Towards Measures of Different and Useful Aspects of Schooling: Why Schools Need Both Teacher Assigned Grades and Standardized Assessments[1]

Alex J. Bowers[2]
Teachers College, Columbia University
Bowers@tc.edu

ABSTRACT:

Summative assessments in classrooms typically result in teacher assigned grades. Grades are well known to be highly predictive of high school graduation, college enrollment and college completion, but there has been little research that explains why. In the psychometrics literature there is a persistent perception that while standardized test scores are objective measures of fundamental academic knowledge, grades are more subjective assessments that may vary school-by-school. This chapter examines the extent to which grades in high school include teacher perceptions of student effort, participation and behavior that is a different and useful measure for schools beyond what can be provided by standardized test scores, and to what extent grades vary between schools. The chapter is organized into three related sections: a review of the literature on the relationship of grades to standardized tests, an example analysis of a large high school student dataset, and, finally, a comparison of the findings from the literature and the analysis to discuss how grades are a useful yet multidimensional assessment of academic knowledge and engaged participation in the schooling process, with the latter being highly related to overall student life outcomes.

INTRODUCTION:

Summative assessments in classrooms typically result in teacher assigned grades. Grades are well-known to be highly predictive of high school graduation, college enrollment and college completion, but there has been little research that explains why. Additionally, in the psychometrics literature, there is a persistent perception that while standardized tests cores are objective measures of fundamental academic knowledge, grades are more subjective assessments that may vary school-by-school. This chapter examines the extent to which grades in high school include teacher perceptions of student effort, participation and behavior that is a different and useful measure for schools and school leadership beyond what can be provided by standardized test scores, and to what extent grades vary between schools. The chapter is organized into three related sections. To provide a discussion of these issues with grades, I first review the literature on the relationship of grades to standardized test scores, the construct validity argument that grades represent a valid measure by teachers of engaged participation, that engaged participation correlates with overall student life outcomes, and how some research has suggested that grades may be "fairer" than standardized tests as grades appear to vary less by student demographics and socio-economic status (SES) than standardized test scores. Across this discussion, I also note how there has been a continual question in the literature about the extent that grades vary by schools, but that there is little evidence that has investigated this issue. Second, I then provide an example of testing these ideas using a hierarchical linear modeling strategy to analyze the large nationally U.S. generalizable sample, the Education Longitudinal Study of 2002 (ELS:2002), which includes almost 15,000 students across hundreds of high schools in the U.S. In this example study, I look to apply the main findings and questions from the literature on grades to examine the relationship between grades and standardized tests, student background and SES, mathematics and English teachers' perception of student participation in class, and how individual student grades vary within and between schools, with a special focus on school-level context and demographics. In the third and final section, I relate the findings from the analysis to the application of the literature to the question of the utility of grades as valid classroom assessments in educational measurement, as the literature and the included study provide an argument that teacher assigned grades are a multidimensional assessment of student work that is a different construct from academic knowledge, and that grades do not seem to be particularly dependent to a large extent on which school a student attends.

Historically, grades have been maligned by psychometricians for their "hodgepodge" nature (Brookhart, 1991), in which when asked what they assign a grade for, teachers respond that grades are assigned for a multitude of outcomes, such as academic knowledge, student participation, effort, and behavior (Cross &

[1] This manuscript is a pre-print of a book chapter published in the book titled *Classroom Assessment as Educational Measurement*. Citation:
Bowers, A.J. (2019) Towards Measures of Different and Useful Aspects of Schooling: Why Schools Need Both Teacher Assigned Grades and Standardized Assessments. In Brookhart, S., McMillian, T. (Eds.) *Classroom Assessment as Educational Measurement*. National Council on Measurement in Education (NCME) Book Series (p.209-223). New York: Routledge.

[2] ORCID: 0000-0002-5140-6428

Frary, 1999; McMillan, 2001), known as “kitchen-sink” grading (Cizek, Fitzgerald, & Rachor, 1995-1996). Some scholars have interpreted this to mean that grades are subjective and unreliable measures of academic performance, and thus must be reformed to align much more to standardized test scores (Brookhart, 1991, 2011). As noted in this research domain, “student’s grades often have little relation to their performance on state assessments (Guskey & Jung, 2012, p.23). But *should* grades have a relation to standardized test performance (Brookhart, 2015)? If test scores are assumed to be an accurate and reliable measure of fundamental academic knowledge, why would schools need another measure of this factor? The purposes of schooling in the U.S. are far from agreed upon (Labaree, 1997) and some have argued that test scores are a poor measure of what the many different stakeholders in schools are looking for schools to instill in their students (Brighouse, Ladd, Loeb, & Swift, 2018; Nichols & Berliner, 2007). Could grades measure different, but important aspects of schooling?

Standardized test scores have historically lacked criterion validity to overall schooling outcomes (Atkinson & Geiser, 2009), to such an extent that many states throughout the U.S., as well as countries globally, have begun to mandate exit and end of course exams (Allensworth, 2005a; Blazer, 2012; Nichols & Berliner, 2007; Warren, Jenkins, & Kulick, 2006) that artificially connect test scores to outcomes through retention, grade promotion and graduation requirements (Maag Merki & Holmeier, 2015). By contrast, teacher assigned grades are strong predictors of overall schooling outcomes, such as graduation or dropping out (Allensworth, 2005b; Barrington & Hendricks, 1989; Battin-Pearson et al., 2000; Bowers, 2010b; Bowers & Sprott, 2012; Bowers, Sprott, & Taff, 2013; Brookhart et al., 2016; Lloyd, 1978) as well as college attendance and graduation (Atkinson & Geiser, 2009; Cliffordson, 2008). In addition, grades are seen as being “fairer” assessments than standardized tests, since grades are not as strongly related to socio-economic status (SES) (Atkinson & Geiser, 2009). As noted by Atkinson & Geiser (2009) “High school grades are sometimes viewed as a less reliable indicator than standardized tests because grading standards differ across schools. Yet although grading standards do vary by school, grades still outperform standardized tests in predicting college outcomes” (p. 665).

