Hitting for Average: Educational Assessment, Unidimensionality, and the Connection to Baseball Hitting Statistics

Alex Romagnoli

Abstract

The traditional points system and subsequent Grade Point Average (GPA) in education perpetuates an evaluation of academic performance which reflects arbitrary weighting of assignments and/or assessments. As such, GPAs which are calculated using a traditional points system are not unidimensional in their design. The baseball batting and slugging percentage, which serves as established metrics for performance evaluations among baseball players, better reflects unidimensionality. In essence, this paper puts forth an analysis and discussion which posits that baseball batting average and slugging percentage can serve as an example for how unidimensionality can become more prevalent in educational assessments, especially as it relates to the traditional points system and GPA.

Keywords: assessment, unidimensionality, baseball, Grade Point Average (GPA), education

Introduction

Teachers across disparate subject areas and through various grade levels consistently employ an assessment practice which is referred to as the traditional points system (Marzano, 2006). Utilizing the traditional points system on assessments is also used to calculate a final Grade Point Average (GPA) at the end of a given marking period and is a sum of the total points received divided by the total possible points. This common practice among teachers, while sensible upon initial consideration, can be an incomplete method to assess student understanding due to difficulty in weighting diverse assessments while maintaining alignment with standards (Brookhart and Nitko, 2019). Examining this style of assessment through the lens of baseball hitting statistics provides insight into both the pitfalls of said assessment style and a way to explore alternatives.

Baseball has utilized statistics as a way to assess player effectiveness and provide needed contextualization since the sport's beginnings in the mid-19th century (Tygiel, 2000). Popular among those statistics is batting average, which is used to assess the production of a hitter. While batting average is a common and oft-cited metric for assessing players, its validity has come under scrutiny (James, 2003; Kenny 2016; Law, 2017). Batting average's calculation of hits/at-bats is similar to the traditional points system in educational assessment as it is a statistic which weighs all hits the same.

This paper highlights the connection between educational assessment—mainly the traditional points system and subsequent GPA—and baseball hitting statistics. Batting average and the traditional points system are first defined and explored as metrics for assessment. Historical context is then provided for further consideration through a literature review of both batting

average and the traditional points system. Finally, this paper posits an alternative to the traditional points system by noting a connection between standards-based assessment and baseball's slugging percentage.

Batting Average as a Metric for Assessing Players

At its core, baseball is about scoring runs. The offense's goal (the hitters) is to hit the baseball into play without the defense (pitchers/fielders) getting the hitter out through various means, primarily through strikeouts (having the batter not make contact with the ball) and putouts. As such, a player's ability to hit the baseball is of paramount importance with much of the sport's statistical focus being placed on hitting.

Batting average is a fairly direct statistic in baseball which takes the number of hits a player has and divides it by the number of at-bats. This provides a percentage of how often a player gets a hit per at-bat. It sounds simple enough, and that statistic has become the single most utilized metric to determine how strong (or weak) a player performs as a batter.

Baseball lore is littered with stories and legends of players having magical seasons rooted in their batting averages. Ted Williams was the last hitter to reach a .400 average in 1941 by playing the final two games of a double header (Bradlee, Jr., 2013). Tony Gwynn nearly hit .338 in 1987 and surprisingly came in second to Andre Dawson in Most Valuable Player (MVP) award voting (Kenny, 2016, p. 216-217).

And then there is the story of Mario Mendoza whose batting average hovered around .200 for so much of his career that a .200 average became known as the "Mendoza line" (Landers, 2018). In a game that is obsessed with numbers and how those numbers tell a story, batting average exists as one of the most utilized metrics.

But batting average represents a limited view of a player's production. Boiling down the production of a particular player to a sole calculation promotes a single factor in a player's overall contribution to the team. Law (2017) posited that, "Batting average doesn't tell you how often a player gets a hit, but how often he gets a hit *ignoring times he drew a walk, gets hit by the pitch, hits a sacrifice fly, makes a successful sacrifice bunt, or reaches via catcher's interference*. Those scenarios don't count as at-bats, but do count as plate appearances" (p.11). That fundamental difference between "at-bats" and "plate appearances" encourages a valuation process that inadequately reflects the full spectrum of a player's experience playing baseball. At-bats are when a hitter is called out or reaches base on a hit, an error, or a fielder's choice. But at-bats do not include walks, being hit by a pitch, sacrifice plays, or catcher interferences. Plate appearances include everything a hitter can possibly do. So not every plate appearance is an at-bat, but every at-bat is a plate appearance. This complicated statistical maneuvering is done to maintain the focus, as much as possible, on what the hitter is responsible for when calculating batting average. However, batting average then limits how a player can be evaluated.

Being able to read pitches and not swing at ones which are more difficult to hit is an important baseball skill, but that is not included in the calculation for batting average. The same goes for how good a player is at making contact to drive in a runner on a sacrifice fly. Hitting a ball high

into the air with a runner on third base, thus giving that runner time to run home and score a run, is another important baseball skill. Again, that is not included in batting average. The larger point here is that the main metric for evaluating a player's production is incomplete.

