Evaluation Plan for a District After-School Program

A Quasi-Experimental Mixed-Methods Approach

Prepared by the District Research Team

Introduction

What follows is a plan for evaluating the district's new after-school program. There are two parts. The formative part is about keeping tabs on things while the program is still getting off the ground, catching problems early so they do not turn into bigger ones. The summative part comes later and asks the tougher question: did the program actually make a difference for students after a full year? Each part covers roughly the same ground. We go through purposes, stakeholders, evaluation questions, methods, how data will be collected and analyzed, logistics, and budget.

Randomly assigning students to either participate or not participate in an after-school program is not really an option here. There are both ethical concerns and practical ones that make it a non-starter. So instead, this plan relies on a quasi-experimental design. For the summative piece, we'll use a nonequivalent comparison group design paired with propensity score matching (PSM), which gets us closer to the rigor of a true randomized controlled trial without requiring us to deny services to any student. The idea is to draw stronger causal conclusions about what the program does while still honoring the fact that participation is voluntary.

The whole evaluation is mixed-methods. Numbers alone won't cut it here. We'll collect quantitative data on academics, attendance, and behavior, but we'll also run interviews, hold focus groups, and send out open-ended surveys to hear directly from students, families, and staff about what the program actually feels like on the ground. Using both kinds of evidence together is what gives the findings real credibility.

Background

So here is the situation. The district is rolling out a brand new after-school program. On paper, it sounds simple enough: academic help, enrichment, a safe place for kids once school lets out. But the people at the top have higher expectations than that. They want to see grades go up, engagement improve, risky behavior go down, and they want working families to actually have somewhere reliable for their kids in the afternoon. The day-to-day programming is supposed to mix homework help with tutoring, STEM stuff, art, physical activity, and social-emotional learning (SEL for short).

The program is expected to serve students across several grade levels and multiple school buildings. Any student can sign up, and participation is voluntary, but the district plans to do targeted outreach to students flagged as at-risk based on grades, attendance, or behavior. This voluntary structure creates the central methodological headache for the evaluation: because students self-select into the program, participants may be systematically different from non-participants in ways that matter. Accounting for that selection bias is the biggest challenge this plan has to address.

The research literature on after-school programs points to a handful of factors that tend to separate effective programs from ineffective ones. Sustained participation, often referred to as dosage, matters a great deal, as do the quality of instruction, how well programming lines up with what students are learning during the school day, and the strength of relationships between staff and students (Durlak et al., 2010; Lauer et al., 2006). This evaluation will look at these implementation factors right alongside student outcome data, so that we end up with a fuller picture of both program quality and impact.

Evaluation Standards

Several established standards and frameworks will guide how this evaluation is designed and carried out. On the program quality side, we are drawing from the Joint Committee on Standards for Educational Evaluation, specifically their Program Evaluation Standards covering Utility, Feasibility, Propriety, Accuracy, and Accountability. The CASEL framework for SEL programming is relevant too, along with Missouri DESE standards for extended learning programs and the district's own curriculum standards and student learning outcomes.

On the research design side, we are leaning on What Works Clearinghouse standards for quasi-experimental work, the American Evaluation Association's Guiding Principles, and the evidence-level guidelines put out by the Institute of Education Sciences. None of these are optional extras. They keep the evaluation honest.

Formative Evaluation Plan

Purposes

The formative piece exists to keep tabs on the program while it is still getting off the ground. If something is not working, we want to know about it early enough to actually fix it. The focus will be on whether activities are being delivered as planned, how well staff are pulling off the program model, whether kids are showing up and staying engaged, and any early warning signs. We are not going to sit around until June to discover that half the sites went off-script in October. Findings go straight into real-time decisions about staffing, scheduling, and content. The main deliverable is a mid-year report with concrete recommendations.

Stakeholders

District administration is the primary client. They're the ones who asked the research team to put this plan together. But the stakeholder list extends well beyond the central office:

- **District Administration:** These are the folks holding the purse strings. They'll be the ones deciding whether the program keeps going, grows, or gets shut down.
- **Program Coordinators and Staff:** The people on the ground every day, actually running sessions and dealing with whatever comes up. Nobody knows what's working and what isn't better than they do.
- **Building Principals:** Principals deal with the logistics side of things: space, scheduling, and the sometimes-messy coordination between the regular school day and after-school hours.
- **Teachers:** Classroom teachers can spot changes that no one else would. If a student starts turning in homework more often or seems more focused, they're the first to know.
- **Students:** Without them, there is no program. Their engagement, satisfaction, and growth are the whole point of doing this evaluation in the first place.
- **Parents/Guardians:** Family buy-in can make or break attendance and retention. Parents need to trust the program enough to keep sending their kids.

Beyond these primary groups, the school board, community partners who contribute supplemental programming, and any grant agencies or funders involved will also have a stake in the findings.

Decisions

Formative findings will feed into several kinds of decisions. If scheduling or staffing ratios need adjusting, that comes from here. Same for targeted interventions when certain students have low attendance or engagement. The findings will also flag professional development needs for program staff, shape how we communicate with families, and call attention to spots where the after-school content does not line up with what students are doing during the school day.

