UCHICAGO Consortium on School Research

RESEARCH REPORT JUNE 2023

Lasting Differences

Math Grades and Gender

John Q. Easton and Briana Diaz

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The authors would like to thank each other for a productive partnership on this paper and a previous one. We have complementary skills and temperaments, making this work enjoyable. We have also worked with several other wonderful research partners over the past several years. These include (in alphabetical order): Silvana Freire, New York University; Naureen Kheraj, Chicago Public Education Fund; and Lauren Sartain, University of North Carolina, Chapel Hill. We would like to give a particular thank you to our qualitative coders, Karlyn Gehring, Rebecca Silverman, and Allison Swimmer. They all played significant roles in the development of this research. We thank our many reviewers at the Consortium and from its Steering Committee, plus Eliza Mueller, Network for College Success, and Barton Dassinger, Principal, Chavez School.

We couldn't have done the work without our partners at Chicago Public Schools, especially LaTanya McDade, former Chief Education Officer, who endorsed the research, and Jared Sell who worked tirelessly to help us obtain the Gradebook data.

Thanks to Jessica Tansey, who provided critical input and feedback throughout the production of this report, Allison Swimmer for her concise write-up of analyses, and Alex Usher and Jessica Puller for thorough and thoughtful review.

This research was generously funded by an anonymous family foundation, the Hewlett Foundation, the Hyman Milgrom Opportunity Fund at the University of Chicago, and the Consortium Investors Council that funds critical work at the UChicago Consortium and whose members include: Brinson Foundation, CME Group Foundation, Crown Family Philanthropies, Lloyd A. Fry Foundation, Joyce Foundation, Lewis-Sebring Family Foundation, Mayer & Morris Kaplan Family Foundation, McCormick Foundation, McDougal Family Foundation, Polk Bros. Foundation, Spencer Foundation, Steans Family Foundation, Square One Foundation, the Chicago Public Education Fund, the Vivo Foundation, and two anonymous foundations.

Cite as: Easton, J.Q. & Diaz, B. (2023). *Lasting differences: Math grades and gender.* Chicago, IL: University of Chicago Consortium on School Research.

This report was produced by the UChicago Consortium's publications and communications staff: Jessica Tansey, Managing Director of Research Communications; and Jessica Puller, Senior Communications Strategist.

Graphic Design: Jeff Hall Design
Photography: Christian Sutter
Editing: Jessica Tansey and Jessica Puller

06.2023/PDF/jh.design@rcn.com

Introduction

Students' course grades matter. Research studies show that students' grades are more predictive than test scores of their future academic success, including high school and post-secondary outcomes.[1] A rigorous national study clearly connected connected students' higher GPAs with higher college graduation rates.[2] Specific to Chicago, several studies from the University of Chicago Consortium on School Research (UChicago Consortium) have shown that students with As and Bs had more positive long-term outcomes than their peers with Cs or lower.[3]

Chicago Public Schools (CPS) recognizes the importance of grades through its longstanding "Bs or Better" campaign. Yet within CPS, boys' grades are consistently lower than girls' grades. For example, ninth-grade young women's math grades were 2.66 vs. young men's 2.33—roughly the difference between a B- and a C+—across all students in this study in 2016–17 and 2017–18. This trend is not isolated to Chicago; as early as 2004, the national average GPA for twelfth-graders was 2.96 for young women vs. 2.72 for young men.[4,5]

CPS leaders and educators want to meet their goal of Bs or better and strong educational outcomes for all students, and they are in conversation about how they're supporting boys. Of particular importance in CPS is the grade performance of Latino and Black boys—together, they comprise more than three-quarters of the boys in CPS, yet they have lower high school and college completion rates than their male peers and girls of other races/ethnicities,[6] which suggests that there is room for educators and leaders to provide more equitable educational experiences districtwide. The CPS Five-Year Vision explicitly discusses the need to improve outcomes for Latino and Black boys.[7]

This study aims to provide insights into this difference in boys' and girls' grades in Chicago by looking first at one grade level and one subject area—ninth-grade math (algebra and geometry)—using two school years of data, 2016–17 and 2017–18.[8] Our findings are specific to math students in these years, and also suggest what may (and may not) be driving differences in other grades and subjects.

Our research questions were informed by informal conversations with and the hypotheses of teachers and school leaders both inside and outside of CPS. While we were not able to test every hypothesis we heard, we tested those we could, given available data, through two core research questions (RQ).

1 Bowen, Chingos, & McPherson (2011); Roderick, Nagaoka, Allensworth, Coca, Correa, & Stoker (2006); Bowers et al. (2013).
2 Bowen et al. (2011).
3 Easton, Johnson, & Sartain (2017); Allensworth, Gwynne, Moore, & de la Torre (2014); Roderick et al. (2006); Allensworth & Clark (2020).
4 DiPrete & Buchmann (2012).
5 Additionally, the new book *Of Boys and Men* by Richard Reeves, contains an entire chapter on educational differences, entitled *Girls Rule, Boys are Behind in Education*. Reeves (2022).
6 DiPrete & Buchmann (2013).
7 Chicago Public Schools (n.d.b.)
8 We limited the scope of our analysis due to the very large size of our database.

RQ1: Are students' demographics, behaviors, and school experiences related to gender differences in grades?

- **RQ1a.** Are gender differences in grades similar across different racial/ethnic groups? (**See Figure 1.**)
- **RQ1b.** Are students' prior achievement, their ninth-grade attendance and/or ninth-grade out-of-school suspensions related to the gender difference in grades? (**See Figures 2 and 4.**)
- **RQ1c.** Are students' survey responses about academic effort and work, social connections, and their math classroom experiences related to the gender difference in grades? (**See Figures 3 and 4.**)

RQ2: How do teachers' grading practices, including choice of grading categories and their weightings, influence differences in grades by gender?

- **RQ2a.** On average, what weights do teachers assign to different grading category families? (**See Figure 5.**)
- **RQ2b.** How does course (algebra or geometry) and level (regular or honors) placement affect category family weighting by gender? (**See Table 2.**)
- **RQ2c.** Do gender differences differ by grading category family (assessments, assignments, behavior, or other)? (**See Figures 6 and 7.**)
- **RQ2d.** How do category family weights (for assessments, assignments, behaviors, or other) relate to the gender difference in final grades?

Preview of findings

This research report is primarily intended for CPS audiences, including teachers, school leaders, administrators, and policymakers. We understand that there are many conversations occurring throughout CPS regarding grading policies and *grading for equity*.[9] The findings in this report do not provide a simple explanation for why young women earned higher grades than young men in ninth-grade math grades. We did eliminate possible drivers of the difference:

- Young women's grades were higher even when we compared ninth-grade young men and women with similar school experiences (as measured by administrative data and survey measures) and previous test scores (**RQ1**).
- Young women's grades were higher in every grading category family (assessments, assignments, behavior, or other), and especially in assignments, when we compared young men and women across all math teachers' ninth-grade grading category families (**RQ2**).
- The weights that teachers applied to different grading categories had a small, but meaningful, influence on the size of the grade difference between young women and men (**RQ2**).
  - > Specifically, young women were more often in honors classes and geometry, where grading category family weights differed from weights in algebra and regular classes.

We hope our findings can inform educators' hypotheses, conversations, and interventions, and serve as a starting place for future researchers. Beyond CPS, other school districts and researchers may also find useful both our findings and our systematic and detailed look at how an electronic gradebook system was used by 398 ninth-grade math teachers.

9 See Chicago Public Schools (2022), which we provide as a resource, not an endorsement. Additional resources for standards-based grading can be found at https://tguskey.com/toms-books/

Data

We analyzed math grades for first-time ninth-graders in the 2016–17 and 2017–18 school years for all CPS students enrolled in traditional public schools (non-charter and non-Options[10] schools). Charter school students' grades are not included because they are not contained in the centralized CPS student information system.[11] Options schools students are not included because they have different graduation requirements and course taking patterns.

We used three distinct but complementary sources of information to help us understand students' grades and to consider why grades of young men and women differ:

Data Source 1:

Administrative data (2016–17 and 2017–18)

This study used administrative records from CPS that included students' race/ethnicity, gender (*see Notes on "gender" on p.5),* eighth-grade test scores, high school attendance, records of out-of-school suspensions, and final ninth-grade math grades (A, B, C, D, or F) for two cohorts of students in SY 2016–17 and SY 2017–18. In addition, we had access to Gradebook data from SY 2016–17 (see *Data Source 2* section for details).

Data Source 2:

5Essentials Survey responses, including supplemental measures (2016–17 and 2017–18):[12]

We examined self-reports of students' "academic effort and work," "social well-being,"[13] and their experiences in their ninth-grade math classes (**see Appendix A** for a listing of the items in all of the measures used in these analyses). These three categories reflected the behaviors and experiences we expected to be most related to the gender difference in ninth-grade math grades.

- **Academic Effort and Work:** Combination of "study habits," "grit," and "academic engagement" measures. The questions in these measures asked students about their own behaviors, for example:
  - > I set aside time to do my homework and study.
  - > I continue steadily toward my goals.
  - > I work hard to do my best in this class.
- **Social Well-Being:** Combination of "emotional health" and "school connectedness" measures. The questions in these measures asked about students' relationships with others and experiences specifically related to their school, for example:
  - > I can always find a way to help people end arguments.
  - > Other students in my school take my opinions seriously.
- **Math Instruction:** Combination of "course clarity," "high-quality math instruction," "course rigor," "academic press," and "academic personalism" measures. These questions in these measures focused on students' experiences in their math class(es), for example:
  - > I know what my teacher wants me to learn in this class.
  - > My teacher encourages students to share their ideas about things we are studying in class.
  - > My teacher wants us to become better thinkers, not just memorize things.
  - > My teacher gives me specific suggestions about how I can improve my work in this class.

10 Options schools are designed to provide an alternative learning environment for students who are not thriving in traditional high schools.

11 Charter school students constituted 31% of first-time ninth-graders in 2016–17 and 2017–18.

12 Student response rates were above 80% in both years.

13 These aggregate measures have been used successfully in previous research. See Jackson, Porter, Easton, Blanchard, and Kiguel (2021).

Data Source 3:

Gradebook data (2016–17 only)

We used the Gradebook electronic teacher grading platform that was in use in CPS in 2016–17; Gradebook data included all information that contributed to a student's final letter grade.[14] With Gradebook, we could thus ascertain whether the differences between young men and young women differed by "grading category families." The Gradebook data set is the single largest set of files that the UChicago Consortium has ever used.

CPS and the Chicago Teacher's Union (CTU) issued guidance[15] around grading, including expectations and best practices in 2017, but teachers had, and continue to have, wide discretion over their grades and gradebook structure. Teachers were able to choose or create these grading categories to reflect the types of work required of students.

