WestEd
WestEd.org

Evaluation of Beast Academy in Mankato Area Public Schools

Technical Report Prepared for

Art of Problem Solving

Thomas Torre Gibney
Juan Carlos Bojorquez
March 2022

Summary of Findings

WestEd was contracted by Art of Problem Solving (AoPS) to evaluate the impact of Beast Academy (BA) use on students' achievement and attitudes in Mankato Area Public Schools (MAPS) in Mankato, Minnesota. The two-year study focused on Beast Academy's effect on mathematics achievement among the population of MAPS students who began the study enrolled in grades 3, 4, and 5 in school year (SY) 2019–20 and remained enrolled in the district through grades 4, 5, and 6 in school year 2020–21. This primary analysis was complemented by survey analysis investigating students' attitudes and perceptions toward math during the 2020–21 school year. The evaluation team sought to answer the following questions:

1. Does two years of participation in Beast Academy have a positive impact on students' math achievement?
2. Does the impact of Beast Academy differ for Cluster and Rising Scholar students?
3. To what extent do students who used Beast Academy in SY 2020–21 report different attitudes and perceptions toward math compared with their peers who did not use Beast Academy?

Results from the evaluation indicated that Beast Academy use in MAPS had a positive, statistically significant effect on math achievement after two years of exposure to the program. The magnitude of the effect varied depending on students' Cluster or Rising Scholar status, with the largest effects seen for Rising Scholar students. Effects were also greater for students with higher use on the Beast Academy Online platform, as measured by the number of lessons attempted over the two-year intervention period. The higher the number of lessons attempted, the larger the gap in math achievement compared to students who had no exposure to Beast Academy Online.

In addition, the evaluation team found that students who reported using Beast Academy in SY 2020–21 had more positive attitudes and perceptions toward math on six of eight attitudinal domains measured on the student survey. Beast Academy students reported higher scale scores on the domains of *motivation to study*, *effort and perseverance*, *interest in mathematics*, *self-concept in mathematics*, *self-efficacy in math*, and *beliefs about mathematics as a learnable subject*. This positive effect on attitudes and perceptions was consistent across domains and grade levels.

The following sections provide technical details related to the methods used to conduct the study, including discussions of the study samples, measures, and analytic approaches for both the achievement and the student survey analyses. Complete analysis results are presented following each section.

Study Methods & Results

Introduction

WestEd's two-year evaluation of the impact of Beast Academy combined two complementary strands of analysis: an impact analysis assessing the effect of Beast Academy use on students' performance in mathematics, and a survey analysis of students' math attitudes and perceptions associated with Beast Academy use. The evaluation was focused on the impact of Beast Academy on students in grades 3, 4, 5, and 6. WestEd presents findings from the study in two formats: a Technical Report and a Summary Memo. This technical report includes detailed descriptions of the two complementary strands of analysis, including details on the methods, analysis, and sample. A separate document containing a streamlined overview of the study and findings, including a discussion and reflections on the findings, can be found in the accompanying Summary Memo (Beast Academy Evaluation Summary Memo).

The impact analysis employed a quasi-experimental design that matched students with a record of Beast Academy use during the two-year study period with an equivalent comparison group of students who were not exposed to the program. The analysis then compared the average outcomes of Beast Academy, or treatment, students with those of their matched comparison group peers. The survey analysis examined the average attitudes of students who self-reported using Beast Academy in the 2020–21 school year and compared them with the average attitudes of students who did not report using Beast Academy.

It bears noting that WestEd's original study of Beast Academy was designed as a one-year study during the 2019–20 school year. Due to COVID-19-related disruptions, the study was modified. Among these modifications were extending the study to a two-year period, changing the performance outcome measure from the state assessment to the district interim assessment, and broadening the study's focus on student outcomes by examining multiple measures of treatment. Additionally, there were changes that occurred during the study period related to students' instruction—including transitioning between in-person and remote learning—and in the district's use of Beast Academy. Changes in learning conditions such as these are likely to have had some influence on student outcomes; however, the evaluation team was limited in its ability to account for such changes. Nonetheless, WestEd made concerted efforts to retain a rigorous design, and supplemented study-collected data about Beast Academy exposure with data from the district and AoPS to ensure fidelity of treatment/comparison designation. Consequently, although WestEd is confident in the rigor of the design and analysis, we advise caution in interpreting the results given these data were collected during an unusual time in students' education.

Below, we describe the achievement and survey analyses in greater detail, including the specific study measures, samples, and analytic approaches used to produce the results. Findings are presented following the technical details of each analysis.

Student Achievement Analysis

Study Measures

The evaluation team used three types of data to conduct the analyses of student achievement.

