POLICY PAPER

Recommendations from the National Panel
on the Future of Assessment Practices

Grading in a Comprehensive and Balanced Assessment System

LEARNING® SCIENCES INTERNATIONAL

Table of Contents

Introduction

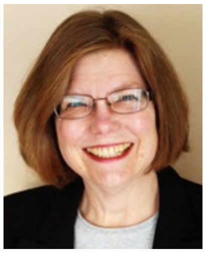
ABOUT THE PANEL

During the 3rd annual Formative Assessment National Conference, **Dylan Wiliam**, **Susan Brookhart**, **Tom Guskey**, and **Jay McTighe** discussed a critical indicator of student learning: grading. Grading should be the bridge between learning first explored in classroom formative assessment and the large-scale assessments that follow. However, much of grading is based on tradition, not evidence. Because of its dependence on teacher judgment, grading has not been well integrated into district assessment for either accountability or teaching and learning. And yet grade-point averages are valued indicators of accomplishment, and used to single out students for honors, awards, and admission to prestigious colleges.

The panel debated important questions about grading: What is the purpose of a grading system? What is the current state of grading practices and policies? Should grading be improved/reformed or replaced with something different? What should the grading system of the future look like? How should grading function within an overall district assessment system?

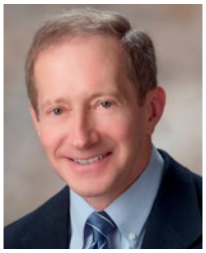
Wiliam, Brookhart, Guskey, and McTighe suggest eight general principals to improve grading practices that should, if followed, go a long way towards transforming it into a functioning component of a balanced and comprehensive assessment system.

Author Bio



SUSAN BROOKHART

Dr. Susan M. Brookhart is Professor Emerita at Duquesne University and an expert consultant with an extensive background working with schools, districts, universities, and states. She studies the role of both formative and summative classroom assessment in student motivation and achievement, the connection between classroom assessment and large-scale assessment, and grading. She is author or co-author of 18 books and more than 70 articles and book chapters and has served as editor for academic journals.



TOM GUSKEY

Dr. Thomas Guskey has served on the National Commission on Teaching & America's Future, the Task Force to develop National Standards for Staff Development. He was named a Fellow in the American Educational Research Association and was also awarded the Association's prestigious Relating Research to Practice Award. He is also the author/editor of 25 award-winning books and more than 250 book chapters and articles. His most recent books include *What We Know About Grading and On Your Mark: Challenging the Conventions of Grading and Reporting*.

JAY MCTIGHE

Jay McTighe brings a wealth of assessment experience from leading classroom formative performance assessment efforts with the Maryland Assessment Consortium, from his work on large-scale performance assessments with the Maryland State Department of Education, and from his many other projects at state and district levels. He is co-author of 15 books, including the award-winning and best-selling *Understanding by Design*® series with Grant Wiggins, and has written more than 40 articles and book chapters.

DYLAN WILIAM

Dr. Dylan Wiliam has helped to successfully implement classroom formative assessment in thousands of schools all over the world. A BBC documentary tracking his work showed how his formative assessment strategies empower students, significantly increase engagement, and shift classroom responsibility from teachers to students. He has written over 300 books, chapters, and articles; his latest book breaks down the gaps between what research tells us works and what we actually do in schools.

Executive Summary

Grading should be a part of a comprehensive and balanced assessment system. Grading sits in the middle of the system and should provide coherent but not redundant information to other system components.

Grades have long been criticized as biased and/or unreliable, and studies have shown teachers vary widely in what they count in grades and how they calculate them (Brookhart et al., 2016). However, when grades are aggregated into GPAs, they are very reliable (Beatty et al., 2015), and predictive of future educational consequences.

In the US, tests and grades convey different information, with grading practices providing a strong measure of learning success in school that does not simply mimic tested achievement (Bowers, 2009). The trick is to maintain and improve the useful information in grades while working to increase reliability, producing a grading system that communicates more clearly.

Grading practices are at least partly rooted in teachers' beliefs about learning, assessment, and the purposes of grading and schooling. "Replacing" grading practices does not replace the beliefs and prior understandings of the teachers. Perhaps the best course of action, then, is to endorse improving or reforming grading practices, with all deliberate speed but with the understanding that progress will be faster some places than others.

Accordingly, the authors offer eight general principles for improving grading practices.

- **Purpose** - All stakeholders, including students and parents, should understand what grades are intended to communicate.
- Goals - Clear learning goals should specify what students should know and be able to do with that knowledge.
- **Learning Continua** – Grades should communicate a student's proficiency level along a continuum based on a collection of evidence rather than judging a student against an aged-based, grade-level standard at a point in time.
- **Criteria** – To ensure greater consistency among teacher judgments of student performance, grades should be based on established sets of evaluative criteria and associated scoring tools (e.g. rubrics) aligned with key standards.
- **Opportunity to Learn** - Students must have had appropriate practice and feedback in advance of graded work.
- **Multiple Measures** - Reporting should be based on multiple measures reflecting different aspects or dimensions of learning.
- **Scale** – The grading scale used should be clearly and frequently communicated to stakeholders.
- **Evidence** - Grades should be based on a collection of evidence assembled over time.