Using a rigorous coding system, our team grouped these grading categories into four mutually exclusive "category families": assessments; assignments; behaviors; and "others." See examples in **Table 1**. Despite great effort, the coding team was ultimately unable to differentiate formative from summative assessments; **Appendix B** describes the coding process. We also examined the weights that teachers assigned to these category families.

Gradebook automatically calculates final grades, using each school's grading rubric (e.g., 80-89=B, 90-100=A, etc.). The calculation of final total points earned is conducted by summing points earned times weight for each category title selected by the teacher.

For example:

Assessments (*Points earned X Weight*) + Assignments (*Points earned X Weight*) + Behavior (*Points earned X Weight*) = Final Total Points

Or in real terms:

Assessments (*90 X .45*) + Assignments (*80 X .45*) + Behavior (*90 X .10*) =

Assessments (*40.50*) + Assignments (*36*) + Behavior (*9*) = 85.5

Notes on datasets:

For RQ1, we use transcript data from SY 2016–17 and SY 2017–18, in addition to demographic and survey data from those years. For RQ1a, we use "points earned" on a scale of 0 to 100 from Gradebook. For RQ1b-c, we use standardized "GPA points": a value between 0 and 4 (A=4.0, B=3.0, C=2.0, D=1.0, F=0). GPA points and survey data were standardized so we could compare values to each other.

For RQ2, we analyzed Gradebook data from only SY 2016–17, because CPS transitioned to ASPEN the following year and we did not have access to the 2017–18 dataset. This portion of the analysis was restricted to algebra and geometry classes and excluded "mastery-based (standards and competency) grading categories."[16] We used "points earned" on a scale of 0 to 100 for RQ2, except for the analysis described in the box titled *Gender differences remained statistically significant even after controlling for different course enrollments*; we used standardized points for that regression model.

TABLE 1

Examples of categories coded into category families

Category family	Assessments	Assignments	Behavior	Other
Gradebook category examples	Quizzes	Homework	Class participation	Content mastery
	Assessments	Assignments	Professionalism	Synthesis
	Tests	Classwork	Organization	Practice/preparation
	Exams	Classwork/homework	Executive function	Accountability

Note: The four category families in this table were used to group 709 unique category titles in the Gradebook dataset; common category titles within each category family are listed as examples. Depending on the section, some categories with similar names were coded into different families. This is because coders looked at specific task names within the grading categories to make final decisions on category families

14 Obtaining this large and complex data set required exceptional effort from both CPS and Consortium staff, and we are grateful to all who made it possible. Our 2022 CPS Gradebook Technical Report provides additional detail. CPS transitioned to ASPEN for electronic grading in 2017–18.

15 Chicago Public Schools (2017).

16 We excluded these grading systems because we did not know what was required of students other than mastering a given state learning standard.

Notes on "gender"

Given the specific focus on gender in this report, it is important to acknowledge two choices:

1. **Terms:** When referring to all CPS students, we refer to *boys* and *girls*; when referring specifically to ninth-grade students, we refer to *young* men and *young* women. In doing so, we intend to use age-appropriate language for the ninth-grade students in our data.
2. **Categories:** Historically, CPS collected data that grouped students into one of two gender categories: male and female. Some students do not fit into one of these categories, but we believe that there are still insights to be gained from analysis of this data. We hope in the future to be able to report data that more fully describes the identities of CPS students. Starting in 2020–21 the gender categories in the CPS demographic questionnaire were: male, female, and non-binary.

Both of our choices are imperfect; we consider the work here a start, not an end for inquiries of the relationship between gender and grades.

Key Findings

Research Question 1

Are students' demographics, past and present behaviors, and school experiences related to gender differences in grades?

1a: Are gender differences in grades similar across different racial/ethnic groups?

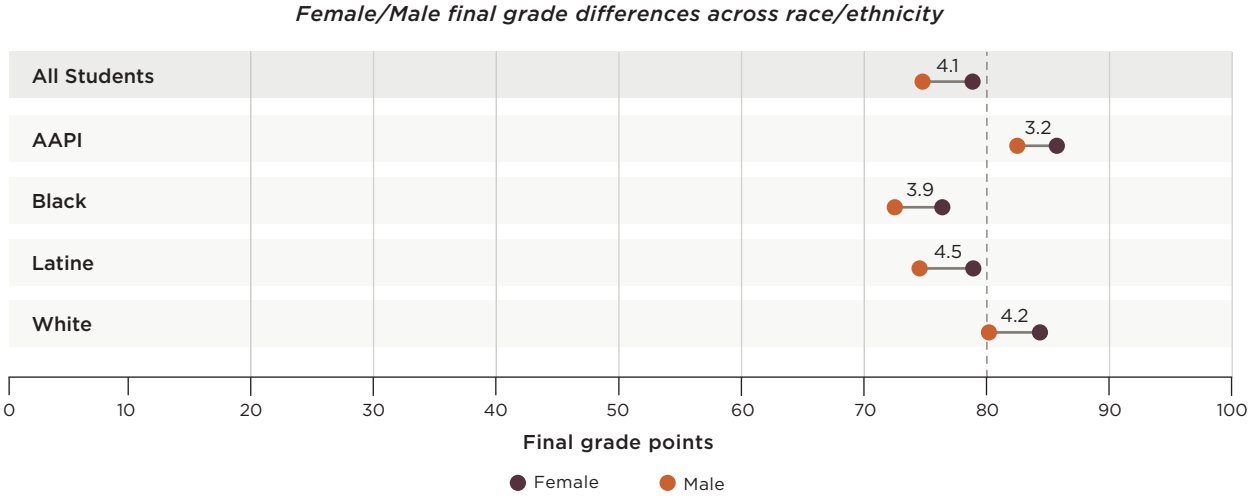
Finding: Gender differences in grades were similar for students across races/ethnicities.

We were often asked, as we worked on this study: "Are gender differences in grades similar or different across race/ethnicity?" CPS educators are acutely aware that the district enrolls primarily Latine and Black students (46.5% and 35.8%, respectively),[17] and are concerned about creating more equitable outcomes for students of color. Here, we found a consistent pattern for the four race/ethnicity categories that had large enough student populations to analyze and report (**see Figure 1**).

The overall gender difference among all students was 4.1 points on the 0 to 100-point grading scale from Gradebook. Young men earned fewer points than young women in all racial/ethnic groups; the difference between young women and young men in each racial/ethnic group ranged from a low of 3.2 among AAPI students, to a high of 4.5 among Latine students. Despite the variation in the gender difference, we did not find a statistically significant difference across

FIGURE 1

Grade differences in ninth-grade math were similar for students of different races/ethnicities



Note: This figure shows gender differences in final grade points (final grade points range from 0-100 and map to a final letter grade of A-F) for all CPS ninth-grade algebra and geometry students in school years 2016–17 and 2017–18 (29,229 students total). The "All Students" includes the total enrollment including students from several racial/ethnic groups that have too few students to separate out (Native American, Multiracial, or students with no race/ethnicity noted) and includes 1,030 students (3.5% of our sample). The four largest groups with which there are large enough student samples to perform statistical tests (Latine, Black, White, and Asian American/ Pacific Islander [AAPI]) are represented in the figure. Our AAPI category combines three CPS data categories—Asian, Pacific Islander/Hawaiian and Asian/Pacific Islander categories—due to the small number of students in the latter two categories.

17 Chicago Public Schools (n.d.a.).

racial/ethnic subgroups. Put differently, this means that the difference in grades between young women and young men across racial/ethnic groups is statistically the same. (**See the ANOVA output in Table A.3 in Appendix A** for details.)

Two points remain notable in **Figure 1**. First, the 4.1-point difference—in young women's scores minus young men's grades—fell near the B- vs. C+ grade for all students, which has implications for students' future outcomes. Second, while the size of the gender difference in Black and Latine students is not statistically different from their White and AAPI peers, in terms of actual grades, Black and Latino young men's grades were lowest among their peers; improving supports for these young men is critical.

1b: Are students' prior achievement, their ninth-grade attendance and ninth-grade out-of-school suspensions related to the gender difference in grades?

Finding: Prior achievement, absences, and out-of-school suspensions do not explain the grades difference.

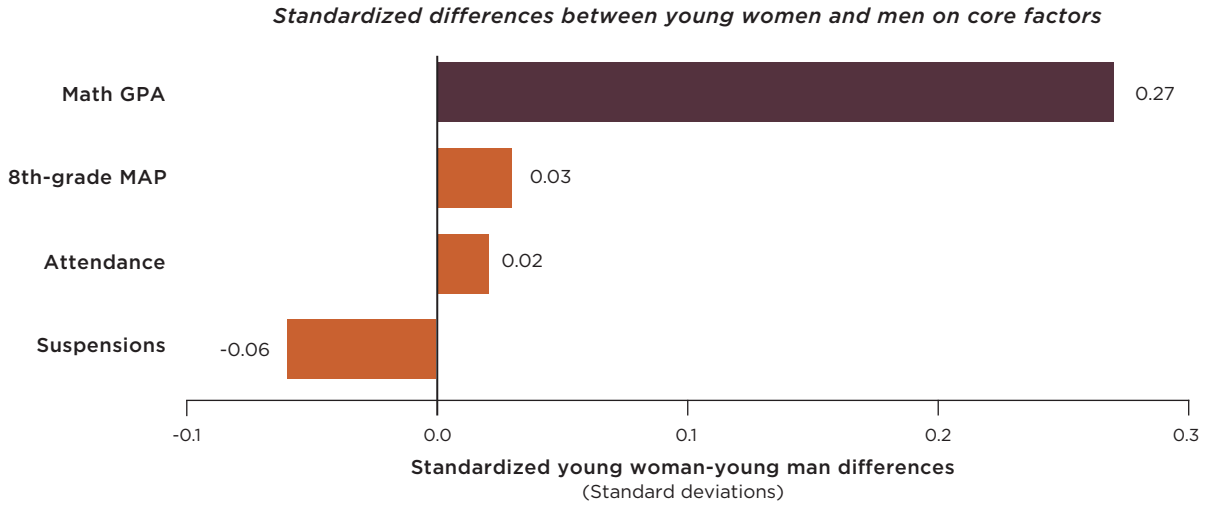
Young women scored 0.27 standard deviation units above young men in ninth-grade math classes (equivalent to 0.30 of a grade point).[18] This is similar to the 0.33 grade point units between ninth-grade young women and young men overall, in cohorts 2016–17 and 2017–18, as noted in the introduction. The numbers aren't identical because the 0.33 grade point difference is based on the full sample, and the 0.27 SD is computed from a smaller, more select sample: students with survey responses.