It is also common to hear baseball fans talk about batting average as an indicator of how well a player is performing or the likelihood of a player getting a hit in their next plate appearance. Every hitter in a baseball team's starting lineup gets at least three plate appearances in a game unless they are substituted out. Just because a player is hitting .250 does not necessarily mean that player will get a hit in the next four at-bats, and a player hitting .300 does not mean they only get on base 30% of the time. Even seasoned baseball fans are prone to thinking like this, and it indicates a fundamental misunderstanding of how batting average works.

GPA and the Traditional Points System as Metrics for Assessing Students

For all students, the main metric for determining progress and evaluating understanding comes down to the grade the student receives for each subject. The traditional points system and Grade Point Average (GPA) is strikingly similar to baseball's batting average. Teachers assign points to a given assessment and then divide the correct answers over the maximum possible number of points which produces an average.

If an assessment has 20 questions with each question being worth one point, and the student correctly answers 14 of those questions, they receive a 70: 14 correct / 20 total questions = 70%. Do this repetitively for an entire academic year, and a teacher can calculate an even larger percentage that reflects the totality of a student's progress. Yes, teachers do get a concrete number which serves as an indicator of how much a student knew on assessments. GPA is a metric for quantifiably identifying what students know or do not know on particular assessments, but herein lies the issue with GPA: it strictly measures the assessments the instructor created and/or administered but not necessarily the totality of a student's experience. Teachers utilize GPA to get a snapshot of a student's progress, and it is *the* established way to gauge the development of a student.

Similar to batting averages, there is also an implied level of success for GPA that is socially constructed and reinforced through constant conditioning. This leads to a nebulous understanding of what constitutes a "good" grade. An "A" is usually universally noted as being the grade a student wants to achieve, but that letter-grade only serves as an abstraction of far deeper machinations that occur within a teacher's gradebook. Blum (2020), in the edited volume titled *Ungrading*, notes in the introduction, "The point is, when we grade, we really convey very little information about what is being assessed" (p. 12). Teachers provide students with rubrics and prompts for assignments which do help with contextualizing what is being assessed, but even the most well-organized and painstakingly created rubric can be incomplete or misinterpreted by both student and teacher. This is due, in part, to the subjectivity that is inherent in grading assignments. Even batting average, which excludes instances of getting on base outside of a hit, is able to articulate its own parameters as a metric; GPA—and grading in general—is not always able to do the same.

History of Grade Point Averages

The necessity and utilization of assessment and grading is logical, because grades provide a way to evaluate competency and or understanding, and grades are often calculated with a traditional points system. There is a natural connection between baseball and numbers, since so much of the game is situated in absolute outcomes: strike or ball, hit or out, play made or error, etc. But as the statistical sample sizes increase, even those outcomes have variances. Regardless, there is an inherent connection between baseball and numbers. With education and assessment, it is not quite so definitive.

At its core, educational assessment attempts to evaluate human understanding and human skill. Robert Marzano has been one of the leading voices on academic assessment and effective pedagogy over the last 20 years. Marzano (2006) defines assessment as, "any planned or serendipitous activity that provides information about students' understanding and skill regarding a specific measurement topic" (p. 35). Through the lens of education, assessment is attempting to measure understanding, and Popham (2020) echoes Marzano's definition saying that assessment, "is a broad and relatively nonrestrictive label for kinds of testing and measuring teachers must do" (p. 13). What becomes important when discussing educational assessment is that a formal test is not always involved.

Popham (2020) highlights the necessary variations in assessment for students saying assessment, "is a label to help remind educators that the measurement of students' status should include far more than paper-and-pencil instruments. Assessment embraces diverse kinds of tests and measurements" (p. 13). By varying the types of assessments, educators are accessing multiple intelligences (Gardner, 1983) and unique ways of learning.

In terms of how understanding has been measured, the history of that is a bit nebulous. Blum (2020) notes that assessment and evaluation have always been a part of the human experience but that specific instances of formal assessment in academic settings are relatively recent. Specifically, Blum cites, "It is only in the eighteenth and nineteenth centuries that written examinations became dominant at institutions such as Oxford and Cambridge, likely because of the increased number of examinees, where the scale made oral examinations impractical" (p. 6). Even as early as the eighteenth and nineteenth centuries, educators were feeling the weight of providing accurate and meaningful assessments for their students and found that utilizing student writing was the best way to balance that professional load. But more importantly, Blum goes on to recognize that, "Our familiar letter-grade system started in 1897 at Mount Holyoke College and has spread to most schools" (p. 7), a surprisingly recent development when looking at the totality of schooling.

