Efficacy Analysis

Retrospective QED using CAASPP data with updated ELSi records

Stepan Oskin • Jiawen Chen, PhD

Prodigy

www.prodigygame.com

Executive Summary

This report describes a set of analyses pertaining to the efficacy of Prodigy as an educational tool. A retrospective quasi-experimental design was used for this study.

Based on Prodigy usage in the 2018-19 school year, school grades in California were selected into either treatment (i.e., high Prodigy usage; 'strong', n = 36; 'weak', n = 71) or control groups (low/no Prodigy usage; n = 145). Their grade-level California Assessment of Student Performance and Progress (CAASPP) scores were retrieved from CAASPP website.

Regression analyses showed that the 'strong' treatment group experienced significantly greater improvement in CAASPP scores from 2018 to 2019 than the control group after controlling for the baseline 2018 CAASPP score and other demographic factors. In addition, there was a significant increase in the percentage of students who met or exceeded expectations in 2019 compared to their 2018 expectations in both the "strong" and "weak" treatment groups in comparison with the control group.

Due to important data quality issues, the findings do not qualify for ESSA Tier II consideration. Nevertheless, the results provide the first piece of evidence indicating Prodigy's effectiveness when comparing high usage students to low/no usage students.

Background

This study used a **retrospective quasi-experimental design** to test the efficacy of Prodigy as an educational product to boost math achievement. Due to small sample size and data quality issues (delineated in later sections), results in this report do not qualify for ESSA Tier II standard. Nonetheless, this study provides the first piece of evidence of Prodigy's efficacy with the inclusion of a control group.

Sample

The publicly available California Assessment of Student Performance and Progress (CAASPP) scores aggregated at the grade-level present a unique opportunity for us to conduct an internal investigation of efficacy using a retrospective quasi-experimental design (QED). **The 2019 CAASPP scores of schools/grades that used Prodigy in the 2018-19 school year (i.e., program year) were compared to schools that did not use Prodigy. The selected schools, both treatment and control schools, were screened for low/no Prodigy usage in the 2017-18 school year (i.e., pre-program year) to maximize the possibility of detecting an effect. The 2018 CAASPP scores were statistically controlled for in the regression models along with other demographic variables.** Selection criteria for the treatment groups are:

- From schools in California as CAASPP is a California based assessment
- In either Grades 3, 4, or 5 as CAASPP starts at Grade 3 and these three grades represent the most active users in Prodigy; in addition, activity patterns of the students in these grades in Prodigy are similar
- A minimum of 20 students within a grade in a school
- High monthly Prodigy usage at the grade level throughout the 2018-2019 school year as defined by high mean percentage of monthly learning students (MLS) in a grade in a school (>= 70%) who answered at least 10 questions in a month
- Low Prodigy usage in the 2017-18 school year as defined by low percentage of students in a grade in a school (<= 10% for the "strong" treatment group; <= 20% for the "weak" treatment group) who answered at least 10 questions in a month

Selection criteria for the control group were similar to those for the treatment group except the control group had low/no Prodigy usage in the 2018-19 school year. In addition, the control grades were selected from the same school districts as treatment grades based on having similar baseline 2018 CAASPP scores as the treatment grades, as evidenced by non-significant t-tests.

A number of issues emerged in the selection process that introduced uncertainties in the quality of the data, such as having more Prodigy users in a grade level in a school than indicated in the public records in the Elementary/Secondary Information System (ElSi), possibly because students had multiple Prodigy accounts and attached all of them to their class. Because we do not collect personally identifiable information (PII), we were unable to resolve such issues in the data cleaning process.

For the analyses, the sample size for the "strong" treatment group was n = 36 school grades, with an average of 84.6% of monthly learning students in a grade in the 2018-19 school year. The sample size for the "weak" treatment group was n = 71 school grades, with an average of 81.5% of monthly learning students in a grade. Prodigy usage as defined by the total number of questions answered by all students in a grade in the 2018-19 school year is shown in Figure 1 below. The sample size for the control group was n = 145 school grades, with an average of 1.7% of monthly learning students in a grade.

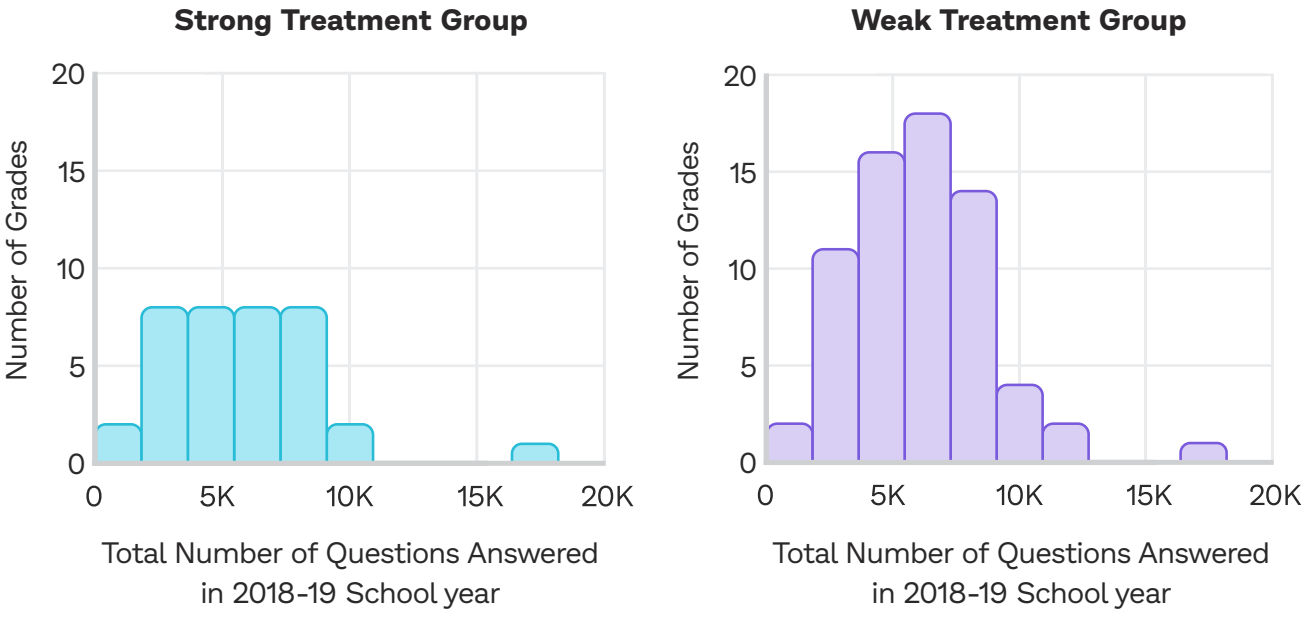


Figure 1. Distribution of total number of questions answered in 2018-19 school year by 'strong' and 'weak' treatment groups.