Young women entered ninth grade with slightly higher math standardized test scores, they attended school more often, and they received out-of-school suspensions considerably less often than young men. These differences are displayed in **Figure 2**, using standard deviations units to make it possible to compare across different metrics on the same scale.

We expected that these differences may explain the grades difference, because we know that attendance, prior achievement, and out-of-school suspensions are related to grades. **Yet when we used statistical techniques to compare young women and young men who had the same rates of prior test scores, attendance, and out-of-school suspensions, young women still had higher math grades.**[19] In fact, the difference after these comparisons was not statistically significantly different from the original difference. **Figure 4** contains the regression coefficients and standard errors for each of the variables we consider here.

FIGURE 2

Young women had slightly higher test scores and attendance rates and had fewer out-of-school suspensions, and much higher math grades than young men



Note: This figure shows the difference in scores between young men and young women (29,229 students) on non-survey measures: eighth-grade MAP performance, attendance, and out-of-school suspensions. Math GPA gender difference is included as a comparison measure. Values above the 0 line indicate higher performance for young women, values below the 0 line indicate higher performance for young men. Notably, for suspensions, this means that young men's suspension rates were higher than young women's.

18 This analysis switches to standard deviation units rather than the actual value for each variable so that we can compare scores or values on different metrics. For example, we cannot compare test scores to attendance rates without putting them onto the same scale. The standard deviation units tell us where each student lies in the overall distribution for each variable (e.g., prior achievement, attendance, and out-of-school suspensions).

19 All regression outputs can be seen in the Appendix, Table A.4.

1c: Are students' survey responses about academic effort and work, social connections, and their math classroom experiences related to the gender difference in grades?

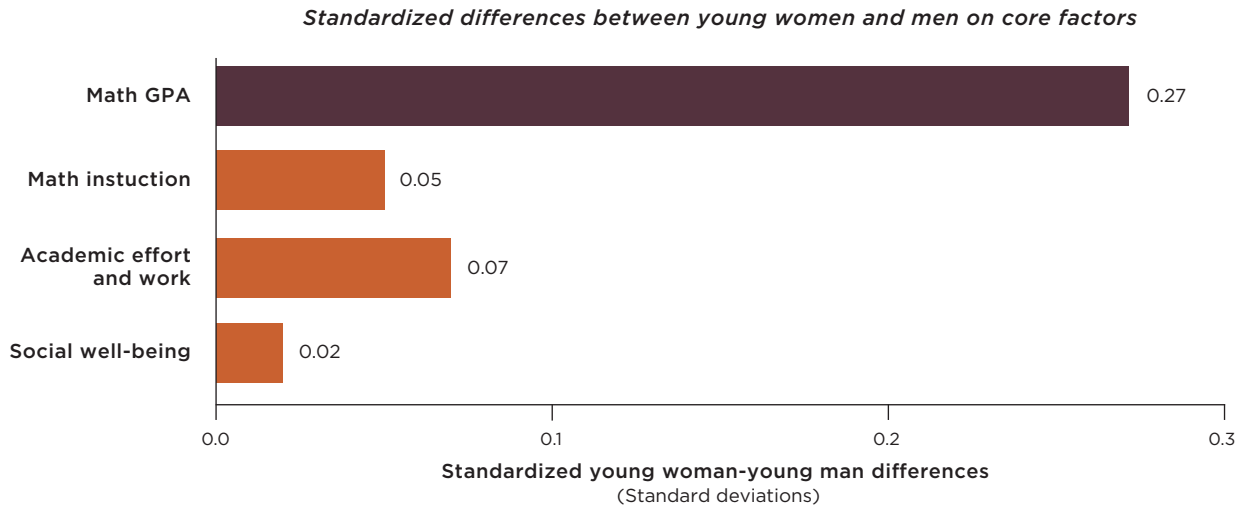
Finding: Student's self-reports of school and math classroom experiences don't explain the difference in grades.

On the annual *5Essentials* Surveys, ninth-grade young women reported better school experiences across all three categories we evaluated—academic effort and work, social well-being, and math instruction (**see Figure 3**). Yet when we used statistical techniques to compare young women and young men who had similar survey reports, young women still had higher math grades (**see Figure 4**). (Again, we used standard deviation units to be able to compare different metrics to each other; **see Table A.4 and Appendix A**.)

This finding is similar to Finding 2. **In fact, when we only compared young women and young men who were similar across all our variables, the difference in their grades was smaller than the overall difference, but the difference remained large and significant (see Figure 4, Models 1 and 7).**

FIGURE 3

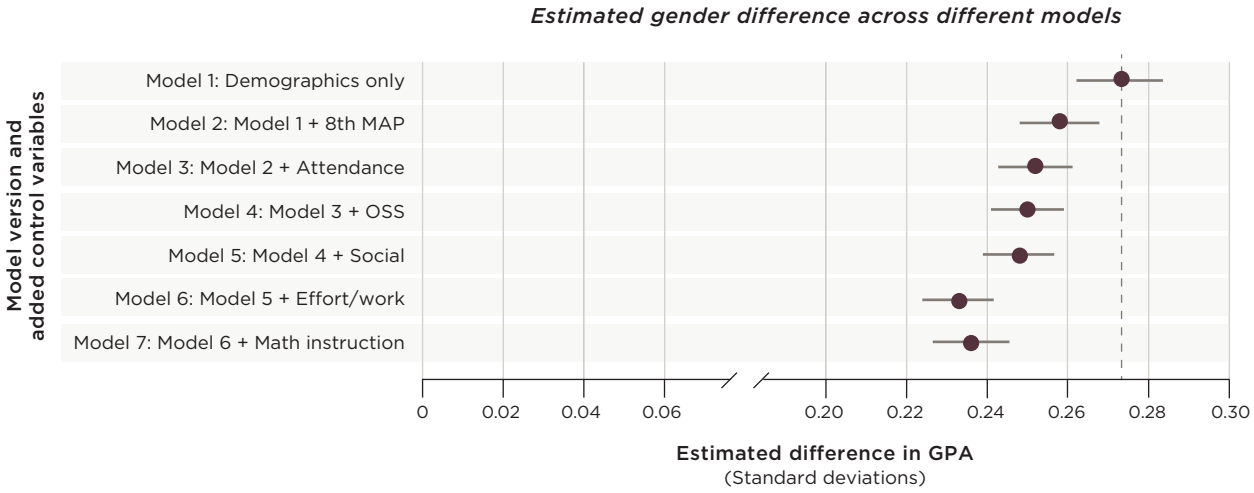
Compared to young men, young women reported stronger positive self-conceptions of work habits, social well-being, and quality of math instruction



Note: This figure shows the difference in scores between young men and young women (29,229 students) on aggregated survey measures: academic effort and work, social well-being, and math instruction. Math GPA is included as a comparison measure. Values above the 0 line indicate higher performance for young women, values below the 0 line indicate higher performance for young men.

FIGURE 4

After adding controls across multiple models, the gender difference became somewhat smaller but still meaningful



Note: This figure shows the coefficient of the male indicator variable (representing the average GPA difference between young women and young men) across different regression models (28,517 students were included across all models, fewer than the 29,229 students in other analyses because of students who did not respond to survey measures used in this analysis.) All models control for student demographics (e.g., race/ethnicity, socioeconomic status, English learner status). Each new model adds a new explanatory variable incrementally. "OSS" refers to out-of-school suspensions. The circle shows the value of the estimated GPA difference, the whisker shows the standard error of the estimate (an approximation of where the true value of the estimate will be 95% of the time). The dashed vertical line is placed at the value of the demographics and gender only model (base model) to aid in comparing other model estimates. See notes in "Data Source 2" on p.3 for additional details.

How to read Figure 4

Figure 4 displays estimates for the gender difference across different statistical models, each of which use an increasing number of statistical controls—incorporating additional variables that we thought might be in part responsible for the gender difference in grades. For ease of comparison, only the estimate of the gender difference is shown for each model. Following the base model (gender and demographics only), each model adds a new variable incrementally: first gender and demographics plus eighth-grade MAP scores; then gender, demographics, and eighth-grade MAP plus attendance, and so on. For example, Model 4: Model 3 + out-of-school suspensions, estimates the gender difference controlling for demographics, eighth-grade MAP scores, attendance, and out-of-school suspensions.

The basic model, only comparing average differences between genders with no controls estimates a gender difference of 0.273 GPA points. The model with all controls, Model 7 estimates a gender difference of 0.236 GPA points. There was a small but statistically significant difference between these model estimates, indicating that adding the control variables explain a small portion of the gender difference, but the difference largely remains.

Research Question 2

How do teachers' grading practices influence differences in grades by gender?

Before diving into RQ2 findings, it is helpful to note that the *The Professional Grading Standards and Grading Practices Guidelines For Chicago Public Schools Teachers* states that:

> *"The primary function of grading is to provide feedback related to student academic achievement expressed through the Illinois Learning Standards and/or learning objectives for each course of study undertaken. Grades are captured through formative and summative assessments and are intended to represent a fair and honest indication of a student's present level of academic mastery at a given point in time... Assignments and assessments are measured using clear criteria that connect with the standards-based objectives...* ***The net result, once grades are entered, is a grade that captures student performance on actual standards or curricular goals and not on disconnected or compliance-oriented tasks."*** [20] (emphasis added)

What goes into creating this "fair and honest indication" of students' learning is nuanced and a discussion topic in many schools. On the one hand, grades reflect multiple factors valued by teachers, and research has clearly shown that it is this multidimensional quality that makes grades good predictors of important outcomes.[21]

Understanding teachers' grading systems and practices is key to understanding how students will fare after they leave the classroom. Teachers know that how they categorize and weight their grades matters. (Weighting here refers to teachers' weights within the electronic grading system—not the weights that may be applied to honors or Advanced Placement courses.) When a teacher puts more emphasis on tests vs. homework vs. in-class discussion participation, it affects different students' grades—in ways that may or may not accurately reflect their learning, and that may open or limit access to future opportunities. Yet the full effects of grading category selection and weighting is largely unknown to many teachers and principals.

2a: On average, what weights do teachers assign to different grading category families?

Finding: Teachers gave high weights to assessments and assignments, and much lower weights to behavior and other grading category families.

Because of the high weights assigned to assessments and assignments in these ninth-grade math classes, these two grading category families contributed the most to final grades for most students and for gender differences in math grades. Contrary to many anecdotes, differences in behavior points earned tended to play little role in the overall gender differences for most students because behavior accounted for so little of most students' grades (**see Figure 5**).[22]

These were average weights applied to different categories of grades across 1,599 sections (individual classrooms) of ninth-grade algebra and geometry classes, and across 398 teachers. However, individual teachers varied greatly in how much weight they placed on different categories for their grades. Within the assignments category family, for example, the weights ranged between 0% and 100%. And although the typical 57% weight of assessments shown in **Figure 5** exceeded the recommendation, that no grading category exceed 50% of the total weight, this category family may be composed of multiple grading categories (e.g., quizzes, tests, formative/summative) that teachers entered in Gradebook. **See Figure A.2 in Appendix A** for full category family distribution.