Stommel (2020), who gives even more context to this discussion of the history of grading, also acknowledges the recency of modern grading by stating, "Letter grades are a relatively recent phenomenon. They weren't widely used until the 1940s…The A-F system appears to have emerged in 1898 (with the *E* not disappearing until the 1930s), and the 100-point or percentage scale became common in the early 1900s" (p. 25). Brookhart et al. (2016), in their substantial, historical review of educational assessment spanning over 100 years, also note that "By the 1940s, more than 80% of U.S. schools had adopted the A-F grading scale" (p. 805). The

assessment and grading style students and their parents/guardians in the United States are most accustomed to has only been popularly utilized for less than 100 years, but its socio-cultural impact is intense with much of the weight of the educational experience being directly impacted by grades. Despite the subjectivity of what an "A" grade means, the letter grade has had a substantial impression on the education of children. This pressure, to academically achieve at high levels, has been found to have a substantial impact on the self-esteem of students (Chemers et al., 2001; Metsapelto et al., 2020).

That "A" is usually informed by the points a student received on assessments, and the history of the letter grade and the points system are intertwined. Stommel (2020) recognizes the points system when discussing the history of assessment, but Marzano pinpoints an exact moment when the points system arose in the United States: the Alpha Test during processing of United States military recruits during World War I. Marzano notes that Darrel Bock (1997) found the 100-point scale being used by the U.S. military to quickly assess recruits since the number of recruits was so overwhelming. Designing the test around a 100-point scale facilitated the entire process (Marzano, p. 34). While the Alpha Test is not exactly the same as the points system that students are familiar with in today's schools, the spirit of quickly assessing progress and understanding based on a 100-point scale has its roots in a military test designed to expedite the processing of recruits.

This abbreviated history of assessment and grading has a common thread stretched across its disparate touchpoints: facilitation of gaining information to recognize students' progress and ease of interpreting said information. In other words, educators are always looking for ways to simplify the assessment process, and that simplification often takes the form of data collection and data presentation. It is all an attempt to get a "snapshot" of a student's progress or to be able to swiftly encapsulate the experience of a student in a singular statistic.

That "snapshot" mentality—the idea of looking at a singular metric or limited number of metrics—is a practice baseball has been utilizing since the 1800s (Tygiel, 2000). But like its educational assessment and GPA counterparts, batting average also has a history of trying to encapsulate the totality of experience into metrics.

History of Batting Average

Baseball's history is also extensive and complicated given its socio-cultural status in the United States and around the world. Historical contexts of statistics (particularly batting average) follow an eerily similar timeline to the history of modern assessment which was outlined earlier. As with the modern letter-grade system, it starts in the nineteenth century.

The origin of baseball statistics comes down to a relatively simple reality that Lewis (2003) notes when conducting his own history of the noted baseball statistician Bill James: statistics were developed in baseball to make sense of the outcomes in a baseball game. While watching a given game, it is difficult to tell who is doing well and who is doing poorly simply by watching. There are instances where one player has a poor at-bat but is generally a strong player (similar to when students do poorly on one assessment but are generally strong students), but that sole observation is not necessarily indicative of the overall performance. Baseball games are also complicated

affairs with so many disparate events occurring that there needed to be a way to succinctly and accurately record their complexities: enter the box score.

Credited to Henry Chadwick (Tygiel, 2001; Lewis, 2003; James, 2003), the box score was created to organize the results of a baseball game in a way that reflected the contributions of the players:

> Going right back to the invention of the box score in 1845, and its subsequent improvement in 1859 by a British-born journalist named Henry Chadwick, there had been numerate analysts who saw that baseball, more than other sports, gave you meaningful things to count, and that by counting them you could determine the value of the people who played the game. (Lewis, p. 69)

Lewis goes on to note that Henry Chadwick was more familiar with the sport of Cricket, so Chadwick's initial box scores viewed baseball through a lens that better reflected the elements of cricket as opposed to baseball. Regardless, the box score became a way to record the details of a baseball game.

Within box scores are numerous statistics which record details ranging from at-bats and hits to innings pitched and runs allowed. This is how batting average was initially developed. Law (2017) asserts as much, saying, "...Henry Chadwick is credited with creating batting average (among many other common baseball stats) in the late 1800s, designing it along the lines of cricket's version of batting average, which is runs divided by outs" (p. 10). With baseball being in its infancy at that time, it made sense to utilize the statistical stylings of a similar game to make sense of a newer one. Law also notes this understanding saying, "At the time, Chadwick's idea had merit…" (p. 10). Looking at the game fundamentally, it does make sense, because it finds the percentage of times a player gets a hit which is the goal of every batter. The drawback with baseball (and all dynamic theaters of experience) is that the sport is a complicated affair with so many outcomes that it requires more statistics.

However, the original metric of batting average remains the premiere measurement to assess a hitter's value. Consequently, a statistic that was originally created in the 1800s by a writer who modeled that statistic after an entirely different game, and which does not reflect all of the possible outcomes of a player's contributions, became the standard.