Achievement data. The primary outcome of interest was elementary-to-middle-grades students' performance in mathematics over the two-year intervention period[1] spanning school years 2019–20 and 2020–21. MAPS provided WestEd with achievement data on students enrolled in grades 2 through 4 in SY 2018–19, the study baseline year (i.e., the year prior to the study intervention period), and achievement data on the same students enrolled in grades 4 through 6 two years later, in SY 2020–21. The data source was the district-administered NWEA Measures of Academic Progress (MAP) interim assessments in mathematics, administered during the spring testing window. Baseline (pre-test) data came from the 2019 spring window and outcome (post-test) data came from the 2021 spring window.[2] WestEd used the RIT (Rasch Unit) scale scores to measure student performance in mathematics over the two-year intervention period.

Administrative data. In addition to student-level achievement data, MAPS provided WestEd with de-identified district administrative data containing information on each student's demographic characteristics and participation in the Beast Academy program. These data were provided with the achievement data at the student-teacher-school level, allowing students to be linked to their school and their math teacher of record. Demographic variables included students' gender, race/ethnicity, English learner status, special education status, and whether the student received free or reduced-priced meals at school (a proxy for economic disadvantage). Demographic variables were used as statistical controls (i.e., covariates) in models; they were not used to conduct any subgroup analyses based on these characteristics. Programmatic variables included indicators for whether students were designated Cluster or Rising Scholar students in each year. The evaluation team used these indicators to separate students into initial treatment and comparison groups. Additionally, the district provided an indicator for whether each student spent the full 2020–21 school year in a distance learning environment due to the COVID-19 pandemic. This measure was also used as a covariate in the impact models.

[1] For the purposes of this study, the intervention period was defined as school years 2019–20 and 2020–21. However, MAPS used Beast Academy prior to the intervention period. Consequently, some students may have used Beast Academy prior to the study's intervention period.

[2] WestEd explored using the state standardized tests, the Minnesota Comprehensive Assessments (MCA), but the level of missing data on these assessments (due to COVID-19-related disruptions) far exceeded that of the NWEA data. Specifically, 56 percent of students within the study population had complete data on the MCA, meaning they had test scores at both the pre-test and post-test time points, compared to 92 percent of students who had complete data on the NWEA assessments. Moreover, since second grade was not a tested grade on the MCA during the study period, the absence of grade 2 baseline data would have prevented the evaluation team from estimating effects for the subsample of students who were enrolled in fourth grade in the 2020–21 school year. For these reasons, the NWEA MAP assessments were selected as the most appropriate data source for the achievement analyses.

Beast Academy usage data. To verify the treatment statuses of each student, and to develop a more nuanced measure of treatment, WestEd received de-identified student-level usage data from MAPS and Art of Problem Solving containing information about students' exposure to the Beast Academy Online platform. Two key variables informed WestEd's conceptualization of exposure to treatment, identified in collaboration with MAPS and AoPS staff: (1) the total number of lessons attempted on the Beast Academy Online platform, and (2) total time spent on math learning activities on the Beast Academy Online platform. Both variables were measured cumulatively over the two-year intervention period: values represented the combined totals for each student over the 2019–20 and 2020–21 school years.

Treatment Fidelity

Treatment was defined as having a record of student Beast Academy use in each of the two school years, as measured by total time spent on math learning activities on the platform. Students with no record of Beast Academy use in either year were identified as potential comparison students. Since the intervention period covered two years, students with a partial record of Beast Academy use—that is, usage in only one of the two years—were excluded from analyses given that they did not meet the definition of treatment *and* could not be selected as comparison students due to the potential for contamination. After identifying treatment and comparison students based on the usage data, WestEd then confirmed each student's status by cross-checking the usage patterns against the district's indicators for Cluster and Rising Scholar status in the administrative data. Once missing data on the outcome variables and demographic measures had been accounted for, this step resulted in three groupings of treatment: students with a two-year record of Beast Academy use who were designated by the district as Cluster students at any time during the same two-year period; students with a two-year record of Beast Academy use who were designated by the district as Rising Scholar students at any time during the same two-year period; and students with a two-year record of Beast Academy use who did not have a corresponding designation as either a Cluster or Rising Scholar student during the same two-year period. A small number (<10) were designated by the district as both Cluster and Rising Scholar students over the two years. These students were included in both the Cluster and Rising Scholar groups. Altogether, WestEd's approach resulted in four groups of students based on their treatment status: (1) Cluster students (n=234); (2) Rising Scholar students (n=87); (3) students who were neither Cluster nor Rising Scholar but who nonetheless used the Beast Academy Online platform during the intervention period (n=255); and (4) the pool of potential comparison students who were eligible to be selected as matches in the quasi-experimental analyses due to their having no record of Beast Academy use during the intervention period (n=459). Groups 1, 2, and 3 comprised the treatment group. Treatment and comparison definitions were strictly limited to the two-year intervention period; the analysis could not account for whether students used Beast Academy prior to the 2019–20 school year.