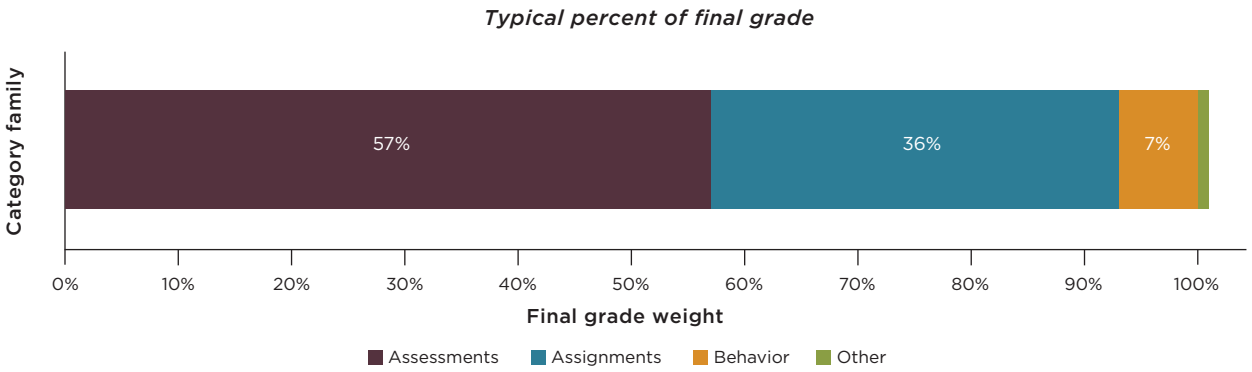
20 See Chicago Public Schools (2017).

21 See Brookhart et al. (2016) for a review of grading research.

22 Teachers in our sample tended to place little weight on behavior for final grades. However, there was a very small number of sections where teachers placed the majority or all of the final grade weight on behavior (see Figure A.2 in Appendix A). For students in these classes, behavior was an outsized proportion of final grades.

FIGURE 5

On average, behavior accounted for 7% of a student's final grade



Note: In addition to having a very low weight (1%), the "other" category family was rarely used by teachers and often contained category titles that our coders could not interpret. 29,229 students are included in this analysis. See Table 1 for examples of what grading categories are included in each grading category family. Component rates, as labeled, do not sum to 100 due to rounding.

2b: How does course (algebra vs. geometry) and level (regular vs. honors) placement affect category family weighting by gender?

Finding: Young men and young women's selection and placement into different math classes created different category grade weights across gender.

Figure 5 shows average patterns across all ninth-grade math courses. But teachers typically placed higher weights on the assessment category family in geometry and honors courses. In contrast, teachers tended to place higher weights on the assignments category family in algebra and regular-level courses. There was no statistically significant difference in the weight placed on behavior category families across classes, although honors and geometry classes had lower weights on behavior than regular and algebra classes did.

Young women's enrollments outnumbered young men's in geometry and honors courses, while young men were relatively overrepresented in algebra and regular-level courses. The dataset did not explain why young women outnumbered young men in the advanced courses, but as we noted previously (**Figure 1**), as measured by standardized test scores, young women tended to enter high school slightly better prepared than young men.

As a result of the different category family weighting and enrollment across courses, young women's final grades were weighted slightly differently on average than young men's final grades. The average weight differences across all young men and young women can be seen in the **Table 2**. See, for example, the assignments category. On average, young men had a weighting of 41.46 applied to their assignment scores; young women had a weighting of 39.03.

The "All Students" column in **Table 2** shows us that young men and young women experienced different weightings in both assessments and assignments, the most used and most heavily weighted category families. Young men had higher weights in assignments than young women because they were most likely to be placed in regular and algebra classes, classes where teachers placed higher weights on assignments. Young women were more likely to be placed in honors and geometry classes, where teachers placed lower weights on assignments. The opposite pattern held with assessments. Young women had higher weightings in this category family because they were more likely to be enrolled in geometry and honors classes that weighted assessments more heavily, and young men were more likely to be enrolled in regular and algebra classes that weighted assessments less highly. These differences in weightings between young women and young men in algebra vs. geometry are shown in the two columns "Algebra" and "Geometry."

While not statistically different because of smaller numbers, behavior was weighted more highly in algebra

and regular classes than in geometry and honors classes. The importance of these weighting differentials is described within RQ2d. Note that the weights for the behavior category differ from **Figure 5 to Table 2**. The average weight for all teachers for the behavior category family was 7%, yet the number here is about twice that. Not all teachers used the behavior category family, but the ones who did assigned higher weights to it—approximately 15.25% overall (not shown). The "other" category family is anomalous—it was rarely used overall (1%), but when used, weighted highly.

TABLE 2

Category family weights differ slightly by gender because of differential placement into math classes and differential weightings by math classes

Category family	Gender	Average category weight toward final grade, with student counts (Out of 100 possible total points)					
		All math classes	All students	Algebra	Algebra students	Geometry	Geometry students
Assessments	Young men	58.73	13,337	56.47	11,566	73.48	1,771
	Young women	60.64	13,585	57.94	11,275	73.85	2,310
Assignments	Young men	41.46	13,545	43.47	12,018	25.63	1,527
	Young women	39.03	13,413	41.45	11,451	24.99	1,962
Behavior	Young men	15.81	6,382	16.32	5,559	12.35	823
	Young women	14.85	6,647	15.31	5,599	12.42	1,048
Other	Young men	48.50	379	49.07	363	35.63	16
	Young women	48.18	265	49.18	248	33.53	17

Note: These weights differ from the average weights shown in Figure 5 for two reasons. First, not all teachers used these four grading category families in their grading—which is why the sum of all listed categories is more than 100. Second, young men and young women are placed differentially into honors vs. regular, and algebra vs. geometry courses. Category families had different weights depending on course level and subject. "All students" represents the count of students who had the category family used in their final grade calculation of 29,229 students represented in this table. The numbers in the table are calculated including only the teachers who used that specific category family in their Gradebook. The "other" family is grayed out to indicate how seldom teachers used category families outside the assessments, assignments, or behavior. We do not show the patterns comparing regular vs. honors because they were so similar to the algebra vs. geometry comparison.

2c: Do gender differences differ by grading category family (assessments, assignments, behavior, or other)?

Finding: Young women outperformed young men in every grading category in both unweighted and weighted points, with one small exception.

Young women outperformed young men in each grading category in unweighted points (**see Figure 6**). ***Unweighted* points are the number of points earned in each grading category family, divided by points possible.**

The difference in unweighted category families between young women and young men were as follows (in descending order):

- Assignments, 6.13 (80.93 vs. 74.8)
- Behavior, 4.74 (87.64 vs. 82.9); and
- Assessments, 3.47 (75.04 vs. 71.57)

The difference was 6.04 in the *other* category (81.55 vs. 76.51)—but this category family was rarely used and can be considered anomalous.

The gender difference in assignments (largest difference) was almost twice the gender difference in assessments (smallest difference; **see Figure 6.**)

This changes when we look at gender differences in *weighted* points in **Figure 7. Weighted points are calculated by multiplying points earned in each grading category by the category weights** (which are then added together for a total score out of 100 possible points). The weighted category points are summed to create total final points, then converted to letter grades.

Figure 7 shows that, on average, young women's grades were 4.11 (77.98 vs. 73.87) higher in total weighted points on a 100-point scale. Specifically, the difference in weighted category families between young women and young men were as follows (in descending order):

- Assessments, 4.06 (42.87 vs. 38.81)
- Behavior, 0.18 (5.86 vs. 5.68)
- Assignments, 0.11 (28.54 vs. 28.43)

The difference in the *other* category was 0.24 (0.71 vs. 0.95), with young men having higher weighted points but as previously noted, it is an anomalous category.

To see a version of this analysis where we used statistical techniques to compare young men and women taking the same classes **see the box titled *"Gender differences remained statistically significant even after controlling for different course enrollments"* on p.16**, which shows that gender differences are smaller but still significant in more advanced courses.)

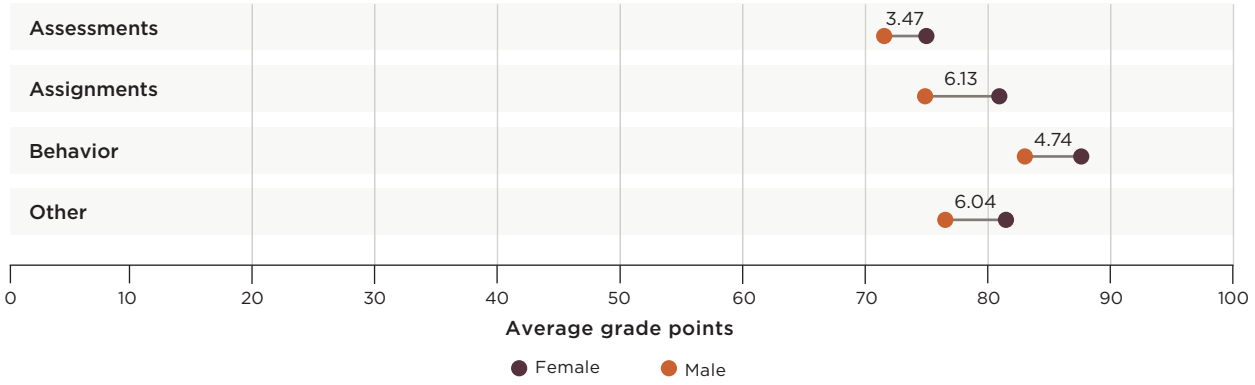
The differences between young women and young men for weighted points (**Figure 7**) vs. unweighted points (**Figure 6**) are smaller for assignments (0.11 vs. 6.13) and behavior (0.18 vs. 4.74)—but larger for assessments (4.06 vs. 3.47). Because of the high weights placed on assessments, the gender difference appears larger than in the unweighted data. Similarly, the lower weights on assignments show a smaller difference between young women and young men than the unweighted points do.

The overall 4.11 difference between young men and young women in total points approximates the 0.33 gender differences in grade points discussed previously (77.98=B- and 73.87 = C+). They differ because the Gradebook sample only contains one cohort of students and is restricted to algebra and geometry and the "grade point sample" includes two cohorts enrolled in all possible ninth-grade math classes.

