homework at a satisfactory level	Complete 90% of homework at a satisfactory level	Complete 80% of homework at a satisfactory level
Receive a mean score of 85% or above on the 3 exams	Receive a mean score of 75% or above on the 3 exams	Receive a mean score of 75% or above on the 3 exams
Complete 3 group projects	Complete 3 group projects	Complete 2 group projects
Complete major project proposal	Complete major project proposal	
Complete major project at an acceptable level of quality		

Source: Adapted from Smith (2003).

How Reliable are Students' Grades?

The answer to this question depends on whether we are talking about single task grades or cumulative grades. When focusing on single task grades, the answer to this question is quite clear. Single task grades are very unreliable. When interpreting this statement, however, it is important to note that the reliability of single task grades is defined in terms of inter-rater reliability (that is, agreement between and among teachers). Also, most early studies focused on the reliability of numerical (or percentage) grades, rather than letter grades.

The landmark studies were conducted by Starch and Elliott (1912, 1913), the first in high school English, the second in high school mathematics. In each study a reasonably large group of teachers was given either an essay (1912) or a worked-out solution to a mathematics problem (1913). They were asked to read the essay or the worked-out solution and assign it a grade from zero to 100. The grades ranged from 50 to 90 for one essay and from 64 to 98 for the other essay. For the

worked-out mathematics problem, the range, unexpectedly, was even larger (from 28 to 92; Starch, 1913).

Almost a century ago, Rugg (1918) published a review of 23 studies published during the previous three years. Among the many conclusions reached by Rugg, two are the most relevant to our discussion. First, "teachers, marking without an objective scale, cannot be expected to mark student work in any subject—mathematics, history, composition, lettering, etc.—within an interval of roughly 8 per cent" (p. 704). Thus, for example, teachers using percentage grading systems cannot reliably differentiate an 83, say, from a 79 or an 87. Second, as one examines the grades given by an individual teacher to the same piece of student work graded at two different times there is "distinct evidence of unreliability of marking" (p. 703). That is, even individual teachers are inconsistent in the grades they assign to the same work sample at different times.

As the evidence of a lack of teacher agreement mounted, both academicians and practitioners began to search for possible explanations. Starch (1913) identified four possible sources of low inter-rater reliability: (1) differences caused by the inability of teachers to "distinguish between closely allied degrees of merit" (p. 630), (2) differences in the criteria used by different teachers (e.g., content, mechanics, and style in grading essays), (3) differences in the quality standards used by different teachers (e.g., what differentiates "excellent" work from "good" work?), and (4) differences in the way that teachers distribute their grades. Over time, each explanation yielded a different solution to the unreliability problem (see Table 6 for a summary).

Table 6
Sources of Unreliability and Proposed Remedies for Low Reliabilities

Source	Historical Solution Proposed
Inability of teachers to differentiate among percentage points	Shift from percentage grades to letter grades
Teachers' use of different criteria	Use standardized scoring rubrics
Teachers' use of different quality standards	Calculate a "correction factor" based on whether teacher was "easy" or "hard" grader and apply the "correction factor" to each teacher's grade
Different grade distributions	Assign a fixed percentage of As, Bs, Cs, Ds, and Fs based on a presumed underlying normal distribution of ability and achievement.

In response to the inability of teachers to make the distinctions required by percentage grading, Rugg (1918) suggested that research "confirms our judgment that five divisions can be handled accurately by teachers" (p. 710). Shortly thereafter, percentage grades were largely replaced by letter grades with five categories: A, B, C, D, and E (later becoming F). Five categories designated by letters A, B, C, D, and F remain the most popular grading system today, with four categories often used in standards-based systems (e.g., Advanced, Proficient, Basic, Below Basic).

To minimize the impact of different teachers using different criteria, Tieje, Sutcliffe, Hillebrand, and Buchen (1915) designed what may have been the first rubric, a rubric designed to evaluate written compositions. In simplest terms, a rubric is a coherent set of criteria for evaluating students' work that includes both the criteria and descriptions of different quality standards for each criterion. The criteria recommended by Tieje and his colleagues ranged from spelling, mechanics, and sentence construction to an ability to reason from premises to conclusions and an "ability to present the argument effectively, that is, with tact and force" (p. 594). Low marks on the "sentence

construction" criterion were given for compositions that had one sentence with a "violent change of construction," or one "straggling sentence," and/or one "unclear sentence." High marks on the "sentence construction" criteria were given to compositions in which none of sentences exhibited any of these problems and met accepted standards of sound sentence structure.

Although rubrics remain popular in grading written compositions, reports, and projects as well as grading performance in the arts (e.g., Panadero & Jonsson, 2013), there is some doubt that rubrics alone will solve the reliability problem. Brimi (2011) conducted a small-scale replication of the Starch and Elliott study in high school English. His sample included 90 teachers who had received seven days of training in the use of a writing rubric developed by the Northwest Regional Educational Laboratory (NWREL). Five days of training took place during the summer with two follow-up days during the school year. At the end of training, the teachers were asked to grade a single essay using a zero to 100 scale. The grades assigned ranged from 50 to 96 (a range similar to that reported by Starch and Elliott more than a century ago.

These findings are consistent with the results of a review of literature conducted by Jonsson & Svingby (2007) who concluded, "rubrics do not facilitate valid judgment of performance assessments *per se*." (p. 130). Rather, if they are to be effective in this regard they must be "complemented with exemplars" or what Wiggins (2013) has referred to as "anchor papers."

Although exemplars and anchor papers may help reduce the problem of teachers holding different quality standards, a very early attempt by Leroy Weld (1917) to solve this problem is particularly noteworthy. Weld designed a system intended to minimize differences in the grades assigned by teachers by assigning each teacher a "correction factor" to compensate for whether a teacher tended, on average, to be a "hard" or an "easy" grader. In other words, his system recognized that teachers held different quality standards, but minimized their impact on the grades that students were assigned by incorporating the appropriate "correction factor."

Finally, an early attempt to solve the problem of substantially different grade distributions across teachers was to encourage teachers to adopt the practice known as "grading on the curve." Simply stated, "grading on the curve" means that a certain percentage of students should receive "A's," a certain percentage should receive "B's," and so on. The recommended percentages were based on the assumption that the distribution of student ability and, hence, achievement approximated a normal (Gaussian) curve. In 1914 the Committee on Standardizing Grades of the American Association for the Advancement of Science (AAAS) recommended that there be "five approximately equal steps of ability, the percentage of students that fall into each group are approximately as follows: Excellent (A), 4 percent; Good (B), 24 percent, Medium (C), 44 percent, Sub-medium (D), 24 percent, and Failure (E), 4 percent" (Ruediger, Henning, & Wilbur, 1914, p. 643). Educators' belief and faith in the normal distribution continued through much of the 20th century.