Sample

The grand sample for the achievement analysis was 1,027 students. They were enrolled in roughly equal proportions across grades 2, 3, and 4 in the baseline year of 2018–19 and grades 4, 5, and 6 two years later in the outcome year of 2020–21. Specifically, 344 students (or about 34 percent) comprised the group of students who began the study in grade 2 and ended the study in grade 4; 356 students (or

about 35 percent) comprised the group of students who began the study in grade 3 and ended the study in grade 5; and 327 students (or about 32 percent) comprised the group of students who began the study in grade 4 and ended the study in grade 6. Of the grand total of 1,027 students across all grades, 55 percent were identified as treatment students. The remaining 45 percent represented the pool of potential comparison students who did not have record of Beast Academy program use at any point during the 2019–20 and 2020–21 school years. Table 1 summarizes the characteristics of the grand sample in terms of average mathematics achievement, exposure to the treatment, and demographics. From this grand sample, WestEd conducted matching procedures to identify the subset of treatment and comparison students who most closely resembled each other in terms of baseline achievement and demographics; table 1 presents the characteristics of the sample prior to conducting the match.

Table 1. Characteristics of the student achievement sample prior to the matching process

Variable	Mean	Standard Deviation	Min	Max
18–19 MAP Math RIT Scale Score *(pre-test)*	208.2	15.89	155	263
20–21 MAP Math RIT Scale Score *(post-test)*	221.6	16.35	162	266
Hispanic/Latino	0.0574	0.233	0	1
American Indian or Alaska Native	0.00292	0.0540	0	1
Asian	0.0166	0.128	0	1
Black or African American	0.0964	0.295	0	1
Native Hawaiian or Other Pacific Islander	0.00195	0.0441	0	1
White	0.781	0.414	0	1
Two or More Races	0.0438	0.205	0	1
English Learner	0.0545	0.227	0	1
Special Education	0.130	0.336	0	1
Free and Reduced Lunch	0.301	0.459	0	1
Female	0.483	0.500	0	1
20–21 Distance Learning Full Year	0.0983	0.298	0	1
Time Math Learning Activities *(hours)*	9.664	17.27	0	224.65
Total Lessons Attempted	103.6	177.7	0	1,294
Beast Academy *(treatment status)*	0.553	0.497	0	1
Rising Scholar	0.0847	0.279	0	1
Cluster	0.228	0.420	0	1

Variable	Mean	Standard Deviation	Min	Max
Beast Academy but not Rising Scholar or Cluster	0.248	0.432	0	1

Note: N = 1,027. Means for all variables except MAP scale scores, time learning math, and total lessons attempted are presented as proportions.

Matching Process

From the grand sample of treatment and potential comparison students, WestEd used propensity score matching to identify ideal matches of students that were well balanced on baseline characteristics in the 2018–19 school year, the year prior to the intervention period. Propensity score matching is a multivariate matching algorithm that identifies one or more comparison students for each treatment student with similar pretest measures of the outcomes and demographic characteristics (Stuart, 2010). The evaluation team matched students on their baseline school, math RIT scale score, and demographic characteristics (race/ethnicity, gender, economic disadvantage, special education status, and English learner status). Each treatment student was matched with one comparison student with the closest propensity score based on the totality of these characteristics (a method known as nearest-neighbor matching). WestEd used matching with replacement to assemble a comparison group that most closely resembled the treatment group and to minimize the potential for bias in subsequent analyses.

The matching process resulted in a sample of 762 students (547 treatment students; 215 comparison students)[3] who were well balanced on the average achievement and demographic characteristics described above. Most importantly, the matched treatment and comparison groups were equivalent at baseline on the pretest measure of achievement, a key criterion for establishing the rigor of the analysis (U.S. Department of Education, 2020). As shown in table 2 (below), in SY 2018–19, the average or mean score on the spring NWEA MAP assessment in mathematics was 212.11 among students who would go on to use Beast Academy the following two years, compared to an average of 212.64 among their matched comparison group peers who did not go on to use Beast Academy the following two years (a difference of .53 scale points). By contrast, in the unmatched sample, the Beast Academy group mean was 213.39, compared to a mean of 201.82 for non-Beast Academy students (a difference of 11.57 scale points). Reducing the unmatched sample to the matched sample thus made the treatment and comparison groups very similar to each other in terms of their prior achievement levels, resulting in a 95.5 percent reduction in bias between the two groups. That is, the two groups would have been significantly different from each other in terms of their starting achievement levels had the match not occurred. The establishment of baseline equivalence on the pretest measure of achievement thus allows

[3] Some comparison students were matched more than once with selected treatment students to achieve the most balanced 1-to-1 matched sample, an artifact of the "matching with replacement" method described in the preceding paragraph. The statistical models upon which all results are based included analytic weights to account for multiple-matched observations. See the Analysis Methods and Results sections for more details.

the evaluation team to have greater confidence that any observed difference in math achievement between the groups two years later is unlikely to be due to differences in their prior achievement levels.