The focus that I aim to address in this chapter is to ask the question: Why? What is it about grades that make them a strong predictor of overall schooling outcomes that adds to the knowledge gained about student learning from standardized test scores? If schools have two measures of different and useful factors about different student outcomes from schooling, then schools should use both sets of measures to inform their practice and decision making (Bowers, 2009, 2011; Brookhart et al., 2016; Farr, 2000).

Examining the Research on Grades in Relation to Standardized Tests

Across K-12 schooling assessment research over the past 100 years, a perennial issue has been the relationship between teacher assigned grades and standardized assessment scores (Brookhart, 2015; Brookhart et al., 2016). As recently reviewed in their literature review of one hundred years of research on grades, Brookhart et al. (2016) discuss the numerous studies that have demonstrated that across multiple contexts, as well as nationally, grades and standardized test scores continually correlate at about 0.5 (Bowers, 2011; Brennan, Kim, Wenz-Gross, & Siperstein, 2001; Duckworth, Quinn, & Tsukayama, 2012; Linn, 1982, 2000; Welsh, D'Agostino, & Kaniskan, 2013). As noted by Brookhart et al. (2016):

> Although some variability exists across years and subjects, correlations have remained moderate but remarkably consistent in studies based on large, nationally representative data sets. Across 100 years of research, teacher-assigned grades typically correlate about .5 with standardized measures of achievement. (p. 882)

This suggests that about 25% of of the variance shared between grades and what is assessed by standardized test scores is academic knowledge (Bowers, 2011).

Grades are also well-known to be strong predictors of overall schooling success (Brookhart et al., 2016). For example, low or failing grades are some of the most accurate predictors of students dropping out of high school (Bowers et al., 2013) in both single time point studies (Allensworth & Easton, 2005, 2007), as well as longitudinal research (Bowers, 2010a, 2010b; Bowers & Sprott, 2012). Additionally, grades are strong predictors of college enrollment and completion (Atkinson & Geiser, 2009; Attewell, Heil, & Reisel, 2011; Cliffordson, 2008) as well as years of schooling and long-term earnings (Jones & Jackson, 1990; Miller, 1998). For example, using the large nationally generalizable NCES High School and Beyond dataset, Miller (1998) showed that for students who were in grade 10 in 1980, their high school grades significantly predicted their annual earnings in 1991, finding a strong independent effect of grades on earnings when controlling for a range of context variables, an effect in addition to years of schooling. Miller (1998) concludes that:

> One might question whether employers are really benefiting from higher grades or from the greater aptitude that is reflected in higher grades. …[this] suggest[s] that it is the actual learning, not aptitude, that matters in predicting longterm productivity. Furthermore, the evidence presented here suggests that some part of the productivity gains might be coming from the soft skills that employers say they want and grades appear to contain. These soft skills of regular attendance, preparation, hard work, and lack of

> disciplinary problems that employers say they value are also valued by schools and reflected in grades. (p. 306-307)

Thus, grades are predictive of overall schooling outcomes, yet only moderately correlate with standardized test scores. A persistent question has thus been, what does the other 75% of grades represent if it is not what is measured in standardized assessment tests (Bowers, 2011; Brookhart, 2015; Brookhart et al., 2016)? In the above quote, Miller (1998) alludes to the idea that perhaps grades are signals of "soft skills", what might be called non-cognitive skills in more recent research (Levin, 2013; West et al., 2016), that include skills that schools and employers highly value that are not included on standardized tests, such as "preparation, hard work, and lack of disciplinary problems".

This issue of what the majority of grades represent has also been a consistent issue in the grading research (Brookhart et al., 2016). As noted throughout this work, this is a question around the validity of grades (Brookhart, 2015). For example, over 70 years ago (Swineford, 1947), in a study of teacher grades and marks for one elementary school, noted "in any event, the data in Table 1 clearly show that the marks assigned by teachers in this school are reliable measures of something, but there is apparently a lack of agreement on just what that something should be" (p.517). Multiple surveys of teachers have shown that teachers award grades for a variety of student behaviors in addition to academic achievement (Brookhart, 1993, 1994; Cizek et al., 1995-1996; Cross & Frary, 1999; McMillan, 2001). For example, McMillan (2001) surveyed over 1,400 teachers in Virginia asking them about their grading practices, and using factor analysis, identified that teachers award grades for a range of behaviors quite similar to those listed above by Miller (2008), behaviors that schools and employers prefer, including effort, ability, improvement, work habits, attention and participation. Thus, rather than teacher grades being subjective and unreliable, as is intimated by the "hodgepodge" and "kitchen-sink" metaphors used in some of the research in this area noted above, it appears that teachers award grades for a variety of student behaviors that are important for overall life outcomes and are valued by students, parents, schools, and future employers (Bowers, 2009). However, much of the survey research asking teachers about their grading practices relies exclusively on teacher perception of their grading practices, rather than on the grades that they actually assign.

A growing set of research studies over the past two decades has focused on the grades that teachers assign. The research has postulated that grades are multidimensional (Bowers, 2011; Brookhart et al., 2016), assessing academic knowledge to a limited extent, but, more importantly, assessing what has been termed a "conative" factor (Willingham, Pollack, & Lewis, 2002), a "common grade dimension" (Klapp Lekholm, 2011; Klapp Lekholm & Cliffordson, 2008, 2009; Thorsen & Cliffordson, 2012), and a "Success at School Factor (SSF)" (Bowers, 2009, 2011). Across these studies, other than academic knowledge, grades appear to measure student engagement through measuring effort, participation and behavior (Brookhart et al., 2016). As recently noted in research examining the relationship of high school grades to college readiness in the state of Alaska (Hodara & Cox, 2016) , the authors note that:

> High school grade point average may be useful because it is not just a measure of cognitive ability; instead, it is a cumulative measure of academic achievement in multiple subjects across a student's high school career and thus may signal a broader range of skills related to college readiness, such as a student's academic tenacity and motivation" (p. i).

