GAVIN PUBLISHERS

Educational Research Applications

Anderson LW. Educ Res Appl 2017: J112.

DOI: 10.29011/2575-7032/100012

Review Article

A Critique of Grades, Grading Systems, and Grading Practices

Lorin W Anderson

Carolina Distinguished Professor Emeritus, University of South Carolina, USA

*Corresponding author: Lorin W Anderson, University of South Carolina, USA. Tel: +803-206-2559; E- mail: anderson.lorinw@gmail.com

Citation: Anderson LW (2017) A Critique of Grades, Grading Systems, and Grading Practices. Educ Res Appl 2017: J112. DOI: 10.29011/2575-7032/100012

Received Date: 12 April, 2017; **Accepted Date:** 15 May, 2017; **Published Date:** 22 May, 2017

Introduction

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 [1]. 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 [2].

These two quotations, written a century apart, are illustrative of 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.org will yield scores of articles with similar points of view. Several educators, most notably [3-9] have published extensive criticisms of grades, grading systems, and grading practices.

Not only are the criticisms timeless, they are widespread. Teachers [2,10] and educational consultants [11] have railed against grading in general and/or specific grading practices. Writing more than a half century ago, Dorothy de Zouche called the giving of grades one of her "ten educational stupidities." Mark Barnes, an education consultant, gave a TED talk in which he addressed the question, "Isn't it time to eliminate grades in education?" [Although I haven't watched the TED talk I'm fairly certain that his answer to the question is "Yes!"]

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, for some, signing grades has value? My purpose in writing this essay is to offer, as the title suggests, a critique of grades, grading systems, and grading practices. I am using the word "critique" as defined in the Merriam Webster Learner's Dictionary, namely, a "careful judgment in which you give your opinion about the good and bad parts of something." To facilitate my critique, I have organized this essay around five basic questions.

- Why do we grade students?
- What do grades mean?
- How reliable are students' grades?
- How valid are students' grades?
- What are the consequences of grading students?

Because most discourse and empirical evidence pertain to secondary schools and postsecondary institutions, the vast majority of my review will focus on these school levels. The lone exception is the discussion of the validity of grades where elementary school research will be included. Also, I will limit my analysis and commentary primarily to the United States, although a few studies of, and commentarieson,grading in other countries are included.

Before addressing these questions, however, let me attend to some definitional matters. "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.

Finally, it should be noted that early writers in the field [12] as well as some British higher education institutions today [13] use the term "marks" rather than "grades" and talk about "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?

Why do we mark pupils at all? 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? [14].

In the above quotation, Campbell suggests two possible reasons for grading students:

- To motivate students to put forth greater effort and
- To provideinformationthat teachers can use to improve their instruction.

- More recently, a third reason for grading has been proffered, namely, to communicateinformation about student learning to a variety of audiences (e.g., parents, employers, members of the media) who want and/or need information about how well students are learning or progressing in order to make decisions about the students [15].

Motivating Students

The belief that grades are inherently motivating is long-standing. Almost a century ago, [16], a junior high school principal, wrote that "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?" (p. 671). Two years later, [14], a high school educator,statedwhile "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" (p. 511). Because these educatorsbelieved that motivation was increased by competition among students, many of these early grading systems were based on rankings among students rather than ratings of the quality of individual student's work or learning [17].

Even if grades do have some motivationalvalue [18], some critics maintain that grades foster the "wrong" kind of motivation. They point out that working harder to achieve better grades is not the same as workingharder to learn more. In fact, the results of several studies suggest that the two "orientations" (working to get good grades vs. working to learn) are inversely related [4]. Furthermore, students who are motivated by grades are less likely to be interested in what they are learning [5], more likely to avoid challenging tasks [19], 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) [20].

Schinske& Tanner (2014) [19] 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

Almost a century ago, [14] wrote that "in practice, the ordinary marking system simply registers relative standing with respect to other pupils in the class. It can be said to give, at most, a general diagnosis of the pupils' relative condition; 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" (p. 510). This statement is as valid today as it was then. Grades typically do not provide sufficiently precise information that can be used by teachers to improve their instruction. To be used for improvement purposes, grades must provide reasonably detailed information about what students, individually and collectively, have and have not learned … know and do not know … can and cannot do. Advocates of "standards-based grading systems" [21]argue that their systems provide the necessary level of detail.

In standards-based grading students are evaluated on the basis oftheir mastery of a clearly articulated set of course objectives (widely known as academic standards, curriculum standards, content standards, or, simply, standards) [22]. Students receive a separate grade for each standard; they may also receive an overall grade for the curriculum unit in which they standards are embedded. (Table 1) contains a sample of a standards-based grade report for a single student in chemistry.

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

Table 1: A Portion of a Standards-Based Report.

The report begins by identifying the five standards associated with a chemistry 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. A cursory examination of the table suggests that the student (Olivia George) is "proficient" or "excellent" in three of the five standards. The two weaknesses are the student's ability to apply the concept of density ("approaches proficiency") and the student's ability to select the appropriate gas law to solve a problem ("well below proficiency"). It is reasonable to assume that information such as this could at the very least help teachers understand where they need to spend additional time and effort. However, the information does not inform teachers as to how they should change their instruction in order for the student to achieve these two standards (Campbell's search for "treatment").