The Connection Between Batting Average and GPA

When calculating a GPA with the traditional points system, a point is a point. There is no distinction between a point on a homework assignment and a point on an exam. To create a "weight," teachers assign more points to a given assessment. A homework assignment can be worth 10 points and an exam worth 500 points. What makes the homework worth 10 points in comparison to an exam of 500 points is determined by the teacher. At the end of the semester, those points are calculated into the same GPA. And though it seems logical to do this, and nearly every teacher has done so, Marzano (2006) asserts, "As intuitively appealing as this system might appear, it has one major flaw - the scores on the various tests are typically not comparable in terms of students' understanding and skill regarding a specific measurement topic" (p 30).

While Jesse Stommel (2020) supports not grading at all, which this article does not suggest, he makes a poignant assertion regarding grading in general:

> Grades are not good feedback. They are both too simplistic, making something complex into something numerical (8/10, 85%), and too complicated offering so many gradations as to be inscrutable (A, A-, A/A-, 85.4%, 8.5/10). (p. 28)

Baseball is refreshingly fair when it comes to its statistics; it is in the interpretation of those statistics where the majority of subjectivity exists. As much as it pains a fan of a particular baseball team to admit, a focused analysis of multiple statistics customarily tells the story of why a team/player succeeded or failed. Not all grades can be considered fully objective because they are often arbitrarily weighted. The reason for this ambiguity, as it relates to the traditional points system, is limited unidimensionality in assessment.

While discussing state standards in educational testing, Marzano (2006) asserts, "Two barriers stand in the way of standards being the focus of effective classroom assessment: (1) too much content and (2) lack of unidimensionality" (p. 13). The first of those barriers, "too much content," is something with which many teachers have struggled. Gallagher's (2009) book, which is ominously titled *Readicide: How Schools are Killing Reading and What You Can Do About It*, compliments Marzano's assertion. While recounting an exhaustive list of standards for 10th grade history students, Gallagher says, "Is it just me, or would it take weeks to teach any one of the preceding standards with any depth?" (p. 9). No, it is not just Gallagher; the amount of time it would take for teachers to cover the necessary amount of content in school is unrealistic. Or, as Marzano (2006) clearly explains, "Another way of looking at this is that schooling, as currently configured, would have to be extended from kindergarten to grade 21 or 22 to accommodate all the standards and benchmarks in the national documents" (p. 13). There is simply too much content to cover given all of the other responsibilities of a classroom teacher.

The second barrier, unidimensionality, is a deeper issue. Marzano's explanation of unidimensionality, which in turn is a commentary on Frederick Lord's research in the 1950s, is particularly poignant as it informs the intersection of GPA and baseball's batting average:

> In simple terms, *unidimensionality* means that a single score on a test represents a single dimension or trait that has been assessed. This concept underpins almost all of measurement theory within education and psychology. To illustrate, in a foundational article on measurement theory, Frederick Lord (1959) explains that a test "is a collection of tasks; the examinee's performance on these tasks is taken as an index of this standing along some psychological dimension." In effect, Lord's comments imply that any test that depicts a student's performance on the test by a single score should, by definition, measure one trait only. (2006, p. 14)

A unit test on *The Great Gatsby* (Fitzgerald, 1925) in a high school English course could include a multitude of measurable standards ranging from literary theory to composition skills, to vocabulary, and more. Within each of those standards are unique ways to assess understanding. Yet, there is still a singular grade which is supposed to reflect overall understanding of a given unit. Think back to the formula for how a traditional, points system GPA is calculated: taking the totality of points accrued over an entire semester and dividing that over the total number of possible points. This results in a singular GPA but is made up of multiple and diverse

assessments each with their own points structures and potentially aligned with disparate standards.

Batting average arguably reflects unidimensionality in a more direct fashion than GPA due to the single statistic being measured. Batting average reflects what it is and what it measures, and it is fair to assert that every at-bat and/or plate appearance for a player is an assessment unto itself. There is a standard (achieving a hit) that is measurable (the player makes contact with the ball resulting in no fielders being able to get the batter out). Even with the simplicity of the statistic being highlighted, batting average is still not entirely unidimensional.

If a right-handed batter is facing a right-handed pitcher, that is one dimension being assessed. The success of the batter against a right-handed pitcher is a different measurable dynamic than if that same batter then faced a left-handed pitcher. In baseball terminology, this is referred to as a batter's splits: splitting the average to show how one hitter does against different-handed pitchers. When evaluating a batter, the player's splits can determine when a manager puts the player in a lineup, so it serves as a vitally important piece of performance data. The batting average can then be even further delineated: average against starters, average against relievers, average at home, average away, etc. The point here is that even in the simplest of baseball metrics, there are variances in how the data is disaggregated.

Law (2017) goes even further to deconstruct the pitfalls of batting average as a viable metric for performance noting, "So if the appeal of batting average as the lord of hitting stats isn't accuracy, or ease of calculation, then what is it?" (p. 17). That is a fair question to ask when batting average alone provides a cloudy lens through which to assess a player's performance. Law goes on to say, "The emphasis on batting average when smarter stats are out there embodies the false dichotomy we've seen in baseball coverage over the last fifteen years…" (p. 17). Like its education counterpart, baseball is prone to overvaluing a single metric to assess hitters' abilities.