Recent research has confirmed that that while grades reflect student self-perception, self-efficacy, and self-control across subjects (Klapp Lekholm & Cliffordson, 2009), these factors are mediated through teacher evaluations of student conduct and homework completion (Duckworth et al., 2012). Thus, these findings indicate that beyond assessment of the academic knowledge reflected in standardized test scores, what teachers assess with grades is student engagement, effort, participation and behavior, which reflect measures of student self-control and self-efficacy. This research postulates that it is these factors that give grades their predictive validity with overall schooling outcomes, since if grades are a valid measure of how well a student can negotiate the non-academic components of the schooling process, then it is these factors that predict later student ability to conform to the institutional expectations that lead to completing high school as well as post-secondary schooling and employment (Bowers, 2011; Brookhart et al., 2016). This issue is exemplified by Kelly (2008), who analyzed data from over 1,500 students across 115 middle school English and language arts classrooms and their teachers in Wisconsin and New York. The study included grading data as well as surveys of students and observation and video data from the classroom, making Kelly (2008) one of the most comprehensive and rich datasets analyzed to date in the grading literature. Using a hierarchical linear modeling framework, the author found that grades were strongly related to student participation and engagement,,and that higher grades appeared to be awarded for engaged participation, rather than "going through the motions". However, there were some differences by student background. As stated by Kelly (2008):

> This study found that in addition to achievement, effort and participation in class are important predictors of the grades that students receive. The chances of an average student receiving a high mark increase dramatically when the student is engaged in class and completes his or her assignments. It is important to note, though, that not every form of participation is rewarded by high marks. Using detailed data on participation in classroom discourse, it is possible to distinguish between procedural engagement ("going through the motions") and substantive forms of engagement... I found that only substantive engagement leads to higher

> grades. This finding suggests that most teachers successfully use grades to reward achievement-oriented behavior and promote a widespread growth in achievement. However, the grading process is not entirely meritocratic. Boys, low-SES students, and Hispanic students all receive lower grades than do other students. (p. XX)

In sum, across this research domain, grades have been shown to be a strong multidimensional assessment of both academic knowledge and student engaged participation in schooling, which then the latter is predictive of overall schooling outcomes (Brookhart et al., 2016). Assessment of engaged participation, then, is through teacher perception of student performance, which is susequently incorporated into grades. Indeed, these findings from the grading literature align well with the broader research on teacher expectations of students. For example, using the Education Longitudinal Study of 2002 (ELS:2002) Gregory and Huang (2013) show that positive teacher expectations predict schooling outcomes, such as college going, and are stronger predictors than many context and background variables (Gregory & Huang, 2013). As another example, in examining the difference between traditional "at-risk" predictors and teacher expectations from the NCES NELS:88 dataset of a nationally generalizable sample of students in grade 8 in 1988, Soland (2013) showed that:

> Generally, teachers were quite accurate at predicting student outcomes... This accuracy appears to have been driven largely by informational asymmetries, because teachers tend to rely on data related to student attitudes, behavior, and effort…(p. 246)
>
> Results concomitantly showed that teachers proved quite accurate in their predictions, often because they relied on academic tenacity data not easily captured in administrative datasets... Teachers naturally collect a huge amount of data, especially related to academic tenacity, simply by observing their students on a daily basis. (p. 259)

Thus, rather than subjective measures of a hodgepodge of factors, this literature clearly demonstrates that grades assess student engaged participation, that grades are predictive of overall outcomes, and that it is important in this research to take teacher perceptions of student performance into account when examining the relationship between grades and test scores. Nevertheless, while this rich literature provides a strong argument for the validity of grades as a multidimensional assessment, one area that has not been explored in depth is the question of the variance in grades across schools. The between school issue is an issue that relates directly to the reliability and validity of grades. For instance, if there is a strong between-school effect on grades, then which school a student attends would then largely determine that student's grades. Conversely, if the variance between schools in student grades is low, then the interpretation would be that the vast majority of schools grade students on similar scales and for similar reasons. One interpretation of a difference in grades at the school level could be the issue of grade inflation. Yet, research that has used the multiple large-scale nationally generalizable NCES decadal surveys has found no grade inflation is evident in K-12 schooling in the US (Pattison, Grodsky, & Muller, 2013). Nevertheless, little of the research on grades has examined the between-school variance in grades to examine the relationship of student background, test scores, and teacher perception of student performance, while controlling for the nested dependent nature of students nested in schools. If a large amount of the variance in grades lies between schools, this could pose a strong validity threat to this literature on the multidimensional validity of grades as useful assessments in schools.

Testing the Claims and Questions from the Literature on Grades

In this section of the chapter, I apply the literature discussed above to examine the extent to which teacher assigned grades are a useful assessment of student engagement, using a large nationally generalizable sample of U.S. grade 10 high school students. This section examines three main aspects of this issue. First, to date, while the standardized grading practices literature claims that grades are unreliable and subjective measures that vary too much across schools to be useful, very little research has been done to examine the extent to which grades actually do vary within and between schools. Second, while critics of standardized assessments note that socio-economic status and ethnicity are strongly associated with test scores, little work has been done to examine the extent to which grades, test scores and SES are related, and to what extent grades may be a fairer, or more "just" assessment that does not vary as strongly by SES or the demographic background of the student as do standardized assessments. Third, once these two main issues are addressed (within/between school variance and student SES/background variables) with control variables, the remaining variance in grades that is not explained by standardized test scores can be examined to show the extent that teacher evaluation of student effort (e.g., participation and behavior) is associated with the grades they assign, and whether this assessment is consistent across schools, and thus perhaps more reliable than previously inferred from the past psychometrics literature.

To examine these issues, I analyzed the restricted use Education Longitudinal Study of 2002 (ELS:2002) dataset. ELS:2002 was originally collected by the National Center for Education Statistics (NCES), in which about 15,400 U.S. grade 10 students across 750 school in 2002 were surveyed on a large array of items concerning their high school experience, as well as collecting demographic information, standardized assessments in mathematics and reading that were aligned to NAEP and PISA, and student report card grades and overall GPA (Ingles et al., 2007). In addition, NCES surveyed the student's English and mathematics teachers from the 2001/2002 academic year asking the teachers about each student's performance in their courses.