Table 2. Baseline equivalence results for 2018–19 MAP math test scores

Sample	BA Mean	BA SD	BA N	Non-BA Mean	Non-BA SD	Non-BA N	Mean diff.	p-value	Effect size
Unmatched sample	213.39	14.85	568	201.82	14.79	459	11.57	0.000	0.781
Matched sample	212.11	13.48	547	212.64	13.11	215	-0.53	0.707	0.040

Note: Reduction in bias from the unmatched to the matched sample was 95.5 percent. *SD* = standard deviation. The effect size is Cohen's d.

In addition to meeting baseline equivalence on the pretest measure of achievement, the matched sample was also equivalent on all other demographic characteristics included in the matching model.

Analysis Methods

WestEd estimated a series of regression models to analyze the impact of Beast Academy on students' math achievement after two years of treatment. Frequency weights were used to account for repeated selection into the comparison group resulting from the matching step, meaning some comparison students were counted more than once in the analyses. Standard errors were clustered at the individual level. WestEd estimated separate models to describe: (1) the effect of any Beast Academy use on students' math achievement; (2) the differential effect of Beast Academy use for Cluster students and Rising Scholar students; and (3) the effect of Beast Academy on students' math achievement at different levels of Beast Academy use. The models included all the baseline characteristics used to conduct the match plus the additional indicator for whether each student spent the full 2020–21 school year in a distance learning environment due to the COVID-19 pandemic.

For the third model examining different levels of Beast Academy use, WestEd created several "use profiles" of treatment students based on the distribution of lessons attempted during the intervention period. The profiles were determined as a function of each student's percentile rank along the distribution of lessons attempted, and thus refer to a given student's level of engagement on the platform relative to that of all matched treatment students. That is, Beast Academy students were assigned a percentile rank based on the way that MAPS students used the Beast Academy Online platform over the two-year intervention period (by contrast, comparison students had no record of use and thus no associated percentile rank). For example, a treatment student with a 50th percentile rank in terms of usage attempted more lessons than half of their treatment peers. In the MAPS study sample, this percentile rank corresponded to attempting 120 lessons over the two years. WestEd defined the 50th percentile student profile as the *average Beast Academy user*, where "average" refers to the

median value of lessons attempted.[4] Profiles were also created for a *low user* (13 lessons attempted; 10th percentile rank), *below average user* (45 lessons attempted; 25th percentile rank), *above average user* (257 lessons attempted; 75th percentile rank), and *high user* (442 lessons attempted; 90th percentile rank). The analyses estimated the predicted increase (or *marginal effect*) in math test scores associated with each use profile. For example, the results for the average Beast Academy user refer to the effect on math performance associated with attempting 120 lessons over a two-year period.

Results

Results from the achievement analyses indicated that Beast Academy students outperformed their matched comparison group peers on the 2021 spring NWEA MAP math assessment by an average of 8.78 scale points (table 3, below). Cluster students scored on average 10.14 points higher than comparison students, while Rising Scholar students scored on average 11.17 points higher than comparison students. The third treatment group, students who were Beast Academy users (based on usage records from the Online platform) but who were not designed as Cluster or Rising Scholar, scored on average 6.74 points higher than comparison students. All these results were statistically significant at the $p < .001$ level.

Table 3. Model results for the student achievement analyses, by Beast Academy status and treatment group

Treatment group	Mean math score, BA students	Mean math score, non-BA students	Mean difference	Standard error	p-value	Treatment N
Any Beast Academy designation	228.28	219.50	8.78***	1.052	0.000	547
Cluster students	229.64	219.50	10.14***	1.382	0.000	211
Rising Scholar students	230.67	219.50	11.17***	1.437	0.000	86
Beast Academy user, neither Cluster nor Rising Scholar	226.24	219.50	6.74***	1.057	0.000	250

Note: Weighted N = 1,094. All results were statistically significant at the p<0.001 level (***).