Although Table 1 focuses on a single student, standards-based systems also permit the identification of learning strengths and weaknesses of an entire class, gradelevel, state, or country (see Table 2). Looking at this aggregated data, we see once again that student achievement relative to the third and fifthstandards is relativelyweak, with one-half of the students failing to reach proficiency (that is, Level 3) in "Applying the Concept of Density" and 85 percent of the students failing to achieve proficiency in "Selecting Appropriate Gas Laws to Solve Problems". Once again, this information can help teachers target their instruction, in this case to an entire class or small 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%

Table 2: A Portion of a Standards-Based Report for a Group of Students. Note: The numbers in the cells represent the Percent 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 ratings) can easily be converted to comparisons between and among them (that is, a student's Ranking Within a Group or Class). Consider, For Example, Olivia George, The Student Depicted in (Table 1). Across the five standards, Olivia has a grading pattern of 4-4-2-4-1. Her achievement in this curricular unit is 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.

Communicatingwitha Variety of Audiences

The primary purpose of grades is to communicate student achievement to students, parents, school administrators, postsecondary institutions, and employers" [15].

This statement, 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 of purpose is quite simple and 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

- There is an exclusive focus on student achievement, and
- The list of audiences is incomplete. Because an exclusive focus on student achievement is intended to enhance the meaning and validity of the assigned grades, more will be said about this issue later.

For now, I will focus on the question of who is missing from the list and why these omitted audiences are important. First and foremost, I would add teachers; not those who assigned the grades to the students, 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?" Second, I would add policy makers (including, but not limited to, elected officials). 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". How does this change level the playing field?How does it benefit transfer students? One verifiable consequence of the change is that approximately 6,000 more students will receive state-supported scholarships to post-secondary institutions, costing the state an additional $50 million over a four-year period.Are the benefits of this change in the grading system worth the cost? Third, I would add members of the media. In doing the research for this paper I came across the following headline from the Washington Post: "Is it becoming too hard to fail? Schools are shifting toward no-zero grading policies" [23]. In essence, these policies discourage teachers from assigning percentage grades lower than 50 if a student makes a "reasonable attempt to complete the work." The primary rationale for this policy is that a grade of zero on a single assignment makes it extremely difficult for a student receiving a grade of zero to overcome this grade and earn a satisfactory grade for an entire marking period (e.g., quarter, semester, or, to a lesser extent, year). 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, want more failing students?

Schneider and Hutt (2013)[24] 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). In this regard, it is instructive to point out that the rapid increase in the use of number and letter grades in the United States corresponded with the passage of compulsory school attendance laws in the late 1800s and early 1900s. One consequence of these laws was a substantial increase in the number of public high schools, from about 500 in 1890 to approximately 10,000 some twenty years later. With more schools came more students and, with more students, a more efficient way of recording and reporting on their academic progress was needed [25].

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 [4], and/or conferences [26] are 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 evident today than in many selective universities in which admissions officers have begun to place a greater value on interviews, essays, and written reports in making admission decisions [27] while, at the same time, theOffice of Communications and Public Affairs at these universi-

ties continues to release to the media the number of valedictorians in, or the mean SAT scores of, the incoming freshman class.

In concluding this section, it is important to emphasize that simply providing grades is not enough. The information inherent in the grades must be used to make decisions about students, teachers, programs, schools, and/or countries. If agrade indicates that a student is not doing well in Algebra I, what should be done about it? If 75 percent of students are failing Algebra I in a particular school, what should be done about it? To move from communication to decision making requires that those making the decisions understand the meaning of the grades.It is to this issue that we now turn.

What Do Grades Mean?

What is a grade? 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 [16].

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 [28].

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 [16] suggested almost a century ago, a student may receive a "good" grade in one teacher's classif 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 in order to achieve a "good" grade. A student may receive a"good" grade if work was handed in on time in one teacher's class [28], 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.

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)

Table 3: A Summary of Differences in What Grades Represent.

As shown in the first row of the table, a grade can represent a student's performance on a single task (e.g., a homework assignment, a quiz or test, an essay, a report). I refer to these as "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, over subjects and teachers (e.g., Grade Point Average). I refer to these as "cumulative grades" (occasionally "course grades" or "Grade Point Average"). 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. As we shall see, the distinction between single task grades and cumulative grades is extremely important when the technical quality of grades (e.g., reliability, validity) is examined.

As shown in the second row, a grade may represent a student's achievement levelat a particular point in time. Alternatively, a grade may represent how much a student has learned over time (that is, how much a student's achievement has improved). Most 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

- It is a difficult thing to measure and
- It is unfair to initially high achieving students who have little if any room to improve [29,30].

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 [31]. 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'[32].

As shown in the third row, a grade may represent academic achievement only [15]. Alternatively, a grade may represent some combination of academic achievement, effort, attendance, class participation, and possibly other factors. A grade representing only academic achievement is typically easier to interpret than a grade representing some combination of factors. As[33]has written, if a teacher, in determining a student's grade, merges "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" (p. 18). Nonetheless, there is some evidence that teachers tend to consider factors other than achievement when assigning grades (e.g., motivation, classroom behavior) [34].

Finally, as shown in the fourth row, a grade may represent student achievement relative to intended learning outcomes (e.g., goals, objectives, or standards). Alternatively, a grade may represent student achievement relative to the achievement of his or her peers.Virtually all grading systems in the early 20th Century were based on comparisons between and among students. Today we would say that these grading systems were "norm-referenced". In 1963, Robert Glaser [35]argued that educators should move away from "norm-referenced" measurement to what he termed "criterion-referenced" measurement. Rather than compare students against each another, criterion-referenced measurement emphasized a student's acquisition of knowledge and skills along a continuum of proficiency, ranging from none to "perfect". By the mid- to late-1980s, the term "standard" had replaced the term "criterion" and "criterion-referenced measurement" gave way to "standards-based grading." Rather than being ranked in terms of their peers, students were to be rated in terms of their learning relative to pre-determined curricular standards or learning expectations.