The "weight" of assignments is something teachers also feel when it comes to assessment, and it exists as an omnipresent precedent. Teachers assess students' work in a similar way to how their own teachers assessed their work when they were students. And those teachers learned from their teachers, and so on, and so on (Lunenberg et al., 2007). Certainly, variations of this exist, but the trend of doing what came before is always there. Scholes' book, *English After the Fall: From Literature to Textuality* (2011), crystallizes this phenomenon through the study of literature. The same texts are read by students year after year from different areas and through different eras (Burke, 2013, p. 15). For Scholes, the reason for questioning the very nature of the term "literature" is of paramount importance: "...I hope to show that we have much to gain by moving from the limiting notion of literature to the more inclusive concept of textuality–a move which I believe can help restore English studies some of the power and pleasure that they have lost in recent years" (Scholes, 2011, p. 31). The "limiting notion of literature" which Scholes refers to can be associated with literary canon which consists of the texts that are deemed to be the most significant.

The power of grades is profound for students as well, and it remains a significant part of the school experience (Pattison et al., 2013), and that power is attributed to the social and academic weight of grades. Grades matter because they impact students throughout their academic careers.

Just like their baseball counterparts, students' careers are not static, and the perceived finality of grades is noteworthy. Guskey (2019) posits, "We must ensure that students and their families understand that grades do not reflect *who* you are as a learner, but *where* you are in your learning journey" (p. 46). Viewing the attainment of knowledge as a journey as opposed to a destination supports a style of grades and assessments which are more formative in nature.

Alternatives in both of these spheres, educational assessment and baseball hitting statistics, have been around for decades. For baseball, those other statistics provide ways to evaluate players in the ongoing discourse of baseball. For education, the conversations on assessment are informed by focusing on standards-based grading.

Slugging Percentage and Weighted Grades

Even if someone is not a devoted baseball aficionado, the formulaic nature of a baseball broadcast is generally known. A hitter steps up to the plate and the announcers will say something akin to, "Hitter McBatterson is hitting .263 this season and driven in 39 with 8 home runs. He has had a tough time hitting lefties this season, though…" Within that fictional and generic statement is a deeper problem: the contextualization is incomplete. McBatterson could be the league leader in walks, or 10 of those runs batted in (RBIs) could have been achieved on sacrifice plays. As was noted at the beginning of this article, there is a difference between "at-bats" and "plate appearances." Kenny, a noted baseball analyst for MLB Network, asserts, "It's just that the batting average alone tells you just a small part of the story" (2016, p. 118). However, there are other ways to measure a baseball player's performance.

The back of most baseball cards has an abundance of statistics. Among the myriad of statistics is the heralded batting average, but there are other statistics too: OPS, OBP, RBI, etc. The extensiveness of the metrics provided for each player is staggering. Even more staggering is a casual visit to the website Baseball Reference (Baseball Reference) which has become the established place to find statistics of any baseball player who ever played professionally. The site has a menagerie of analytics which range from the simple (HR or Home Runs) to the wildly complex (wOBA or weighted on-base Average). Complicated statistics utilize what are called "linear weights" (Law, 2017) to better articulate how valuable a given event in baseball is to the overall outcome of a game.

Baseball's recognition of linear weights makes the sport an interesting case study for how related events can be valued differently within the context of an endeavor. As was stated before, this is something baseball does very well in its advanced metrics. In reference to baseball, Law (2017) describes how linear weights work in the sport's statistics:

> So it might make more sense to think about the events that lead to run scoring in fractional terms for the hitter, which is the philosophy behind "linear weights" methods of evaluating offense. With linear weighting, any event from a hitter is worth, on average some fraction of a run, so if you assign the fractional amount of a run to each of those events and add them all up, you'll get a number that measures in runs the value of all of that hitter's actual production for a given time period. (p. 37)

This makes sense when thinking about a baseball game; if a player has one at-bat and hits a home run, the value of that at-bat is far higher than if the player were to only get a single. But

with batting average, both of the events—a single and a home run—statistically look the same because they are both hits.

One popular metric that does utilize weighted statistics is Slugging Percentage which Kenny (2016) expertly explains: "Slugging percentage is total bases divided by at-bats, a measure of a player's power" (p. 322). The value of each hit a player achieves produces different results. Spatz (2013) provides the formula for slugging percentage: "The formula is singles + doubles*2 + triples*3 + home runs*4, divided by total at-bats" (p. 319). A single base hit is valued as one, a double as two, a triple as three, and a home run as four. But even in that logical distinction, there are subtleties that cannot always be accounted for: was it a two-run homerun, or a solo homerun, a double with two runners on, or a double with no runners on, etc.? Regardless, slugging percentage serves as an additional way to determine a player's productivity outside of mere batting average which values all hits the same.