Data

The 2019 (outcome variable) and 2018 (baseline control variable) grade-level CAASPP scores for both the treatment and control groups were downloaded from CAASPP website. Prodigy usage was determined by aggregating Prodigy usage data for the 2018-19 school year and the 2017-18 school year. Demographic variables included the percentage of students in a school that were eligible for the free or reduced priced lunch program (FARL ratio) as an indicator of socioeconomic status.

Results

Cross cohort comparisons

The following two sets of regression models compared the study sample's 2019 achievement outcomes with the outcomes from the same grade level in the same school in 2018. For example, the achievement of a Grade 4 in the sample in 2019 was compared with the achievement of Grade 4 from the same school in 2018 to determine how much improvement was made by the Grade 4 students in the study sample compared to last year's Grade 4 students.

Table 1 below shows the descriptive statistics of the study variables by experimental conditions and school grades. **On average, the control group had statistically significantly higher baseline 2018 CAASPP score (M = 2452.83) than both the 'strong' (M = 2436.32; $t(55.22) = 2.29$, $p = 0.03$) and 'weak' (M = 2438.87; $t(145.12) = 2.49$, $p = 0.01$) treatment groups. This difference was diminished in the 2019 CAASPP scores. On the other hand, the control group had significantly lower FARL ratio in their schools than both the 'strong' and 'weak' treatment schools.** Figure 2 below shows the distribution of 2019 and 2018 CAASPP scores for the control and the two treatment groups.

Table 1
Sample sizes and study variable means by school grade and experiment condition.

Grade	Condition	Sample Size	2019 CAASPP	2018 CAASPP	Δ CAASPP	FARL ratio
Grade 3	control	61	2430.70	2428.37	2.33	60.43%
	'strong' treatment	12	2423.36	2408.23	15.13	68.20%
	'weak' treatment	25	2424.85	2414.09	10.76	69.68%
Grade 4	control	48	2473.81	2469.86	3.95	57.40%
	'strong' treatment	13	2457.13	2444.99	12.15	71.15%
	'weak' treatment	27	2458.27	2447.92	10.35	68.72%
Grade 5	control	36	2474.06	2471.59	2.46	71.98%
	'strong' treatment	11	2475.74	2456.72	19.02	79.80%
	'weak' treatment	19	2475.65	2458.63	17.03	79.05%
All Grades	control	145	2455.73	2452.83	2.90	62.30%
	'strong' treatment	36	2451.56	2436.32	15.24	72.81%
	'weak' treatment	71	2451.15	2438.87	12.28	71.82%

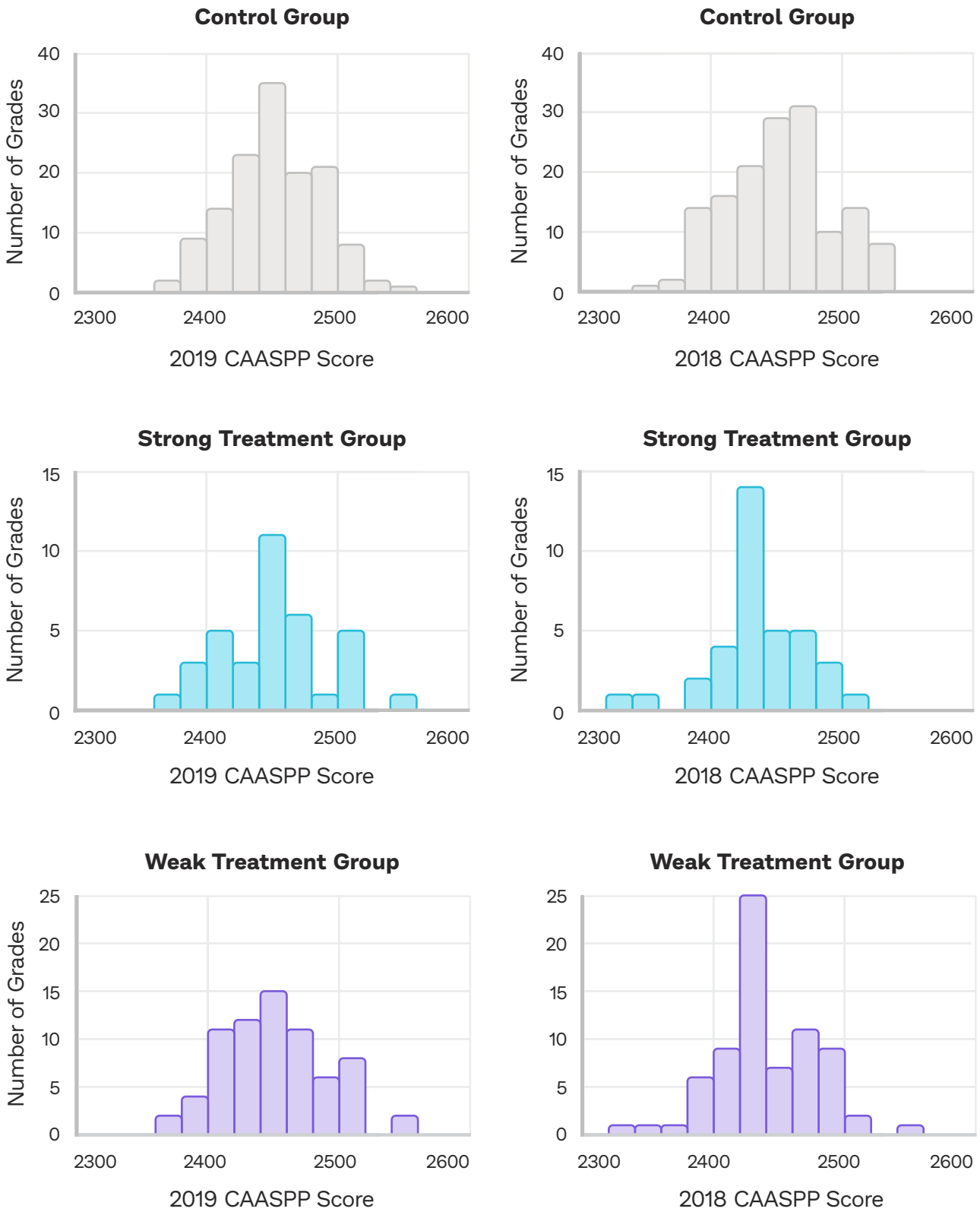


Figure 2. Distributions of CAASPP scores by school year and experimental condition.