As noted in Table 1, for this analysis I included the non-cumulative grade point average across all courses for students in grade 10 as well as grade 10 mathematics and reading standardized tests scores and a range of student and school background variables as well as teacher ratings of student engagement. In addition, because ELS:2002 is not a simple random sample, but is a probabilistic complex sample, I applied the sampling weights to allow for generalization to all three million students who were in grade ten in the U.S. in 2002. Due to the restricted nature of the data, all sample sizes are rounded to the nearest ten.

For my variable selection I drew on the literature in this domain reviewed above, particularly relying on previous research on teacher perception and grades using the ELS:2002 dataset, such as Gregory and Huang (2013). At the student level I included perceptions from both English teachers and mathematics teachers as the previous research in this area has shown that while these perception variables are moderately related at about a 0.5 correlation, they performed well independently in the previous research when loaded into the same equation (Gregory & Huang, 2013). At the school level, previous research has indicated that grades may be related to school-level factors, such as student demographics and school size (Roderick & Camburn, 1999). For the analysis, to examine the issues outlined above in grades across schools I used Hierarchical Linear Modeling (HLM) (Hox, 2010; Raudenbush & Bryk, 2002) in SPSS (Heck, Thomas, & Tabata, 2012) to examine two models with fixed effects. For both HLM analyses, the dependent variable is non-cumulative grade 10 GPA, which is the average of a student's grades across all subjects from only grade 10. In each model I control for student and school context and background variables, as well as student mathematics and reading achievement. In the second model, I add teacher perception of student performance using the variables outlined in Table 1.

The analysis resulted in three main findings. First, while the unconditional HLM indicated that there is a statistically significant amount of variance in grades between schools (Wald Z = 13.390, $p<0.001$), the interclass correlation coefficient (ICC) shows that only 16.52% of the variance in grade 10 GPA is between schools. This indicates that less than a fifth of the variance in grades is between schools as indicated by the variables in the data base. As noted in the literature review and framing above, if there is a large effect on grades depending on which school a student attends, the hypothesis would be that how teachers grade students is related to which school those teachers and students are in, which would throw into doubt the literature on the usefulness of grades as assessments of engaged participation in schooling since this difference would manifest through between-school variance. The ICC result suggests that there is a small amount of variance in grades between schools. This indicates that while there is some relationship between which school a student attends and the grades that the student receives, the vast majority of the variance (83.48%) is at the student, rather than school level.

Second, Table 2 presents the results of the two HLM analyses. For each coefficient for each model, I first present the coefficient for each variable (Coeff.), followed by the standardized coefficient (β), which can be interpreted as the effect size, followed by the standard error (SE). In Model A, only student mathematics and reading achievement, student background, and school-level background and context variables are included, which account for 36.83% of the variance at the student level and 45.54% of the variance at the school level. In Model B, English and mathematics teacher ratings of student effort, participation and behavior explained an additional 33.17% of the variance in grade 10 GPA at the student level and an additional 13.49% at the school level (subtract Model B variance explained from Model A at each level). These results indicate that controlling for test scores, and background and demographic variables at the student and school level, teacher evaluations of student effort, participation and behavior make up a significant portion of what grades represent.

Third, in examining the individual parameter estimates in the full final Model B in Table 2, the only significant ethnicity variable is Native American, and the standardized coefficient (beta) for SES is relatively small, in stark contrast to the literature on these variables as they relate to standardized test scores. In contrast to previous research (Kelly, 2008), I find no evidence that Hispanic students have significantly lower grades controlling for the other variables in Model A or Model B. The estimates of multiple other variables are of interest. As an example, in replication of multiple studies in the grading literature (DiPrete & Buchmann, 2013; Kelly, 2008; Lewis & Willingham, 1995; Thorsen & Cliffordson, 2012), females received higher grades on average than males (0.108 grade points) controlling for the other variables in the model. For teacher perceptions of student performance for both English and mathematics teachers, these variables confirm much of the literature on student engaged participation being strongly related to student grades. Strong positive predictors were "student works hard for good grades", "how often student completes homework", and "how often student is attentive in class". Interestingly, for English teachers, "how often student is tardy" and "how often student is disruptive in class" were not significantly related to grades, whereas both of these variables were significantly related to grades for mathematics teachers. Mathematics teacher perception of tardiness for mathematics classes was negatively related to student grades as expected, however, student disruptions were positively related with a small effect size.

While Model B explained 70% of the 83.5% of the variance at the student level, Model B also explained over half (59%) of the 16.5% of the variance at the school level. At the school level, context and demographics of the student body were significantly related to individual student grades. For negative relationships, students in schools with a higher percentage of minority students, and larger enrollment schools receive lower grades.