Teachers can do the same thing when looking at grades, but weighting with traditional points can be far more arbitrary and random. The number of points assigned to a given assessment, unless it is a standardized test, is at the discretion of the instructor. Additionally, the type of assignment (essay, group project, multiple choice test, etc.) can yield varied weights which, in turn, can all be aligned with the same standards.

For teachers, weighting assessments reflects valuations placed on particular skills students need to exhibit or specific information that is necessary for mastery of a given standard. A typical practice in education is to weight assessments according to increased points. As was noted earlier, the traditional points system for grading utilizes the maximum number of points as a denotation of what is weighted more:

- Homework assignment - 10 points
- Essay - 50 points
- Final exam - 100 points

In the above example, the final exam is the most substantial assessment because it has the most points, while the homework assignment—with the smallest number of points—is least important. Brookhart and Nitko (2019) dispel this common usage of weighting saying, "Such incompatibilities make a *simple sum of the marks* an invalid basis for grading. You will need to grade each assessment in a way that makes scales compatible" (p. 339). The theory behind increasing the maximum total of points for a given assessment in order to give it more weight is both simple and logical: if the assessment is more important, it should have more points. On the surface, that makes complete sense. Where that theory loses its edge is when calculating the final grades and GPA.

Think back to baseball batting average and how that is calculated. Using the fictional player from earlier, Hitter McBatterson, consider a scenario where the player has played 10 games and had three at-bats each game for 30 total at-bats. These were the player's stats during that stretch:

- 6 singles
- 2 doubles
- 2 home runs
- 3 strikeouts
- 7 runs batted in (RBIs)

Using the simple equation of batting average (hits / at-bats), we get a batting average of .333 (10/30). Within that average, there is no weighting of anything; the only metric that matters is hits. That can be problematic for determining a player's effectiveness because it measures only one element of a player's contribution: hits.

If that same stretch of 10 games incorporated Hitter McBatterson's slugging percentage, there is an entirely different way of calculating his effectiveness as a hitter. As a reminder, slugging is a weighted metric that applies values to different types of hits:

- 1*singles + 2*doubles + 3*triples + 4*home runs / at-bats

For that stretch of 10 games, McBatterson has a slugging percentage of .600 which is much different than the batting average of .333. However, both statistics—batting average and slugging percentage—still reflect the same performance. The GPAs and grades that teachers calculate are usually determined by the traditional points system that was mentioned before, but that system is only utilizing one dynamic: the points a student accrues. And within those points is natural variance in what is valued.

Link (2018) conducted an extensive survey of nearly 3,000 K-12 teachers on their perceptions of grading, and the results reflected how diverse and varying educators' grading practices are. Link's findings highlight the natural differences in educators' philosophies when grading, but those differences are notably divergent between elementary teachers (grades K-6) and secondary teachers (grades 7-12). According to Link, secondary teachers are more likely to account for behavior and effort when grading student work and are more willing to give grades of "zero." However, what was most profound in Link's findings was, "Middle/high school teachers also preferred using their own grading procedures as compared to elementary teachers" (p. 70). This is vitally important in the discussion of utilizing points when grading because of how elementary and secondary teachers construct their assessments.

Elementary and Secondary Assessments

Since elementary education is focused on building the skills of students to prepare them for increasingly more difficult work, elementary teachers are usually obligated to have more controlled assessments and adhere to programs which the school district requires teachers to use (Schwartz, 2019). One such program is Lexile scoring which assesses reading abilities.

The popular publisher, Scholastic, has one such program which utilizes Lexile scores. According to Scholastic's website, their program "serves two functions: it is the measure of how difficult a text is OR a student's reading ability level" (Scholastic). The score is a letter (A-Z) with "A" being the lowest and "Z" being the highest. Scholastic provides the texts for the teachers which are designed to increase in difficulty to assist the student in growing his/her literacy. It is a strong system that is backed by valid and reliable assessments which support student learning (American Institutes for Research, 2019). Accompanied with the specifically designed texts are standardized assessments which complement the texts students read. Those assessments are part of the Scholastic Reading Inventory (SRI), and they provide both baseline information for educators to begin their reading instruction (where the students start) and measure growth over time (how students' literacy skills and vocabulary grow).

Within the Lexile scores is a determination of where students stand in regards to their reading abilities. There are, as always, exceptions and conditions when it comes to those Lexile scores and the grade levels of students. Some students in fourth grade may read at a fifth grade level and vice-versa. However, the statistics are reflective of what students' abilities are. Compare this with secondary education which, as Link (2018) noted, is more often an open forum for teachers to grade how they want to grade. And before the repercussions of that are analyzed, it is important to deconstruct why that trend is understandable.

Moreso than elementary education, secondary education is rooted in content as opposed to skills which include reading comprehension (Wexler, 2019). This is not to say skills are not taught at the secondary level, because that would be utterly preposterous to suggest. Distinguishing between skills and content, while acknowledging that the two inform one another, is an important dynamic to address.