Regression Models Predicting 2019 CAASPP Score

Table 2 below shows the regression coefficients from two linear regression models, one that compared the 'strong' treatment group with the control group and another that compared the 'weak' treatment group with the control group. These regression models were used to examine the effectiveness of Prodigy in improving math achievement. The regression equation is as follows:

$$\text{2019 CAASPP score} = \beta_0 + \beta_1 * \text{treatment condition} + \beta_2 * \text{baseline 2018 CAASPP score} + \beta_3 * \text{school grade} + \beta_4 * \text{FARL ratio} + \beta_5 * \text{ethnic composition}$$

Table 2
Regression models predicting 2019 CAASPP score.

	'Strong' treatment	**'Weak' treatment**
Treatment condition	**9.87****	**7.04****
2018 CAASPP score	.75***	.77**
Grade 4a	9.18*	8.68*
Grade 5a	15.45***	15.77***
FARL ratio	-39.51***	-35.92***
% white students	-22.41*	-18.90*

Note. *$p < .05$; **$p < .01$; ***$p < .001$.
[a]Grades 4 and 5 were compared to Grade 3.

After controlling for baseline 2018 CAASPP score, school grades that used Prodigy experienced significantly greater improvement in 2019 CAASPP score than school grades that did not use Prodigy for both the 'strong' and 'weak' treatment groups. The effect sizes in both conditions are small, $f^2 = .04$ and $f^2 = .03$, respectively. In addition, Grades 4 and 5 scored significantly higher than Grade 3. School grades with higher FARL ratio (i.e., more students eligible for free or reduced priced lunch) scored significantly lower in 2019 CAASPP. Figure 3 below shows the average difference in CAASPP scores from 2018 to 2019 by experimental condition and at the state level across all schools in California.

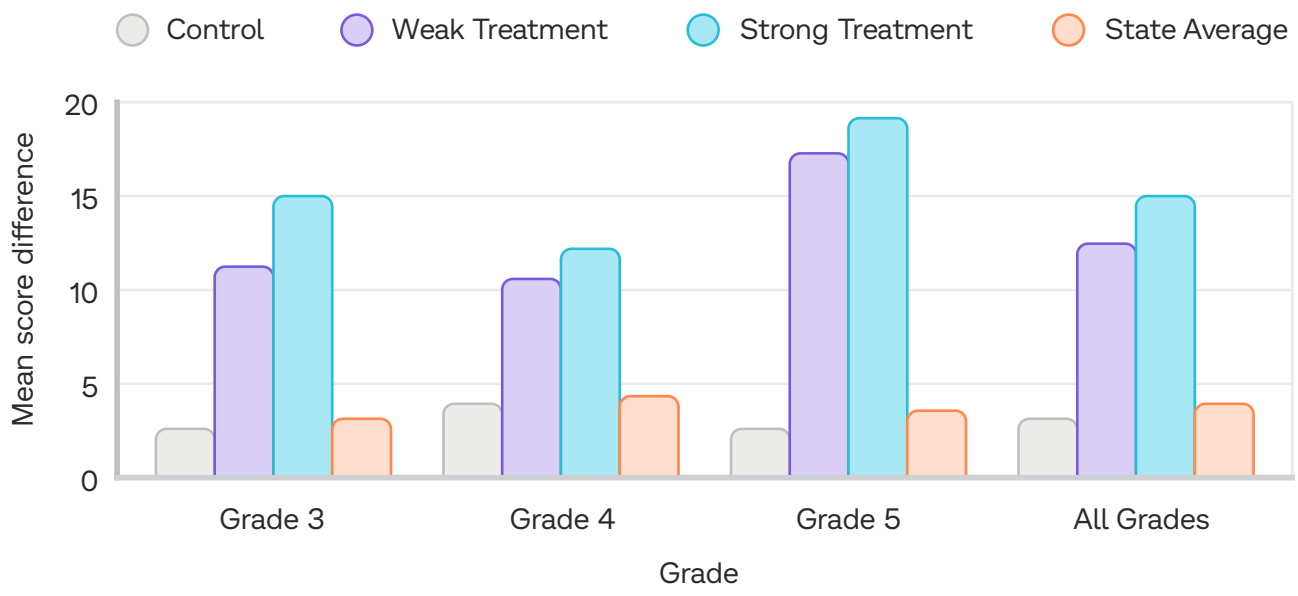


Figure 3. Bar chart of differences in CAASPP scores between 2019 and 2018.

Regression Models Predicting Meeting or Exceeding Expectations in 2019

Another achievement outcome measured in this study was the percentage of students in a grade who met or exceeded expectations in 2019. Table 3 shows the difference in the percentage of students in a grade who did not meet, nearly met, met, or exceeded expectations between 2018 and 2019. Two regression models similar to the ones shown above were estimated with CAASPP scores in 2018 and 2019 exchanged for the percentage of students who met or exceeded expectations in 2018 and 2019. The regression coefficients are shown in Table 4 below. **After controlling for baseline 2018 percentage of students meeting or exceeding expectations, school grades that used Prodigy experienced significantly greater increase in the percentage of students who met or exceeded expectations in 2019 than school grades that did not use Prodigy for both the 'strong' and 'weak' treatment samples.** Grades 4 and 5 students did not experience greater change compared to Grade 3 students. School grades with higher FARL ratio were more likely to show decrease in the percentage of students who met or exceeded expectations in 2019. Figure 4 below shows the change in the percentage of students in each expectations category by grade.