Unfortunately, the distributions of grades assigned by teachers at that time were not normally distributed (Rugg, 1918) and this lack of normality of assigned grades continues (Office of Research, 1994). Of the several hundred grade distributions that Rugg examined fewer than 10% could be described as "perfectly symmetrical;" furthermore, "not more than two or three in 100 of all those examined has been found to be approximately normal" (p. 705). With respect to the data reported as part of the 1998 National Educational Longitudinal Study (NELS:88) by the U.S. Office of Research (1994), almost 70% of eighth grade students in their national sample reported receiving "mostly A's" or "mostly B's."

There is a great deal of evidence that the reliability of single task grades is virtually non-existent. Can the same thing be said about cumulative grades? Most of the studies that address this question include Grade Point Average (GPA) as the primary cumulative grade. A student's GPA is

computed by aggregating individual task grades across the courses in which the student is enrolled during a particular semester (e.g., all courses completed during the most recent Spring semester) or for an entire academic career (that is, all courses leading to the award of a high school diploma or a bachelor's degree). Typically, an A grade is worth 4 points, a B grade is worth 3 points, and so on. In contrast with the studies of the reliability of single task grades, these studies focus on the stability of GPAs over courses and over time (see, for example, Bacon & Bean, 2006; Etaugh, Etaugh, & Hurd, 1972).

One of the more recent studies, conducted by Saupe & Eimers (2012) at the University of Missouri, illustrates both the procedure and the results. The study began with the collection of the end-of-fall-semester GPAs of 5,000 freshmen students. GPAs were collected each subsequent semester, with slightly smaller sample sizes each semester, the result of students leaving the University. Alpha reliability coefficients were computed for two semesters, four semesters, six semesters, and eight semesters, four alpha coefficients in all. Because alpha coefficients represent the percent of variance in GPAs that can be attributed to differences among students, rather than differences across semesters, the larger the coefficient, the more reliable the GPAs are over time. The alpha coefficients were 0.72 (for two semesters), 0.84 (for four semesters), 0.86 (for six semesters), and 0.91 (for eight semesters). Similar findings have been reported by Etaugh, Etaugh, & Hurd (1972), Willingham, Pollack, & Lewis, (2000), and Bacon & Bean (2006).

When attempting to answer the question of the reliability of grades, then, we have a conundrum. Single task grades are not reliable at all whereas cumulative grades (at least in the case of GPAs) tend to be quite reliable. At the same time, however, we know that cumulative grades are determined to some extent by aggregating students' single task grades. How can this inconsistency be explained?

To answer this question, let us consider an example of how "unreliable" single task grades and "reliable" cumulative grades can co-exist. The data presented in Table 7 are quite similar to the data collected by Starch and Elliott. There is a single student (that is, one row) who has written an essay that is scored by five teachers (that is, five columns). The entry in each cell is the numerical score assigned by each teacher. They range from 30 to 90, with a mean of 60. The logical conclusion from these data (and the conclusion reached by Starch and Elliott) is that the grades assigned are quite unreliable (that is, quite inconsistent across teachers).

Table 7
Teacher numerical grades of one student's written composition

Student	Tchr 1	Tchr 2	Tchr 3	Tchr 4	Tchr 5	Mean
A	80	60	30	40	90	60

In Table 8 a second student has been added (that is, an essay written on the same topic by a different student). The same teachers assign grades to the second essay. If we focus only on the second student the pattern of inconsistency is quite similar to that found for the first student. The numerical grades range from 10 to 70 with a mean of 46. The range of grades, 60, is identical to the range for the first student.

Table 8
Teacher numerical grades with two hypothetical students

Students	Tchr 1	Tchr 2	Tchr 3	Tchr 4	Tchr 5	Mean
A	80	60	30	40	90	60
B	70	40	10	30	80	46

Rather than focusing on each student individually, let us compare them in terms of their grades. All five teachers assigned higher grades to the first student's essay; the overall mean score differs by 14 points. Even with the lack of agreement across teachers on each individual student's essay, then, it is quite clear that the teachers consistently favor Student A's essay over Student B's essay.

If we add more students, replace teachers with semesters, and replace numerical grades with GPAs in the cells of the table, we are able to simulate a portion of the data from the Saupe and Eimers' (2012) study (see Table 9). The data in the columns of the table suggest there are, in fact, differences in GPAs across the eight semesters. A focus on the rows of the table, however, indicates that students 1 through 3 consistently have lower GPAs (with means of 1.94, 2.25, 2.31, respectively) than students 8 through 10 (with means of 3.37, 3.43, 3.50, respectively). The alpha coefficient for the entire data set represented in Table 9 is approximately 0.90 (which compares quite favorably with Saupe and Eimers' coefficient of 0.91). That is, approximately 90% of the variation in GPAs can be attributed to differences among students, not semesters.

Table 9
Students x GPAs

Student ID	Sem 1	Sem 2	Sem 3	Sem 4	Sem 5	Sem 6	Sem 7	Sem 8
0001	1.5	1.0	1.5	2.5	2.5	1.5	2.0	3.0
0002	2.0	2.0	3.0	2.0	1.5	1.5	3.0	3.0
0003	2.0	1.5	2.0	3.0	2.0	3.0	2.0	3.0
0004	2.5	2.0	2.5	2.5	2.0	2.0	3.0	3.0
0005	2.5	3.0	3.0	2.0	2.5	2.5	1.5	2.5
0006	2.5	2.5	2.5	3.0	2.0	3.0	3.0	4.0
0007	3.0	2.0	3.0	3.0	2.5	2.5	3.0	4.0
0008	3.0	3.0	4.0	3.0	3.0	3.5	3.5	4.0
0009	3.0	2.5	3.0	4.0	3.5	3.5	4.0	4.0
0010	3.0	3.0	3.5	3.5	3.5	3.5	4.0	4.0

As this example illustrates, it is quite possible to have cumulative grades that are quite stable over time even when single task grades reflect a great deal of teacher disagreement. Teachers may have different quality standards that cause them to differ from one another in the grades they assign to student work; at the same time, however, these quality standards are such that these teachers can still agree that some work is superior to other work.

How Valid are Student Grades?