FIGURE 6

Young women outperformed young men in all category families in unweighted points

Average unweighted points by category family

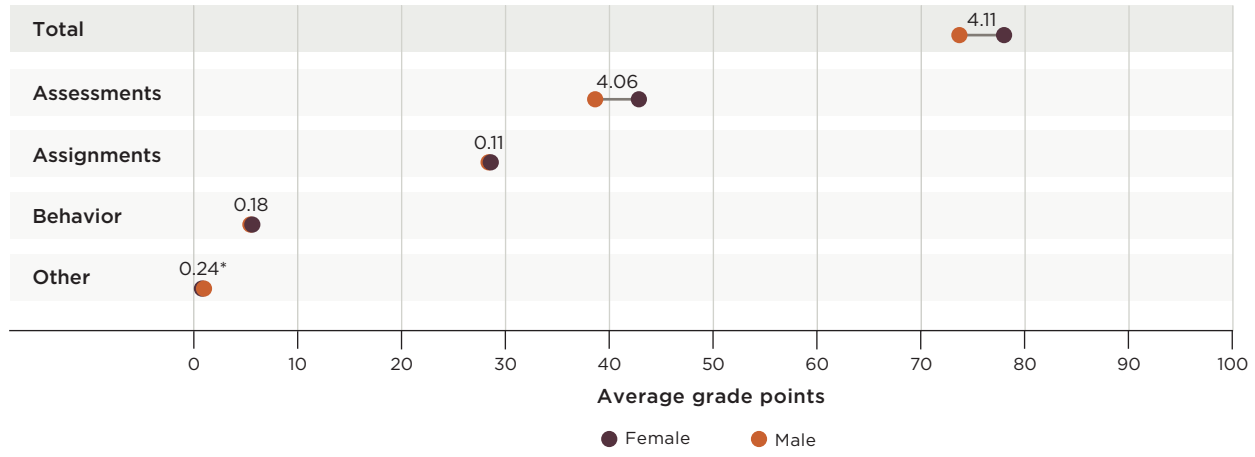


Note: This figure shows the typical unweighted points for each category family for young women and young men (29,229 students); the difference between them is labeled. This figure can be interpreted as: on average, how many points did a student earn in each category family before weights were applied? There is no "total points" category because each separate grading category ranges from zero to 100 points. Total points are only meaningful once the grading category weights are applied to the unweighted points, as in Figure 7.

FIGURE 7

With weights applied, the gender gap in assessments was larger

Average weighted points by category family



Note: This figure shows the typical weighted points for each category family for young women and young men (29,229 students); the difference between them is labeled. This figure can be interpreted as: on average, how much did points in each category family contribute to a student's final grade? Not all teachers used all category families, and there was wide variation in grade weights applied to category families as displayed in Figure A.2 in Appendix A. * The difference in the *other* category was 0.24 (0.71 vs. 0.95), with young men having higher weighted points but as previously noted, it is an anomalous category.

Gender differences remained statistically significant even after controlling for different course enrollments

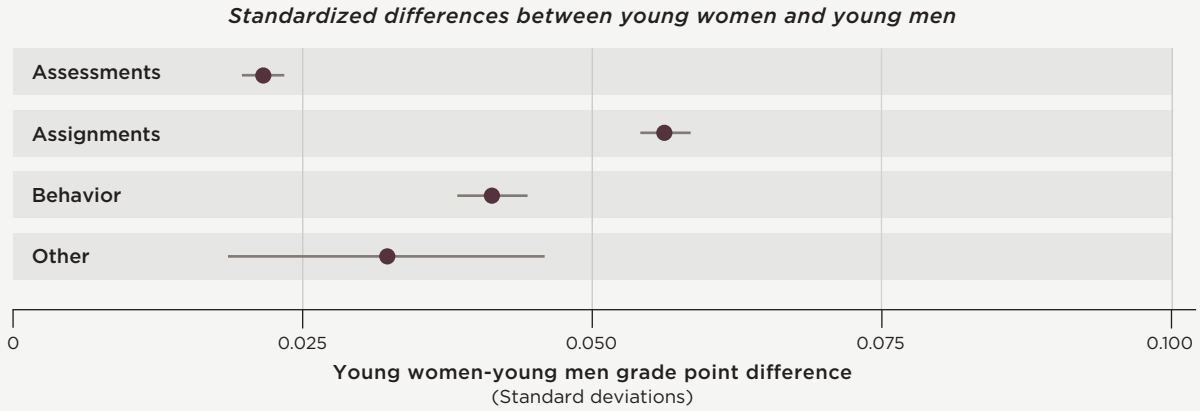
Gender differences are still significant when we control for course and level: if we compare young men to young women with the same prior achievement, taking the same course and level, there are still differences between them in points earned in the three major category families.

The analyses discussed in Figures 1 and 6 showed that young women outperformed young men in their final math grades (**Figure 1**) and in all Gradebook categories within their ninth-grade math classes (**Figure 6**). Young women outperformed young men in all category families after weights were applied for final grades, as well (**Figure 7**). But as we've noted, young men and young women were enrolled in different classes with different final grade weightings. So it is also important to ask: *What about young men and women enrolled with the same eighth-grade test scores, in the same ninth-grade math classes with the same weightings?*

We ran a series of regression equations that controlled for differences across classroom enrollments. The results of the unweighted points models are shown in **Figure A**. The dot shows the value of the estimated GPA difference, the dotted "whisker" line shows the standard error of the estimate (an approximation of where the true value of the estimate will be 95% of the time). For example, the assessments model estimated the size of the difference to be 0.0216 standard deviations in favor of young women, with a 95% confidence interval of +/-0.002, indicating that **even comparing very similar students, young women's math grades were still higher by a small but statistically significant amount.** Controlling for course enrollment patterns, the smallest gender difference in points earned is in the assessment category family. The largest difference is in the assignment category family.

FIGURE A

Young women's math grades were still higher by a small but statistically significant amount when comparing similar students



Note: This figure displays the coefficients of a male indicator variable (i.e., if the student was male) for different points-only regression models. In addition to the male indicator, models controlled for math course type (algebra/geometry) and math course level (honors/regulars) enrollment. Each model was run separately, but they are displayed together for comparison purposes. The dot represents the estimated value of the coefficient, and the dotted "whisker" line shows the error of the estimate (95% confidence interval). A smaller whisker line indicates less error than a larger whisker. 29,229 students are included in this analysis.

TABLE A

Ninth-grade math enrollment, by level and course

	Young Men				Young Women		
	Algebra	Geometry	Level total		Algebra	Geometry	Level total
Regular	8,323	299	8,622	**Regular**	7,319	394	7,713
Honors	3,353	1,491	4,844	**Honors**	4,145	1,936	6,081
Course total	11,676	1,790	**13,466**	**Course total**	11,464	2,330	**13,794**

Note: A single enrollment record includes either algebra or geometry, at either a regular or honors level. If a student was in double math (taking both algebra and geometry), they would have two records, one for each course.

2d: How do category family weights (for assessments, assignments, behaviors, or other) relate to the gender difference in final grades?

Finding: When higher weights were placed on assessments, it had the effect of shrinking, but not eliminating the GPA difference.

Throughout this report we have noted that young women outperformed young men in both unweighted and weighted points.[23] The difference between unweighted and weighted points is subtle, but important in determining final grades. Ultimately, teacher-assigned weights play an important role in determining students' final grades.

To summarize, we found that students' final grades were influenced by **three quasi-independent factors:**

1. The course and level students were assigned to;
2. The number of points the students earned in each grading category (percent of points possible in each category); and
3. The grading categories teachers chose and the amount of weight they assigned to each.

We saw the smallest difference between genders in unweighted[24] points in the assessments family (**see Figure 6 on p.15**). Yet once the weightings were applied, the assessments category displayed the largest gender difference (**see Figure 7 on p.15**). And when comparing similar students in similar classes, young men's points were closest to young women's in assessments (**see Figure A on p.16**).

On the other hand, young women most outperformed young men in unweighted points in the assignments family (**see Figure 6**). However, once the weightings were applied, the gender difference in assignments was reduced to near zero (**see Figure 7**).

Ultimately, young women took more advanced classes (geometry and honors) and their grades were lower than they would've been if they'd taken algebra and regular classes, effectively shrinking a potentially larger gender-grades difference. On the other hand, if more young men were in the geometry and honors classes that gave higher weights to assessments, we might see smaller gender differences, given the findings in **Figure A**.

23 There is one exception—the weighted "other" category. But the sample size is very small; few teachers used this category (see Table 2), so we do not highlight it here.

24 Unweighted = points earned divided by points possible.

Implications

Ultimately, we found that **gender differences in points earned and in final grades were similar across students of different races/ethnicities and were not driven by prior achievement, attendance, out-of-school suspensions, or school and math class experiences for the first-time ninth-graders in this study.** Course placement (algebra vs. geometry and regular vs. honors) and differential grading category family weightings played a small and complex role in gender differences in final grades. So, what do we do now?

CPS may benefit from renewing the *Professional Standards* guidelines created jointly with the CTU. Weighting decisions have subtle but important impacts on final grades and on gender differences. This research shows the great variability among teachers in weighting decisions. As we reported in an earlier paper,[25] few teachers follow the default grading categories and weights. Given the great interest in CPS for alternative grading systems (standards-based, mastery-based, and equity-focused grading) widespread discussion of these topics would be appropriate at this time to provide a common understanding of basic expectations for grading practices. Students may benefit from better understanding how their grades are affected by their teachers' grading categories and weights.

School communities districtwide may benefit from rich discussions about how to improve young men's experiences and outcomes in school. This research confirmed many anecdotal experiences: young women generally view their experiences in school more positively than young men. Not only do they enter ninth grade better prepared, but they have slightly better attendance and are much less likely to have out-of-school suspensions than young men. Young women report working harder on their schoolwork than young men, they feel somewhat more socially connected to schools, and they report better experiences in their math classes than young men.

How can we help young men experience school and their math classes better, and ultimately improve their outcomes? These may be longstanding questions, but they still prevail. A system-wide examination about school-by-school variability may be a helpful start, followed by considering what changes schools and educators can make. At the individual school level, instructional leadership teams could evaluate available data (e.g., *5Essentials* Survey results, *Cultivate* Survey results, grades, attendance, etc.) to understand the experiences of young men in greater detail, bring in teachers and students to discuss, and consider potential changes.