Evaluation Questions

Four questions drive the formative evaluation. First, is the program actually being run the way it was drawn up? We need to know if planned activities are really happening at every site or if some have quietly gone off-script. Are there enough qualified staff? Has the schedule held steady?

Second, are students showing up and staying engaged? That means looking at daily and weekly attendance, total hours per week, and whether kids seem checked in or just physically present.

Third, what are the people involved saying about all of this? We want to hear from students, parents, and staff about what is working and what is frustrating.

Fourth, can we spot any early signs that things are headed in the right direction? Homework completion going up, for instance, or a shift in school-day attendance.

Participant Sample

We'll collect formative data from every student enrolled in the after-school program. All program staff will be surveyed as well. Since the staff is small enough, there's no need to sample. For parents and guardians, a stratified random sample will be drawn and surveyed, with the goal of hitting at least a 30% response rate so the results actually mean something. Observations will happen at each participating school site.

Methods

Data collection for the formative piece will draw on a mix of quantitative and qualitative strategies:

Data Collection Method	Type	Source	Frequency
Program attendance logs	Quantitative	Program staff	Daily
Structured program observations	Qualitative	Research team	Bi-weekly per site
Student engagement surveys	Quantitative	Students	Monthly
Staff reflection logs and interviews	Qualitative	Program staff	Monthly
Parent satisfaction surveys	Mixed	Parents/Guardians	Mid-year
Fidelity checklists	Quantitative	Research team / Staff	Weekly

For observations, the research team will use a structured protocol that looks at instructional quality, how engaged students are, how well the space is managed, and whether activities line up with program goals. The protocol will be adapted from established instruments, most likely the Out-of-School Time (OST) observation tool or something comparable. Staff reflection logs, meanwhile, give program staff a place to jot down weekly notes about what's going well, what's not, and any adjustments they've made on the fly.

Logistics

Formative work kicks off the moment the program launches and runs through the end of the first semester. One member of the research team will serve as the primary point of contact for program sites. Having a single liaison keeps communication cleaner. Observation schedules will go out ahead of time and rotate across sites so no one building

feels singled out or overly disrupted. Each month, the research team will sit down with program coordinators to go over preliminary findings and talk through what might need to change. By the close of the first semester, the team will deliver a mid-year formative report to district administration.

Budget

Item	Estimated Cost
Personnel (research team, ~100 hours @ $35/hr)	$3,500
Survey platform (e.g., Qualtrics)	$500 (or in-kind)
Observation protocol materials and printing	$200
Incentives for parent survey completion	$300
Data management software	$250
Contingency	$250

Total Estimated Formative Evaluation Budget: $5,000

Summative Evaluation Plan

Purposes

Where the formative evaluation asks "how is this going so far?," the summative evaluation asks the harder question: "did it work?" The goal here is to figure out whether the after-school program actually moved the needle on the things it was supposed to improve: academic achievement, school engagement, attendance, and student behavior. The evidence that comes out of this phase will be what district leaders rely on when deciding whether to keep the program running, expand it, retool it, or shut it down.

Stakeholders

The stakeholder list is largely the same as for the formative phase, but the audience shifts somewhat. District administration and the school board are front and center here, since they're the ones making the big resource allocation and continuation decisions. Program staff, teachers, families, and community partners will also receive summative findings. And if grant funding is part of the picture, funders will expect to see this evidence as proof of whether their investment paid off.

Decisions

Several high-stakes decisions hinge on what the summative evaluation finds. The most basic one: should the program continue, expand, or get shut down? Beyond that, decision-makers will want to know if there is a statistically meaningful impact on academics, which specific components seem to matter most, whether it makes sense to invest in scaling to more schools, and what would need to change before the next cycle.

Evaluation Questions

Five questions anchor the summative evaluation. First, did the program actually move the needle on academics? We are looking at GPA gains relative to matched counterparts, whether MAP scores tell a different story, and whether more hours in the program tracks with larger gains. That dose-response piece is particularly important.

Second, what about attendance and chronic absenteeism? Were participants coming to school more often than matched non-participants? Did chronic absenteeism drop?

Third, behavior and engagement. Fewer disciplinary referrals would be a good sign. So would students reporting that they feel more connected to school and more invested in what is happening there.

Fourth, how do the people closest to the program perceive its value? We want to hear from students about what felt most worthwhile, from parents about the effects they noticed at home, and from staff about which pieces seemed to matter most.

Fifth, was the program carried out the way it was supposed to be? Did higher-fidelity sites get better results? And which specific implementation factors seem most closely linked to success?

Research Design: Quasi-Experimental Approach

We can't randomly assign students to participate or not, and that's just not on the table for this kind of program. So the next best thing is a quasi-experimental nonequivalent comparison group design. It's the most rigorous option available when true randomization isn't possible. The What Works Clearinghouse (WWC) recognizes this type of design as capable of producing evidence that "meets standards with reservations," provided the right analytic adjustments are in place.