> Given an average school system with … forty to forty-eight pupils under the care of one teacher, (how can we) organize a plan of grading and promotion, and outline a course of study (for the two must go together) that will enable and assist each pupil to progress as rapidly as possible and still secure the necessary education usually comprised in the elementary and high school courses. (Dempsey, 1912, p. 373, emphasis mine)

Answering the validity questions is more difficult than answering the question of reliability. As was true of reliability, there are different types of validity. Similarly, as was true of the reliability of single task grades, there are recognized threats to the validity of grades. The increased difficulty stems from

the need to accept several assumptions when examining the validity of grades (e.g., that the plan of grading and promotion is consistent with the course of study).

Different Types of Validity

The validity of grades can be examined by answering two questions. First, do students who learn more get better grades? If they do, the grades, in a descriptive sense, are reasonably valid. This is the type of validity implied by Dempsey (above). Second, are students who receive better grades more successful in subsequent grade levels, school levels, or life in general? If they are, the grades, in a predictive sense, are reasonably valid (Thorsen & Cliffordson, 2012). The data most frequently used to answer both questions come from studies of course grades and grade point averages, both examples of cumulative grades. No studies of the validity of single task grades were located.

Threats to Validity

There are two generally recognized threats to the validity of grades. The first is the difference in the grades assigned by teachers in different schools, particularly schools with radically different student populations. The results of the aforementioned National Educational Longitudinal Study of 1988 (NELS:88) are instructive in this regard (U.S. Office of Research, 1994). In the study, eighth-grade students who were selected as part of a nationally representative sample were asked to indicate the grades they typically received (e.g., mostly A's, mostly B's). Next, students were divided into two groups: those who attended high poverty schools and those who attended more affluent schools. Within each group, the students' reported grades were compared to their NELS:88 scores. Students in high poverty schools who received "mostly A's" in English had about the same NELS:88 reading scores as did the "C" and "D" students in the more affluent schools. On the NELS:88 mathematics test, the scores of "A" students in the high poverty schools most closely resembled the scores of "D" students in the more affluent schools. Similar results have been reported by Simmons, Brown, Bush, & Blyth (1978) and Willingham, Pollack, & Lewis (2000).

The second threat to validity is grade inflation, a somewhat more recent phenomenon (Rojstaczer & Healy, 2010). Grade inflation can be defined as the tendency to award progressively higher academic grades for work that would have received lower grades in the past. It is important to note that higher grades in themselves do not prove grade inflation; it is also necessary to demonstrate that the grades are not deserved. Slavov (2013) describes the negative impact of grade inflation on the validity of grades assigned by teachers in higher education institutions. "Because grades are capped at A or A+, grade inflation results in a greater concentration of students at the top of the distribution. This compression of grades diminishes their value as an indicator of student abilities. Without grade inflation, a truly outstanding student might be awarded an A, while a very good student might receive a B+. With grade inflation, both students receive A's, making it hard for employees and graduate schools to differentiate them" (p. 2).

Evidence Pertaining to the Validity of Grades

Studies investigating descriptive validity (or what used to be called "concurrent validity") typically examine the relationship between cumulative grades and test scores. The interpretation of the results of these studies in terms of the validity of grades is based on two fundamental assumptions. First, test scores accurately reflect student achievement. Second, students with higher test scores have learned more.

The correlations between cumulative grades, broadly defined, and test scores in these studies range from 0.30 to 0.75. Lower correlations are found in studies of the relationship between students' overall GPAs and their composite scores on comprehensive test batteries (e.g., McCandless, Roberts, & Starnes, 1972). The correlations increase when grades in specific subject

matter (e.g., reading, mathematics) are related to scores on subject-specific tests (Farr & Roelke, 1971; Lekholm & Cliffordson, 2008). Finally, the correlations are the strongest when a study investigates the relationship between students' scores on tests aligned with the content and objectives of a specific course (so-called "end-of-course" tests) and the grades that students receive in that course (e.g., Algebra I; Boykin, 2010).

When we turn to studies of predictive validity, most of the available studies address the question: "How well does high school grade point average (HSGPA) predict success in postsecondary institutions?" "Success" typically is defined in terms of college grade point averages, occasionally in terms of receiving/not receiving an undergraduate degree.

The results of these studies are quite positive. HSGPA is consistently the strongest predictor of college grades, with college entrance examination scores improving the prediction by a small but statistically significant amount (Zahner, Ramsaran, & Steelde, 2014). More specifically, the correlation coefficients of HSGPA with college GPA tend to range from 0.35 to 0.55. When these coefficients are corrected for (1) restriction of range of HSGPA, (2) differences in the college courses in which students are enrolled, and (3) differences in instructors' grading standards, there is a substantial increase in their magnitude. Ramist, Lewis, & McCamley-Jenkins (1994), for example, reported an increase from 0.36 to 0.69 when these three corrections were made.

Quite importantly, the strength of these coefficients remains virtually unchanged over the student's college career. In fact, Geiser & Santelices (2007) found that the predictive weight associated with HSGPA accounted for a greater proportion of variance in cumulative fourth-year GPA than did first-year college grades. Finally, there is some evidence (although sparce) that HSGPA predicts the likelihood that a student will receive a college degree. Astin, Tsui, & Avalos (1996), for example, reported that two-thirds of students with HSGPAs of "A" graduated from college as opposed to one-fourth of students with HSGPAs of "C."

Although almost all of the predictive validity studies have focused on college success, two additional studies are worthy of mention. Kurlaender & Jackson (2012) conducted a five-year longitudinal study of slightly more than 13,000 students in three large California school districts. The study began when the students were in seventh grade and ended the year they were expected to graduate from high school. In addition to GPAs, their data set included race/ethnicity, gender, special education placement, free lunch status, and standardized test scores. Based on a series of analyses, the authors concluded that "seventh grade GPA is consistently a significant predictor of high school completion, controlling for a variety of other characteristics" (p. 16). Furthermore, receiving even one F on the eighth grade report card increased the likelihood that a student would not complete high school.

In another longitudinal study, Arnold (1995) followed 81 high school valedictorians who graduated from high school in the spring of 1981, for 14 years. Among the major results of the study are that the valedictorians "continued to do well in college with an overall GPA of 3.6" (p. 310). Also, they had careers in fields such as accounting, medicine, law, engineering, and education.

In summary, then, the available evidence tends to support both the descriptive and predictive validity of cumulative grades. Specifically, cumulative grades tend to be positively related to (1) achievement test scores, (2) the likelihood of receiving a high school diploma, (3) college grades over multiple years, and (4) the likelihood of earning a college degree.