A strong relationship should exist between instruction and learning, classroom formative assessment, and grades. This relationship should be based on the learning outcomes and criteria that should be common to all three. Building this relationship, or clarifying it where it exists only in part, should be the first step in improving the functioning of grades in an overall district assessment system.

Grading in a Comprehensive and Balanced Assessment System

What is the place of grading in a comprehensive and balanced district assessment system? A previous white paper (Brookhart, McTighe, Stiggins, & Wiliam, 2019) shows how grading is one of the components of district assessment of student learning. That document described five components of a comprehensive and balanced assessment system. From assessments collected nearest to learning outward, they are: short-cycle classroom formative assessment, medium-cycle formative assessment, classroom summative assessment (grading), long-cycle formative assessments, district-level summative assessments, and annual state accountability assessments. Grading sits in the middle of the system and is closer to the learning than the large-scale assessments that come later. As such, grading needs to bridge between the learning first explored in classroom formative assessment and the learning measured later with large-scale summative assessment, providing coherent but not redundant information to those other components.

Historically, the treatment of grading has been somewhat two-faceted. On the one hand, grading has been separated from other indicators of student learning, treated differently because of its dependence on teacher judgment (Brookhart, 2013a), and not well integrated into district assessment for either accountability or teaching and learning. On the other hand, grade-point averages are valued indicators of accomplishment, and students with high GPAs are singled out for honors and awards and admission to prestigious colleges.

This white paper addresses the place of grading in a comprehensive and balanced assessment system by building on a series of questions. What is the purpose of a grading system? What is the current state of grading practices and policies? Should grading be improved/reformed or replaced with something different? What should the grading system of the future look like? How should grading function within an overall district assessment system?

WHAT IS THE PURPOSE OF A GRADING SYSTEM?

Grades are symbols (numbered categories, letters, or descriptors) assigned to student work or aggregated into composite measures for reporting. The primary goal of grading and reporting is communication (Guskey & Bailey, 2010), providing information to interested stakeholders in such a way that they can understand it and use it. Most authors would agree that the main information communicated by grades should be students' current achievement status on intended learning outcomes (Guskey & Bailey, 2010; O'Connor, 2018), although a few would argue that the main information should be student growth or student performance relative to peers.

Some argue that grades should serve to motivate students and use that argument to support practices like taking off points for late work. This kind of purpose for grading is, at best, an extrinsic motivator, best for motivating behavior and not learning. The students for whom this would work best are students who already put forth effort in their schoolwork. We understand the need to teach students work habits, but engineering points in a grade is not the best way to do that (Guskey, 2011) because it results in a grade where the measure of learning is mixed with a measure of behavior. Many teachers will not give up the "carrot and stick" aspect of grading until they have some other strategies in their repertoire. If grading reforms are to succeed, they must be accompanied by reforms in strategies to support classroom management and student work habits.

Many teachers will not give up the "carrot and stick" aspect of grading until they have some other strategies in their repertoire. If grading reforms are to succeed, they must be accompanied by reforms in strategies to support classroom management and student work habits.

The achievement reflected in grades is typically school-based achievement, that is, the construct is different from the kind of general achievement reflected in standardized test results. Grades have long been criticized as biased and/or unreliable, and studies have shown teachers vary widely in what they count in grades and how they calculate them (Brookhart et al., 2016). However, the difference between graded and tested achievement is more than can be explained by unreliability. When grades are aggregated into GPAs, they are very reliable (Beatty et al., 2015), and they predict future educational consequences like dropping out of school (Bowers, Sprott, & Taff, 2013), and college admission and success (Atkinson & Geiser, 2009; Bowers 2010; Westrick et al., 2015).

Bowers (2009) has termed the portion of grades not accounted for by standardized tests of achievement as a "success at school factor" (p. 623), which may be related to differences in both learning and assessment contexts between school and standardized testing. Westrick and colleagues (2015) showed that high school grades were a slightly better predictor of first- and second-year college grades than ACT scores, although both were valuable predictors, also supporting the notion that high school grades include some sort of success at school factor.

Galla and colleagues (2019) investigated the relationship of both tested and graded achievement to on-time graduation from college. They showed that the incremental predictive validity of high school grades was associated with student self-regulation, while the incremental predictive validity of test scores was associated with cognitive ability. This study is important because it adds a theoretical basis (student self-regulation) to the previous practical construct of success at school. Tested achievement and graded achievement are related, and despite the very legitimate complaints lodged against grading, grading offers information that is not measured elsewhere. To find out why that is, we turn to a description of current grading practices and policies.

Tested achievement and graded achievement are related, and despite the very legitimate complaints lodged against grading, grading offers information that is not measured elsewhere.

WHAT IS THE CURRENT STATE OF GRADING PRACTICES AND POLICIES?

Historically, grading practices have been notably variable (Brookhart et al., 2016; Guskey & Brookhart, 2019). While most of the variation in grades reflects differences in achievement (Brookhart et al., 2016), some teachers put some weight on "academic enablers" like effort and participation (McMillan, 2001; McMillan, Myran, & Workman, 2002), while others may include contributing factors such as behavior and attendance. This tendency has usually been explained as the teacher trying to mitigate or engineer appropriate social consequences for hard-working students or use grading as a means of control. Recently, Bonner and Chen (Bonner & Chen, 2009; Chen & Bonner, 2017), have also shown that the use of academic enablers in grading is also partly related to teachers' views of learning (constructivist vs. behaviorist) and to the type of learning outcome (knowledge vs. skill). However, grading qualities like effort, behavior, and participation can easily be biased and has not been shown to actually improve motivation (Bonner & Chen, 2019).