When using the more refined measure of treatment—students' usage levels on the Beast Academy Online platform, as measured by the number of lessons attempted over the two-year intervention

[4] WestEd also explored using alternative measures to group students according to different use patterns, such as using standard deviations from the arithmetic mean and using quantile-based groupings of usage data. Ultimately, the percentile-rank framework was chosen as the most appropriate measure given the skewed distribution of the data, which featured a high number of observations at the lower end of the distribution (e.g., in the 2-to-50 lessons attempted range) and comparatively fewer observations at higher levels of usage.

period—Beast Academy students also outperformed their matched comparison group peers (table 4, below). Specifically, the average Beast Academy user in MAPS (i.e., the student at the 50th percentile) attempted 120 lessons on the platform, a level of usage that resulted in a 4.94-point average increase in MAP math scores over the comparison group mean. A student who met the profile for a low user (10th percentile) scored an average of 2.63 points higher than comparison students; a below average user (25th percentile) scored 3.32 points higher than comparison students; an above average user (75th percentile) scored 7.90 points higher than comparison students; and a high user (90th percentile) scored 11.89 points higher than comparison students. These results were statistically significant at the $p < .001$ level. In addition, WestEd conducted sensitivity analyses to determine whether the results were affected by the inclusion of students with atypical use patterns (e.g., students who could be considered outliers due to exceptionally high levels of engagement on the platform). These sensitivity tests indicated that results did not materially change when removing outliers. Thus, the results presented below are based on the full matched sample, without censoring the data to remove outliers.

Table 4. Model results for the student achievement analyses, by levels of Beast Academy use

Use profile	Usage level (in lessons attempted)	Corresponding percentile rank	Marginal effect	Predicted math score	Standard error	p-value
Low user	13	10th	2.63***	222.08	0.705	0.000
Below average user	45	25th	3.32***	222.78	0.635	0.000
Average Beast Academy user	**120**	**50th**	**4.94*** **	224.40	0.499	0.000
Above average user	257	75th	7.90***	227.36	0.441	0.000
High user	442	90th	11.89***	231.35	0.765	0.000

Note: Weighted N = 1,094. All results were statistically significant at the p<0.001 level (***). The "average Beast Academy user" is defined as a student with the median number of lessons attempted during the intervention period. Corresponding percentile rank refers to where a given student fell on the distribution of lessons attempted. The marginal effect refers to the predicted change in the coefficient on math performance associated with the corresponding number of lessons attempted, and the predicted math score is the regression-adjusted score associated with that number of lessons attempted. The model includes all matched treatment and comparison students, without censoring the data to remove students with atypical use patterns.

Student Survey Analysis

Study Measures

To develop the student survey, WestEd collaborated with MAPS and Art of Problem Solving to identify a set of measurable constructs related to students' attitudes and perceptions toward math. Drawing on the existing literature on this topic, the evaluation team identified eight discrete domains whose associated measures have been validated through prior research. The language of some items on two

constructs was modified slightly to specify "math" content. These items are identified with an * in table 5 below. The eight domains are defined in their respective sources as follows:

- *Motivation to study* is a composite attribute drawn from the Computer Attitude Questionnaire (CAQ) that captures students' motivation/persistence and study habits (Christensen & Knezek, 2014). Items include statements such as, *"When I don't understand a problem, I keep working until I find an answer."*
- *Effort and perseverance*, drawn from the Organisation for Economic Co-operation and Development's (OECD) Student Approaches to Learning (SAL) instrument, refers to the "volitional process" governing students' approaches to learning, that is, the degree of priority invested in a learning task (Marsh et al., 2006). Items include statements such as, *"When I study for math, I put forth my best effort."*
- *Interest in mathematics* refers to students' intrinsic motivation to learn math, often arising from feelings of involvement, stimulation, or personal significance (Marsh et al., 2006). Items include statements such as, *"When I learn math, I sometimes get totally absorbed."*
- *Cooperative learning* refers to students' preferences for learning in groups (Marsh et al., 2006). Items included statements such as, *"I learn most when I work with other students."*
- *Self-concept in mathematics* refers to students' "cognitive appraisals and judgments" of their own abilities in math (Marsh et al., 2006). Items include statements such as, *"I do well in math."*
- *Growth mindset* is defined as "the belief that intelligence and academic ability are not fixed but are malleable and can be increased through effort and learning" (Snipes & Loan, 2017). Items include statements such as, *"My intelligence is something that I can't change."*
- *Self-efficacy in math* captures a combination of students' effort, initiative, and perceptions that they can complete math tasks (Blazar & Kraft, 2017). Items included statements such as, *"Even when math is hard, I know I can learn it."*
- *Beliefs about mathematics as a learnable subject* refers to the extent to which students believe that math can be learned, in contrast to the perception that math is only accessible to high-ability students (De Corte, 2015). Items included statements such as, *"If I work hard, then I will understand math."*

To measure each of the domains described above, WestEd constructed scale composites of the survey items that made up each domain, consistent with their use in prior studies. Alpha coefficients were computed for each composite measure to assess the reliability of each scale (see below, "Scale Reliability"). Each student in the dataset received a composite score for each domain, representing their average attitudes or perceptions with respect to the items in that scale. See table 5 for the complete list of items in each domain.