What are the Consequences of Grading Students?

> The meaning of numbers can determine the fate of one's future, especially in education. A grade is more than a number; it's a quality of life. (Mathews, 2016, front cover)

It is quite true that the grades can and do impact the quality of students' lives. It is important to point out, however, that these impacts can be positive or negative. Unfortunately, most of the critics focus only on the negative. Kohn (1999, 2011), for example, has compiled a list of negative consequences of grading students using letters or numbers. Included on the list are the following:

- Grades tend to reduce students' interest in the learning itself.
- Grades distort the curriculum.
- Grades spoil teachers' relationships with students.
- Grades spoil students' relationships with each other.

As one peruses this list, it seems reasonable to ask whether other words or phrases could be substituted for "grades" in these statements without changing the accuracy of the statement. Consider the following:

- Boring teachers, activities, and tasks reduce students' interest in the learning itself (Baurelein, 2013)
- Federal and state mandates distort the curriculum (Robelen, 2011).
- Negative teacher behavior spoils teachers' relationships with students (Banfield, Richmond, & McCroskey, 2006).
- Pecking order, cliques, and self-segregation spoil students' relationships with each other (McFarland, Moody, Diehl, Smith, & Thomas, 2014).

These rewritten statements are not intended to suggest that grades are not harmful to some students. To the contrary, there is ample evidence to suggest that they are (Areepattamannil & Freeman, 2008; Bacon, 2011). Rather, the revised statements are intended to show that grades are no more or less harmful than many other aspects of schooling.

More importantly, however, the available evidence suggests that the negative effects of grades on students tend to accumulate over time. More than 40 years ago, Kifer (1975) conducted a quasi-longitudinal study of students at four grade levels (2, 4, 6, and 8). At each level, two groups of students were identified. Group A included students who had been in the top 20% of their class each year. Group B included students who had been in the bottom 20% of their class each year. Students in both groups were administered an academic self-concept (ASC) scale. For the second grade students, the two groups did not have significantly different ASC scores. By the eighth grade, however, the differences between the two groups were both substantial and statistically significant. Furthermore, the graphs prepared by Kifer showed quite clearly that although the mean ASC scores of Group A did not change much from grade to grade, for Group B there was almost a linear decline.

Forty years ago, I wrote, "The verb 'to fail' refers to the inability of an individual to attain success with respect to a particular goal. 'Failure' is a noun, which refers to a person who, having failed to attain a series of related goals, perceives himself as incapable of success in the future.… Failing is (or can be) beneficial for individuals, whereas failure is virtually always detrimental" (Anderson, 1976, p. 1). Consistently receiving low grades (e.g., mostly D's and F's) is likely to transform "failing" into "failure."

How does this transformation happen? Unlike single task grades, which pertain to individual pieces of student work, cumulative grades at some unknown point in a student's school career begin to apply to the students themselves. For example, when a student writes a series of "A" essays over time or consistently receives "A" grades on quizzes or tests, he or she becomes an "A" student. On the other hand, a student who consistently prepares a series of poorly written essays or has consistently poor performances on tests can easily be labeled a "D" or "F" student.

The debate about the negative effects of grades has been going on for decades and will likely continue in the foreseeable future. To provide some perspective to this debate, I would like to conclude this section with something Stanley S. Marzolf (1955) wrote almost 60 years ago:

> There is a rumor going about that assigning school marks is in conflict with principles of mental health. … [Those who are spreading the rumor] suggest that marking is a persistent evil that the prospective teacher [should] learn to circumvent or at least palliate. … It is my contention that <u>many of the evils of marks and marking are unnecessary and arise from ignorance, incompetence, and spite</u>.… If one is to learn, one must have knowledge of results. (p. 10, emphasis added)

Discussion

> The power of grades to impact students' future (lives) creates a responsibility for giving grades in a fair and impartial way. (Johnson & Johnson, 2002, p. 249)

In 1902 Herbert Mumford authored a bulletin entitled "Market Classes and Grades of Cattle with Suggestions for Interpreting Market Quotations." Over the past century, great strides have been made in the grading of cattle (see, for example, Hale, Goodson, & Savell, 2013). Unfortunately, the same cannot be said of the way that students are graded. What needs to be done to move us forward? I offer five recommendations.

Recommendation 1

We must fully integrate concerns about grading into discussions on how best to improve our education system and achieve educational excellence.

> Grading must be raised from its present status as just another chore to its real function as … evaluation of pupil accomplishment and the efficiency of our educational institutions. (Cureton, 1971, p. 8)

Over the past half century, there have been numerous recommendations as to the best ways to reform public education in the United States. These recommendations tend to include the need to increase the rigor of the curriculum, employ highly qualified teachers, provide more personalized learning opportunities for students, integrate technology into the instructional program, and improve school-community relations. Notably absent from these lists is anything to do with the way students are graded. Concerns about grading, when they do arise, seem to lie outside the important components of the educational system. It should not be surprising, then, that many of the changes made in grading policies and practices over the past quarter century have been rather superficial (e.g., shifting from a 7-point scale to a 10-point scale, advocating standards-based reports, requiring numerical grades on individual assignments to be at least 50).

Because grading systems, like school calendars, are ingrained within educational system, however, substantive changes in grading policies and practices are neither easily made nor easily adopted. After a committee of parents, teachers, and administrators in Evanston, Illinois, spent four years designing a new system for report card grades, the proposed system was not approved by the school board (*Chicago Tribune*, 2003).

Recommendation 2

We must design grading systems and implement grading practices that are models of integrity and are perceived by all parties as fair.

> During the past ten years it has been increasingly evident that one of the contributory causes of 'failure' in the public schools has been a bad administration of the marking system. (Rugg, 1918, p. 701)

If our grading system is, in fact, "one of the contributory causes of 'failure' in the public schools," it is not sufficient to put "duct tape" (Crowley, 2015) on our current grading policies and practices. The way in which students are graded and the way in which those grades are reported must be re-examined and, ultimately, reconceptualized. This reconceptualization would benefit if attention was paid to two issues: grade integrity and fairness.

Grade integrity is "the extent to which each grade awarded … is strictly commensurate with the quality, breadth and depth of a student's performance" (Sadler, 2009, p. 807). What are some of the features of grading systems with integrity?