In light of this discussion it seems reasonable to conclude that grades do not have (nor have they ever had) universal meanings. The standardization sought by Rorem almost a hundred years ago has not come to fruition and, quite likely, never will. Rather, the meaning of a grade is context- or situational-specific. As a consequence, it is virtually impossible to compare a grade in Ms. Davis's biology class with the "same" grade in Ms. Crawford'sbiology class, let alone compare a grade in Ms. Crawford's biology class with the "same" grade in Mr. Herzog's geometry class.

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 action should be taken given this state of affairs? Almost a century ago, [12] emphasized the need to achieve universal meanings of grades. It seems quite clear that neither the number nor the letter has carried a common meaning to the student, to the teacher, and to the administrator alike. We must not forget that these three groups have to 'think together' concerning school marks, and our aim is to build a marking system simple and economical to administer, and yet one which will enable these three sets of minds to agree on the marks to be put on the results of instruction" (p. 713).Unfortunately, as mentioned above, very little progress has been made in this regardduring the intervening years; furthermore, little, if any, progress is foreseen.

A more reasonable alternativewould beto recognize and embrace the context- or situational-specific nature of grades. This would require that each teacher (or group of teachers) communicates clearlythe meaning of each grade. (Table 4) illustrates one attempt to communicate the meaning of letter grades [36]. Note that it is possible (and, in some cases, may be desirable) to provide 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").

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

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

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 in order 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 signs a contract in which the teacher is committed to award the contracted grade if the student meets or exceeds thoselevels [37].

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		

Table 5: A Sample Contract System [38].

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 meaning of grades. Without such attempts, the interpretation of a grade rests solely with the recipient of the grade, typically, the student. When this happens, we are left with an entire classroom, school, or educational system composed of Humpty Dumptys.

How Reliable are Students' Grades?

As suggested earlier, 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 the question of reliability is quite clear. Single task grades are veryunreliable. 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 grades (also known as percentage grades since the scale ranges from 0 to 100).

Two of the landmark studies were conducted by [39,40], the first in high school English, the second in high school mathematics. In each study a fairly large group of teachers wasgiven the same student's written response to a task (i.e., two written essays in the first study, a worked-out solution to a mathematics problem in the second)and asked to assign a grade from 0 to 100 to each written response. For the two essays the grades ranged from 50 to 90 and from 64 to 98. For the worked-out mathematics problem, the range of grades was even larger (28 to 92) [41].

Rugg (1918) [12] conducted a systematic review of 23 studies published during the previous three years. Two of theconclusions reached by Rugg are quite 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 byan individual teacherto the same piece of student workgraded at two different times there is "distinct evidence of unreliability of marking" (p. 703). That is, individual teachers are not consistent 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) [41] identified fourpossible sources of low inter-rater reliability:

- Differences caused by the inability of teachers to "distinguish between closely allied degrees of merit" (p. 630).
- Differences in the criteria used by different teachers (e.g., content, mechanics, and style in grading essays).
- Differences in the quality standards used by different teachers (e.g., what differentiates "excellent" work from "good" work?).
- 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).

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.

Table 6: Sources of Unreliability and Proposed Remedies for Low Reliabilities.

In response to the inability of teachers to make the small distinctions required bypercentage grading, [12]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 populargrading 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, [42] 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 standardsfor 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 was 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 [43], there is some doubt that the introduction and use of rubrics alone will solve the reliability problem. Brimi (2011) [44] 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 0 to 100 scale. The grades assigned ranged from 50 to 96 a range similar to that of [39].

These findings are consistent with the results of a review of literature conducted by [45] who concluded that "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 others have referred to as "anchor papers" [46]. Anchor papers are intended to help teachers gain a more complete understanding of the meaning of both the criteria and, perhaps more importantly, the quality standards (both of which are described quite concisely on the rubric).

Although anchor papers may help reduce the problem of teachers having different quality standards, a very early attempt by [25] 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 had different quality standards, but minimized their impact on the grades that students were assigned by giving each teacher a correction factor based on previous grades assigned by the teacher and then using this correction factor to adjust each student's grade accordingly.

Finally, one of the early attempts 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 1914the Committee on Standardizing Grades of the American Association for the Advancement of Science, for example,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" [47].Educators' belief and faith in the normal distribution continued through much of the 20thCentury.

Unfortunately, the distributions of grades assigned by teachers at that time were not normally distributed. In his review of research, [12] concluded that "there is enough evidence that teachers' marks tend to be 'skewed' to the high end of the scale" (p. 704). Of the several hundred grade distributions he examined, Rugg found that fewer than 10 percent could be described as "perfectly symmetrical" and that "not more than two or three in a hundred of all those examined by been found to be approximately normal" (p. 705).Data reported as part of the National Educational Longitudinal Study (NELS:88) suggests that the distribution of grades remains negatively skewed [48]. Almost 70 percent 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 cumulative grade. A student'sGPA 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 Spring, 2016, semester) or for an entire academic career (that is, all courses leading to the award of a bachelor's degree). Typically, an A grade is worth 4 points, a B grade is worth 3 points, and so on. The studies focus on the stability of GPAs over courses and over time [49,50]. Notice that in contrast with the reliability of single task grades, the reliability of cumulative grades is defined in terms of stability.

One of the more recent studies, conducted by [51] at the University of Missouri, illustrates both the procedure and the results. The study began by collecting the end-of-fall-semester GPAs of5,000freshmen student's GPAswere collected each subsequent semester, with slightly smaller sample sizes each semester due to 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 [49,50,52]. Based on the available data, then, it seems reasonable to conclude that, unlike single task grades, cumulative grades are quitereliable.

When asked about 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) are very reliable. At the same time, however, we know that cumulative grades are determined to some extent by students' single task grades which are unreliable.How can this inconsistency be explained?

As mentioned earlier, we are dealing with two different forms of reliability: consistency across teachers in the case of single task grades, and consistency across teachers and subjects over time in the case of cumulative grades. Consider the data presented in (Table 7), which are similar to the data collected in the two Starch and Elliot studies.