Table 5. Attitudinal domains, component items, and associated response scale contained in the student survey

Component items	Associated response scale
Domain: Motivation to study	
I do my math work by myself without anyone making me do it.*	1 = Strongly Disagree 2 = Disagree 3 = Agree 4 = Strongly Agree
If I do not understand something, I keep thinking about it.	
When I don't understand a problem, I keep working until I find the answer.	
I review my math work every day.*	
I try to finish whatever I begin.	
Sometimes, I change my way of doing my math work.*	
I enjoy working on a difficult problem.	
I think about many ways to solve a difficult problem.	
I never forget to do my math work.*	
I like to work out problems which I can use in my life every day.	
If I do not understand my teacher, I ask him/her questions.	
I listen to my teacher carefully.	
If I fail, I try to find out why.	
I work hard.	
When I do a job, I do it well.	
Domain: Effort and perseverance	
When I learn math, I work as hard as possible.*	1 = Almost Never 2 = Sometimes 3 = Often 4 = Almost Always
When I learn math, I keep working even if the math is difficult.*	
When I learn math, I try to do my best to learn the math taught.*	
When I study for math, I put forth my best effort.*	
Domain: Interest in mathematics	
When I learn math, I sometimes get totally absorbed.	1 = Disagree 2 = Disagree Somewhat
Because learning math is fun, I wouldn't want to give it up.	

Component items	Associated response scale
Math is important to me personally.	3 = Agree Somewhat 4 = Agree
Domain: Cooperative learning	
I like to work with other students.	1 = Disagree 2 = Disagree Somewhat 3 = Agree Somewhat 4 = Agree
I learn most when I work with other students.	
I like to help other people do well in a group.	
It is helpful to put together everyone's ideas when working on a project.	
Domain: Self-concept in mathematics	
I do well in math.	1 = Disagree 2 = Disagree Somewhat 3 = Agree Somewhat 4 = Agree
Math is one of my best subjects.	
I have always done well in math.	
Domain: Growth mindset	
My intelligence is something that I can't change.	1 = Not True at All 2 = A Little True 3 = Somewhat True 4 = Mostly True 5 = Completely True
Challenging myself won't make me smarter.	
There are some things I cannot learn.	
If I am not naturally smart in a subject, I will never do well in it.	
Domain: Self-efficacy in math	
I have pushed myself hard to completely understand math.	1 = Totally Untrue 2 = Mostly Untrue 3 = Neither True nor Untrue 4 = Mostly True 5 = Totally True
If I need help with math, I make sure that someone gives me the help I need.	
If a math problem is hard to solve, I often give up before I solve it.	
Doing homework math problems helps me get better at doing math.	
Math is too hard.	
Even when math is hard, I know I can learn it.	
I can do almost all the math in this class if I don't give up.	
I'm certain I can master the math skills taught.	
When doing math, I focus on learning more than I focus on how long it takes.	
I have been able to figure out the most difficult problems in math.	

Component items	Associated response scale
Domain: Beliefs about mathematics as a learnable subject	
Making mistakes is an important part of learning math.	1 = Totally Disagree 2 = Disagree 3 = Somewhat Disagree 4 = Somewhat Agree 5 = Agree 6 = Totally Agree
My math teacher thinks that mistakes are okay and can be helpful for learning.	
My math teacher appreciates it when we have tried hard, even if I make mistakes.	
My math teacher tries to make math interesting.	
Math is used by a lot of people in their daily life.	
I can understand even the most difficult math problems.	
If I work hard, then I will understand math.	
Anyone can learn math.	

Note: Items marked with an asterisk (*) were modified slightly to specify "math" content.

Sample

The survey was administered in the spring of 2021 to the population of MAPS 4th, 5th, and 6th graders whose parents had previously consented to their participation in the study. Students were asked to confirm their assent prior to answering the survey questions. WestEd received 1,280 responses, 88 of which (or 6.8 percent) were removed prior to analysis due to students opting out at the point of survey administration. From the remaining sample of 1,192 assenting respondents, 71 students did not complete the survey after the assent question; these observations were removed due to missing data. An additional 60 students did not answer the item asking respondents if they had used Beast Academy in the past year; these observations were also removed since the evaluation team could not estimate results for students with missing treatment statuses. For the purposes of the survey analysis, treatment status was determined by students' self-reported use of Beast Academy. The resulting sample comprised a total of 1,061 students who assented to participate in the study, completed the survey, and whose treatment status could be confirmed.