Letter grades and GPA have been a major touchstone of this article, and that assessment / grading practice reflects a student conveying understanding through multidimensional and multimodal means. In a given secondary class, a student will complete a multitude of assessments (e.g., essays, exams, worksheets, projects, labs). As was discussed earlier, all of these disparate types of assessments provide an approximate quantification to gauge student understanding, and that quantification is frequently the points system. That points system is then amalgamated by combining all of the unique and disparate assessments to produce a final letter grade.

What supports all of these assessments are the standards with which schools are obligated to align their instruction and assessments, and all educators are well-acquainted with the reality of needing to reference standards when developing lessons, writing syllabi, or preparing assessments. However, even the standards reflect the *content over skills* reality.

Marzano, Norford, and Ruyle (2019) highlight a Mid-Continent Research for Education and Learning study (2014) which notes a distinction between what are "declarative" standards and "procedural" standards. Marzano, Norford, and Ruyle state that standards, "that begin with the word *knows* or *understands* are examples of declarative knowledge. Those that begin with the word *executes* are examples of procedural knowledge" (p. 21). Using those terms, they analyze the standards to highlight the trend that most sets of standards are predominantly "declarative" in nature. Or, as was stated earlier, standards often focus on content and not necessarily skills.

While all teachers are focused on providing students with the building blocks of skills necessary to engage in more difficult work, secondary teachers are content experts and provide students with knowledge and skills that will be required in both higher education and the workforce. Secondary teachers devote a significant amount of instructional time to content in order to align with standards which are predominantly "declarative" in nature.

The number of standards—both declarative and procedural—that all teachers need to align their lessons and assessments with is extensive. As Popham (2020) asserts, "Teachers tend to give less than careful attention to the whole array of standards" due to the sheer number. However, by

focusing assessments and grades on standards—both declarative and procedural—a way to make assessments more unidimensional arises: standards-based grading.

Standards-Based Grading and Unidimensionality

Some schools practice what is referred to as "standards-based grading" where, in lieu of letter grades or a points system, students are graded on whether or not they meet specific standards. In the scope of educational assessment, standards-based grading is a relatively new trend in schools (Brookhart et al., 2016, p. 805). There is no letter grade on the students' report cards, and the delineation of subjects is not anchored solely by content. Instead, students are assessed "Beginning," "Approaching," "Meeting," or "Above" in a multitude of standards (Iamarino, 2014). Students may still receive point grades for assignments and assessments throughout the school year, but the final "grade" of the semester is not a "grade" at all; it is a measurement of growth. While standards-based grading is not always quantifiable with points, it is still a direct method to assess for understanding and/or competency in skills or content. In other words, it is a method of assessment that is more unidimensional.

Iamarino's (2014) analysis of how standards-based grading can be beneficial to both students and teachers again highlights the pitfalls of continuously relying on the points system to grade. It all goes back to the idea that, "Final grades are sourced from gradebook figures (points), and there is often no comprehensive system in place to determine the integrity of this method through which those figures are collected" (p. 3). And Vatterott (2015) echoes the importance of developing a grading and assessment system that focuses on the standards that students need in order to succeed instead of arbitrarily weighting assignments with points.

The link between standards-based grading and baseball statistics is closer than it appears because of what this article has highlighted: unidimensionality. Standards-based grading focuses on how a student performs on a particular standard (synthesizing information in a non-fiction text, effectively utilizing supporting information to substantiate a stance in an essay, finding the solution to a multiplication problem using the distributive property, etc.). With baseball statistics, similar attention is given to measuring performance which is contextualized in order to provide the most useful data possible (batting average against left-handed or right-handed pitchers, slugging percentage, etc.) Both teachers and baseball coaches want the same thing: to give their students or players the best chance to succeed using data which best reflects strengths and weaknesses.

Conclusion

There is no ultimate assessment or statistic, in either education or baseball, which will provide a definitive metric that succinctly and infallibly relays understanding and/or achievement. The innumerable factors that impact assessments make such an endeavor impossible, and this article does not argue for the existence of such a metric. No matter how many factors are taken into account, there will always be varying degrees of influence and new/arising factors that complicate the creation of such an assessment.

What baseball does effectively, from which education can learn, is focusing its assessments to be as unidimensional as possible to maximize the aim of said assessments. For all of the shortcomings of baseball's batting average, it is still unidimensional in its focus on players getting hits. Slugging percentage expands on that dimension to include weighted metrics which further contextualize how valuable those hits are; but again, it is a focus on the dimension of hits and nothing else. The points system and subsequent GPA in education reflects baseball's batting average, but the noise of all the disparate dimensions used to calculate a GPA (homework, essays, projects, exams, class participation, etc.) complicate the very nature of that average.

References

American Institutes for Research. (2019). Scholastic reading measure: Reliability and validity study. Scholastic. https://teacher.scholastic.com/education/researchpdf/technical_reports/SCHL_Scholastic_Reading_Measure_Reliability_and_Validity_Study.pdf

Baseball Reference. (n.d.). *Baseball Reference*. Retrieved August 21, 2023, from https://www.baseball-reference.com/.