1. The tasks given to students for the purpose of assigning grades should be representative of the essential intended learning outcomes.
2. The quality standards used by teachers at the same grade level or teaching the same course should be as similar as possible.
3. Students should be given sufficient information so that they understand the bases for the grades they receive. If this is done well, students can improve their own ability to make reasonable judgments about the quality of their work.
4. Representatives of a variety of audiences (also known as stakeholder groups) should be asked to provide input into the grading systems and practices and to review a final draft of the systems and practices before they are published.

In many respects, fairness, it seems, is like beauty. That is, whether something is fair or unfair lies in the eyes of the beholder. Teachers are quite inconsistent in their beliefs about fair and unfair grading practices (Green, Johnson, Kim, & Pope, 2007). For example, 57% of the teachers who were surveyed believed it was fair to include student performance on homework in the calculation of report card grades; 43% believed it to be unfair. Similarly, 48% of the teachers believed it was fair to grade an essay test knowing the identity of the student who wrote the essay; 52% believed it was unfair.

Students, on the other hand, seem to be much more in agreement when it comes to the issue of the fairness (Alm & Colnerud, 2015). In general, students perceive grading and grades to be unfair when teachers:

1. fail to follow the guidelines of the current grading system;
2. assign grades based on unreliable information;
3. allow themselves to be influenced by irrelevant factors; and
4. are ambiguous or unclear in the explanations they give for the grades they assign.

Issues of fairness are particularly important when the focus of attention turns to students with special needs. As Munk & Bursuck (2003) have written, "many students with disabilities receive inaccurate and unfair grades that provide little meaningful information about their achievement" (p. 38). To be fair to students with special needs, grading systems must (1) start with clear purposes in mind, purposes that take into consideration the information needs of parents and other teachers; (2) incorporate adaptations for special needs students that are workable and promote access to and success with the general curriculum, and (3) include opportunities for individualized grading (similar to that provided by contract grading as described earlier).

In combination, integrity and fairness provide a sound basis for setting the criteria used to evaluate grading policies and practices. Finally, rather than advocating for one particular grading system (e.g., Scriffiny, 2008), we need to design policies and practices that achieve the purpose(s) for which grades are assigned and meet the information needs of the audiences to whom the grades will be reported.

Recommendation 3

We must find ways to communicate grades so that the information needs of a variety of audiences are met.

> We need to show where a kid is in relation to the standards. We have to explain if a kid is meeting the standards, exceeding them, or below them. … Standards are a tool that lets teachers and parents monitor the rigor of the work children are expected to do. (A principal quoted in Kreider & Caspe, 2002)

> Last quarter I got this report that says 'he's meeting the standard' or 'he's not meeting the standard' or 'he's exceeding the standard.' These report cards don't even tell you if your kid is really doing okay. … I don't know if he's doing 'A' work, 'B' work, or 'C' work. (A mother quoted in Kreider & Caspe, 2002)

My purpose of including these two excerpts is to illustrate the point that educators do not always know best. Educators may believe that standards-based grading systems provide the best information for parents, but as the mother's quote clearly indicates, such is not the case. Rather than assume they understand the information needs of various audiences, educators would be wise to ask them. For example, Sorian & Baugh (2002) reported the results of telephone interviews with 292 policymakers, randomly selected from all 50 states. The questions focused on their use of information as well as their attitudes toward various types of information. Only one-fourth of the respondents reported reading material they received in detail; about one-half reported skimming for general content. They reported being more likely to read material carefully if they found it to be "relevant." "Irrelevant" material was (1) too long, dense, or detailed, (2) full of jargon, and (3) seen as overly subjective or biased.

Engaging members of various audiences in ongoing dialogues about grade reports seems a much wiser approach than assuming that we, as educators, know what they need. With respect to parents, for example, Munk (2003) developed a survey that can be used to determine what parents want and need from the grades their children receive (see Table 10). Similar surveys can be developed for each stakeholder group. Once the needs of each audience are identified, a collaborative effort can be made to design reporting systems that meet those needs.

Table 10
Survey of Parents' Perceptions of the Purposes of Grades

Directions: Rank these purposes in order of importance by writing a number from 1 (most important) to 13 (least important) next to each purpose. Use each number only once.	
1. Tell me whether my child has improved in his/her classes.	Rank___
2. Tell me how to help my child plan for his/her future.	Rank___
3. Tell me how hard my child is trying.	Rank___
4. Help me plan for what my child will do after high school.	Rank___

Table 10 cont.
Survey of Parents' Perceptions of the Purposes of Grades

5. Tell me what my child needs to improve on to keep a good grade.	Rank___
6. Tell me how well my child works with classmates.	Rank___
7. Tell me what my child is good at and not so good at.	Rank___
8. Tell colleges and employers what my child is good at.	Rank___
9. Tell me how much my child can do on his/her own.	Rank___
10. Tell me how my child's performance compares to other children's.	Rank___
11. Tell me how to help my child improve.	Rank___
12. Tell me what classes my child should take in high school.	Rank___
13. Motivate my child to try harder.	Rank___

Recommendation 4

We need to ensure that prospective teachers are prepared to design and implement defensible grading practices when they enter their classrooms; furthermore, we need to incorporate discussions about grading systems and practices into continuing professional development.

> There is very little interest today [in problems inherent in grading students]. A survey of measurement textbooks is discouraging. Worse than this, the vast majority of states do not even require measurement courses for teacher certification. (Cureton, 1971, p. 7)

More than four decades later, Cureton's statement holds true. Teacher certification programs in most states require students to pass a course with measurement, assessment, and/or evaluation in the title. An examination of three of the most popular textbooks used in these courses, however, suggests that a single chapter is devoted to grading students, a chapter consistently placed at or near the end of the book. The bulk of these texts focus on practical and technical issues surrounding tests and assessment.

With respect to in-service teachers, professional development sessions (perhaps organized by subject matter areas in high schools) can be be used to discuss issues pertaining to grading policies and practices. Question such as the following can be used as prompts for the discussion.

1. What factors do you include when you grade students?
2. What information do you obtain for each factor (e.g., achievement, effort)?
3. How do you differentiate among the various letter grades (e.g., A, B, C, D, F)?
4. How do you combine individual task grades into a cumulative grade?

Ideally, discussions over time could lead to more standardized, uniform grading policies and practices (such as that envisioned by many of the early writers in the field).

Recommendation 5

We need to conduct thoughtfully designed, well implemented studies of grades, grading systems, and grading practices that provide greater understanding of the problems as well as practical ways of solving the problems once they are fully understood.