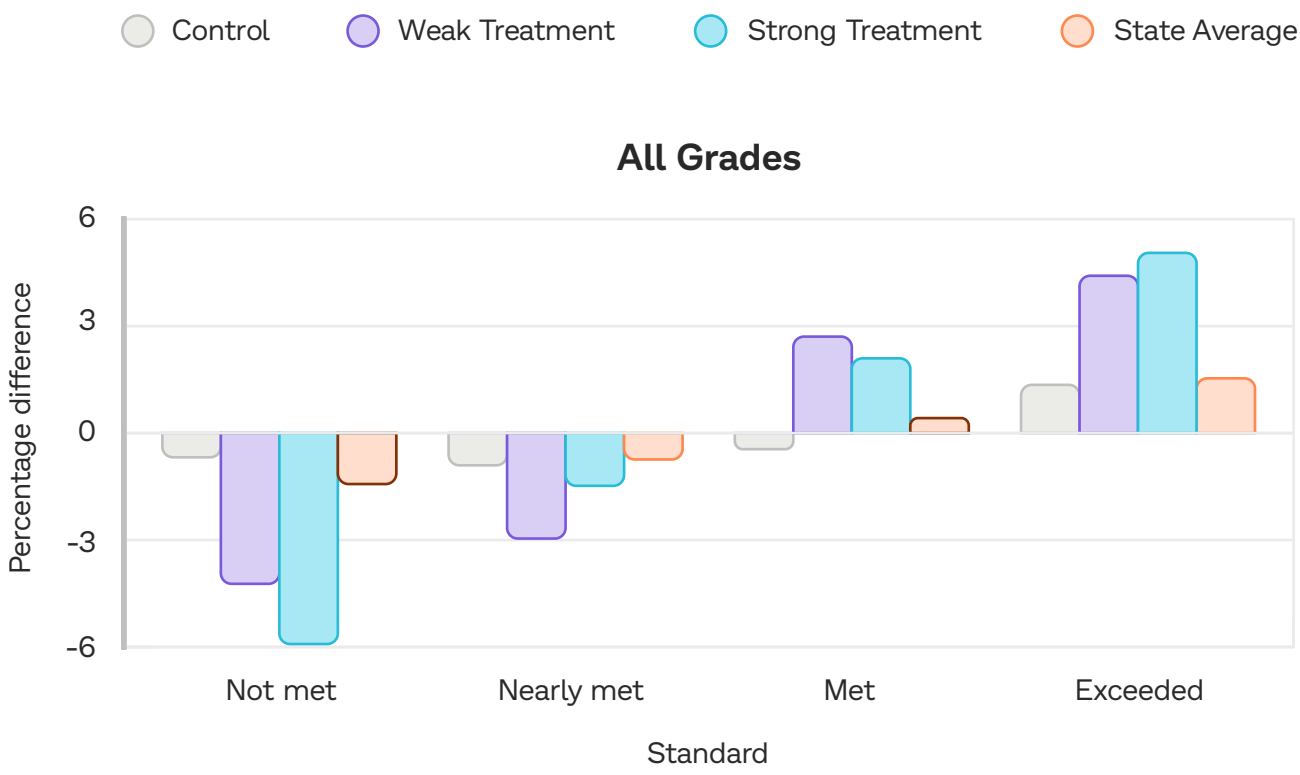
Table 3
Sample sizes and study variable means by school grade and experiment condition.

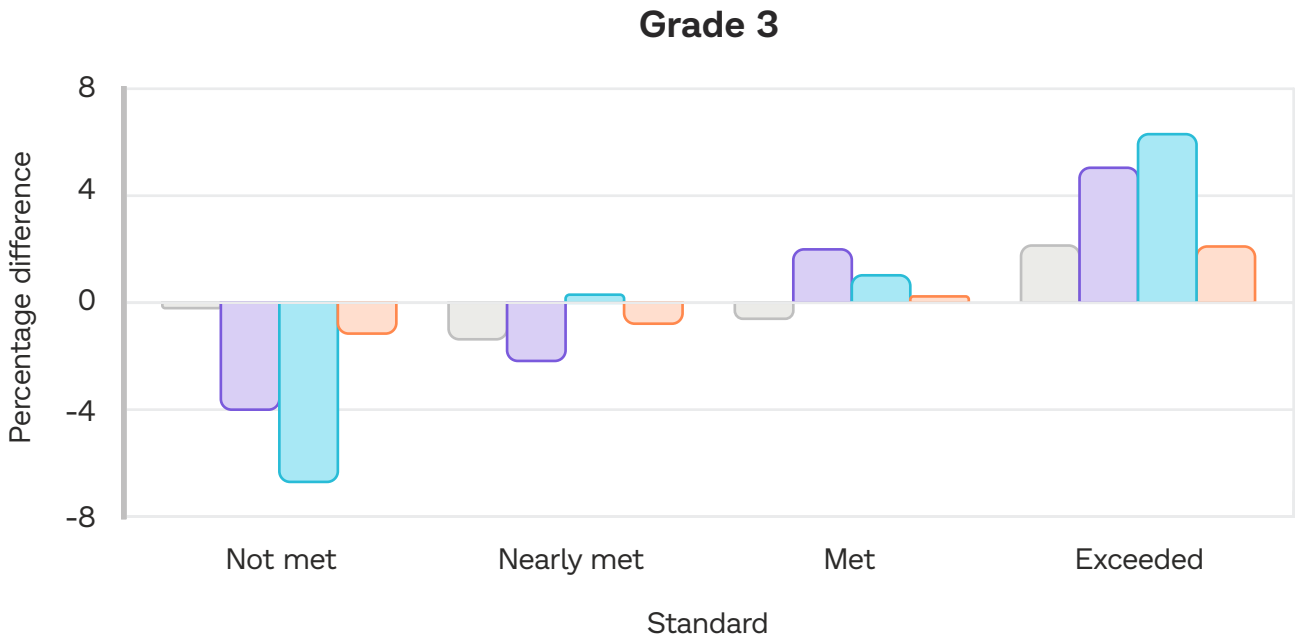
Grade 3	control	-0.17%	-0.95%	-0.39%	1.32%
	'strong' treatment	-6.90%	0.24%	0.43%	6.23%
	'weak' treatment	-3.83%	-1.89%	0.99%	4.73%
	state average	-0.80%	-0.49%	-0.06%	1.35%
Grade 4	control	-0.76%	-0.35%	-0.54%	1.65%
	'strong' treatment	-4.74%	-0.29%	2.02%	3.01%
	'weak' treatment	-3.25%	-2.33%	2.60%	2.99%
	state average	-1.52%	-0.49%	0.46%	1.54%
Grade 5	control	-0.67%	-0.56%	0.68%	0.56%
	'strong' treatment	-5.88%	-4.53%	4.61%	5.81%
	'weak' treatment	-5.70%	-4.61%	5.33%	4.99%
	state average	-1.67%	-0.27%	0.42%	1.52%
All Grades	control	-0.41%	-0.65%	-0.17%	1.24%
	'strong' treatment	-5.81%	-1.41%	2.28%	4.94%
	'weak' treatment	-4.11%	-2.79%	2.76%	4.14%
	state average	-1.33%	-0.42%	0.27%	1.47%

Table 4
Regression models predicting meeting or exceeding expectations in 2019.