There is also evidence that teachers weigh effort, behavior, and participation more heavily for low-achieving students, such that with high effort and excellent behavior, a student with low achievement receives at least a passing grade (Randall & Engelhard, 2010). The teacher's intention is usually to achieve some feeling of equity. However, over-emphasizing effort and behavior for lower-achieving students may work against the equity the teacher may be trying to achieve for her students. For example, giving bonus points to low-achieving students for things unrelated to status on learning goals leads to grades that convey a false message about a student's actual achievement. Thus, a student who received a C in a math class may be moved to the next level of math class when in fact he does not have the prior knowledge to benefit from that class and in the long run will fall further behind. A U.S. Department of Education study (1994) based on a nationally-representative sample of students found that students in high poverty schools (more than 75% free or reduced price lunch) who received mostly A's in English got about the same reading score as C and D students in affluent schools. In the same study, students receiving A's in

A U.S. Department of Education study (1994) based on a nationally-representative sample of students found that students in high poverty schools (more than 75% free or reduced price lunch) who received mostly A's in English got about the same reading score as C and D students in affluent schools.

mathematics from high-poverty schools most closely resembled D students in affluent schools. However, because low-SES students may have lower opportunity to learn, these grades may still be good predictors of college success.

U.S. grading practices that treat students of differing abilities differently also result in less of a relationship between intelligence test scores and grades than if grading practices were based on test scores alone. Historically, the correlation between IQ and grades in the U.S. is about .40 (Bowers, 2019). However, in the United Kingdom, where secondary school grades come from the General Certificate of Secondary Education exams, the correlation between IQ and grades is about .81 (Deary, Strand, Smith, & Fernandes, 2007), very similar to the correlation in the U.S. between IQ and the ACT, another test of achievement (.77, Koenig, Frey, & Detterman, 2008). The fact that in the US tests and grades are more different supports a strong view of classroom learning and also supports the importance of a success in school measure (grades) that does not simply mimic tested achievement (Bowers, 2009).

Taken together, then, the research evidence suggests grades report school-based achievement that includes: partly general achievement of the sort also measured by standardized tests; partly contextual achievement or "success in school" (because the assessments are linked to classroom assignments that draw some of their meaning from the instructional setting, and because learning enablers count); plus perhaps more unreliability than one might like at the level of individual grades. The reliability issue improves when grades are aggregated into GPAs, but it is worth pointing out that the individual grades on assignments are the grades most important for students to understand their learning. The trick will be to maintain and improve the useful information in grades while working to increase the reliability, producing a grading system that communicates more clearly.

Much of grading is based on tradition, not evidence, and there is still much room for improvement.

In short, both research and the experience of the authors suggest that much of grading is based on tradition, not evidence, and that there is still much room for improvement. While some teachers carefully connect each grade with explicitly stated learning outcomes and agreed-upon criteria, others rely on assessments that do not appropriately match the intended learning outcomes and apply idiosyncratic grading methods. One of the authors (JM), for example, remembers an elementary school principal who reported an interesting meeting with the parents of triplets. Each was in a different fourth grade class. One teacher graded English Language Arts by averaging spelling tests and worksheet results, while the other two used running records and a whole-language approach. And in social studies, one teacher used project-based learning and the other two based grades primarily on tests of facts. The parents were understandably confused about whether their triplets were all learning the same fourth grade curriculum and how they were performing.

This highlights an assessment system problem. The district has a detailed, standards-based curriculum, available to parents on the district website. If it also had a comprehensive assessment system, of which grading were a component, the assessment system would be based on a consensus about the meaning of the learning goals, the assessment evidence used to determine the extent to which students have achieved them, and how student learning will be represented through grades. The fact that three students of comparable abilities received different grades based on very different assessments, which in turn were based on differing teachers' interpretations of the curriculum standards and their own preferred pedagogical and assessment approaches, reveals the inconsistency.

SHOULD GRADING BE IMPROVED/REFORMED OR REPLACED WITH SOMETHING DIFFERENT?

This is actually a more complicated question than it seems. One would think that a district could adopt a brand-new grading program, much like they adopt other new programs, and simply change things – but it turns out that's not as straightforward as it sounds. Grading practices are at least partly rooted in teachers' beliefs about learning, assessment, and the purposes of grading and schooling. "Replacing" grading practices does not replace the beliefs and prior understandings (and, in some cases, emotional responses) of the teachers who are asked to change their practices. As one of the authors (DW) noted, "We can't just tell them they're doing it wrong. We need evidence." The studies reviewed above have shown that current grading practices are often idiosyncratic.

"Replacing" grading practices does not replace the beliefs and prior understandings (and, in some cases, emotional responses) of the teachers who are asked to change their practices. As one of the authors (DW) noted, "We can't just tell them they're doing it wrong. We need evidence."

There is some (e.g., Pollio & Hochbein, 2015), but not yet strong and compelling, evidence that standards-based grading or other suggested reforms may be better than conventional grading practices. Standards-based grading principles call for reporting academic achievement separately from behavior and non-cognitive factors; basing both assessment and grading on standards; prioritizing the most recent evidence of learning for report card grades; allowing students to revise and resubmit work; using proficiency-based rubrics and decision rules for aggregating them that take their ordinal nature into account; and using quality formative (ungraded) assessments as well as summative (graded) assessments (Guskey & Bailey, 2010; Peters et al,, 2017).