> As a matter of fact, we are forcing each other into all sorts of vague compromises just because no one has facts. … I am not in favor of all the traditions which are stoutly maintained, but I wish to say with equal emphasis that I am not in favor of adopting radical suggestions just because they are offered with persistence. (Judd, 1910)

At present, grading policies and practices are grossly under-researched fields. As was true in Judd's time, we continue to lack facts. If you read articles written during the first two decades of the 20th century you will likely be impressed by two things. First, there is an emphasis on solving practical problems. Second, data are used to inform decisions about these problems. A century ago, then, this seemed to be common practice.

Today's educators seemed to have moved away from empirical investigations to the comfort of Op Ed pieces. These pieces tend to go in one of two directions. Either the author advocates for a particular approach to solving an identified grading problem (typically *sans* data) or the author demonizes grading, typically ending the piece with a call to eliminate grading all together. Unfortunately, this latter group of authors fail to appreciate the fact that grading, like school calendars and group instruction, is part of the very fabric of formal schooling. As long as there is formal schooling, teachers will assign grades.

If we are to move forward, then, we need fewer opinion and advocacy pieces and more empirical evidence and thoughtful dialogue. And, as we move forward, we would be wise to conduct "practical" research studies, keeping in mind Judd's call for facts, rather than "radical positions … offered with persistence."

References

Alm, F., & Colnerud, G. (2015). Teachers' experiences of unfair grading. *Educational Assessment, 20*(2), 132-150.

Anderson, L. W. (1976). Should students fail? *Education Report, 19*(1), 1, 4.

Andersson, A. (1998). The dimensionality of the leaving certificate. *Scandinavian Journal of Educational Research, 42*, 25-40.

Areepattamannil, S., & Freeman, J. G. (2008). Academic achievement, academic self-concept, and academic motivation of immigrant adolescents in the greater Toronto area secondary schools. *Journal of Advanced Academics, 19*(4), 700-743.

Arnold, K. (1995). *Lives of promise: What becomes of high school valedictorians: A fourteen-year study of achievement and life choices* San Francisco: Jossey-Bass.

Astin, A., Tsui, L., & Avalos, J. (1996). *Degree attainment of American colleges and universities: Effect of race, gender, and institutional type*. Washington, DC: American Council on Education.

Bacon, L. C. (2011, Summer). *Academic self-concept and academic achievement of African-American students transitioning from urban to rural schools*. Retrieved from http://ir.uiowa.edu/cgi/viewcontent.cgi?article=2582&context=etd.

Bacon, D. R., & Bean, B. (2006). GPA in research studies: An invaluable but overlooked opportunity. *Journal of Marketing Education, 28*(1), 35-42.

Bailey, J. M., & McTighe, J. (1996). Reporting achievement at the secondary level: What and how. In T. R. Guskey (Ed.), *Communicating student learning: 1996 Yearbook of the ASCD* (pp. 119–140). Alexandria, VA: ASCD.

Balingit, M., & St. George, D. (2016, July 5), *Is it becoming too hard to fail? Schools are shifting toward no-zero grading policies*. Retrieved from https://www.washingtonpost.com/local/education/is-it-

becoming-too-hard-to-fail-schools-are-shifting-toward-no-zero-grading-policies/2016/07/05/3c464f5e-3cb0-11e6-80bc-d06711fd2125_story.html.
Banfield, S. R., Richmond, V. P., & McCroskey, J. C. (2006). The effect of teacher misbehaviors on teacher credibility and affect for the teacher. *Communication Education, 55*, 63-72.
Barnes, M. (2014, November 19). *How four simple words can solve education's biggest problem*. Retrieved from https://www.youtube.com/watch?v=5-NykI2jOZw.
Bauerlein, M. (2013, September 20). *Boredom in class*. Retrieved from http://educationnext.org/boredom-in-class.
Boykin, A. S. (2010, March). *The relationship between high school course grades and exam scores, E & R Report No. 09-39*. Retrieved from http://files.eric.ed.gov/fulltext/ED564393.pdf.
Brimi, H. M. (2011, November). *Reliability of grading high school work in English*. Retrieved from http://pareonline.net/getvn.asp?v=16&n=17.
Bull, B. (2013, April 10). *5 common reasons for the importance of letter grades*. Retrieved from http://etale.org/main/2013/04/10/5-common-reasons-for-the-importance-of-letter-grades.
Campbell, A. L. (1921). Keeping the score. *School Review, 29*(7), 510-519.
Chicago Tribune. (2003, May 11). Should they get an A for effort? Retrieved from http://articles.chicagotribune.com/2003-05-11/news/0305110292_1_meeting-standards-pupil-school-board.
Crowley, B. (2015, February 25). *Grading: A duct-taped system in need of an overhaul?* Retrieved from http://www.teachingquality.org/content/blogs/brianna-crowley/grading-duct-taped-system-need-overhaul.
Cureton, L. W. (1971). The history of grading practices. *NCME Measurement in Education*, 2(4), 1-9.
Davis, B. G. (1993). *Tools for teaching*. San Francisco: Jossey-Bass.
Dempsey, C. H. (1912). Flexible grading and promotions. *Journal of Education, 75*(14), 373-376.
de Zouche, D. (1945). The wound is mortal: Marks, honors, unsound activities. *The Clearing House, 19*(6), 339-344.
Esty, W. W., & Teppo, A. R. (1992). Grade assignment based on progressive improvement. *The Mathematics Teacher, 85*, 616-618.
Etaugh, A. F., Etaugh, C. F., & Hurd, D. E. (1972). Reliability of college grades and grade point averages: some implications for the prediction of academic performance. *Educational and Psychological Measurement, 32*(4), 1045-1050.
Everett, M. (2013, October 18). *A conundrum: Grading for improvement versus grading against a standard*. Retrieved from https://thereformingtrombonist.wordpress.com/2013/10/18/a-conundrum-grading-for-improvement-versus-grading-against-a-standard.
Farr, R., & Roelke, P. (1971). Measuring subskills of reading: Intercorrelations between standardized reading tests, teacher ratings, and reading specialists' ratings. *Journal of Educational Measurement, 8*(1), 27-32.
Ferriter, B. (2015). *If grades don't advance learning, why do we give them?* Carrboro, NC: Center for Teaching Quality. [http://www.teachingquality.org/content/blogs/bill-ferriter/if-grades-don-t-advance-learning-why-do-we-give-them]
Finkelstein, I. E. (1913). The marking system in theory and practice. *Educational Psychology Monographs, 10*.
Frisbie, D. A., & Waltman, K. K. (1992), Developing a personal grading plan, *Educational Measurement: Issues and Practice, 11*(3), 35-42.
Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist* 18 (8): 519–522.