	‘Strong’ treatment	‘Weak’ treatment
Treatment condition	**5.06****	**4.56****
% met or exceeded expectations in 2018	.73***	.75***
Grade 4a	-2.25	-.2.04
Grade 5a	-2.06	-1.09
FARL ratio	-23.80***	-22.01***
% white students	-11.48*	-10.71*

Note. *$p < .05$; **$p < .01$; ***$p < .001$.
[a]Grades 4 and 5 were compared to Grade 3.


Control
Weak Treatment
Strong Treatment
State Average
All Grades
Percentage difference
6
3
0
-3
-6
Not met
Nearly met
Met
Exceeded
Standard


Grade 3
Percentage difference
8
4
0
-4
-8
Not met
Nearly met
Met
Exceeded
Standard

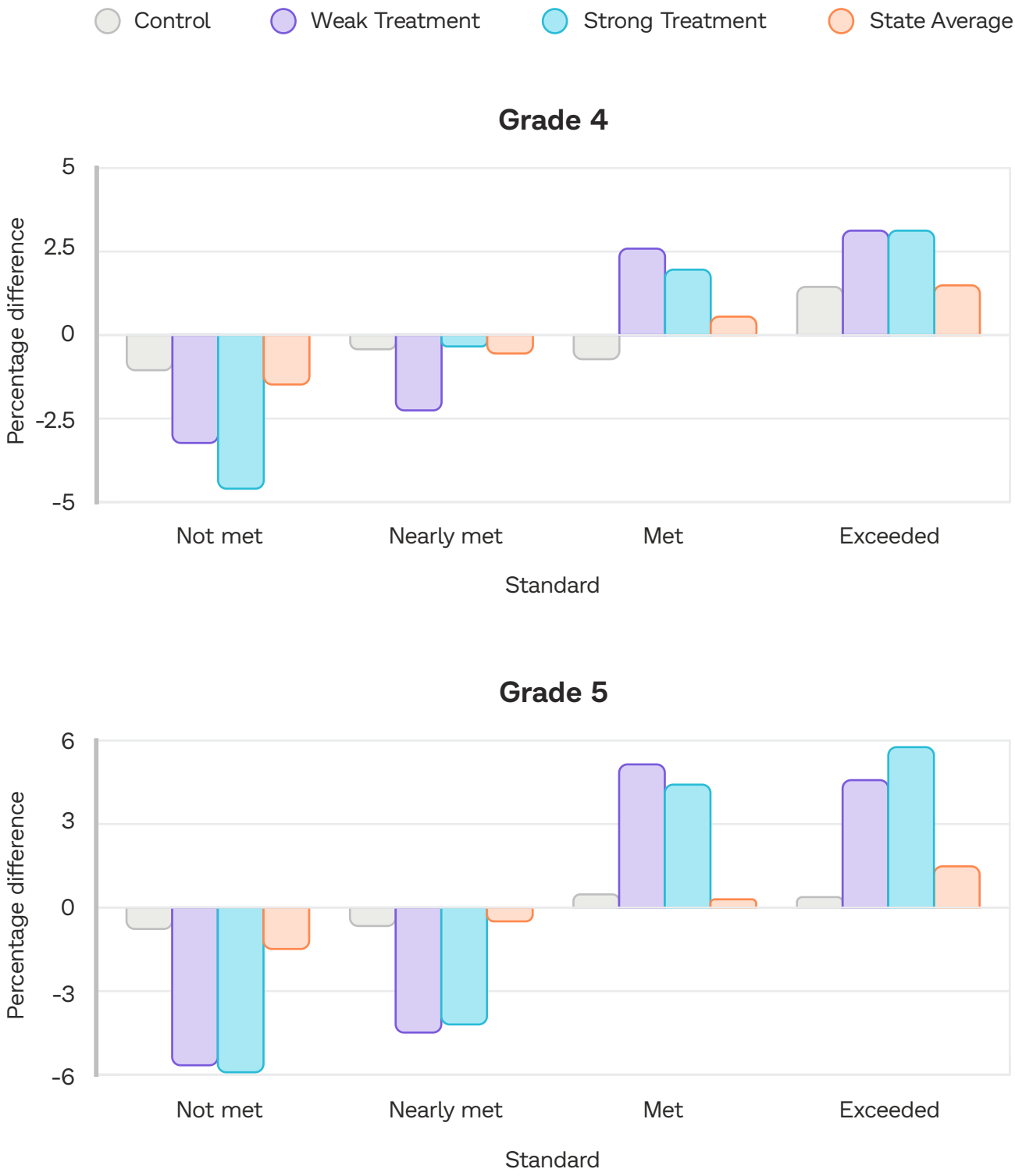

Control
Weak Treatment
Strong Treatment
State Average
Grade 4
Percentage difference
5
2.5
0
-2.5
-5
Not met
Nearly met
Met
Exceeded
Standard
Grade 5
Percentage difference
6
3
0
-3
-6
Not met
Nearly met
Met
Exceeded
Standard

Figure 4. Bar charts of differences in the percentage of students in a grade who did not meet, nearly met, met, or exceeded expectations between 2018 and 2019.

Within cohort comparisons

The next two sets of regression models compared the study sample's 2019 achievement outcomes with their outcomes from 2018. For example, the achievement of a Grade 4 in the sample in 2019 was compared with the same group's achievement in 2018 when they were in Grade 3 to determine how much improvement the group of students has made over one school year. **Because CAASPP testing starts at Grade 3 level, the following analyses only included Grades 4 and 5 from the study sample since there are no Grade 2 scores to be compared with for the Grade 3 sample, resulting in a very sample size. Consequently, the regression results should be taken only as rough estimates of Prodigy's impact, not precise indicators.**

Table 5 below shows the descriptive statistics of the study variables by experimental conditions and school grades. **On average, the control group had higher baseline 2018 CAASPP score (M = 2446.63) than both the 'strong' (M = 2431.40) and 'weak' (M = 2431.27) treatment groups. In addition, the control group had lower FARL ratio than both the 'strong' and 'weak' treatment schools.**

Table 5

Sample sizes and study variable means by school grade and experiment condition.