If practiced as described, these standards-based reforms would improve the function of grading in a comprehensive and balanced assessment system. However, even if replacing traditional grading practices with standards-based grading practices were the strategy selected, it is not likely that replacement would actually be a complete, "new leaf" replacement of grading policies and practices. Olsen and Buchanan (2019) looked at the changes in understanding and grading practices of teachers involved in year-long professional development designed to reform grading practices in two secondary schools. Change did occur, but it was partial and did not necessarily have intended effects. Based on their own individual grading histories, teachers expressed confusion and tensions around whether grades should reflect academic achievement only or effort and behavior as well. More to the point of this discussion, teachers sometimes adapted recommended strategies and then refined them to fit their classroom context, in the process using a belief system that the grading reforms meant to change. For example, a teacher in that study who disagreed that grades should be based on achievement (which she called "content") argued that it didn't value the different ways people learn; then she justified giving a C to a student who had learned to persevere and behave appropriately but had not mastered the content because "he held onto a C's worth of the parts of this course that I care about, but it wasn't based on all the tests" (p. 22). In other words, she used the language of the reform to justify not reforming. Similarly, Townsley, Buckmiller, and Cooper (2019) described principals' perceptions of teacher resistance to grading reform as an obstacle to moving toward standards-based grading.

Perhaps the best course of action, then, is to endorse improving or reforming grading practices, with all deliberate speed but with the understanding that progress will be faster some places than others, except in unusual circumstances (for example, when a new alternative school is founded and can build its assessment system from scratch). The goal, then, is for districts to move as quickly as they can to

> The goal is for districts to move as quickly as they can to implement grading reforms that would at the least result in grades that communicated clearer information about student standing on intended learning outcomes.

implement grading reforms that would at the least result in grades that communicated clearer information about student standing on intended learning outcomes and at best embody the suggestions below and even other improvements discovered as the work progressed.

WHAT SHOULD THE GRADING SYSTEM OF THE FUTURE LOOK LIKE?

Suggestions for improving a grading system can be derived from the research discussed above, the authors' work and experiences, and other published summaries and recommendations (Guskey & Brookhart, 2019; McTighe & Curtis, 2019; McTighe & Willis, 2019). Accordingly, we offer the following general principles for improving grading practices.

- Agreement among stakeholders (educators, parents, and students) about the purpose of grades and the meaning they are intended to communicate

- Clear learning goals specifying both what students should know (content) and be able to do with that knowledge (cognitive processes and skills), further clarified as the general goals described in standards are instantiated in unit goals and instructional objectives and daily learning targets for students

- The use where appropriate of learning continua linked to grade-level standards to support accurate reporting for all students, not just those on grade level

- Clear criteria and models of good work to help students understand what it is they are trying to learn and how they will know they are making progress

- Appropriate practice and feedback in advance of graded work to allow students to learn before they are evaluated

- Separate reporting of different kinds of achievement and performance: Product (current status on intended learning outcomes, indicated by quality of work on well-designed assessments); Process (learning enablers like practice work on ungraded assessments, homework, class work, collaboration, responsibility, self-regulation, effort, and so on); and Progress (amount of gain or growth) (Guskey, 2015; Guskey & Bailey, 2010)

- The use of a grading scale with shared meaning among stakeholders, and clear communication about that scale

- Basing grades on a collection of evidence assembled over time and aggregated using appropriate methods (whether judgments or calculations) that result in the most accurate estimate of students' current achievement status, progress, or learning process, depending on what is being measured and reported

GENERAL PRINCIPLES FOR IMPROVING GRADING PRACTICES

Purpose – All stakeholders, including students and parents, should understand what grades are intended to communicate.

Goals – Clear learning goals should specify what students should know and be able to do with that knowledge.

Learning Continua – Grades should communicate a student's proficiency level along a continuum based on a collection of evidence rather than judging a student against an aged-based, grade-level standard at a point in time.

Criteria – To ensure greater consistency among teacher judgments of student performance, grades should be based on established sets of evaluative criteria and associated scoring tools (e.g. rubrics) aligned with key standards.

Opportunity to Learn – Students must have had appropriate practice and feedback in advance of graded work.

Multiple Measures – Reporting should be based on multiple measures reflecting different aspects or dimensions of learning.

Scale – The grading scale used should be clearly and frequently communicated to stakeholders.

Evidence – Grades should be based on a collection of evidence assembled over time.

It is worth noting that most of the grading software programs available to educators today are not based on our knowledge of better practice but rather on tradition and most common practice. As such they pose significant obstacles to educators' efforts to institute reforms. That obstacle notwithstanding, we discuss each of these general principles in more detail below.

PURPOSE.

All stakeholders, including students and parents, should understand what grades are intended to communicate. The purpose should be more than a vague notion of rating students' accomplishment in school. Educators, parents, and students should know the specific purpose of their grading and reporting system: what it will communicate, what it does not communicate, and what additional information is available. Setting the purpose of a district grading system is ultimately the responsibility of the school board and, usually by delegation, the central office administrators. However, a wise district will engage teachers, parents, and students in explicit conversations about grading and use their input to make sure the grading system provides the kind of information everyone will use and that multiple purposes, when they exist, are understood.