School administrators could examine grading practices within their buildings. One example of school-wide grading practices that school administrators could refer to the is the Chavez Grade Audit Report, developed jointly by a CPS school principal and a data specialist.[26] This report is available on the CPS network, accessible for CPS staff only, at https://co-ps-chavez-sites-w01.cps.k12.il.us/reports/

25 Diaz & Easton (2022).

26 Dassinger & Langworthy (2023).

It could be useful to examine and discuss course placement policy and effects, since it is a district goal to provide challenging and ambitious instruction to all students. A clear and transparent understanding of how and why placement decisions are made, at both a district- and school-level, may be helpful in understanding how to improve boys' school experiences and learning. A sizable number of young men may be prepared and able to succeed in a more advanced class. However, research has shown that simply assigning students to advanced coursework to increase equity in course-taking can actually lead to negative long-term outcomes—thus increasing advanced course enrollment for boys is a goal, providing supports, for both students and teachers, to ensure students are successful in those more difficult courses will be an important accompanying strategy.

Finally, researchers could provide additional quantitative and qualitative examinations of gender differences in grades. This study only looked at ninth-grade math in Chicago; other questions for investigation could include:

- What about other grade levels, subject areas, and districts?
- What does this look like in schools—what could interviews, focus groups, and observations tell us about the potential drivers and interventions for these differences?
- Do matches between students' and teachers' genders affect the patterns we found?
- Are boys' and girls' grades closer to one another in some schools?
  - If so, might we be able to point to particular practices within, or characteristics of, those schools to uncover promising ideas for other schools?

References

Allensworth, E.M., & Clark, K. (2020)
High school GPAs and ACT scores as predictors of college completion: Examining assumptions about consistency across high schools. *Educational Researcher, 49*(3), 198-211

Allensworth, E.M., Gwynne, J.A., Moore, P., & de la Torre, M. (2014)
Looking forward to high school and college: Middle grade indicators of readiness in Chicago Public Schools. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Bowen, W.G., Chingos, M.M., & McPherson, M. (2011)
Crossing the finish line: Completing college at America's public universities. Princeton, NJ: Princeton University Press.

Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M.T., & Welsh, M.E. (2016)
A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803-848.

Chicago Public Schools. (n.d.a.)
Demographics. Retrieved from https://www.cps.edu/about/district-data/demographics/

Chicago Public Schools. (n.d.b.)
Five year vision 2019-2024. Retrieved from https://www.cps.edu/link/7e1e31ba540540f49145e6227a388105.aspx

Chicago Public Schools. (2017)
Professional grading standards and professional practices *guidelines for Chicago Public School teachers.* Chicago, IL: Chicago Public Schools. Retrieved from https://www.cps.edu/globalassets/cps-pages/about-cps/policies/administrative-hearings/professional_grading_standards.pdf

Chicago Public Schools. (2022)
Grading for equity. Retrieved from https://www.cps.edu/sites/equity/tools/me/grading-for-equity/

Dassinger, B., & Langworthy, R. (2023)
Chicago Public Schools network. Retrieved from https://co-ps-chavez-sites-w01.cps.k12.il.us/reports/

Diaz, B., & Easton, J.Q. (2022)
CPS Gradebook technical report. Chicago, IL: University of Chicago Consortium on School Research.

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, NY: Russell Sage Foundation.

Easton, J.Q, Johnson, E., & Sartain, L. (2017)
The predictive power of ninth-grade GPA. Chicago, IL: University of Chicago Consortium on School Research.

Jackson, C.K., Porter, S.C., Easton, J.Q., Blanchard, A., & Kiguel, S. (2020)
School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights, 2*(4), 491-508.

Reeves, R. (2022)
Of boys and men. Washington, DC. Brookings Institution Press.

Roderick, M., Nagaoka, J., Allensworth, E., Coca, V., Correa, M., & Stoker, G. (2006)
From high school to the future: A first look at Chicago Public Schools graduates' college enrollment, college preparation, and graduation from four-year colleges. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Appendix A

Survey and Statistical Details

This appendix includes *5Essentials* Survey item details, as well as statistical tables and figures that provide additional information and support for the findings shared in the main text of the paper. We hope that this information is useful for readers who wish to delve more deeply into the data and analysis. While most of the tables and figures included share the results of statistical procedures, the last figure is purely descriptive, sharing the proportional breakdown of category families across all math sections we analyzed.

TABLE A.1

***5Essentials* Survey items included in analyses**

<table>
<tr><th colspan="2">Math class</th></tr>
<tr><td>Course clarity</td><td>Academic press</td></tr>
<tr><td>• I know what my teacher wants me to learn in this class.
• I learn a lot from feedback on my work.
• It is clear what I need to do to get a good grade.
• The homework assignments help me learn the course material.
• The work we do in class is good preparation for the tests.</td><td>The teacher for this class:
• Expects me to do my best at all times.
• Expects everyone to work hard.
In this class, how often:
• Are you challenged?
• Does the teacher ask difficult questions on tests?
• Do you have to work hard to do well?
The teacher for this class:
• Wants us to become better thinkers, not just memorize things.</td></tr>
<tr><td>Math instruction</td><td>Academic personalism</td></tr>
<tr><td>In your MATH class this year, how often do you do the following:
• Apply math to situations in life outside of school.
• Discuss possible solutions to problems with other students.
• Explain how you solved a problem to the class.
• Explain how you solved a problem to the class.
• Solve a problem with multiple steps that takes more than 20 minutes.
• Write a few sentences to explain how you solved a math problem.
• Write a math problem for other students to solve.</td><td>The teacher for this class:
• Notices if I have trouble learning something.
• Is willing to give extra help on schoolwork if I need it.
• Helps me catch up if I’m behind.
• Gives me specific suggestions about how I can improve my work in this class.
• Explains things in a different way if I don’t understand something in class.</td></tr>
<tr><td>Classroom rigor</td><td></td></tr>
<tr><td>The teacher for this class:
• Encourages students to share their ideas about things we are studying in class.
• Encourages students to share their ideas about things we are studying in class.
• Often requires me to explain my answers.
• Encourages us to consider different solutions or points of view.
• Doesn’t let students give up when the work gets hard.
• In my class, we talk about different solutions or points of view.</td><td></td></tr>
</table>

TABLE A.2

Supplemental measures and items from *5Essentials* Survey included in analyses

Social well-being	Academic effort and work
Emotional health	**Study habits**
• I can always find a way to help people end arguments • I listen carefully to what other people say to me • I'm good at working with other students • I'm good at helping other people.	• I always study for tests • I set aside time to do my homework and study • I try to do well on my schoolwork even when it isn't interesting to me • If I need to study, I don't go out with my friends.
School connectedness	**Grit** (perseverance facet) (Duckworth et al., 2006)
• I feel like a real part of my school • People here notice when I'm good at something • Other students in my school take my opinions seriously • People at this school are friendly to me • I'm included in lots of activities at school	• I finish whatever I begin • I am a hard worker • I continue steadily towards my goals • I don't give up easily
	Academic engagement
	• The topics we are studying are interesting and challenging • I usually look forward to this class • I work hard to do my best in this class • Sometimes I get so interested in my work I don't want to stop

TABLE A.3

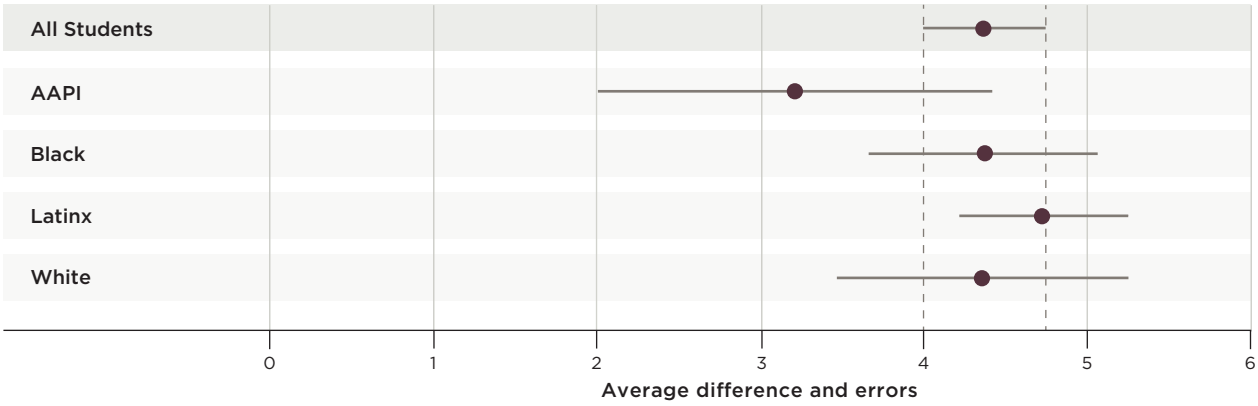
ANOVA show significant gender differences, significant race differences, but no race × gender interaction in total points earned

Source of variation	Adjusted sum of squares	Degrees of freedom	Mean square	F statistic	Prob > F
Total model	358,808.82	7	51,258.40	219.87	0.0000
Gender	58,852.27	1	58,852.27	252.45	0.0000
Race/ethnicity	230,747.41	3	76,915.81	329.93	0.0000
Gender X Race/ethnicity	862.03	3	287.34	1.23	0.2960
Residual	6,125,372.50		233.13		
Total	**6,484,181.40**	**26,282**	**246.72**		
Number of observations = 26,283			R squared = 0.0553		

Note: There are four race categories: AAPI (Asian American/Pacific Islander), Black, Latine, and White. Even though race and gender have within-group statistical differences, the overall model explains only 5.5% of the variation in total points earned. The 26,283 students included here are those for whom we had both gender and race/ethnicity data.

FIGURE A.1

Graphical representation of ANOVA output for racial/ethnic differences



Note: This figure shows the estimated difference between young men and young women in final grades across all students and within racial/ethnic groups. The figure contains the same information as Table A.3. "All students" is the baseline, dotted lines display the error of the estimate. If another estimate is within the bounds of the all-students error, the difference between the estimates is not considered to be statistically significant (29,229 total students).