Blum, S. D. (2020). Introduction: Why ungrade? Why grade? In S. D. Blum (Ed.), *Ungrading: Why rating students undermines learning (and what to do instead)* (pp. 1–22). West Virginia University Press.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement, Issue and Practice, 16*(4), 21-23.

Bradlee Jr., B. (2013). *The Kid: The immortal life of Ted Williams*. Little, Brown and Company.

Brookhart, S., Guskey, T., Bowers, A., Mcmillan, J., Smith, J., Smith, L., & Stevens, M. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4). https://doi.org/10.3102/0034654316672069

Brookhart, S. M., & Nitko, A. J. (2019). Educational assessment of students (8th ed.). Pearson.

Burke, J. (2013). *The English teacher's companion: A completely new guide to classroom, curriculum, and the profession* (4th ed.). Penguin

Chemers, M. M., Hu, L., & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology, 93*(1), 55–64. https://doi.org/10.1037/0022-0663.93.1.55

Fitzgerald, F. S. (1925). *The Great Gatsby*. Charles Scribner's Sons.

Gallagher, K. (2009*). Readicide: How schools are killing reading and what you can do about it*. Stenhouse Publishers.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. Basic Books.

Guskey, T. R. (2019). Grades versus comments: Research on student feedback. *The Phi Delta Kappan, 101*(3), 42-47.

James, B. (2003). *The new Bill James historical baseball abstract*. Free Press.

Kenny, B. (2016). *Ahead of the curve: Inside the baseball revolution*. Simon & Schuster.

Landers, C. (2018, May 22). How did Mario Mendoza become a shorthand for batting futility? *Cut 4 by MLB.com*. https://www.mlb.com/cut4/how-did-the-mendoza-line-become-an-mlb-term-c277392972

Law, K. (2017). *Smart baseball*. Harper Collins.

Law, K. (2020). *The inside game: Bad calls, strange moves, and what baseball behavior teaches us about ourselves*. William Morrow.
Lewis, M. (2004). *Moneyball*. W. W. Norton & Company.
Link, L. (2018). Teachers' perceptions of grading practices: How pre-service training makes a difference. *Journal of Research in Education*, *28*(1), 62-91.
Lord, F. M. (1959). Problems in mental test theory arising from errors of measurement. *Journal of the American Statistical Association*, *54*(286), 472-479. https://doi.org/10.2307/2281785
Lunenberg, M., Korthagen, F., & Swennen, A. (2007). The teacher educator as a role model. *Teaching and Teacher Education*, *23*(5), 586–601. https://doi.org/10.1016/j.tate.2006.11.001
Marzano, R. J. (2006). *Classroom assessment and grading that work*. Association for Supervision and Curriculum Development.
Marzano, R. J., Norford, J. S., & Ruyle, M. (2019). *The new art and science of classroom assessment*. Solution Tree.
Metsäpelto, R. L., Zimmermann, F., Pakarinen, E., Poikkeus, A. M., & Lerkkanen, M. K. (2020). School grades as predictors of self-esteem and changes in internalizing problems: A longitudinal study from fourth through seventh grade. *Learning and Individual Differences*, *77*, 1-10. https://doi.org/10.1016/j.lindif.2019.101807
Mid-Continent Research for Education and Learning. (2014). *Content knowledge - online edition: The process of this work*. McRel International. http://www2.mcrel.org/compendium/docs/process.asp
Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky falling? Grade inflation and the signaling power of grades. *Educational Researcher*, *42*(5), 259-265.
Popham, W. J. (2020). *Classroom assessment: What teachers need to know* (9th ed.). Pearson.
Scholastic. (n.d.). Lexile levels: What parents need to know. *Scholastic*. Retrieved August 21, 2023, from https://www.scholastic.com/parents/books-and-reading/reading-resources/book-selection-tips/lexile-levels-made-easy.html#:~:text=What%20Is%20a%20Lexile%20Level,a%20student's%20reading%20ability%20level.
Scholes, R. (2011). *English after the fall: From literature to textuality*. University of Iowa Press.
Schwartz, S. (2019, December 3). *The most popular reading programs aren't backed by science*. Education Week. https://www.edweek.org/teaching-learning/the-most-popular-reading-programs-arent-backed-by-science/2019/12
Spatz, L. (2013). *Historical dictionary of baseball*. Rowman & Littlefield.
Stommel, J. (2020). How to ungrade. In S. D. Blum (Ed.), *Ungrading: Why rating students undermines learning (and what to do instead)* (pp. 25–41). West Virginia University Press.
Tygiel, J. (2000). *Past time: Baseball as history*. Oxford University Press.
Vatterott, C. (2015). *Rethinking grading: Meaningful assessment for standards-based grading*. Association for Supervision and Curriculum Development.
Wexler, N. (2019, August). The radical case for teaching kids stuff. *The Atlantic 324*(2), 20-23.

Alex Romagnoli (aromagno@monmouth.edu) is the Interim Associate Dean of the School of Education at Monmouth University in West Long Branch, NJ.