There is no perfect grading system. All systems involve some trade-offs involving specificity, recency, and precision of information. For example, some standards-based grading systems report only on selected "power standards," for the sake of having a concise and actionable report card. When this is the case, stakeholders should know it, and they should know how and where they can get information about other standards if they desire. The authors have found that in most districts and schools, the primary purpose of grading is to communicate students' current status on the learning outcomes in curriculum and standards. There are often secondary purposes, and even secondary variables reported on the report card (for example, a learning skills scale), but absent a particular reason to do otherwise, information about current achievement of current learning goals typically provides the most actionable information.

GOALS.

Clear learning goals specify what students should know and be able to do with that knowledge. Clear learning goals unite curriculum, instruction, and assessment (McTighe & Willis, 2019) and are the basis of a sound grading system.

Goal clarity by teachers enables them to provide appropriate instruction and use assessments that enable valid inferences about student learning. Teachers can then help their students understand the learning goals, e.g., by using "I CAN…." statements, previewing assessments and co-developing "success criteria." When students are clear about goals, they will be better able to regulate their own learning, e.g., setting a goal and working toward it, monitoring their understandings, and adjusting their work as they go.

Assessments of student performance, and the associated grades that result, should be closely aligned to targeted goals. As an example, the Next Generation Science Standards (NGSS) (www.nextgenscience.org/) present three inter-related learning goals: disciplinary core ideas, crosscutting concepts, and science and engineering practices. Together, the essential structure of the NGSS supports the intention that students should be involved in doing science, i.e., applying science content using the practices. Performance expectations, associated assessments and resulting grades should be based on all three goals. Thus, if a science teacher working in an NGSS district based students' grades exclusively on tests of factual knowledge, her grading system would be out-of-sync with the intentions of the NGSS.

Educators have asked us what we think about grading social-emotional learning (SEL). SEL goals should not be evaluated and graded in a traditional (e.g., ABCDF) manner, because SEL goals are hard to measure and most measures are easily gamed (Duckworth & Yeager, 2015). However, both teachers and students can collect evidence of SEL-type goals, and students can reflect, self-assess, and set personal goals for them. Teachers can give feedback on SEL goals, as well, noting progress and giving suggestions for improvement.

LEARNING CONTINUA.

One of the challenges of basing grades on grade-level standards is that these may not be appropriate for everyone. For example, students having an Individualized Education Plan (IEP) are typically working toward modified standards and would be assessed and graded accordingly. Moreover, they would need a modified report card to properly communicate their achievement levels (Jung & Guskey, 2011).

When considering standards-based grading, it is important to recognize that grade-level standards typically specify learning goals and performance expectations that are deemed appropriate for the "average" student in that grade. However, we know that learners don't come to school at identical readiness levels or learn and progress at identical rates. This reality could be addressed by using learning (or proficiency) continua as the reporting frame for some areas of the curriculum. Proficiency continua already exist in English language arts and writing (e.g., Bonnie Campbell-Hill http://www.bonniecampbellhill.com/Handouts/English/Reading%20&%20Writing%20Continuums%20-%20Color.pdf) and world languages (e.g., ACTFL proficiency guidelines https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012).

The authors can envision a report card framework that would communicate a student's proficiency level (e.g., on narrative writing or proportional reasoning) along a continuum based on a collection of evidence. Rather than judging (and grading) a student against an aged-based, grade-level standard at a point in time, such a report could communicate both where a student is now, as well as how she had progressed over time, irrespective of their age or grade level. Such a grading and reporting system based on proficiency continua aligns well with a competency- or mastery-based educational approach. We already do this for Karate (via colored belts) and swimming (using the Red Cross levels). Of course, such a framework can also produce normative information, and parents may want additional information about what is considered normal for a grade level. If that is the case a range of proficiency levels, rather than a single one, could communicate this information.

Before this strategy could be employed, standards, curriculum, and instruction would have to be organized around the same continua. This strategy is not, however, outside the realm of possibility for the future and could solve the current problems of grade-referencing and modifying. The authors would love to see more work done on continua and then experimentation with communication according to where a student is, not what he did or didn't "get," at the end of a report period.

CRITERIA.

Learning is a relatively permanent change in understanding and skill (Soderstrom & Bjork, 2015). The work students do, their performances, can be an unreliable index of whether such long-term learning has taken place. However, the work students do is what one can observe, and what one can hold students accountable for doing. An ideal grading system should regularly assess students on things they were taught days or weeks ago, to see if what was taught was learned. Because we can only judge learning through performance, performance some time after instruction is evidence of learning; performance straight after the end of the instruction is not. Another important aspect of measuring learning is to apply criteria to student work that are most likely to indicate the underlying learning, as opposed to surface features of the work unrelated to the learning, or indicators of following directions.