Grade	Condition	Sample size	2019 CAASPP	2018 CAASPP	ΔCAASPP	FARL ratio
Grade 3	control	58	2471.84	2431.44	40.41	58.72%
	'strong' treatment	13	2457.13	2418.69	38.45	71.15%
	'weak' treatment	24	2458.27	2419.14	39.13	68.72%
Grade 4	control	54	2487.76	2462.95	24.80	63.03%
	'strong' treatment	11	2475.74	2446.42	29.32	79.80%
	'weak' treatment	18	2474.08	2449.47	24.61	80.21%
All Grades	control	112	2479.52	2446.63	32.88	60.80%
	'strong' treatment	24	2465.66	2431.40	34.26	75.12%
	'weak' treatment	45	2464.59	2431.27	33.32	73.32%

Regression Models Predicting 2019 CAASPP Score

Table 6 below shows the regression coefficients from two linear regression models, one that compared the 'strong' treatment group with the control group and another that compared the 'weak' treatment group with the control group.

Table 6
Regression models predicting 2019 CAASPP score.

	'Strong' treatment	'Weak' treatment
Treatment condition	**2.45**	**.28**
2018 CAASPP score	.77***	.79***
Grade 5a	-4.91	-5.90
FARL ratio	-40.53***	-39.69***

Note. ***p < .001.
[a]Grades 5 was compared to Grade 4.

After controlling for 2018 CAASPP score, there was no significant difference in school grades that used Prodigy compared to ones that did not in 2019 CAASPP score for the 'strong' and 'weak' treatment sample. School grades with higher FARL ratio scored significantly lower in 2019 CAASPP. Ethnic composition again was not a significant predictor of 2019 CAASPP score. Figure 5 below shows the average change in CAASPP scores from 2018 to 2019 by experimental condition and at the state level across all schools in California.

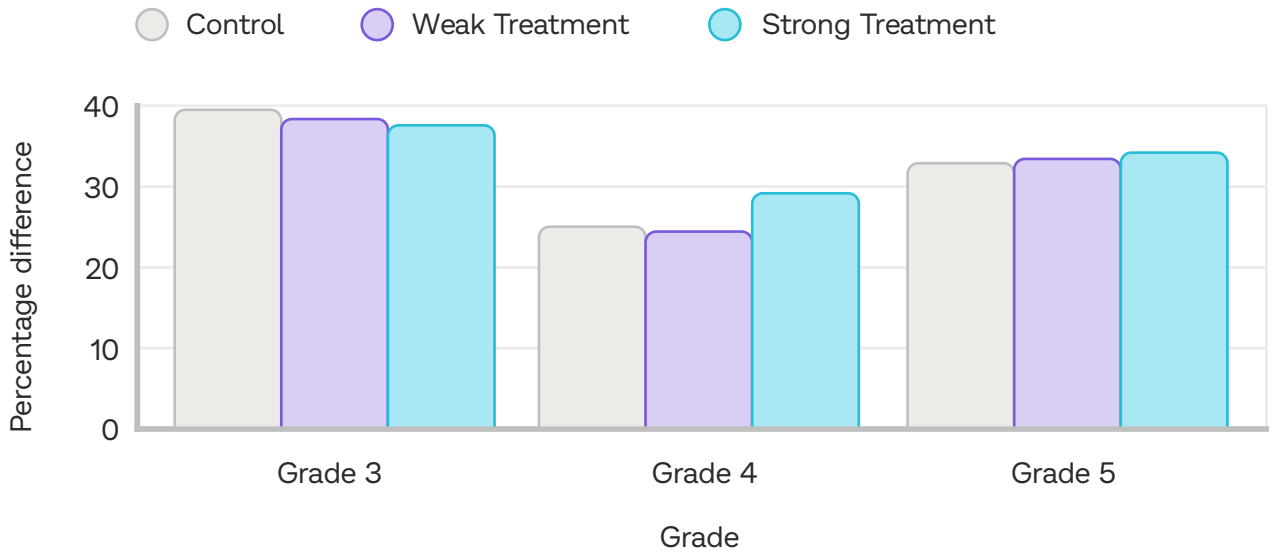


Figure 5. Bar chart of change in CAASPP scores between 2019 and 2018.

Regression Models Predicting Meeting or Exceeding Expectations in 2019

Table 7 shows the change in the percentage of students in a grade who did not meet, nearly met, met, or exceeded expectations between 2018 and 2019**. For Grade 4, there were increases in the percentage of students who nearly met expectations and decreases in the percentage of students who did not meet, met, or exceeded expectations. For Grade 5, there were increases in the percentage of students who did not meet expectations and exceeded expectations and decreases in the percentage of students who nearly met or met expectations.** Two regression models estimated the impact of Prodigy use on the percentage change in meeting or exceeding expectations from 2018 to 2019. The regression coefficients are shown in Table 8 below. **After controlling for baseline 2018 percentage of students meeting or exceeding expectations, only the FARL ratio negatively predicted the outcome in both the 'strong' and 'weak' treatment samples. Being in the treatment groups was not a significant predictor.**

Grades 4 and 5 students did not differ in their percentage change from 2018 to 2019. Ethnic composition was not a significant predictor. Figure 6 below shows the change in the percentage of students in each expectations category by grade.

compared the 'strong' treatment group with the control group and another that compared the 'weak' treatment group with the control group.