Table 1: Descriptive statistics from analyses of ELS data

	Mean	(SD)	Min	Max	ELS:2002 variable label and description
GPA for all 10th grade courses	2.67	0.87	0	4	F1GPA10: Non-cumulative grade 10 GPA all courses
Grade 10 Mathematics	50.71	9.91	19.38	86.68	BYTXMSTD: Grade 10 mathematics stand. T-score
Grade 10 Reading	50.53	9.89	22.57	78.76	BYTXRSTD: Grade 10 reading stand. T-score
SES	0.03	0.74	-2.12	1.87	F1SESR: Student socio-economic status
Female	0.50	0.50	0	1	BYSEX = 1 (male ref. group)
African American	0.17	0.38	0	1	BYRACE2 = 1
Student is Hispanic	0.15	0.35	0	1	BYS15 = 1
Asian	0.13	0.33	0	1	BYRACE3 = 1
Hawaiian/Pacific Islander	0.02	0.14	0	1	BYRACE4 = 1
Native American	0.04	0.21	0	1	BYRACE5 = 1
English is native language	0.83	0.38	0	1	BYSTLANG = 1
Non-Traditional family	0.41	0.49	0	1	BYFCOMP > 1: Both birth parents not present in home
English Teacher rating					
Student works hard for good grades	0.69	0.46	0	1	BYTE04: 0=no, 1=yes
How often student completes homework	3.01	1.01	0	4	BYTE13: 0=never, 1=rarely, 2=some of the time, 3=most of the time, 4=all of the time
How often student is absent	1.16	0.72	0	4	BYTE14: (same as previous)
How often student is tardy	0.63	0.84	0	4	BYTE15: (same as previous)
How often student is attentive in class	2.95	0.88	0	4	BYTE16: (same as previous)
How often student is disruptive in class	0.59	0.87	0	4	BYTE17: (same as previous)
Mathematics Teacher rating					
Student works hard for good grades	0.68	0.47	0	1	BYTM04: 0=no, 1=yes
How often student completes homework	2.99	1.02	0	4	BYTM13: 0=never, 1=rarely, 2=some of the time, 3=most of the time, 4=all of the time
How often student is absent	1.15	0.70	0	4	BYTM14: (same as previous)
How often student is tardy	0.58	0.80	0	4	BYTM15: (same as previous)
How often student is attentive in class	2.96	0.89	0	4	BYTM16: (same as previous)
How often student is disruptive in class	0.55	0.84	0	4	BYTM17: (same as previous)
School-level variables					
Urban	0.34	0.47	0	1	URBAN = 1 (rural ref. group)
Suburban	0.34	0.47	0	1	URBAN = 2 (rural ref. group)
% Free Lunch	24.51	19.13	0	96.2	CP02PLUN
% Minority students	34.36	31.20	0	100	CP02PMIN
Student teacher ratio	16.62	4.25	4.39	40	CP02STRO
Enrollment (in thousands)	1.27	0.84	0.02	4.64	CP02STEN/1000

Table 2: Hierarchical linear models explaining grade 10 GPA of ELS data

	Model A			Model B		
Parameter	Coeff.	β	SE	Coeff.	β	SE
Student-level variables						
Grade 10 Mathematics	0.032 ***	0.371	0.001	0.021 ***	0.235	0.001
Grade 10 Reading	0.015 ***	0.168	0.001	0.009 ***	0.103	0.001
SES	0.166 ***	0.142	0.011	0.085 ***	0.073	0.010
Female	0.303 ***	0.175	0.013	0.108 ***	0.062	0.012
African American	-0.066 **	-0.029	0.021	-0.013		0.020
Hispanic	-0.019		0.027	0.039		0.025
Asian	0.088 *	0.034	0.034	0.054		0.032
Hawaiian/Pacific Islander	-0.062		0.054	-0.054		0.057
Native American	-0.092 **	-0.022	0.030	-0.064 *	-0.015	0.027
English is native language	-0.147 ***	-0.064	0.026	-0.015		0.025
Non-Traditional family	-0.133 ***	-0.076	0.014	-0.054 ***	-0.031	0.012
English Teacher rating						
Student works hard for good grades				0.208 ***	0.111	0.018
How often student completes homework				0.153 ***	0.179	0.009
How often student is absent				-0.088 ***	-0.074	0.010
How often student is tardy				0.008		0.009
How often student is attentive in class				0.055 ***	0.055	0.010
How often student is disruptive in class				-0.008		0.008
Mathematics Teacher rating						
Student works hard for good grades				0.163 ***	0.088	0.018
How often student completes homework				0.144 ***	0.169	0.009
How often student is absent				-0.077 ***	-0.062	0.010
How often student is tardy				-0.028 **	-0.025	0.009
How often student is attentive in class				0.064 ***	0.066	0.010
How often student is disruptive in class				0.030 **	0.029	0.008
School-level variables						
Urban	-0.076		0.046	-0.051		0.042
Suburban	-0.023		0.036	-0.006		0.032
% Free lunch	0.004 **	0.086	0.001	0.004 ***	0.096	0.001
% Minority students	-0.002 *	-0.062	0.001	-0.002 **	-0.084	0.001
Student Teacher ratio	0.006		0.004	0.011 **	0.053	0.004
Enrollment in thousands	-0.099 ***	-0.096	0.022	-0.081 ***	-0.078	0.021
Intercept	0.325		0.083	-0.230 **		0.087
Percentage of variance explained						
at student level	36.83			70.00		
at school level	45.54			59.03		
BIC	22000.13			9211.69		

However, there were also two significant positive findings, with students in schools with higher percentages of free and reduced price lunch students receiving higher grades as well as students who attend schools with larger student teacher ratios. While the effect sizes are small, these two positive relationships perhaps indicate that teachers in poorer schools and schools with larger student teacher ratios give slightly higher grades.

The Utility of Grades as Valid Classroom Assessments in Educational Measurement

As noted in the first section, throughout the literature and from the analysis discussed in this chapter, teacher assigned grades include assessment of student engaged participation as well as academic knowledge. However, also noted in the literature, is a lack of attention to the question of the extent to which grades vary across schools (do your grades depend to a large part on which school you attend?), how grades may vary based on school context and demographics (such as do richer schools give higher grades?), how student demographics relate to grades (such as do grades vary by demographics like test scores?), and finally, how teacher perceptions of student classroom performance relate to grades (testing the engaged participation component of grades). Overall, across the literature and the analyses presented in this chapter, the evidence suggests that teacher assigned grades are a useful and consistent measure of student engaged participation across schools, with little variance between schools in grades, a perhaps fairer distribution in relation to student demographics and SES in comparison to standardized tests, and that teacher perceptions of engaged participation account for a large percentage of what grades assess. I consider each issue in turn throughout this final section of the chapter.