Green, S. K., Johnson, R. L., Kim, D.-H., & Pope, N. S. (2007). Ethics in classroom assessment practices: Issues and attitudes. *Teaching and Teacher Education, 23*, 999-1234.

Guskey, T. R. (2002). *How's my kid doing? A parent's guide to grades, marks, and report cards*. San Francisco: Jossey-Bass.

Guskey, T. R. (2011). Five obstacles to grading reform. *Educational Leadership, 69*(3), 16-21.

Hale, D. S., Goodson, K., & Savell, J. W. (2013). *USDA beef quality and yield grades*. College Station, TX: Texas A & M Department of Animal Science.

Hoover, E. (2012). High school class rank, a slippery metric, loses its appeal for college. *The Chronicle of Higher Education*, 59(14), A1, A5.

Johnson, D. H., & Johnson, R. T. (2002). *Meaningful assessment: A manageable and cooperative process*. New York: Pearson.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review, 2*(2), 130-144.

Judd, C. H. (1910). On the comparison of grading systems in high schools and colleges. The *School Review, 18*(7), 460-470.

Kifer, E. (1975). Relationships between academic achievement and personality characteristics: A quasi-longitudinal design. *American Educational Research Journal, 12*, 191-210.

Kohn, A. (1999). From degrading to de-grading. *High School Magazine, 6*(5), 38-43.

Kohn, A. (2011). The case against grades. *Educational Leadership, 69*(3), 28-33.

Kreider, H., & Caspe, M. (2002). *Defining "fine"—Communicating academic progress to parents*. Cambridge, MA: Harvard Family Research Project. Retrieved from http://www.hfrp.org/family-involvement/publications-resources/defining-fine-communicating-academic-progress-to-parents.

Kurlaender, M., & Jackson, J. (2012). Investigating middle school determinants of high school achievement and graduation in three California school districts. *California Journal of Politics and Policy, 4*(2), 1-24.

Lekholm, A. K., & Cliffordson, C. (2006). Discrepancies between school grades and test scors at individual and school level: Effects of gender and family background. *Educational Research and Evaluation, 14(*2), 181-199.

Marzolf, S. S. (1955). Mental hygiene aspects of school marks. *The Yearbook of the National Council on Measurements Used in Education,* (12), 10-12. Retrieved from http://www.jstor.org/stable/41862777.

Mathews, A. (2016). *The new epidemic grading practice: A systematic review of America's grading policy*. Bloomington, IN: Xlibris Corporation.

McCandless, B. R., Roberts, A., & Starnes, T. (1972). Teachers' marks, achievement test scores, and aptitude relationships with respect to social class, race, and sex. *Journal of Educational Psychology, 63*, 153-158.

McFarland, D. A., Moody, J., Diehl, D., Smith, J. A., & Thomas, R. J. (2014). Network ecology and adolescent social structure. *American Sociology Review, 79*, 1088-1121.

McKeachie, J. W. (1999). *Teaching tips: Strategies, research and theory for college and university teachers* (10th ed.). Boston: Houghton Mifflin.

Mumford, H. W. (1902). *Market classes and grades of cattle with suggestions for interpreting market quotations*. Bulletin Number 78. Urbana, Illinois: Agricultural Experiment Station.

Munk, D. D. (2003). *Solving the grading puzzle for students with disabilities*. Whitefish Bay, WI: Knowledge by Design.

Munk, D. D., & Bursuck, W. D. (2003). Grading students with disabilities. *Educational Leadership, 61*(2), 38-43.

Office of Research. (1994). *What do student grades mean? Differences across schools*. Washington, DC: U. S. Department of Education.

Panadero, E., & Jonnson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129-144.

Pitler, H. (2016, June 1). *My problems with letter grades in school*. Retrieved from http://inservice.ascd.org/my-problems-with-letter-grades-in-school.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups*. College Board Report No. 93-1. Retrieved at http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1993-1-student-group-differences-predicting-college-grades.pdf

Robelen, E. (2011, December 8). *Most teachers see the curriculum narrowing, Survey finds*. Retrieved from http://blogs.edweek.org/edweek/curriculum/2011/12/most_teachers_see_the_curricul.html.

Rojstaczer, S., & Healy, C. (2010, March 4). *Grading in American colleges and universities*. Retrieved from http://www.gradeinflation.com/tcr2010grading.pdf.

Rorem, S. O. (1919). A grading standard. *School Review, 27*(9), 671-679.

Ruediger, W. C., Henning, G. N., & Wilbur, W. A. (1914). Standardization of courses and grades. *Science, 40*(1035), 642-643.

Rugg, H. (1918). Teachers' marks and the reconstruction of the marking system. *Elementary School Journal, 18*(9), 701-719.

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*, 807-826.

Saupe, J. L., & Eimers, M. T. (2012, June). *Alternative estimates of the reliability of college grade point averages*. Paper presented at the Annual Forum of the Association for Institutional Research, New Orleans, Louisiana.

Schinske, J., & Tanner, K. (2014). Teaching more by grading less (or differently). *Life Sciences Education, 13*(2), 159-166.

Schneider, J., & Hunt, E. (2013). Making the grade: A history of the A–F marking scheme, *Journal of Curriculum Studies*, DOI: 10.1080/00220272.2013.790480.

Schwartz, B. & Sharpe, K. (2011, January 9). *Do grades as incentives work?* Retrieved from https://www.psychologytoday.com/blog/practical-wisdom/201101/do-grades-incentives-work.

Scriffiny, P. (2008). Seven reasons for standards-based grading. *Educational Leadership, 66*(2), 70-74.

Simmons, R. G., Brown, L., Bush, D. M., & Blyth, D. A. (1978). Self-esteem and achievement of black and white adolescents. *Social Problems, 26*, 86-96.

Slavov, S. (2013, December 26). *How to fix college grade inflation*. Retrieved from http://www.usnews.com/opinion/blogs/economic-intelligence/2013/12/26/why-college-grade-inflation-is-a-real-problem-and-how-to-fix-it

Smith, K. (2003). *Contract grading rubric, Civil Engineering 4101*. Minneapolis, MN: University of Minnesota Center for Writing.

Sorian, R,. & Baugh, T. (2002). Power of information: Closing the gap between research and policy. *Health Affairs*, *21*, 264-273.