The full sample of 1,061 students had low overall levels of missing data (6 percent of all items missing, on average), but non-response patterns varied substantially by item. To minimize the potential for bias due to missing data, scale composites were constructed using the subsamples of observations that had complete responses to all the items in each scale, resulting in eight different samples (one for each scale). Reliability tests were conducted on these non-missing samples as well as the samples that contained missing data.

Scale Reliability

Results from the reliability tests indicated that all scales except growth mindset held together as expected, meaning their underlying response patterns reliably captured the constructs of interest. All constructs had reliability estimates above the generally accepted threshold, with the exception of

growth mindset. (Results for growth mindset are discussed below following table 6.) Consequently, evaluators could be confident that the scales were measuring what they were intended to measure. The results in table 6, below, display the properties of the scales constructed using the non-missing data samples of observations. As an additional sensitivity check, WestEd also conducted reliability tests on the samples with missing data, revealing negligible differences in the alphas within each pair of scales. This confirmed that the scales based on the non-missing samples were preferrable to the scales with missing data, since the results were less likely to be affected by non-response bias and there was no appreciable loss in reliability by restricting the samples to non-missing observations.

Table 6. Reliability test results for the student survey scales

Scale	Number of items	Average inter-item covariance	Cronbach's alpha	Sample size
Motivation to study	15	0.1633809	0.8509	977
Effort and perseverance	4	0.2718259	0.7872	1,049
Interest in mathematics	3	0.5208319	0.7787	1,027
Cooperative learning	4	0.3203166	0.7211	1,035
Self-concept in mathematics	3	0.6484448	0.8581	1,045
Growth mindset	4	0.1477337	0.5546	69
Self-efficacy in math	10	0.4281949	0.8495	980
Beliefs about mathematics as a learnable subject	8	0.3958709	0.8118	1,023

Note: Growth mindset was not included in subsequent analyses due to high levels of missing data and failure to meet the generally accepted reliability threshold of alpha = .70 or higher.

Responses to the four items that comprised the growth mindset scale featured high levels of missing data. Fully 94 percent, or 992 students out of the full sample of 1,061, were missing at least one response to this set of items; 32 percent, or 340 students, failed to respond to all four items. In addition, scale reliability for this domain (alpha = .55) fell below the generally accepted threshold of .70 (Marsh et al., 2006), indicating that this composite measure was not capturing the underlying construct of growth mindset. As a result, the evaluation team considered estimates related to growth mindset to be unreliable. Diagnostic tests of the growth mindset models confirmed that these outcomes could not be estimated with such high levels of missing data. Growth mindset outcomes are therefore not included in this report.

Analysis Methods

To analyze the differences in math attitudes and perceptions between students who reported using Beast Academy in the 2020–21 school year and those who did not, WestEd regressed an indicator for treatment status on the scale composite scores for each of the seven attitudinal domains (excluding growth mindset). The regression models included school fixed effects to control for unobserved differences in students' school settings; this approach compared Beast Academy and non-Beast Academy students within the same school. Models calculated the average scale scores for each group within their schools and then averaged these estimates across all schools in the sample. The resulting output identified significant differences between the average responses of Beast Academy students and those of their non-Beast Academy peers.

Results

As shown in Table 7, below, Beast Academy students reported more positive perceptions and attitudes related to math on six of the eight attitudinal domains: *motivation to study*, *effort and perseverance*, *interest in mathematics*, *self-concept in mathematics*, *self-efficacy in math*, and *beliefs about mathematics as a learnable subject*. The largest positive association between Beast Academy use and math attitudes was found for the *self-concept in mathematics* measure (.71 average difference between Beast Academy students' scale scores and non-Beast Academy students' scale scores). The second-largest positive association was found for the *self-efficacy in math* measure (.47 average scale-score difference). These results were followed by *interest in mathematics* (.30), *effort and perseverance* (.29), *beliefs about mathematics as a learnable subject* (.26), and *motivation to study* (.24) in order of descending magnitude of differences between the two groups. No significant differences were found between Beast Academy and non-Beast Academy students on the *cooperative learning* domain (*growth mindset* was not included in analyses). These results were consistent when disaggregating the analyses by grade level, with only fractional differences observed in the magnitude of the differences and their associated p-values.