In considering the issue of the extent that grades vary between schools, while there is a statistically significant proportion of variance in grades at the school level, it is relatively small. As noted in the literature in the first section, an area that has lacked attention in the grading literature has been the issue of examining between-school variance. If a large amount of the variance in grades is between schools, then which school you attend determines to some extent student grades. I find that there is weak evidence at best for this hypothesis. It does not appear that which school a student attends determines to a large extent the student's grades. In comparison, the proportion of variance between schools for standardized test scores has long been reported to be around 25% (Borman & Dowling, 2010; Coleman, 1990; Rumberger & Palardy, 2005). This suggests that the vast majority of the variance in grades is at the student or classroom level. Indeed, I recommend further research in this area, as research in the grading literature has indicated variability at the classroom level. For instance, Kelly (2008) notes:

> I found a strong contextual effect of classroom achievement level on grades, where a student's chances of receiving a high grade improve if she or he is in a lower-achieving class. This frog-pond type effect of being high achieving compared to one's classmates is quite strong. For both high- and low-achieving students, being in a classroom where students are low achieving substantially increases the chances of receiving an A. A likely explanation for this phenomenon is that grading is a relativistic process; teachers' expectations of students' performance are conditioned by experiences in the classroom. (p. 45)

This quote is a strong indication that additional research is needed in this area, as perhaps a three-level model would provide additional information on this issue, nesting students in classrooms in schools. If there is a strong classroom effect, across multiple classrooms and averaged into a single GPA, this effect might wash out and not be detectable using a two level model of students in schools as presented in section two here, limited to the data that available in ELS:2002.

Nevertheless, I do identify four variables at the school level that are weakly related to grades, with small effect sizes. In contrast to the individual-level parameters which shows that higher SES students receive somewhat higher grades, controlling for the other variables in the model, students who attend poorer schools (as defined by higher percentages of free and reduced price lunch students) and students in schools with larger student teacher ratios receive slightly higher grades on average. These results may be an indication of the "frogpond" effect above, or perhaps are a weak indication of grade inflation for students attending under-resourced schools, or schools in historically disadvantaged contexts. I encourage future research in this domain.

At the student level, the analysis in the second section provides a good example of the effects noted in the literature. As with the previous literature discussed above (Brookhart et al., 2016), grades are a multidimensional assessment of both student academic achievement and engaged participation. In the analysis of the ELS data, both the mathematics and reading standardized assessment scores were significantly related to grade 10 GPA in the final model. Interestingly, for Model B, including teacher perception of student effort and participation explained about as much of the variance in grades as did test scores and demographics combined. Teacher perception of how hard a student works for good grades and how often the student completes homework had comparable magnitude of effect sizes to the mathematics and reading standardized assessments, a core component of grades noted throughout the literature.

However, how tardiness and disruption relate to grades is discussed much less in the literature. Of note, in the analyses reported here, for English teachers, perceptions of student tardiness and disruption to the classroom were not significantly related to student grades, while both of these variables were significantly related to grades for mathematics teachers. However, the disruptive to class variable for mathematics teachers was positive, which was unexpected. Perhaps, when controlling for the variance explained by all of the other variables in Model B, disruption may have a positive effect uniquely in mathematics, as mathematics achievement, working hard, completing homework, absences, tardiness and attentiveness are already controlled for. I encourage future research in this area.

And finally, I turn to the issue of how student demographics relate to grades, discussed in the literature and examined in Model B of the analyses presented in this chapter. First, for SES, the analyses replicate and agree with the previous research showing that teacher perceptions are stronger than SES when it comes to grading (Gregory & Huang, 2013), as the magnitude of the effect size for SES on grades is smaller than the teacher perception variables. However, there is a large reduction in the effect size for SES on grades depending what variables are included in the analyses. For example, some of the variance in grades that is explained by SES in Model A is taken up within the teacher perception variables in Model B. A much more profound example of this is demonstrated with African American and Asian students. In Model A, the coefficient for African American students is negative, while it is positive for Asian students, controlling for other variables in the model. When controlling for teacher perception of student performance in Model B, these two variables are no longer significant. I interpret this in two ways. First, it may be that teacher perception is in effect an equalizer, making grades "fairer" than test scores, as test scores are strongly related to student demographics, even when controlling for internal school and teacher processes and perceptions (Rumberger & Palardy, 2005). Alternatively, a second explanation may be that the variance that was contributing to the negative coefficient for African American students on grades and the positive coefficient for Asian students in Model A can then be attributed to teacher perception in Model B. Indeed, there is a long-running debate in education research on teacher expectations and self-fulfilling prophecies (Madon, Jussim, & Eccles, 1997; Raudenbush, 1984). It may be that if there is a significant bias in teacher perceptions of students based on student ethnicity, then the results of this study may indicate that this bias perhaps acts through teacher perception of student hard work, homework completion, absences, tardiness, attentiveness, and disruption in class. I encourage future research in this area.

Conclusion and Implications

While some of the past literature has claimed that grading is "hodgepodge," in this chapter I have discussed the literature and an analysis framework that demonstrates that teacher assigned grades include student engaged participation that does not vary extensively by school. Additionally, of the variance within and between schools, the variables nominated in the literature that I included in the analysis in this chapter explain the vast majority of the variance in grades, both at the student level and between schools. This leads me to three main implications. First, it appears that in comparison to standardized tests cores, less of the variance in grades is between schools (16.5% here) than it is for tests (usually reported to be around 25% in the literature). Thus, in comparison to standardized tests, for grades it matters even less which school a student attends. Overall there does not appear to be strong evidence for "easy grading" or "hard grading" schools. However, as noted in both sections above, the classroom level may be a different story, as individual classes may have very skewed grading ranges (such as honors high school English). But overall, I interpret these findings to suggest that teachers are fairly consistent in how they grade in the aggregate across schools in the U.S. This can be seen as an argument for the reliability of grades.