Starch, D. (1913). Reliability and distribution of grades. *Science, 38,* 630-636.

Starch, D., & Elliott, E. C. (1912). Reliability of grading of high school work in English. *School Review, 20*, 442-457.

Starch, D., & Elliott, E. C. (1913). Reliability of the grading of high school work in mathematics. *School Review, 21*, 254-259.

Taylor, H. (1980). *Contract grading*. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation.

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation, 18*, 153-172.

Tieje, R. E., Sutcliffe, E. G., Hillebrand, H. N., & Buchen, W. (1915). Systematizing grading in freshman composition at the large university. *English Journal, 4*(9), 586-597.

Tinney, J. (2014, January 12). *What do letter grades actually mean?* Retrieved from http://www.jordantinney.org/ what-do-letter-grades-actually-mean.

Tomlinson, C., & McTighe, J. (2006). *Integrating differentiated instruction and understanding by design*. Alexandria, VA: ASCD.

University of Liverpool. (2015). *Code of practice on assessment, Appendix A: University Marks Scale, Marking Descriptors and Qualification Descriptors*. Retrieved from https://www.liverpool.ac.uk/media/livacuk/tqsd/code-of-practice-on-assessment/appendix_A _2011-12_cop_assess.pdf.

Weld, L. D. (1917). A standard of interpretation of numerical grades. *School Review, 25*, 412-421.

Wiggins, G. (2013, January 17). *Intelligent vs. thoughtless use of rubrics and models, Part 1*. Retrieved from https://grantwiggins.wordpress.com/2013/01/17/intelligent-vs-thoughtless-use-of-rubrics-and-models-part-1.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*(1), 1–37. doi:10.1002/j.2333-8504.2000.tb01838.x

Zahner, D., Ramsaran, L. M., & Steedle, J. T. (2014). *Comparing alternatives in the prediction of college success*. New York: Council for Aid to Education.

About the Author

Lorin W. Anderson
University of South Carolina (Emeritus)
anderson.lorinw@gmail.com
Lorin W. Anderson is a Carolina Distinguished Professor Emeritus at the University of South Carolina, where he served on the faculty from August, 1973, until his retirement in August, 2006. During his tenure at the University he taught graduate courses in research design, classroom assessment, curriculum studies, and teacher effectiveness. He received his Ph.D. in Measurement, Evaluation, and Statistical Analysis from the University of Chicago, where he was a student of Benjamin S. Bloom. He holds a master's degree from the University of Minnesota and a bachelor's degree from Macalester College. Professor Anderson has authored and/or edited 18 books and has had 40 journal articles published. His most recognized and impactful works are *Increasing Teacher Effectiveness, Second Edition*, published by UNESCO in 2004, and *A Taxonomy of Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, published by Pearson in 2001. He is a co-founder of the Center of Excellence for Preparing Teachers of Children of Poverty, which is celebrating its 14th anniversary this year. In addition, he has established a scholarship program for first-generation college students who plan to become teachers.

About the Guest Editors

Lorin W. Anderson
University of South Carolina (Emeritus)
anderson.lorinw@gmail.com
Lorin W. Anderson is a Carolina Distinguished Professor Emeritus at the University of South Carolina, where he served on the faculty from August, 1973, until his retirement in August, 2006. During his tenure at the University he taught graduate courses in research design, classroom assessment, curriculum studies, and teacher effectiveness. He received his Ph.D. in Measurement, Evaluation, and Statistical Analysis from the University of Chicago, where he was a student of Benjamin S. Bloom. He holds a master's degree from the University of Minnesota and a bachelor's degree from Macalester College. Professor Anderson has authored and/or edited 18 books and has had 40 journal articles published. His most recognized and impactful works are *Increasing Teacher Effectiveness, Second Edition*, published by UNESCO in 2004, and *A Taxonomy of Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, published by Pearson in 2001. He is a co-founder of the Center of Excellence for Preparing Teachers of Children of Poverty, which is celebrating its 14th anniversary this year. In addition, he has established a scholarship program for first-generation college students who plan to become teachers.

Maria de Ibarrola
Department of Educational Research, Center for Research and Advanced Studies
mdeibarrola@gmail.com
Maria de Ibarrola is a Professor and high-ranking National Researcher in Mexico, where since 1977 she has been a faculty-member in the Department of Educational Research at the Center for Research and Advanced Studies. Her undergraduate training was in sociology at the National Autonomous University of Mexico, and she also holds a master's degree in sociology from the University of Montreal (Canada) and a doctorate from the Center for Research and Advanced

Studies in Mexico. At the Center she leads a research program in the politics, institutions and actors that shape the relations between education and work; and with the agreement of her Center and the National Union of Educational Workers, for the years 1989-1998 she served as General Director of the Union's Foundation for the improvement of teachers' culture and training. Maria has served as President of the Mexican Council of Educational Research, and as an adviser to UNESCO and various regional and national bodies. She has published more than 50 research papers, 35 book chapters, and 20 books; and she is a Past-President of the International Academy of Education.

D. C. Phillips
Stanford University
d.c.phillips@gmail.com
D. C. Phillips was born, educated, and began his professional life in Australia; he holds a B.Sc., B.Ed., M. Ed., and Ph.D. from the University of Melbourne. After teaching in high schools and at Monash University, he moved to Stanford University in the USA in 1974, where for a period he served as Associate Dean and later as Interim Dean of the School of Education, and where he is currently Professor Emeritus of Education and Philosophy. He is a philosopher of education and of social science, and has taught courses and also has published widely on the philosophers of science Popper, Kuhn and Lakatos; on philosophical issues in educational research and in program evaluation; on John Dewey and William James; and on social and psychological constructivism. For several years at Stanford he directed the Evaluation Training Program, and he also chaired a national Task Force representing eleven prominent Schools of Education that had received Spencer Foundation grants to make innovations to their doctoral-level research training programs. He is a Fellow of the IAE, and a member of the U.S. National Academy of Education, and has been a Fellow at the Center for Advanced Study in the Behavioral Sciences. Among his most recent publications are the *Encyclopedia of Educational Theory and Philosophy* (Sage; editor) and *A Companion to John Dewey's "Democracy and Education"* (University of Chicago Press).

SPECIAL ISSUE
Historical and Contemporary Perspectives on Educational Evaluation

education policy analysis archives

Volume 26 Number 49 April 16, 2018 ISSN 1068-2341

Please send errata notes to Audrey Amrein-Beardsley at Audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.