Criteria are guidelines, rules, or principles by which student learning and performance are judged. Judgment-based evaluation can be made more reliable when it is based on clear and appropriate criteria. Since grading typically involves teacher judgment, having established criteria, aligned to targeted learning goals, is critical to a reliable grading system. Ideally, districts would establish sets of evaluative criteria and associated scoring tools (e.g. rubrics) aligned with key standards. Having such well-developed evaluation tools would make it more likely that teacher judgments of student performance, and the concomitant grades they assign, will be more consistent with the judgments of other teachers. In the absence of well-developed and agreed-upon criteria, some teachers may focus on "surface-level" features of student work (e.g., neatness, number of words) rather than the essential qualities that reveal student understanding and skill (Brookhart, 2013b), thus rendering their grades misaligned to standards and less consistent with their colleagues.

Criteria can be communicated to students in many ways, including as lesson-by-lesson "look-fors" or success criteria, in rubrics and other evaluative tools, in scripts and other thinking guides (Panadero, Tapia, & Huertas, 2012), in examples of work, or in some combination of these. It is the tasks and criteria that operationalize the learning goal for students and teachers, as well. Criteria contribute to the clarity of the goals, for both students and teacher. Interestingly, it seems to be much more difficult for teachers to design and communicate clear criteria to students than it is to write a learning goal statement. Recommending the use of clear, learning-focused criteria in grading is perhaps setting a higher bar than it may sound like to the readers. This is very difficult to do well.

FAIRNESS AND OPPORTUNITY TO LEARN.

In order for grades to report what students have learned after they have been given a fair opportunity to learn it, students must have had appropriate practice and feedback in advance of graded work. That is, learning precedes reporting what has been learned. The feedback that students receive on ongoing work should be based on the same criteria as will be used for grading. As above, those criteria should be learning-focused rather than about following directions or about surface features of the assignments. Effective feedback fuels the formative learning cycle, guides student self-regulation of learning, and helps students connect the practice and learning work they do with the grades they receive. As such, it also supports a view of learning as something students control.

MULTIPLE MEASURES.

Reporting should be based on multiple measures reflecting different aspects or dimensions of learning. These dimensions are often categorized as Product, Progress, and Process (Guskey, 2015; Guskey & Bailey, 2010). Product measures report students' current status on intended learning outcomes, indicated by quality of work on well-designed assessments, performances, or demonstrations. These are the grades this report recommends as primary, and perhaps the only marks that should be called "grades" on a report card. Product grades should summarize a student's current status on learning goals, be based on well-designed assessments, be evaluated with learning-focused criteria, and be accurately summarized using decision rules or computations that maintain intended meaning when combining component grades into the report card grade.

Progress measures report amount of gain or growth, usually understood as change from one time point to another, say the beginning of the report period to the end. Parents and students ultimately do want progress information, but the authors do not recommend progress indicators be the main grades on a report card, for three reasons. First, the amount of progress depends on the starting point. The higher students' achievement was to begin with on measures of the specified learning goals for the course, the less growth they will be able to show. Second, some standards (at least, as standards are currently written) have higher ceilings than others. Some standards are mostly about comprehension of facts and concepts in one area, while others require original thinking and making connections. So in some standards, there is much less potential progress to make. Third, progress is difficult to measure with grades because appropriately equivalent, comparable scores often do not exist in classroom measures (the "apples to oranges" situation). For this reason, progress is more accurately indicated with tested, not graded, achievement, and with longer time intervals than typical report card periods (e.g., nine weeks). Other components of the district's assessment system are better suited to communicate progress than report card grades.

Process measures report learning enabling behaviors like turning in homework and class work, collaboration, effort, and so on. They also may report important social and emotional learning goals such as empathy, resilience, responsibility, habits of mind, and the like. Process measures often are part of report cards and are often reported in a Learning Skills or Citizenship section on the report card. Because learning enabling skills are mostly behaviors (albeit learning-related ones), many of them can be well assessed with a frequency scale (e.g., "Completes homework: Usually, Often, Sometimes, Rarely"). Such skills, whether called Learning Skills or by some other term, communicate important information that is different from the product grades. Select skills to report that have general application across learning in the disciplines for which product grades are reported.

We also would like to suggest the use of a quality dimension with process indicators. For example, there could be a way to distinguish the student who completes the homework assignment but does it incorrectly versus the student who completes only half but it is all done well, or, for example, a way to distinguish the student who participates every day in class discussions but adds little to the conversation versus the student who participates rarely but shows deep thinking about the topic.

Report card grades should not be the only communication vehicle in a comprehensive and balanced assessment system. Each component of a balanced assessment system (Brookhart, McTighe, Stiggins, & Wiliam, 2019) can and should be communicated to each of its primary information users.

SCALE.

In the context of educational assessment and grading, scale refers to the number and type of points or gradients used to judge and report on student performance. Familiar scales used on schools include 5 points (A, B, C, D, F), 4-point rubrics, and 1–100 percentage points.

Every grading scale requires trade-offs. Shorter scales with fewer categories of performance are generally recommended over longer scales with more categories, like the percentage scale, because it is easier for graders to make reliable judgments using fewer categories (Guskey & Brookhart, 2019). As the number of categories goes up, inter-rater reliability goes down. The trade-off there is that with fewer categories, it is not possible to report nuanced differences in performance. As Wiliam (2000, p. 109) observed, "If scores and percentages are prone to spurious precision, then grades are prone to spurious accuracy.

Another appealing feature of shorter scales (e.g., ABCDF, Advanced/Proficient/Approaching/Beginning) is that they solve the "zero" problem – or more generally, the problem with any low score disproportionately pulling down the mean when scores are averaged – noted for the 0-100 scale. One can also solve the zero problem with a percentage scale, but it requires truncating the scale (e.g., the lowest score that can be awarded is a 59).