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

Table 7: Teacher Numerical Grades of one Student's Written Composition.

We have one 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 or percentage score assigned to the essay by each teacher. As can be seen in (Table 7), the grades range from 30 to 90, with a mean of 60. The logical conclusion based on these data (and the conclusion reached by Starch and Elliott) is that the grades assigned are quite unreliable (that is, quite inconsistent across teachers).

But, what would happen if we added a second student (that is, an essay written on the same topic by a second student) and we ask the same teachers to grade that student's essay (see Table 8). If we focus attention only on the second student (Student B), a similar pattern of inconsistency emerges. The grades given by the teachers to this student's essayrange from 10 to 70 with a mean of 46. Note that the range of grades assigned to the two students, 60 points, is identical.

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

Table 8: Teacher Numerical Grades with Two Hypothetical Students.

If rather than focusing on each student separately, we compare these students in terms of the grades assigned to them, a different picture emerges. All five teachers assigned higher grades to the first student's essay; theoverall mean percentage score differs by 14 points.Even with the lack of agreement among the teachers on each individual student's essay, then, it is quite clear that the grades assigned by the teachers reliably differentiate student A's essay from student B's.

If we add more students, replace teachers with semesters, and replace percentage grades with GPAs in the cells of the table, we can represent the data from the [51]. (Table 9) contains a simulated portion of Saupe and Eimer's data set (10 teachers x 8 semesters).An examination of the columns of the table suggests that there are differences in GPAs across semesters. However, only one of these semester-to-semester differences is as large as 1.5, with almost one-half of these differences being zero or one-half of a grade point. If we examine the rows of the table, students 0001 through 0003 consistently have lower GPAs (with means of 1.94, 2.25, 2.31, respectively) than students 0008 through 0010 (with means of 3.37, 3.43, 3.50, respectively). Thus, even though there is some variation in GPAs from semester to semester, this "between semester" variation is quite small relative to the "between student" variation. The alpha coefficient across all eight semesters for the data in Table 9 is approximately 0.90, as compared with Saupe and Eimers' coefficient of 0.91. That is, more than 90 percent of the variation in GPA can be attributed to differences among students, rather than differences among semesters.

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

Table 9: Students x GPAs.

In conclusion, 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, in fact, have different quality standards that cause them to differ from one another in the grades they assign to student work while, at the same time, allowing them to agree that some student work is superior to other work.

How Valid are Student Grades?

Our problem may be stated somewhat as follows: Given an average school system with … forty to forty-eight pupils under the care of one teacher, (how) to organize a plan of grading and promotion, and to 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 [53].

Answering the validity questions is somewhat more difficult than answering the question of reliability. As was true of reliability, there are different types of validity. Similarly, as was true ofthe 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 test scores accurately reflect student achievement, that college grades accurately reflect college success).

Different Types of Validity

The validity of student 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) and hasbeen labeled concurrent validity [54]. 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; that is, they are said to possess a reasonable degree of predictive validity [55]. The data needed to answer these questions come from studies ofcourse 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 differencein 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 [48]. 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, the 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 mostly closely resembled the scores of "D" students in the more affluent schools. Similar results have been reported by [52,56]. The evidence from studies such as these clearly suggest that, as [57] put it, "an A is not an A is not an A" (p. 294).

The second threat to validity is grade inflation, a somewhat more recent phenomenon [58]. 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) [59] provides an example of 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 employers and graduate schools to differentiate (between) them (p. 2).

Notice that in his analysis Slavov addressed both concurrent and predictive validity. Both students received a grade of A despite the fact that, according to Slavov's perspective, the first student (who is truly outstanding) learned more than (or demonstrated greater ability than) the second (who is only very good). Thus, according to Slavov's analysis, grade inflation is likely to diminish the concurrent validity of grades. Furthermore, because of this failure to differentiate between the two undergraduate students, employers and graduate school admissions officers have a more difficult time selecting between them (if such a selection is necessary). The logical conclusion of this argument is that when the wrong choice is made, the predictive validity of grades is likely to be lessened.

Evidence Pertaining to the Validity of Grades

Studies investigating the concurrent validity typically examine the relationship between cumulative grades and test scores. The argument underlying the design of these studies is quite straightforward. If test scores accurately reflect student achievement and if students with higher test scores receive higher grades, then students who receive higher grades have learned more.

The available evidence from the available studies suggests that the correlations between cumulative grades, broadly defined, and test scores range from 0.30 to 0.75. The smallest correlations can be found in studies of the relationship between students yearly

average grades and their composite scores on comprehensive test batteries [60]. When the focus shifts to grades in a specific subject matter in relation to scores on subject-specific tests (e.g., reading, mathematics), the correlations range from 0.50 to 0.75 [54,61-63]. Finally, when the 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), the correlations are at the upper end of the aforementioned range[64]. When the results of the available studies are combined, then, it seems reasonable to conclude that theshared variation of subject-specific grades awarded and subject-specific test scores is somewhere between 25 and 50 percent. The magnitude of this shared variation is particularly impressive in light of the restricted range of assigned grades as a result of the severe negative skewness of the grade distribution (as mentioned earlier).

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 beingdefined in terms of earning an undergraduate degree.

As was true for the studies of concurrent validity, the results of the predictive validity 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 [65-67]. 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 the restriction of range of HSGPA, differences in the college courses in which students are enrolled, and differences in instructors' grading standards, there is a substantial increase in their magnitude. [68], for example, reported an increase from 0.36 to 0.69 when these corrections were made. If corrections for the restriction of range of college GPAs (resulting from grade inflation) were made, these correlations likely would be even stronger.