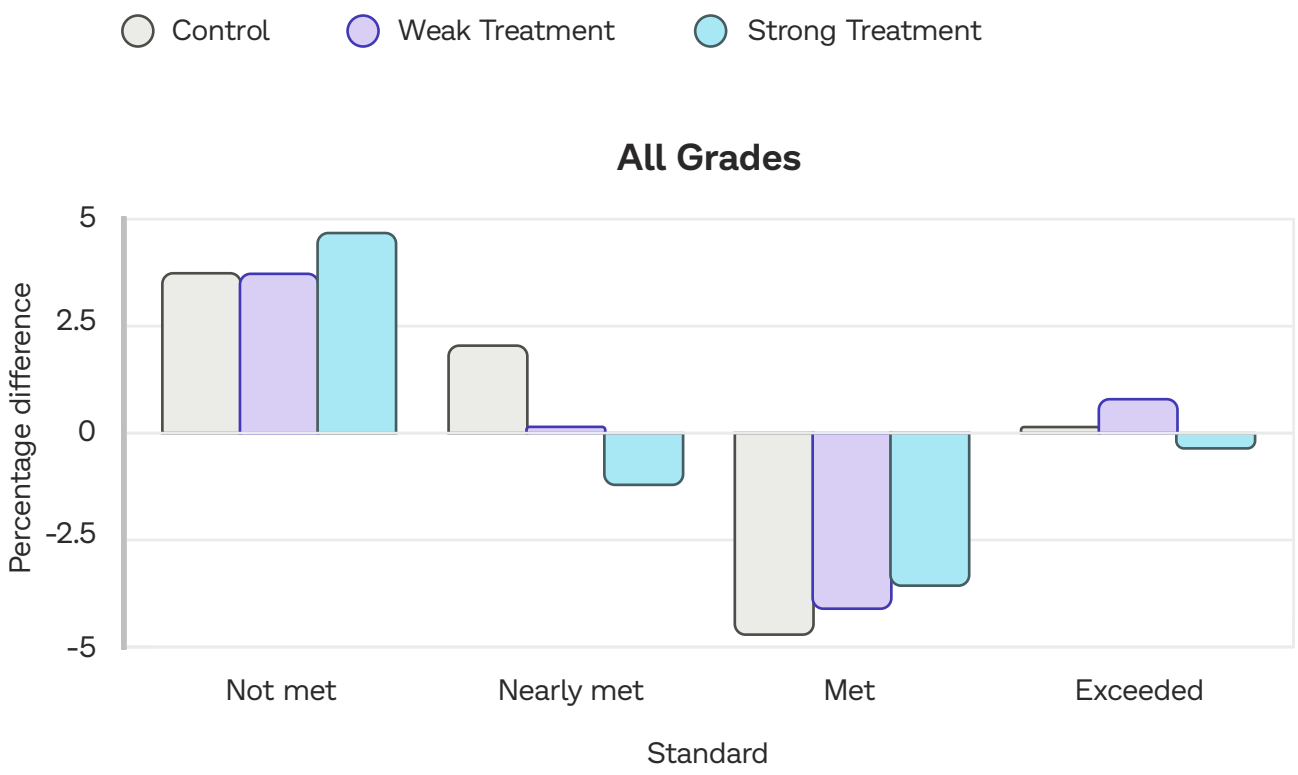
Table 7
Sample sizes and study variable means by school grade and experiment condition.

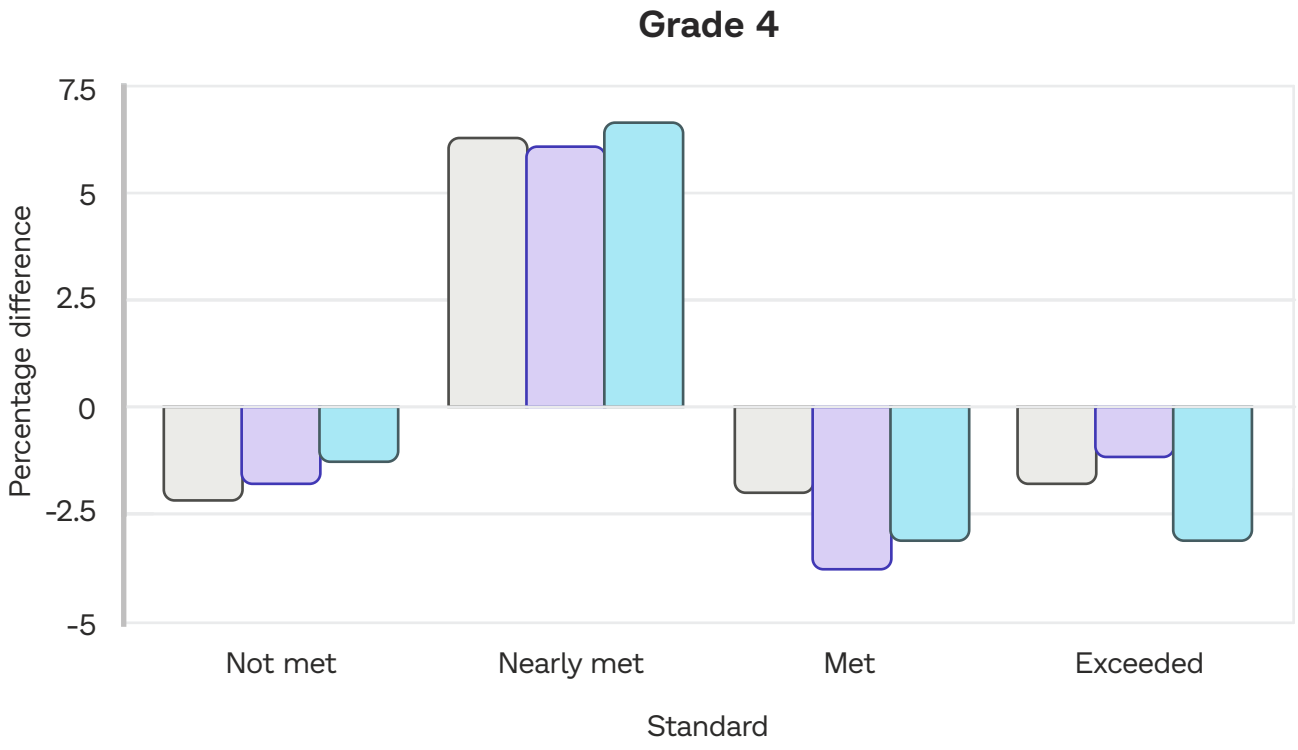
Grade	Condition	Δ(Not met expectations)	Δ(Nearly met expectations)	Δ(Met expectations)	Δ(Exceeded expectations)
Grade 3	control	-2.27%	6.62%	-2.25%	-2.09%
	'strong' treatment	-1.18%	6.98%	-2.85%	-2.95%
	'weak' treatment	-1.96%	6.52%	-3.33%	-1.24%
Grade 4	control	10.55%	-5.16%	-7.59%	2.19%
	'strong' treatment	11.71%	-10.58%	-4.21%	3.08%
	'weak' treatment	12.70%	-9.81%	-5.69%	2.80%
All Grades	control	3.91%	0.94%	-4.82%	-0.03%
	'strong' treatment	4.73%	-1.07%	-3.48%	-0.18%
	'weak' treatment	3.90%	-0.01%	-4.27%	0.38%

Table 8
Regression models predicting meeting or exceeding expectations in 2019.

	'Strong' treatment	'Weak' treatment
Treatment condition	**1.52**	**0.98**
% met or exceeded expectations in 2018	.68***	.70***
Grade 5a	-1.41	-1.16
FARL ratio	-23.12***	-24.04***

Note. ***p < .001.
[a]Grades 5 was compared to Grade 4.