Further, it is not necessary to use the same scale for reporting the results of individual pieces of work, projects, tests, and so on as the one used for reporting achievement at the end of a marking period. While shorter scales (ABCDF or proficiency scales) are recommended for final grades, individual grades could differ during a marking period. For example, one could report percentages on tests, proficiency ratings or progress along a learning progression for performance assessments, and so on, and then, at the end of the marking period, these would be aggregated to provide an overall grade for the whole marking period.

Whatever grading scale is used should be clearly communicated to stakeholders, over and over – it takes longer to create shared meaning than one might think. Traditional grading scales like ABCDF and the percentage scale have the disadvantage that people think they know what they mean, even though they often do not. For example, a grade of C is generally considered "average" when, in fact, the average grade given in the U.S. is a B (OERI, 1994), or perhaps even higher by now. Most district grading policies recommend the use of criterion-referenced grading methods; if in a given class all students do very well, they could all get As, which would mean A was average. The concept of "average" does not fit with the assessment and grading practices we are recommending – and, indeed, which are already advocated in most districts. However, old habits die hard, especially when they become entrenched in popular culture (for example, using ABCDF to rate restaurants or movies). It is very difficult for stakeholders to "unlearn" things they think they know.

It would seem that substituting a newer scale, for example a proficiency scale (Advanced/Proficient/Approaching/Beginning), might avoid the problem of over-familiarity and give districts a chance to start from scratch communicating what each of the scale levels means. However, even in this case, there will be people who want to connect the new scale with their previous understanding of grades, for example thinking that "Advanced is like an A, Proficient is like a B," and so on. Some districts are reforming grading systems by using a newer scale for elementary school grades and allowing secondary school grades to be reported on the traditional ABCDF scale to facilitate the calculation of grade-point averages. This could be a useful compromise if the grades were assigned using the principles we are recommending in this section.

The authors have not found any magic bullet for developing shared meaning around a grading scale, but there are things that could and should be done. Easily readable material should be prepared for students, parents, and district staff, and circulated widely. Attention should be drawn to this material at regular intervals, stakeholders should have multiple opportunities to ask questions, and the scale should be concisely defined on report cards. Conversations about grading will benefit greatly by referencing illustrative samples of student work (also known as anchors) associated with the grading scale. The anchors serve to clarify the differences in the performance levels of the scale. The grading scale should be used as intended (so, for example, Advanced doesn't end up essentially meaning an A, and so on). For students who aspire to be Advanced, the performance level descriptions and illustrative work samples should clarify what that looks like, so that students who are capable of performing as described can aim for that kind of performance.

EVIDENCE.

Grades should be based on a collection of evidence assembled over time. As with all assessment, grading is an evidentiary process. The quality of the evidence makes a great deal of difference. Each piece of evidence, whether student work on an assignment or teacher observation of what a student does or says, should support valid conclusions about whatever it is being used to grade, for example a particular learning outcome or learning skill and should be interpreted accurately and without bias.

In other words, the rules of argument apply. Each piece of evidence should support the conclusion (the grade) reached by the grader. The days of ascribing extra points to an academic grade for returning field trip permission slips or being the classroom helper are—or at least should be—over.

HOW SHOULD GRADING FUNCTION WITHIN AN OVERALL DISTRICT ASSESSMENT SYSTEM?

Components of a comprehensive and balanced district assessment system include daily short-cycle formative assessment, medium-cycle formative assessment, grading, benchmark/ interim assessments, and district-level summative and accountability assessment. Historically, researchers and educators have looked for grading to be related to summative, tested achievement. There is some evidence that if certain effective grading practices are followed, grades will be moderately related to tested achievement (Welsh, D'Agostino, & Kaniskan, 2013; Willingham, Pollack, & Lewis, 2002; Pollio & Hochbein, 2015), although there is also ample evidence that such grading practices are not always followed. Grades will never be highly related to standardized achievement measures, for all the reasons described above. A moderate relationship should be the goal. Predicting tested, accountability achievement is not the only, or even the main, function grades should serve in a district assessment system. The main system function of grades – that is, classroom summative assessment – is to summarize student status on learning goals. The reported learning, in turn, has been informed by, in some senses created by, the short- and medium-cycle classroom formative assessment. In an overall assessment system, grades function as the certifying mechanism that show the outcome of all that formative assessment and the learning involved in it.

A strong relationship should exist between instruction and learning, classroom formative assessment, and grades. This relationship should be based on the learning outcomes and criteria that should be common to all three.

This means a strong relationship should exist between instruction and learning, classroom formative assessment, and grades. This relationship should be based on the learning outcomes and criteria that should be common to all three. Building this relationship, or clarifying it where it exists only in part, should be the first step in improving the functioning of grades in an overall district assessment system.

SUMMARY

Grading should be a part of a comprehensive and balanced assessment system. Grading should be based on clear learning outcomes/targets, appropriate assessments of those outcomes, and a reporting system that clearly communicates a summary of student achievement. That summary should be a synthesis of evidence reflecting students' current level of learning or accomplishment, not students' average of performance over time. Where students were at the beginning or halfway through a learning sequence doesn't matter. How many times they fell short during that sequence doesn't matter. What matters is what they have learned and are able to do currently or "at this time." In this paper, we have offered general principles that should operate within a grading system to accomplish this, regarding setting purpose, describing learning goals, using learning continua, establishing criteria, considering fairness and opportunity to learn, considering multiple measures, constructing an appropriate scale, and using evidence. Attention to these matters should go a long way towards transforming grading, which often is not well integrated into a district's assessment system, into a functioning component of a balanced and comprehensive assessment system.