Quite importantly, the strength of these coefficients remains virtually unchanged over the student's college career. In fact, [69] found that the "predictive weight associated with HSGPA increased after the freshman year, accounting for a greater proportion of variance in cumulative fourth-year than first-year college grades" (p. 2). Finally, there is some evidence (although sparce) that HSGPA predicts the likelihood that a student will receive a college degree. Astin, Tsui& Avalos (1996) [70], 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". About half of the students with HSGPAs of "B" received a college degree.

Although almost all studies of the predictive validity of grades focus on college success, there are two additional studies that are noteworthy. Kurlaender & Jackson (2012) [71] 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 7th 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.

Karen Arnold (1995) [72] conducted a 14-year longitudinal study of 81 high school valedictorians who graduated from high school in Spring, 1981. Although she presented a host of data in her book-length report, she provided a concise summary of her results. The valedictorians "continued to do well in college with an overall GPA of 3.6. Most went on to work in conventional careers such as accounting, medicine, law, engineering, and education" (p. 310).

In summary, then, the available evidence tends to support both the concurrent and predictive validity of cumulative grades. Specifically, there is some evidence that cumulative grades are positively related to

- Achievement test scores.
- The likelihood of receiving a high school diploma.
- College grades over multiple years, and
- 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" [73].

It is quite true, as Mathews points out, that the grades students receive can and do impact the quality of their lives. It is important to point out, however, that the impacts can be either positive or negative. In addition, factors other than grades can and do impact students' lives.

Most of the critics of grading tend to focus only on the negative effects. Alfie Kohn (1999,2011)[4,5], for example, has compiled a list of eight negative consequences of grading students using letters or numbers. They are:

- Grades tend to reduce students' interest in the learning itself.
- Grades tend to reduce students' preference for challenging tasks.
- Grades tend to reduce the quality of students' thinking.

- Grades distort the curriculum.
- Grades waste a lot of time that could be spent learning.
- Grades encourage cheating.
- 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. In this regard, consider the following:

- Boring teachers, activities, and tasks reduce students' interest in the learning itself [74].
- Federal and state mandates distort the curriculum [75].
- Negative teacher behavior spoils teachers' relationships with students [76].
- Pecking order, cliques, and self-segration spoil students' relationships with each other [77].

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 are [78,79]. Rather, the revised statements are intended to show that grades are no more or less harmful than many other aspects of schooling (e.g., boring tasks, federal mandates, negative teacher behaviors, and cliques).

More importantly, however, the evidence suggests that the negative effects of grades on students tend to accumulate over time. To use terminology introduced earlier, the impact of a single task grade is likely to be minimal. The impact of a cumulative grade, on the other hand, can be very severe.

A study conducted by [80] is instructive in this regard. Kifer conducted a quasi-longitudinal study of students at four grade levels (2, 4, 6, and 8). At each of these grade levels, two groups of students were identified. Group A consisted of students who had been in the top 20 percent of their class each year. Group B consisted of students who had been in the bottom 20 percent of their class each year. To students in both groups he 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, the differences between the two groups were both substantial and statistically significant. Furthermore, the graphs prepared by Kifer showed quite clearly that the ASC scores of Group A did not change much from grade to grade. For Group B, however, 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 determinantal" [81]. Consistently receiving low grades (e.g., mostly D's and F's) is likely to transform "failing" into "failure."

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. Thatis, 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 (in the teacher's eyes) an "A" student. On the other hand, a series of poorly written essays or poor performances on tests can prompt a teacher to view the student as a "D" or "F" student.

Once a grade is transferred from the student's work to the student himself or herself, it can influence the way a teacher grades subsequent work by that student. Suppose, for example, a student perceived by the teacher as an "A" student submits an essay that receives a grade of "B." It would not be uncommon for the teacher to write a note on the essay when it is returned to the student. "I know you can do better. Please revise and resubmit." On the other hand, this same essay written by a "D" student might be greeted with surprise since the work is "better than expected." If the disparity is sufficiently large, cheating may be suspected.

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 a fairly extensive quotation from a paper written by Stanley S. Marzolf almost sixty 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. … That reporting marks is often a cause of much anxiety is undoubtedly true. It is my contention that many of the evils of marks and marking are unnecessary and arise from ignorance, incompetence, and spite. Though present practices are by no means ideal, there are nevertheless some values in marks which can and must contribute to mental health. These values must be preserved in any modification of present practices. If one is to learn, one must have knowledge of results" [82] (emphasis mine).

Discussion

The power of grades to impact students' future (lives) creates a responsibility for giving grades in a fair and impartial way (Johnson and Johnson, 2002, p. 249 [83].

As I reflect on the reading I have done in the preparation of this paper, my overall impression is that we have a long way to go before we fulfill our "responsibility for giving grades in a fair and impartial way." In 1902 Herbert Mumford, [84] a professor at the University of Illinois, 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 [85]. Unfortunately, the same cannot be said of the way that students are graded. What needs to be done to move forward? I offer fiverecommendations.

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" [17].

Since the publication of A NationatRiskin 1983, there have been numerous recommendations as to the best ways of reforming public education in the United States, particularly preK-12 education. There seems to be some consensus that we need to increase the rigor of the curriculum, employ highly qualified teachers, provide more personalizedlearning opportunities for students, integrate technology into the instructional program, and improve school-community relations (including parent education and involvement). Typically, the reformers argue that these improvementsmust be seen as interrelatedparts of a conceptual framework intended to substantially improve the entire educational system [86]. Notably absent from these conceptual frameworksis anything to do with grades, grading systems, or grading practices. As a consequence, concerns about grading, when they arise, are seen to lie outsidethe educational system. Such a view permits changes in grading policies and practices to be made without considering the impact of these changes on the educational system as a whole (or vice versa).