TABLE A.4

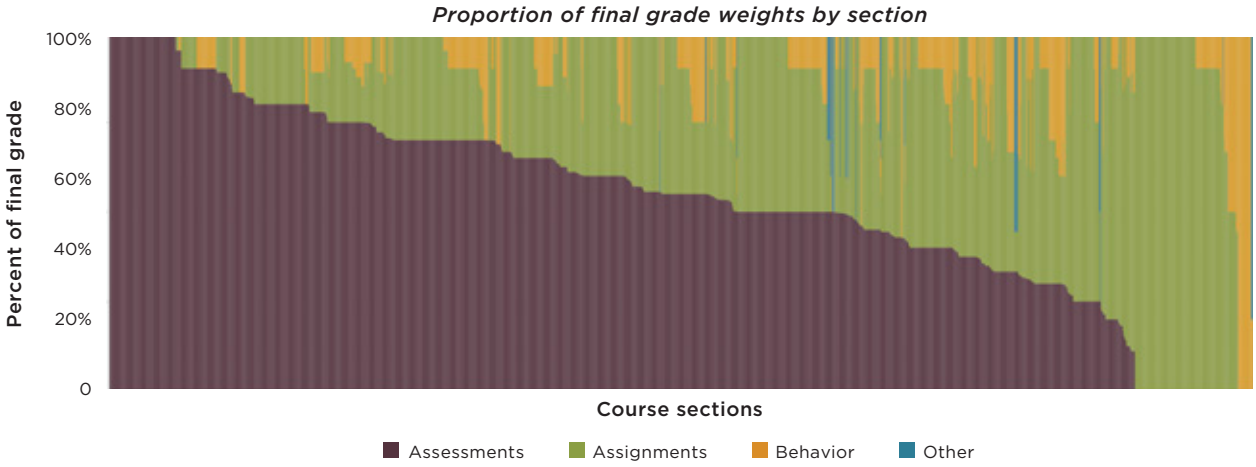
Survey and observables regression output

Male coefficient	-0.273	-0.258	-0.252	-0.250	-0.248	-0.233	-0.236
Standard error	0.011	0.010	0.009	0.009	0.009	0.009	0.010
Demographics	x	x	x	x	x	x	x
Prior achievement		x	x	x	x	x	x
Attendance			x	x	x	x	x
Discipline				x	x	x	x
Social well being index					x	x	x
Work habits index						x	x
Math instruction							x
R^2	0.020	0.195	0.302	0.303	0.306	0.330	0.331

Note: This figure shows the coefficient of the male indicator variable (representing the average GPA difference between young women and young men) across different regression models (28,517 students were included across all models; fewer than the 29,229 students in other analyses because of students who did not respond to survey measures used in this analysis.) All models control for student demographics (e.g., race/ethnicity, socioeconomic status, English learner status). Each new model adds a new explanatory variable incrementally. "OSS" refers to out-of-school suspensions. The circle shows the value of the estimated GPA difference, the whisker shows the standard error of the estimate (an approximation of where the true value of the estimate will be 95% of the time). The dashed vertical line is placed at the value of the demographics and gender only model (base model) to aid in comparing other model estimates. See notes in "Data Source 2" on p.3 for additional details.

FIGURE A.2

There was a substantial amount of variation around how teachers applied final category family weights



Note: Each bar represents one of the 1,599 sections (classrooms) in the Gradebook analysis. The y-axis displays the proportion of final grade weight for a given category family. If a bar is one color that means 100% of the final grade weight was applied to that single category family.

Appendix B

Re-coding process and goals

Given the importance of coding teachers' grading categories into families, here we provide detail on our process. In addition, Appendix B explains our attempt to differentiate formative from summative assessments and explains why we failed.

Re-coding process

Coder training and goal identification: We used a more rigorous procedure and employed three former teachers familiar with electronic grading platforms during our additional round of gradebook validation. Across a series of four rounds of coding, they were trained to achieve consistency and reliability. The coding team first talked through the different types of tasks and classroom activities to align on definitions, and then worked to code the same dataset of Gradebook tasks to ensure alignment on them. We identified two main goals of re-coding: **1)** reducing the 11% of category titles labeled as Other/unclassifiable in our first technical report on Gradebook,[27] and **2)** distinguish between formative and summative tasks, as initially assigned under different grading category titles. Coders continued to meet to discuss and resolve difficult coding issues and discrepancies.

Table B.1 shows the final agreement rates amongst coders for the large category families, including a measure of Fleiss' Kappa. Fleiss' Kappa is a measure of coder inter-reliability that specifically measures how different responses are from random assignment. A Kappa value of 0 shows no difference between categories being randomly assigned, a value of 1 indicates high alignment between coders and a low likelihood of alignment being due to chance, and a value of -1 indicates low alignment between coders and a low likelihood of the nonalignment being due to chance. Generally speaking, coders looked for greater, positive Kappa values, as low Kappa values were a signal to regroup and re-norm around those coded items.

TABLE B.1

Final Kappa agreement rates across coders

Category family	Kappa	z	p.value
Assignments	0.708	41.872	0
Assessments	0.869	51.415	0
Behavior	0.588	34.798	0
Other/unclassified	0.045	2.639	0.008
Overall grade	0.365	21.595	0
Standards-based	0.596	35.24	0

A note on terminology: Coders examined category titles that were initially assigned by teachers and looked at the corresponding day-to-day classroom activities associated with each category title. These day-to-day gradebook entries will be referred to as "tasks" for the remainder of this guide. Coders then used those tasks to infer teachers' use of each category title, which were then assigned to an updated "category family." Once category families were assigned, this allowed for further analysis.

Coders also broke the category families out into different levels, with each level leading to increased precision. Coders noticed that teachers initially organized their gradebooks into four overarching categories: Traditional (which includes assessments, assignments, and behavior), weekly, standards-based, or "other." Upon closer inspection, coders identified that what was initially labeled as "weekly" were sections where teachers assigned an overall grade for the semester. These coding schemes were misidentified as weekly, as the tasks were entered into the gradebook each week, but then overwritten by the next week's grade. Coders were able to eliminate weekly and mastery-based from Level 1 and categorize those Gradebook category titles into more meaningful categories. The majority of this paper will focus on the traditional category family from Level 1, with additional category families in Level 2 under this traditional bucket.

27 Diaz & Easton (2022).

Tables B.2, B.3, and B.4 define the different category families that coders used. Level 1 is the broadest set of category families, with Level 2 getting more granular to allow for increased coding precision. **Table B.3** shows the ultimate traditional category family buckets, and **Table B.4** shows additional traditional category families that coders devised for themselves to help with evaluating each category title.

Methodology: To assign all category titles to a Level 2 category family, coders **1)** looked at the category titles used by a teacher, **2)** identified the task(s) associated with that given teacher and category title, and **3)** selected the appropriate category family to label it. An example of coding procedures and discussions are outlined below.

A category title labeled "homework" was usually coded into the category family assignment: learning process. There was, however, discussion among coders that homework assignments were actually used to reinforce and evaluate the concepts taught in class that day, so some coders chose to label the category title "homework" as assignment: check for understanding, instead. After discussion, coders were usually able to ensure alignment in how to conceptualize this sort of variation moving forward.

TABLE B.2

Level 1 coding guide

Category family	Definition
Traditional	Type of task
Mastery-based	Common Core standards or other topics
Overall grade	Single value for quarter or semester
Other	Can't tell

TABLE B.3

Level 2 coding guide (Traditional category family only)

Category family	Definition
Behavior	Task that is not evaluating a physical work product
Assessment	Evaluation of mastery of content
Assignment	Task that is part of the learning process
Other	Can't tell

TABLE B.4

Level 2 coding guide (Traditional category family—more specific)

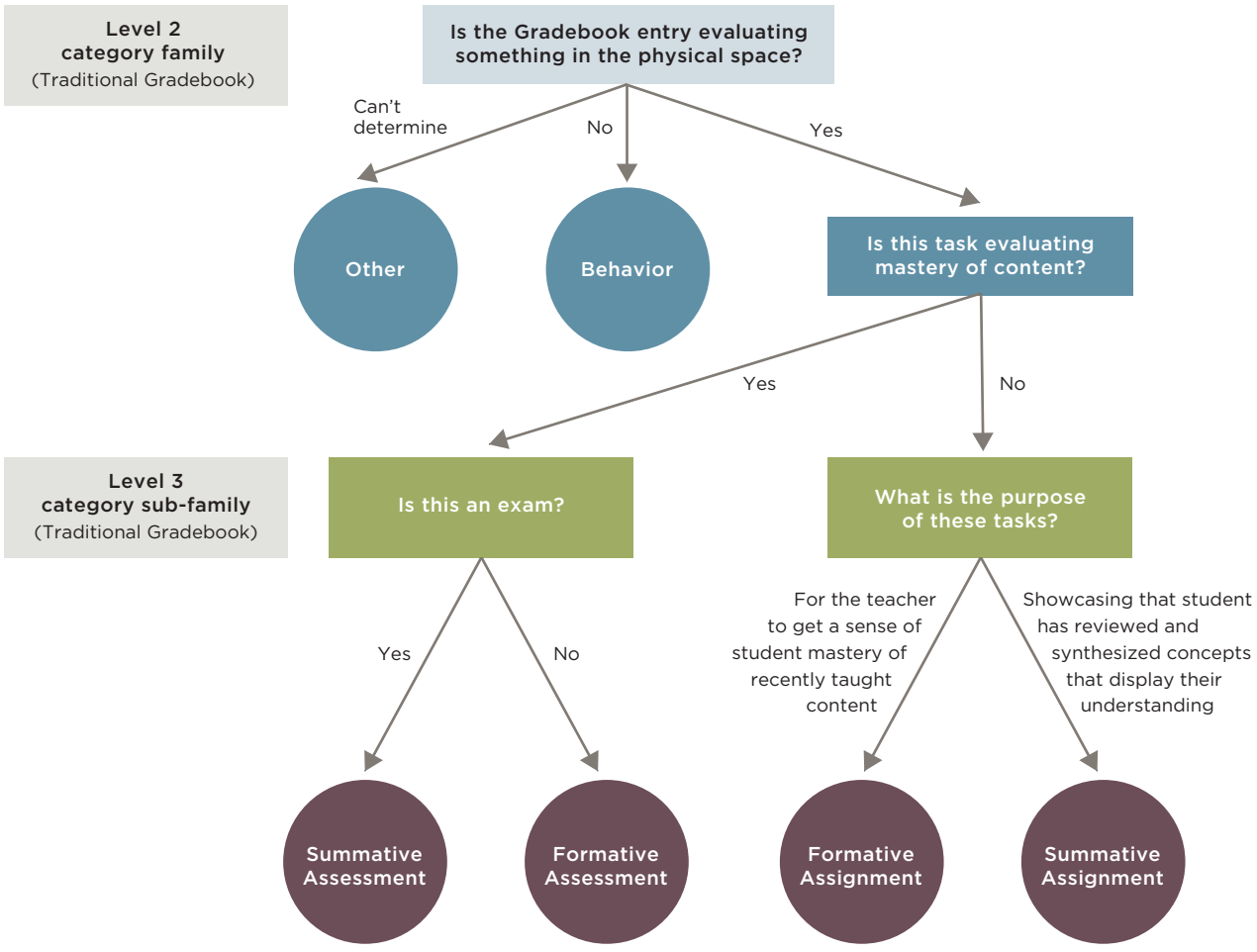
Category family	Definition
Assessment: formative	An evaluation of mastery of content, though not a unit assessment
Assessment: summative	An evaluation of mastery of content that is a unit assessment
Assessment: general	An evaluation of mastery of content, but cannot distinguish between formative and summative
Assignment: check for understanding, i.e., formative	Classroom task that allows the teacher to get a sense of student mastery of recently taught content
Assignment: learning process	Classroom task that allows students to practice recently taught tasks and reinforce concepts
Assignment: project, i.e., summative	Assignment that showcases that students have reviewed and synthesized concepts that display their understanding
Assignments: general	Classroom task, but cannot distinguish between formative and summative
Behavior	Task that is not evaluating a physical work product but rather student conduct and preparedness

There were other category titles that initially seemed more straightforward to code than they were in practice. For example, a category title "class participation" was usually coded into the category family behavior, though there were some category titles labeled "class participation" that had an associated task labeled "bell ringer." This led to further discussion amongst coders about the purpose of such tasks and how teachers conceptualize them—some teachers may use bell ringers as comprehension checks of the previous day's content and therefore should be coded as assignments: check for understanding, whereas other teachers may use bell ringers as a way to get students prepared for class and establish class professionalism, which might be more aligned with behavior. When such variation arose, each coder explained why they coded the given task as such, and coders then voted on how they would update their category family assignment. After such discussions, coders were usually able to align on how a given task should be coded into a category family and use those frameworks to ensure alignment moving forward.