References

References

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college admissions tests. *Educational Researcher, 38*, 665–676.

Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice*, *34*, 31-40.

Bonner, S. M., & Chen, P. P. (2009). Teacher candidates' perceptions about grading and constructivist teaching. *Educational Assessment*, *14*, 57-77.

Bonner, S. M., & Chen, P. P. (2019). The composition of grades: Cognitive and non-cognitive factors. In T. R. Guskey & S. M. Brookhart (Eds.) *What we know about grading: What works, what doesn't, and what's next* (pp. 57-83). Alexandria, VA: ASCD.

Bowers, A. J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration*, *47*, 609-629.

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment Research and Evaluation*, *15*(7), 1–18. Retrieved from http://pareonline.net/pdf/v15n7.pdf

Bowers, A. J. (2019). Report card grades and educational outcomes. In T. R. Guskey & S. M. Brookhart (Eds.) *What we know about grading: What works, what doesn't, and what's next* (pp. 32-56). Alexandria, VA: ASCD.

Bowers, A. J., Sprott, R., & Taff, S. (2013). Do we know who will drop out? A review of the predictors of dropping out of high school: Precision, sensitivity and specificity. *The High School Journal*, *96*, 77–100.

Brookhart, S. M. (2013a). The use of teacher judgment for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice*, *20*, 69-90.

Brookhart, S. M. (2013b). *How to create and use rubrics for formative assessment and grading.* Alexandria, VA: ASCD.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, *68*, 803-848.

Brookhart, S., McTighe, J., Stiggins, R., & Wiliam, D. (2019). *The future of assessment practices: Comprehensive and balanced assessment systems*. West Palm Beach, FL: Learning Sciences International.

Chen, P. P., & Bonner, S. M. (2017). Teachers' beliefs about grading practices and a constructivist approach to teaching. *Educational Assessment*, *22*, 18-34.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). *Intelligence and educational achievement*, *35*, 13-21.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*, 237-251.

Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., D'Mello, S. K., Finn, A. S., & Duckworth, A. L. (2019). Why high school grades are better predictors of on-time college graduation than are admissions test scores: The roles of self-regulation and cognitive ability. *American Educational Research Journal* online first.

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

Guskey, T. R. (2015). *On your mark: Challenging the conventions of grading and reporting*. Bloomington, IN: Solution Tree.

Guskey, T. R., & Bailey, J. M. (2010). *Developing standards-based report cards*. Thousand Oaks, CA: Corwin.

Guskey, T. R., & Brookhart, S. M. (Eds.) (2019). What we know about grading: *What works, what doesn't and what's next*. Alexandria, VA: ASCD.

Jung, L. A., & Guskey, T. R. (2011). *Grading exceptional and struggling learners*. Thousand Oaks, CA: Corwin.

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence*, 36, 153-160.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20–32.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *Journal of Educational Research*, *95*, 203–213.

McTighe, J., & Curtis, G. (2019). *Leading modern learning: A blueprint for vision-driven schools* (2nd ed.). Bloomington, IN: Solution Tree.

McTighe, J., & Willis, J. (2019). Upgrade Your Teaching: Understanding by Design meets neuroscience. Alexandria, VA: ASCD.

O'Connor, K. (2018). *How to grade for learning: Linking grades to standards* (4th ed.). Thousand Oaks, CA: Corwin.

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

Olsen, B., Buchanan, R., (2019). An investigation of teachers encouraged to reform grading practices in secondary schools. *American Educational Research Journal*, online first DOI: 10.3102/0002831219841349

Panadero, E., Tapia, J. A., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and individual differences*, *22*, 806-813.

Peters, R., Kruse, J., Buckmiller, T., & Townsley, M. (2017). "It's just not fair!" Making sense of secondary students' resistance to a standards-based grading. *American Secondary Education, 45*(3), 9-28.

Pollio, M., & Hochbein, C. (2015). The association between standards-based grading and standardized test scores as an element of a high school reform model. *Teachers College Record*, *117*, 110302.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education*, *26*, 1372–1380.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on psychological science*, 10(2), 176-199.

Townsley, M, Buckmiller, T., & Cooper, R. (2019). Anticipating a second wave of standards-based grading implementation and understanding the potential barriers: Perceptions of high school principals. *NASSP Bulletin*, 1-19. DOI: 10.1177/0192636519882084.

Welsh, M. E., D'Agostino, J. V., & Kaniskan, R. (2013). Grading as a reform effort: Do standards-based grades converge with test scores? *Educational Measurement: Issues and Practice*, *32*(2), 26-36.

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT scores, high school grades, and SES. *Educational Assessment*, *20*, 23-45.

Wiliam, D. (2000). The meanings and consequences of educational assessments. *Critical Quarterly*, *42*, 105-127.

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*, 1-37.

LEARNING®
SCIENCES
INTERNATIONAL

LearningSciences.com | 1-800-979-3316