It should not be surprising, then, that many of the changes that have been made in grading policies and practices over the past quarter century have been rather superficial. For example, a shift from a 7-point scale (A = 93-100) to a 10-point scale (A = 90-100) advantages some students (e.g., those with scores of 90 who now receive an A), but disadvantages others (e.g., those with scores of 98 whose accomplishment is diminished somewhat by virtue of more students receiving A's). In the larger scheme of things, however, the change makes little, if any, difference. Based on the studies of interrater reliability reviewed earlier we know that a grade of 90 is easier to earn in Ms. Smith's class than in Ms. Phillips' class, in journalism than in calculus, and on true-false tests than on analytical papers. Until more substantive problems inherent in grading are resolved, there is nothing uniform about so-called Uniform Grading Scales. Just because the range of numbers is standardized does not mean that the quantity or quality of work within each score range is identical. It is unlikely that more substantive problems will be resolved until grading policies and practices are considered to be an integral part of the educational system.

Because grading systems, like school calendars, are fundamental components of the educational system, changes in grading systems are not 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 [87]. I would argue, however, that if the grading system changes were embedded in changes that are perceived as needed or desired in order to improve the educational system as a whole, the likelihood of accepting these changes might have been greater.

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" [12].

Rugg's statement provides a nice transition from the first recommendation to the second. If our grading system is, in fact, "one of the contributory causes of 'failure' in the public schools," it isnot sufficient to put duct-tape [88] on our current grading systems and practices. The current grading systems and practices (i.e., purposes, methods, and meaning) must be re-examined and reconceptualized. Two concepts that would be useful in both are-examination and reconceptualization are grade integrity and fairness.

Grade integrity can be defined as "the extent to which each grade awarded, either at the conclusion of a course or [unit] of study, or for an extended response to an assessment task, is strictly commensurate with the quality, breadth and depth of a student's performance" [89]. Grade integrity, then, is the "correspondence between the actual level of achievement and what the assigned grade is assumed to stand for, as judged from either explicitly stated intentions, or inferences from practice and usage" [90]. Grading systems and practices are more likely to possess integrity if they adhere to principles such as these:

- The tasks given to students for the purpose of assigning grades should be representative of the essential intended learning outcomes. As a corollary, not all tasks should be graded.
- The quality standards used by teachers at the same grade level or teaching the same course should be as similar as possible. This will ensure that the grades represent degrees of accomplishment, rather than the "whims" of a particular teacher.
- 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.
- 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 seen as fair or unfair lies in the eyes of the

beholder. Teachers are quite inconsistent in their beliefs about fair and unfair grading practices [66]. For example, 57% of the teachers responding to the survey believed it was fair to include student performance on homework in the calculation of report card grades, whereas 43% believe 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, whereas 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 of grading practices and the grades they [91]. In general, students perceive grading and grades to be unfair when teachers:

- Fail to follow the guidelines of the current grading system;
- Assign grades based on unreliable information;
- Allow themselves to be influenced by irrelevant factors; and
- 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 [92] have written, "many students with disabilities receive inaccurate and unfair grades that provide little meaningful information about their achievement" (p.38). The authors suggested that to be fair to students with special needs, grading systems must

- Start with clear purposes in mind, purposes that take into consideration the information needs of parents (see Figure 1which follows) and other teachers.
- Incorporate adaptations for special needs students that are workable and promote access to and success with the general curriculum, and
- Include opportunities for individualized grading (similar to that provided by contract grading as described earlier).

Grade integrity and fairness, then, form a sound basis for a complete review of current grading systems and practices. In combination, these two concepts provide a framework for setting goals to achieve as new grading systems and practices are designed. Finally, as we work toward designing new systems and practices, advocating for one particular system or set of practices, as numerous educators have done [21,93], should be avoided. Rather, we shoulddesign a system and related set of 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. This leads nicely to our third recommendation.

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. That's why standards were developed in the first place. You can tie your 'A' to standards. Standards are a tool that lets teachers and parents monitor the rigor of the work children are expected to do" [94].

"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 mean they moved my son up a level, which is great. But we're also a little worried about that because I don't know if he's doing 'A' work, 'B' work, or 'C' work", A mother quoted in [94].

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 know what's in the best interest of parents (or legislators or journalists), educators would be wise to ask them what they need to know and understand about their children's learning and progress.

Sorian& Baugh (2002) [95], for example, reported on the results of telephone interviews with 292 policymakers, randomly selected from all fifty states. The questions focused on their information practices as well as their attitudes toward various types of information. The results indicated that only one-fourth of the respondents read material they received in detail withabout one-half of the respondents skimming for general content. The respondents reported being more likely to read material carefully if they found it to be "relevant".The material was seen as "not relevant" if it was

- Too long, dense, or detailed.
- Full of jargon, and
- Seen as overly subjective or biased.

Unfortunately, based on my decades of experience, educators like to provide a great amount of detail and use jargon (including oft-used acronyms).

Yogi Berra was quoted as saying "You can observe a lot just by watching." I would suggest that educators can learn a lot about the information needs of various audiences just by listening. Engaging members of the various audiences in an ongoing dialogue about grade reporting is a much wiser approach than assuming that we, as educators, know what they need. With respect to parents, for example, Munk (2003)[96]developed a survey that can be used to determine what parents want and need from the grades their children receive (see Table 10).

	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
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

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

Similar surveys can be developed for each stakeholder group. Once the needs of each audience are identified, a collaborative effort to design reporting systems that meet those needs can be made.

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 [17].