Figure B.1 is indicative of how coders used the criteria to think through and assign each category in a traditional gradebook to a category family. The initial technical report included Level 2 categories, but this updated coding scheme breaks these Level 2 categories out into more specific ones to allow for increased coding precision. Coders did not ultimately use these more granular categories in their final coding schema, but they did allow coders to norm on the purpose of different kinds of tasks and how they are used in classrooms to increase coding precision and alignment.

FIGURE B.1

Coding criteria and process



Re-coding goals

1: Reducing the number of category titles in the *other* category

A key goal of the additional round of validation was to reduce the number of Gradebook category titles that coders in the earlier paper assigned into the Level 1 category family "other." To accomplish this, our raters sought to understand how teachers were using these category titles and if they could be re-assigned to a different category family to increase precision. Examples of gradebook titles initially coded as other are "content mastery," "synthesis," "practice/preparation," and "accountability." Coders were generally successful at assigning tasks initially labeled as other into more specific category families that allowed for further analysis.

The final Level 1 rates of agreement among coders are in **Table B.5**.

TABLE B.5

Level 1 coder agreement rates

Level 1 category family	Frequency	Percent	Cumulative
Traditional	7,447	91.64%	91.64
Standards	538	6.62%	98.26
Overall grade	74	0.91%	99.17
N/A | Other	67	0.82%	100
Total	8,126	100%	

Table B.6 shows the percentage of category titles coded into each family from the original report.

TABLE B.6

Percent of category titles from original report

Category family	Percent
Assessments	43%
Assignments	27%
Behavior	10%
Mastery-based	7%
Weekly	3%
Other/unclassifiable	11%

Note: 8,126 total distinct section/categories; 2,223 total distinct sections.

Creating a new category scheme

Additionally, to account for the wide variety of Gradebook category titles used by teachers, coders devised a new coding scheme with Level 2 category families beyond the default six category titles (assignments, homework, class participation, quizzes, exams, and projects), as seen in **Figure B.1**. Coders were able to more accurately assign each task an accurate category family with this new scheme.

Despite the increased structure and accuracy that these additional category families provided, there were still many tasks that were difficult to code due to lack of perceived alignment between the category title and the tasks assigned within those titles. For example, a category title called "college and career (cw, participation, etc.)" included tasks such as "wkst: u1 h p. 48," which coders interpreted as an in-class worksheet, but the same category title by the same teacher also included the tasks "class participation" and "binder check 2." There was consensus that the in-class worksheet would fall into the assignment: learning process category family, whereas the "class participation" and "binder check 2" would fall into the behavior category. **See Table B.4on p.28** where these new category families are defined.

During these agreement check-in and norming conversations, coders also discussed that high-level completion of the worksheet could be aligned with behavior if the student was working diligently and productively during class time and had used the content they learned in class to produce high-level work product. In this sense, because of the nature of various types of tasks that fell into a singular category, coders sometimes interpreted a given task differently. This led to occasional disagreement and the need for discussion and re-alignment, but once coders had norming conversations, alignment usually ensued. The broad nature of these category titles, as conceptualized by some teachers, made coding difficult. This is to say that, sometimes, the task under a category title did not appear to outsiders as consistent with the meaning of the category title. It is very likely that teachers had a much better understanding of their intent, but we outsiders could not read their minds, only try to interpret terse and cryptic task names.

2: Failure to distinguish between "formative" and "summative" assessments

The CPS/CTU grading guidelines define formative assessments as assessments that are "frequent and inform instructional decision-making throughout a marking period of time…What makes an assessment 'formative' is not the design of a test, technique, or self-evaluation, per se, but the way it is used - i.e., to inform in-process teaching and learning modifications." Summative assessments, on the other hand, are defined by the guidelines as being "used to evaluate student learning skill acquisition, and academic achievement at the conclusion of a defined instructional period… Summative assessments are often recorded as scores or grades that are then factored into a student's permanent academic record."

Based on the CPS/CTU definitions of formative and summative assessments, coders were not always able to distinguish between the two for certain tasks. Based on the definition provided for formative assessment, it seemed to coders as if these are meant to assess the process of learning, rather than the mastery of content which, in coders' experience, was more aligned with something like classwork and homework than a quiz. The former two types of tasks evaluate the learning process with checks for understanding and practicing concepts taught during classroom instruction, whereas the latter is more of an evaluation of mastery and is an assessment rather than assignment.

In their attempt to differentiate summative from formative, the coders devised two new category families: Formative assignment and summative assignment. These turned out to be very useful for discussion among the coders, but they were not ultimately part of the final coding category families, as coders found it easier to differentiate formative and summative assessments than assignments writ large. We tried to differentiate assignments based on whether or not they were a check for understanding (summative) or part of the learning process (formative), but ultimately coders were unable to distinguish meaningfully between these two because the purpose of these tasks within the classroom is not always singular. Coders independently identified that they were able to see more fine-grained levels of these category families, but when it came to norming, agreement rates among coders were low since these tasks can serve multiple purposes in the classroom. There were not high enough agreement rates to keep these two different assignment sub-categories.

Further examination of how such classroom tasks are used in other subject areas such as social studies rather than math leaves room for future analysis as well. Additional discussion is required to understand how teachers are using these category families in their Gradebooks, specifically with regards to the weights associated with them as well. Ultimately, coders were only moderately successful with distinguishing between formative and summative student work. **While some consensus was reached, the resulting families had too little variation to use in our current paper and led to an increased number of category families. Coders ultimately decided to return to the broader category families for their final coding schema.**

The final Level 2 rates of agreement among coders are in **Table B.7**.

TABLE B.7

Level 2 coder agreement rates

Level 2 Final	Frequency	Percent	Cumulative
Assignments	3,561	47.82%	47.82
Assessments	2,871	38.55%	83.37
Behavior	961	12.90%	99.27
Other	54	0.73%	100
Total	7,447	100%	

ABOUT THE AUTHORS

JOHN Q. EASTON is Senior Advisor in the Institute for Policy Research at Northwestern University. Prior to his work at Northwestern, he served as Senior Fellow at the University of Chicago Consortium on School Research. Earlier in his career he was Deputy Director and Executive Director at the UChicago Consortium. In addition to working at university-based research centers, Easton also has held governmental and foundation positions. He was Director of the Institute for Education Sciences in the U.S. Department of Education and Vice President for Programs at the Spencer Foundation. He is actively involved in several Advisory Boards at non-profit organizations, including the Illinois Economic Security Advisory Board and the Research Visiting Panel at the Educational Testing Service. He is Chair of the Illinois Workforce and Education Research Collaborative Advisory Board and Chair of the Early Childhood Research Alliance for Chicago Launch Committee. He just completed a two-year term on the Chicago Public Schools Accountability Redesign Committee.

BRIANA DIAZ is a graduate student at the Harris School of Public Policy and was a Research Analyst at the UChicago Consortium at the time this research was conducted. Prior to joining the UChicago Consortium, Briana worked in New Orleans public education for nine years, serving as a middle- and high-school science teacher, school district data analyst and database architect, and policy researcher. Her research interests include educational accountability systems, district budgets and finance, and school leadership.

This report reflects the interpretation of the authors. Although the UChicago Consortium's Steering Committee provided technical advice, no formal endorsement by these individuals, organizations, or the full Consortium, should be assumed.

UCHICAGO Consortium on School Research

Steering Committee

GREG JONES
Co-Chair
The Academy Group

REBECCA VONDERLACK-NAVARRO
Co-Chair
Latino Policy Forum

Institutional Members

BOGDANA CHKOUMBOVA
Chicago Public Schools

STACY DAVIS GATES
Chicago Teachers Union

SARAH DICKSON
Chicago Public Schools

SHANNAE JACKSON
Chicago Public Schools

TROY LARAVIERE
Chicago Principals and Administrators Association

Individual Members

NANCY CHAVEZ
Slalom

JAHMAL COLE
My Block, My Hood, My City

ACASIA WILSON FEINBERG
The Cleveland Avenue Foundation for Education

MEGAN HOUGARD
Chicago Public Schools

PRANAV KOTHARI
Revolution Impact, LLC

AMANDA LEWIS
University of Illinois at Chicago

LUISIANA MELENDEZ
Erikson Institute

SHAZIA MILLER
Mathematica

KAFI MORAGNE-PATTERSON
UChicago Inclusive Economy Lab

CRISTINA PACIONE-ZAYAS
Illinois State Senate

PAIGE PONDER
Discovery Partners Institute

CARLA RUBALCAVA
Mikva Challenge

ELLEN SCHUMER
COFI

PAM WITMER
Golden Apple Foundation

JOHN ZEIGLER
DePaul University

UCHICAGO Consortium on School Research

1313 East 60th Street
Chicago, Illinois 60637

T 773.702.3364
F 773.702.2010

@UChiConsortium
consortium.uchicago.edu

OUR MISSION With the goal of supporting stronger and more equitable educational outcomes for students, the UChicago Consortium conducts research of high technical quality that informs and assesses policy and practice in the Chicago Public Schools. We seek to expand communication among researchers, policymakers, practitioners, families, and communities as we support the search for solutions to the challenge of transforming schools. We encourage the use of research in policy action and practice but do not advocate for particular policies or programs. Rather, we help to build capacity for systemic school improvement by identifying what matters most for student success, creating critical indicators to chart progress, and conducting theory-driven evaluation to identify how programs and policies are working.

THE UNIVERSITY OF CHICAGO | UEI URBAN EDUCATION INSTITUTE

ISBN 978-0-9814604-1-3