Second, teacher perception of student engaged participation makes up a large portion of grades. When I define engaged participation as the teacher's perception of how hard students work for good grades, homework completion, absence and tardiness, attentiveness, and class disruptions, these account for more than half of the variance explained in grade 10 GPA. These components of engaged participation mirror those that teachers note across the surveys discussed earlier in this chapter when teachers are surveyed about what they award grades for.
Together, these results mirror recent findings from over 100,000 students' grades in Chicago Public Schools (Allensworth & Luppescu, 2018), in which the authors looked primarily at the relationship of attendance (as a proxy for participation) and test scores to grades. As noted by Allensworth and Luppescu (2018):

> School-level variance is almost completely explained by observable factors. This suggests some degree of consistency in assigning grades among education professionals; the standards for grades across schools may not be as arbitrary as is often believed. Rather than finding large unexplained differences in grades based on which school a student attends, or which teacher they have, we find there are observable factors that systematically explain most of the differences in the grades that students receive in different types of schools, and with different teachers... the factors that are most strongly associated with differences in students' GPAs are their course attendance and tested skills. (p. 31)

Thus, given this literature and the analysis in this chapter, I argue for the usefulness of grades as accurate assessments of classroom engaged participation. In combination with standardized test scores, grades provide a valuable means to understand both student academic achievement as well as their levels of engaged participation in the schooling process. In the work of schools in helping to promote student success and transitions throughout primary, secondary and post-secondary schooling and into careers, ensuring that grades and test scores are included together in a balanced conversation about supporting student performance and success is vital to ensuring that schools promote a focus on both academic achievement and engaged participation.

Suggested Citation:

Bowers, A.J. (2019) Towards Measures of Different and Useful Aspects of Schooling: Why Schools Need Both Teacher Assigned Grades and Standardized Assessments. In Brookhart, S., McMillian, T. (Eds.) *Classroom Assessment as Educational Measurement.* National Council on Measurement in Education (NCME) Book Series (p.209-223). New York: Routledge.

REFERENCES:

Allensworth, E. M. (2005a). Dropout rates after high-stakes testing in elementary school: A study of the contradictory effects of Chicago's efforts to end social promotion. *Education Evaluation and Policy Analysis, 27*(4), 341-364. doi:10.3102/01623737027004341

Allensworth, E. M. (2005b). Graduation and dropout trends in Chicago: A look at cohorts of students from 1991 through 2004. Retrieved from www.consortium-chicago.org/publications/p75.html

Allensworth, E. M., & Easton, J. Q. (2005). *The on-track indicator as a predictor of High School graduation*. Retrieved from www.consortium-chicago.org/publications/p78.html

Allensworth, E. M., & Easton, J. Q. (2007). *What matters for staying on-track and graduating in Chicago public high schools: A close look at course grades, failures, and attendance in the freshman year*. Retrieved from Chicago: http://www.consortium-chicago.org

Allensworth, E. M., & Luppescu, S. (2018). *Why do students get good grades, or bad ones? The influence of the teacher, class, school, and student* Retrieved from Chicago: IL: https://consortium.uchicago.edu/sites/default/files/publications/Why%20Do%20Students%20Get-Apr2018-Consortium.pdf

Atkinson, R. C., & Geiser, S. (2009). Reflections on a Century of College Admissions Tests. *Educational Researcher, 38*(9), 665-676. doi:10.3102/0013189x09351981

Attewell, P., Heil, S., & Reisel, L. (2011). Competing Explanations of Undergraduate Noncompletion. *American Educational Research Journal, 48*(3), 536-559. doi:10.3102/0002831210392018

Barrington, B. L., & Hendricks, B. (1989). Differentiating characteristics of high school graduates, dropouts, and nongraduates. *Journal of Educational Research, 82*(6), 309-319.

Battin-Pearson, S., Abbott, R. D., Hill, K. G., Catalano, R. F., Hawkins, J. D., & Newcomb, M. D. (2000). Predictors of early high school dropout: A test of five theories. *Journal of Educational Psychology, 92*(3), 568-582. doi:10.1037/0022-0663.92.3.568

Blazer, C. (2012). *National trends in end-of-course assessment programs*. Retrieved from

Borman, G. D., & Dowling, M. (2010). Schools and Inequality: A Multilevel Analysis of Coleman's Equality of Educational Opportunity Data. *Teachers College Record, 112*(5), 1201-1246.

Bowers, A. J. (2009). Reconsidering grades as data for decision making: More than just academic knowledge. *Journal of Educational Administration, 47*(5), 609-629. doi:10.1108/09578230910981080

Bowers, A. J. (2010a). 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.

Bowers, A. J. (2010b). Grades and Graduation: A Longitudinal Risk Perspective to Identify Student Dropouts. *Journal of Educational Research, 103*(3), 191-207. doi:10.1080/00220670903382970

Bowers, A. J. (2011). What's in a grade? The multidimensional nature of what teacher-assigned grades assess in high school. *Educational Research and Evaluation, 17*(3), 141-159. doi:10.1080/13803611.2011.597112

Bowers, A. J., & Sprott, R. (2012). Examining the Multiple Trajectories Associated with Dropping Out of High School: A Growth Mixture Model Analysis. *Journal of Educational Research, 105*(3), 176-195. doi:10.1080/00220671.2011.552075

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*(2), 77-100.

Brennan, R. T., Kim, J., Wenz-Gross, M., & Siperstein, G. N. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review, 71*(2), 173-215.

Brighouse, H., Ladd, H., Loeb, S., & Swift, A. (2018). *Educational Goods: Values, Evidence, and Decision-Making*. Chicago, IL: University of Chicago Press;.

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice, 10*(1), 35-36.

Brookhart, S. M. (1993). Teachers' Grading Practices: Meaning and Values. *Journal of Educational Measurement, 30*(2), 123-142. doi:10.1111/j.1745-3984.1993.tb01070.x

Brookhart, S. M. (1994). Teachers' Grading: Practice and Theory. *Applied Measurement in Education, 7*(4), 279-301. doi:10.1207/s15324818ame0704_2

Brookhart, S. M. (2011). *Grading and Learning: Practices that support student achievement*. Bloomington, IN: Solution Tree Press.

Brookhart, S. M. (2015). Graded Achievement, Tested Achievement, and Validity. *Educational Assessment, 20*(4), 268-296. doi:10.1080/10627197.2015.1093928

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research, 86*(4), 803-848. doi:10.3102/0034654316672069

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1995-1996). Teachers' assessment practices: Preparation, isolation and the kitchen sink. *Educational Assessment, 3*(2), 159-179.