More than four decades later, there have been a few changes Teacher certification programs in most states require students to pass a course with measurement, assessment, and/or evaluation in its title. An examination of three of the most popular textbooks used in these courses, however, suggests that a single chapter (equivalent to between 5 and 10 % of the total number of pages in the book) is devoted to grading students. In contrast, almost one-third of the books deal with technical issues surrounding testing (e.g., multiple-choice, short-answer completion) and assessment (e.g., performance assessment, portfolio assessment). Furthermore, in all three textbooks, the chapter on grading is at or near the end of the book. In light of the many issues and concerns addressed in this paper, I would suggest that this is quite insufficient for prospective teachers to gain a knowledge of grading that they can use to design and implement defensible grading practices.

With respect to in-service teachers, I would suggest professional development sessions (perhaps organized by subject matter areas in high schools) in which teachers describe their grading systems and practices. A common set of questions can be developed so that comparisons across the various approaches can easily be made. Examples of such questions would include:

- What tasks do you assign for grading purposes?
- How do you grade student work on the tasks you assign?
- How do you differentiate among the various letter grades (e.g., A, B, C, D, F)?
- How to you combine task grades into a cumulative grade for a grading period or course?

A chart summarizing the answers to these questions (and others) can be created. Once created, similarities and differences, as well as strengths and weaknesses, can be identified and discussed. Ideally, discussions over time could lead to a more standardized approach to grading systems 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. Who knows regarding this particular matter of languages whether we are slavishly following traditions or fighting for a really good? Who knows whether the conservatives or the radicals are right? What is more, who can know under existing conditions? Personally, 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 [97].

At present, grades, grading systems, and grading practices are grossly under-researched fields. If you read the articles written during the first two decades of the 20th Century that are included in the references, you will likely be impressed by two things. First, there is an emphasis on solving practical problems. Second, data are used to inform decisions related to solving these problems. A century ago, this, apparently, was common practice. As a prime

example, [12] was able to locate 39 references over a three-year period that "dealt rather directly with the problem of the standardization of teachers' marks" (p. 701). If you were to spend the next several weeks searching various databases on the Internet, you quite likely will not be able to compile a list of 39 studies that provide data that would help you solve any of the problems associated with grades, grading systems, and grading practices.

It seems as though today's educators 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 solvingan 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 opinions 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" as we attempt to design grading systems and practices that are models of integrity, are seen as fair and impartial, and convey meaningful and useful information about individual students and entire educational systems.

References

1. Finkelstein IE (1913) The marking system in theory and practice. Educational Psychology Monographs 10.
2. Ferriter B (2015) If grades don't advance learning, why do we give them? Carrboro, NC: Center for Teaching Quality.
3. Kohn A (1993) Punished by rewards: The trouble with goal stars, incentive plans, A's, praise, and other bribes. Houghton Mifflin,Boston.
4. Kohn A (1999) From degrading to de-grading. High School Magazine 6: 38-43.
5. Kohn A (2011) The case against grades. Educational Leadership 69: 28-33.
6. Guskey TR, Bailey JM (2001) Developing grading and reporting systems for student learning. Thousand Oaks, CA: Corwin Press.
7. Guskey TR(2002) How's my kid doing? A parent's guide to grades, marks, and report cards. San Francisco: Jossey-Bass.
8. Guskey TR (2013) The case against percentage grades. Educational Leadership 71: 68-72.
9. Guskey TR (2014) Class rank weighs down true learning. Phi Delta Kappan 95: 15-19.
10. de Zouche D (1945) The wound is mortal: Marks, honors, unsound activities. The Clearing House 19: 339-344.
11. Barnes M (2014) How four simple words can solve education's biggest problem.
12. Rugg H (1918) Teachers' marks and the reconstruction of the marking system. Elementary School Journal, 18: 701-719.
13. University of Liverpool (2015) Code of practice on assessment, Appendix A: University Marks Scale, Marking.
14. Campbell AL (1921) Keeping the score. School Review 29: 510-519.
15. Bailey JM, McTighe J (1996) Reporting achievement at the secondary level: What and how. TR Guskey. Communicating student learning, Alexandria, VA: 119-140.
16. Rorem SO (1919) A grading standard. School Review 27: 671-679.
17. Cureton LW (1971) The history of grading practices. NCME Measurement in Education 2: 1-9.
18. Bull B (2013) 5 common reasons for the importance of letter grades.
19. Schinske J, Tanner K (2014) Teaching more by grading less (or differently). Life Sciences Education 13: 159-166.
20. Schwartz B, Sharpe K (2011) Do grades as incentives work?
21. Scriffiny PL (2008) Seven reasons for standards-based grading.Educational Leadership 66: 70-74.
22. Tomlinson C, McTighe J (2006) Integrating differentiated instruction and understanding by design. ASCD, Alexandria, VA.
23. BalingitM, StGeorge D (2016) Is it becoming too hard to fail? Schools are shifting toward no-zero grading policies.
24. Schneider J, HuntE (2013) Making the grade: A history of the A-F marking scheme. Journal of Curriculum Studies 46: 201-224.
25. Weld LD (1917) A standard of interpretation of numerical grades. School Review 25: 412-421.
26. Pitler H (2016) My problems with letter grades in school.
27. Hoover E (2012) High school class rank, a slippery metric, loses its appeal for college. The Chronicle of Higher Educatio 59: A1-A5.
28. Tinney J (2014). What do letter grades actually mean?
29. Davis BG (1993) Tools for teaching. Jossey-Bass,San Francisco.
30. McKeachie JW (1999) Teaching tips: Strategies, research and theory for college and university teachers.In: 10.Houghton Mifflin,Boston.
31. Esty WW, Teppo AR (1992) Grade assignment based on progressive improvement. The Mathematics Teacher 85: 616-618.
32. Everett M (2013) A conundrum: Grading for improvement versus grading against a standard.
33. Guskey TR (2011) Five obstacles to grading reform. Educational Leadership 69: 16-21.
34. Andersson A (1998) The dimensionality of the leaving certificate. Scandinavian Journal of Educational Research 42: 25-40.
35. Glaser R (1963) Instructional technology and the measurement of learning outcomes. American Psychologist 18: 519-522.
36. Frisbie DA, Waltman KK (1992) Developing a personal grading plan. Educational Measurement: Issues and Practice 11: 35-42.