Control
Weak Treatment
Strong Treatment
All Grades
Percentage difference
5
2.5
0
-2.5
-5
Not met
Nearly met
Met
Exceeded
Standard


Grade 4
Percentage difference
7.5
5
2.5
0
-2.5
-5
Not met
Nearly met
Met
Exceeded
Standard

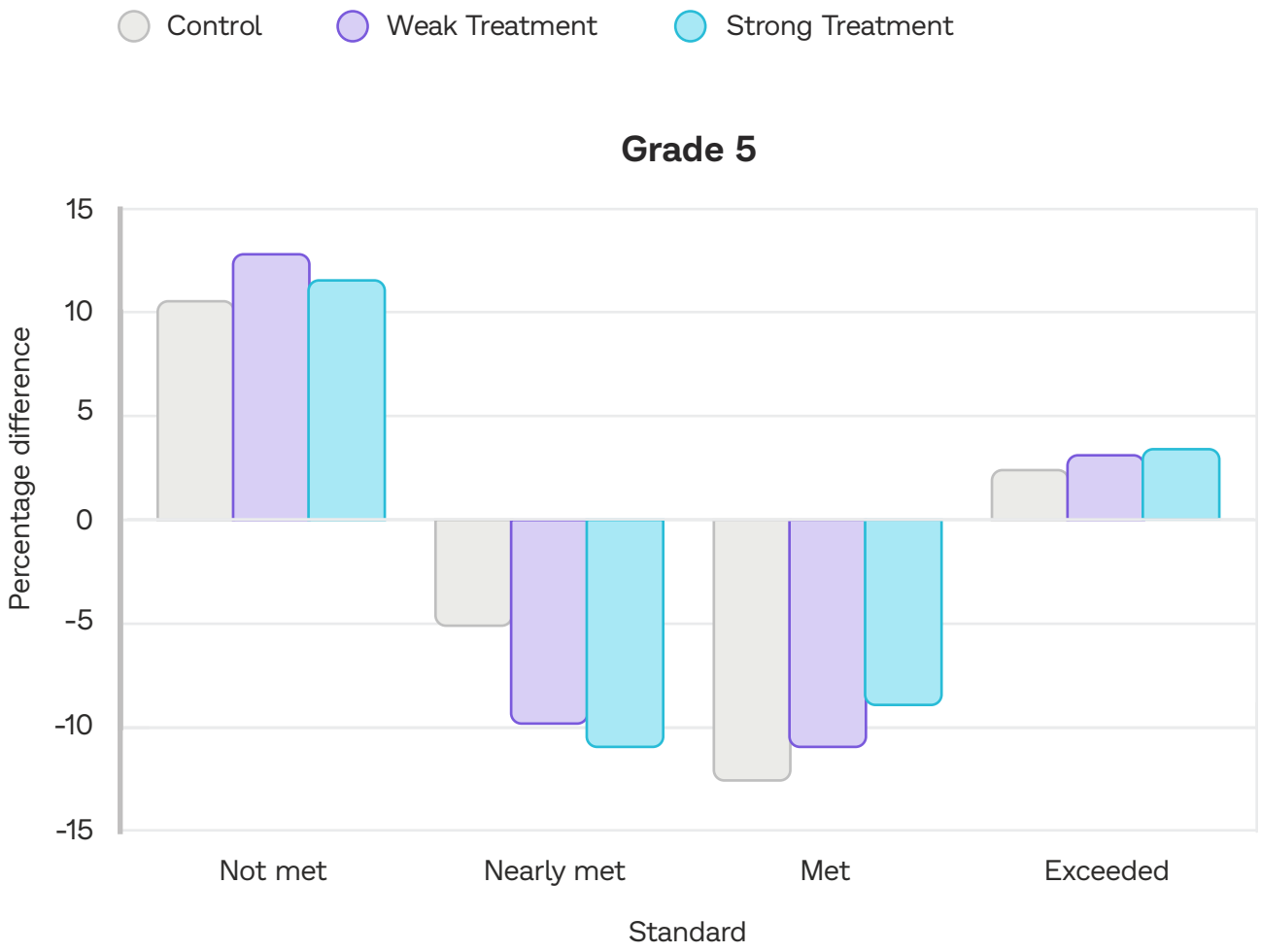


Figure 6. Bar charts showing changes in the percentage of students in a grade who did not meet, nearly met, met, or exceeded expectations between 2018 and 2019.

Data Quality Issues

Despite the encouraging findings, there were significant data quality issues that disqualify this study from ESSA Tier II consideration. First of all, to estimate high vs. low Prodigy usage in the 2018-19 school year for the purpose of identifying treatment and control groups, several Prodigy datasets had to be combined by linking schools to teachers to classes to students. Because personally identifiable information was not collected when a new Prodigy account was created, we were not able to verify if a teacher account was created by a real teacher or if multiple student accounts were created by the same student. This led to potential problems such as a class created in Prodigy was not an actual class or a student had multiple accounts with various levels of usage attached to the same class. As an example, data quality check spotted numerous instances where the percentage of monthly learning students in a school grade was over 100%, which should not be possible. There could be a number of reasons for this, including unreliable or outdated grade enrolment information in ELSi from which school records were obtained, erratic teacher-school links, or erratic teacher or student activities. Second, we were not able to track changes to teachers and classes over time. As a result, we could not determine if a teacher added new students from a new school semester to his/her old Prodigy class or if a teacher, after getting a new job at another school, kept using the same class that was attached to the old school. In addition, because the study used a retrospective QED design, we cannot ensure that the control group did not use other educational programs which might have impacted their achievement. These data quality issues must be addressed through improved data collection process and alternative study designs in order to meet ESSA requirements.

Conclusion

This study represents the first foray into using a quasi-experimental design for our efficacy research. By comparing treatment to control groups, findings from this study provide preliminary evidence suggesting that Prodigy is effective in improving math achievement. Specifically, Californian students in school grades that used Prodigy throughout the 2018-19 school year showed greater improvement in their CAASPP scores at the grade-level aggregate than students who did not use Prodigy. This was shown by treatment grades having lower 2018 CAASPP scores but similar 2019 CAASPP scores compared to control grades. Students in the treatment grades were able to make up the ground, with Prodigy being a potential contributing factor. In addition, there was a significant increase in the percentage of students who met or exceeded expectations in 2019 compared to their 2018 expectations in both treatment groups in comparison with the control group. However, because the data quality issues discussed above, the findings from this study do not qualify for ESSA Tier II standard. The regression coefficients should not be generalized or used for extrapolation, but be treated as an approximation of the effect of Prodigy. Overall, the findings from this study are encouraging. Better quality data and analyses at the student level would paint a clearer picture of Prodigy efficacy.