Cliffordson, C. (2008). Differential Prediction of Study Success Across Academic Programs in the Swedish Context: The Validity of Grades and Tests as Selection Instruments for Higher Education. *Educational Assessment, 13*(1), 56-75. doi:10.1080/10627190801968240

Coleman, J. S. (1990). *Equality and achievement in education*. San Francisco: Westview Press.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72.

DiPrete, T. A., & Buchmann, C. (2013). *The Rise of Women: The Growing Gender Gap in Education and What it Means for American Schools* New York: Russell Sage Foundation.

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What *No Child Left Behind* leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology, 104*(2), 439-451. doi:10.1037/a0026280

Farr, B. P. (2000). Grading practices: An overview of the issues. In E. Trumbull & B. Farr (Eds.), *Grading and reporting student progress in an age of standards* (pp. 1-22). Norwood: Christopher-Gordon Publishers.

Gregory, A., & Huang, F. (2013). It Takes a Village: The Effects of 10th Grade College-Going Expectations of Students, Parents, and Teachers Four Years Later. *American Journal of Community Psychology, 52*(1-2), 41-55. doi:10.1007/s10464-013-9575-5

Guskey, T. R., & Jung, L. A. (2012). Four steps in grading reform. *Principal Leadership, 13*(4), 22-28.

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2012). *Multilevel modeling of categorical outcomes using IBM SPSS*. New York, NY: Routledge

Hodara, M., & Cox, M. (2016). *Developmental education and college readiness at the University of Alaska*. Retrieved from Washington, DC: http://ies.ed.gov/ncee/edlabs/projects/project.asp?projectID=393

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (Second ed.). New York: Routledge.

Ingles, S. J., Pratt, D. J., Wilson, D., Burns, L. J., Currivan, D., Rogers, J. E., & Hubbard-Bednasz, S. (2007). *Education longitudinal study of 2002: Base-year to second follow-up data file documentation*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Jones, E. B., & Jackson, J. D. (1990). College Grades and Labor Market Rewards. *The Journal of Human Resources, 25*(2), 253-266. doi:10.2307/145756

Kelly, S. (2008). What Types of Students' Effort Are Rewarded with High Marks? *Sociology of Education, 81*(1), 32-52. doi:10.1177/003804070808100102

Klapp Lekholm, A. (2011). Effects of School Characteristics on Grades in Compulsory School. *Scandinavian Journal of Educational Research, 55*(6), 587-608. doi:10.1080/00313831.2011.555923

Klapp Lekholm, A., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: effects of gender and family background. *Educational Research and Evaluation, 14*(2), 181-199.

Klapp Lekholm, A., & Cliffordson, C. (2009). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation, 15*(1), 1-23. doi:10.1080/13803610802470425

Labaree, D. F. (1997). Public goods, private goods: The American struggle over educational goals. *American Educational Research Journal, 34*(1), 39-81. doi:10.3102/00028312034001039

Levin, H. M. (2013). The Utility and Need for Incorporating Noncognitive Skills Into Large-Scale Educational Assessments

The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), (pp. 67-86): Springer Netherlands.

Lewis, C., & Willingham, W. W. (1995). The Effects of Sample Restriction on Gender Differences. *ETS Research Report Series, 1995*(1), i-57. doi:10.1002/j.2333-8504.1995.tb01648.x

Linn, R. L. (1982). Ability testing: Individual differences, prediction, and differential prediction. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 335-388). Washington DC: National Academy Press.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Lloyd, D. N. (1978). Prediction of school failure from third-grade data. *Educational and Psychological Measurement, 38*(4), 1193-1200.

Maag Merki, K., & Holmeier, M. (2015). Comparability of semester and exit exam grades: long-term effect of the implementation of state-wide exit exams. *School Effectiveness and School Improvement, 26*(1), 57-74. doi:10.1080/09243453.2013.861353

Madon, S., Jussim, L., & Eccles, J. (1997). In search of the powerful self-fulfilling prophecy. *Journal of Personality and Social Psychology, 74*(2), 791-809. doi:10.1037/0022-3514.72.4.791

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

Miller, S. R. (1998). Shortcut: High School Grades As a Signal of Human Capital. *Educational Evaluation and Policy Analysis, 20*(4), 299-311. doi:10.3102/01623737020004299

Nichols, S. L., & Berliner, D. C. (2007). *Collateral Damage: How High Stakes Testing Corrupts America's Schools*. Cambridge: Harvard Education Press.

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. doi:10.3102/0013189x13481382

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology, 76*(85-97).

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Roderick, M., & Camburn, E. (1999). Risk and recovery from course failure in the early years of High School. *American Educational Research Journal, 36*(2), 303-343. doi:10.3102/00028312036002303

Rumberger, R. W., & Palardy, G. J. (2005). Test scores, dropout rates, and transfer rates as alternative indicators of high school performance. *American Educational Research Journal, 42*(1), 3-42. doi:10.3102/00028312042001003

Soland, J. (2013). Predicting High School Graduation and College Enrollment: Comparing Early Warning Indicator Data and Teacher Intuition. *Journal of Education for Students Placed at Risk (JESPAR), 18*(3-4), 233-262. doi:10.1080/10824669.2013.833047

Swineford, F. (1947). Examination of the Purported Unreliability of Teachers' Marks. *The Elementary School Journal, 47*(9), 516-521. doi:10.2307/3203007

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation, 18*(2), 153-172. doi:10.1080/13803611.2012.659929

Warren, J. R., Jenkins, K. N., & Kulick, R. B. (2006). High School Exit Examinations and State-Level Completion and GED Rates, 1975 Through 2002. *Educational Evaluation and Policy Analysis, 28*(2), 131-152. doi:10.3102/01623737028002131

Welsh, M. E., D'Agostino, J. V., & Kaniskan, B. (2013). Grading as a Reform Effort: Do Standards-Based Grades Converge With Test Scores? *Educational Measurement: Issues and Practice, 32*(2), 26-36. doi:10.1111/emip.12009

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F. O., & Gabrieli, J. D. E. (2016). Promise and Paradox. *Educational Evaluation and Policy Analysis, 38*(1), 148-170. doi:doi:10.3102/0162373715597298

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