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

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

39. Starch D, Elliott EC (1912) Reliability of grading of high school work in English. School Review 20: 442-457.

40. Starch D, Elliott EC (1913) Reliability of the grading of high school work in mathematics. School Review 21: 254-259.

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

42. Tieje RE, Sutcliffe EG, Hillebrand HN, Buchen W (1915) Systematizing grading in freshman composition at the large university. English Journal 4: 586-597.

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

44. Brimi HM (2011) Reliability of grading high school work in English.

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

46. Wiggins G (2013) Intelligent vs. thoughtless use of rubrics and models, Part 1.

47. Ruediger WC, Henning GN, Wilbur WA (1914) Standardization of courses and grades. Science 40: 642-643.

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

49. Etaugh AF, Etaugh CF, Hurd DE (1972) Reliability of college grades and grade point averages: some implications for the prediction of academic performance. Educational and Psychological Measurement 32: 1045-1050.

50. Bacon DR, Bean B (2006) GPA in research studies: An invaluable but overlooked opportunity. Journal of Marketing Education 28: 35-42.

51. Saupe JL, Eimers MT (2012) 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.

52. Willingham WW, Pollack JM, Lewis C (2000) Grades and test scores: Accounting for observed differences. ETS Research Report Series 2000: 1-177.

53. Dempsey CH (1912) Flexible grading and promotions. Journal of Education 75: 373-376.

54. Hopkins KD, George CA, Williams DD (1985) The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. Journal of Educational Measurement 22: 177-182.

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

56. Simmons RG, Brown L, Bush DM, Blyth DA (1978) Self-esteem and achievement of black and white adolescents. Social Problems 26: 86-96.

57. Durm MW (1993) An A is not an A is not an A: A history of grading. The Educational Forum 57: 294-297.

58. Rojstaczer S, Healy C (2010) Grading in American colleges and universities: 1-6.

59. Slavov S (2013) How to fix college grade inflation.

60. McCandless BR, 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: 153-159.

61. 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: 27-32.

62. Pedulla JJ, Airasian PW, Madauss GF (1980) Do teacher ratings and standardized test information yield the same information. American Educational Research Journal 17: 303-307.

63. Lekholm AK,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: 181-199.

64. Boykin AS (2010) The relationship between high school course grades and exam scores, E & R Report No. 09-39.

65. Camara WJ, Echternacht G (2000) The SAT (R) and high school grades: Utility in predicting success in college. College Entrance Examination Board New York: 1-15.

66. Green SK, Johnson RL, Kim DH, Pope NS (2007) Ethics in classroom assessment practices: Issues and attitudes. Teaching and Teacher Education 23: 999-1011.

67. Zahner D, Ramsaran LM, Steedle JT (2014) Comparing alternatives in the prediction of college success. Council for Aid to Education, New York: 1-18.

68. Ramist L, Lewis C, McCamley-Jenkins (1994) Student group differences in predicting college grades: Sex, language, and ethnic groups: 1-45.

69. Geiser S, Santelices M V (2007) Validity of high-school grades in predicting student success beyond the freshman year. Center for Studies in Higher Education, University of California, Berkeley: 1-35.

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

71. 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: 1-24.

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

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

74. Bauerlein M (2013) Boredom in class.

75. Robelen E (2011) Most teachers see the curriculum narrowing, Survey finds.

76. Banfield SR, Richmond VP, McCroskey JC (2006)The effect of teacher misbehaviors on teacher credibility and affect for the teacher. Communication Education 55: 63-72.

77. McFarland DA, Moody J, Diehl D, Smith JA, Thomas RJ(2014)Network ecology and adolescent social structure. American Sociology Review 79: 1088-1121.

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

79. Bacon LC (2011) Academic self-concept and academic achievement of African-American students transitioning from urban to rural schools.

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

81. Anderson LW (1976) Should students fail? Education Report 19: 1-4.

82. Marzolf SS (1955) Mental hygiene aspects of school marks. The Yearbook of the National Council on Measurements Used in Education 12: 10-12.

83. Johnson DH, Johnson RT (2002) Meaningful assessment: A manageable and cooperative process. Pearson, New York.

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

85. Hale DS, Goodson K, Savell JW (2013) USDA beef quality and yield grades. A & M Department of Animal Science. College Station, TX.

86. Partnership for the 21st Century (2016) Building your roadmap for 21st Century learning environments.

87. Chicago Tribune (2003) Should they get an A for effort?

88. Crowley B (2015) Grading: A duct-taped system in need of an overhaul?

89. Sadler DR(2009) (a) Grade integrity and the representation of academic achievement. Studies in Higher Education 34: 807-826.

90. Sadler DR (2009) (b) Conceptualizing and setting parameters for grade integrity. Griffith Institute for Higher Education, Griffith University, Brisbane, Australia:1-17.

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

92. Munk DD, Bursuck WD (2003) Grading students with disabilities. Educational Leadership 61: 38-43.

93. Iamarino DL (2014) The benefits of standards-based grading: A critical evaluation of modern grading practices. Current Issues of Education 17: 1-12.

94. Kreider H, Caspe M (2002) Defining fine-Communicating academic progress to parents. Cambridge, MA: Harvard Family Research Project.

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

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

97. Judd CH (1910) On the comparison of grading systems in high schools and colleges.The School Review, 18: 460-470.