Table 7. Model results for the student survey analyses

Domain	Mean for Beast Academy students	Mean for non-Beast Academy students	Mean difference	Standard error	p-value	Domain sample size
Full sample (All grades)						
Motivation to study	3.178	2.942	0.236***	0.0283	0.000	977
Effort and perseverance	3.570	3.278	0.292***	0.0370	0.000	1,049
Interest in mathematics	3.056	2.761	0.295***	0.0521	0.000	1,027
Cooperative learning	3.297	3.300	-0.003	0.0431	0.946	1,035

Domain	Mean for Beast Academy students	Mean for non-Beast Academy students	Mean difference	Standard error	p-value	Domain sample size
Self-concept in mathematics	3.496	2.790	0.706***	0.0520	0.000	1,045
Self-efficacy in math	2.938	2.472	0.466***	0.0449	0.000	980
Beliefs about mathematics as a learnable subject	5.326	5.065	0.261***	0.0444	0.000	1,023
Grade 4 sample						
Motivation to study	3.188	2.918	0.27***	0.0490	0.000	364
Effort and perseverance	3.581	3.246	0.335***	0.0605	0.000	405
Interest in mathematics	3.149	2.788	0.361***	0.0842	0.000	396
Cooperative learning	3.345	3.271	0.074	0.0663	0.266	398
Self-concept in mathematics	3.431	2.745	0.686***	0.0846	0.000	405
Self-efficacy in math	2.914	2.422	0.492***	0.0762	0.000	372
Beliefs about mathematics as a learnable subject	5.342	5.031	0.311***	0.0733	0.000	397
Grade 5 sample						
Motivation to study	3.184	2.932	0.252***	0.0404	0.000	396
Effort and perseverance	3.568	3.292	0.276***	0.0565	0.000	410
Interest in mathematics	3.077	2.779	0.298***	0.0797	0.000	402
Cooperative learning	3.383	3.403	-0.020	0.0636	0.752	403
Self-concept in mathematics	3.475	2.706	0.769***	0.0798	0.000	409
Self-efficacy in math	2.964	2.498	0.466***	0.0652	0.000	388
Beliefs about mathematics as a learnable subject	5.330	5.122	0.208**	0.0619	0.001	401
Grade 6 sample						
Motivation to study	3.200	2.956	0.244**	0.0708	0.001	217
Effort and perseverance	3.584	3.285	0.299**	0.0910	0.001	234
Interest in mathematics	2.914	2.630	0.284*	0.1310	0.031	229
Cooperative learning	3.040	3.145	-0.105	0.1210	0.385	234

Domain	Mean for Beast Academy students	Mean for non-Beast Academy students	Mean difference	Standard error	p-value	Domain sample size
Self-concept in mathematics	3.736	2.971	0.765***	0.1290	0.000	231
Self-efficacy in math	3.080	2.452	0.628***	0.1090	0.000	220
Beliefs about mathematics as a learnable subject	5.401	4.991	0.41**	0.1220	0.001	225

Note: Asterisked figures indicate results are significant at the p<0.05 level (*), the p<0.01 level (**), or the p<0.001 level (***). "Mean for Beast Academy students" refers to the regression-adjusted treatment group mean. "Domain sample size" refers to the subsample of non-missing observations for all items in that domain.

Conclusion

This technical report concludes with a brief list of finding highlights obtained from WestEd's evaluation of Beast Academy in Mankato Area Public Schools. For a complete discussion and reflection on these findings, please see the accompanying Summary Memo.

- The evaluation team found that Beast Academy had statistically significant, positive effects on mathematics achievement.
- The impact of Beast Academy on mathematics achievement varied by treatment group, with the largest effects seen for Rising Scholar students.
- The impact on achievement also varied by the level of engagement on the Beast Academy Online platform, as measured by the number of lessons attempted over the two-year intervention period.
- The evaluation team found that Beast Academy students had more positive attitudes and perceptions toward math on six of the eight domains measured.
- Beast Academy students were more likely than non-Beast Academy students to agree with positive statements related to their interest in and enjoyment of math in the current school year.

References

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis, 39*(1), 146–170.

Christensen, R., & Knezek, G. (2014). Comparative measures of grit, tenacity and perseverance. *International Journal of Learning, Teaching and Educational Research, 8*(1), 16–30.

De Corte, E. (2015). Mathematics-related beliefs of Ecuadorian students of grades 8–10. *International Journal of Education Research, 72*, 1–13.

Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6*(4), 311–360.

Snipes, J., & Loan, T. (2017). *Growth mindset, performance avoidance, and academic behaviors in Clark County School District* (REL 2017–226). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from http://ies.ed.gov/ncee/edlabs.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, *25,* 1−21. https://10.1214/09-STS313

U.S. Department of Education. (2020). What Works Clearinghouse: Standards handbook (Version 4.1). https://ies.ed.gov/ncee/wwc/handbook

Suggested citation: Torre Gibney, T., & Bojorquez, J. C. (2022). *Evaluation of Beast Academy in Mankato Area Public Schools: Technical Report*. WestEd.