Propensity Score Matching (PSM)

The fundamental worry with a voluntary program like this one is selection bias: students who choose to show up may be different in important ways from those who don't. Propensity score matching (PSM) is the tool we'll use to deal with that. In short, a propensity score captures the likelihood that a given student would participate in the program, based on a set of observed baseline characteristics. By matching each participant to a non-participant with a similar propensity score, we create two groups that look much more alike on the variables we can measure, which makes any comparison between them far more credible.

The covariates going into the propensity score model will include:

- Baseline GPA (prior year)
- Baseline standardized test scores (prior year MAP scores)
- Prior-year attendance rate
- Number of prior-year disciplinary referrals
- Free/reduced lunch eligibility (as a proxy for socioeconomic status)
- Grade level
- Gender
- Race/ethnicity
- English Learner (EL) status
- Special education (IEP) status
- School building

Once the matching is done, the research team will run balance diagnostics to make sure the two groups really are comparable on all the observed covariates. The benchmark here is a standardized mean difference below 0.25 on each covariate, which is consistent with WWC guidelines. If one-to-one nearest-neighbor matching doesn't get us there, we'll explore alternatives like caliper matching, stratification on propensity scores, or inverse

probability of treatment weighting (IPTW) until we find an approach that produces adequate balance.

Supplementary Design Elements

PSM alone is good, but layering on a few additional design elements makes the causal argument considerably stronger:

- **Pretest-Posttest Design:** We'll collect baseline measures of every academic and behavioral outcome before the program starts. That way, instead of just comparing groups at the end of the year, we can compare how much each group changed, and that goes a long way toward ruling out preexisting differences as an explanation.
- **Dose-Response Analysis:** Within the treatment group, we'll look at whether students who participated more (in terms of total hours) had better outcomes than those who participated less. If more participation tracks with better results, that's one more piece of evidence pointing toward a real program effect.
- **Difference-in-Differences (DiD):** This is another analytic layer. A DiD approach compares the pre-to-post change in outcomes for the treatment group against the same change for the comparison group. It's particularly useful because it accounts for any stable, unmeasured differences between the two groups that don't change over time.
- **Sensitivity Analysis:** Rosenbaum bounds or a similar technique will help us gauge how sensitive our findings are to potential unmeasured confounders, things like motivation or family support that we can't directly observe. This won't eliminate the limitation, but it makes us transparent about how much hidden bias would have to exist to overturn the results.

Data Collection

Answering the summative questions requires pulling together a broad set of data, both numbers and narratives. The table below lays out what we plan to collect, where it comes from, when it gets collected, and which evaluation question each data source speaks to.

Data Type	Specific Measure	Source	Timing	Evaluation Question(s)
Academic Achievement	Cumulative GPA	District SIS	Pre (prior year) & Post (end of year)	1
Academic Achievement	MAP / standardized test scores	State assessment data	Pre & Post	1
Academic Achievement	Course grades (core subjects)	District SIS	Quarterly	1
Attendance	School-day attendance rate	District SIS	Pre & Post (daily)	2

Attendance	Chronic absenteeism indicator	District SIS	Pre & Post	2
Behavior	Disciplinary referrals / suspensions	District SIS	Pre & Post	3
Engagement	Student engagement survey	Students	Pre & Post	3, 4
Program Participation	Daily attendance logs (dosage)	Program staff	Daily	1, 2, 3, 5
Perceptions	Student experience interviews/focus groups	Students (sample)	End of year	4
Perceptions	Parent satisfaction survey	Parents/Guardians	End of year	4
Perceptions	Staff interviews	Program staff	End of year	4, 5
Implementation	Fidelity observation scores	Research team	Ongoing	5

Quantitative Data

The backbone of the quantitative data is the district's own Student Information System (SIS). GPA, course grades, standardized test scores, daily attendance, and disciplinary records are all already being collected as part of normal district operations, so there's no extra burden on students or staff. Prior-year values for each of these measures serve as our pretests; current-year values become the posttests.

Program staff will track participation data day by day using a standardized sign-in and sign-out system. This includes which activities each student attends and how many total hours they accumulate over the year. Getting this dosage data right is essential for the dose-response analysis described earlier.

Both program participants and comparison students will take a validated student engagement survey (something like the Student Engagement Instrument, or a version the district adapts for its own context) at the start and end of the school year. The survey measures school connectedness, cognitive engagement, and the quality of students' relationships with peers and teachers.

Qualitative Data

After the program year ends, the research team needs to hear directly from the people involved. That means semi-structured interviews and focus groups with a purposive sample of students, parents, and staff. On the student side, we are thinking four to six focus groups with six to eight kids each, sorted by grade level and school site. The idea is to create space for honest conversation about what the experience was actually like. We will also sit down individually with eight to ten parents and six to eight staff members. Every session will be audio-recorded and transcribed.

Data Analysis Strategies

Quantitative Analysis

- **Propensity Score Estimation:** A logistic regression model will generate propensity scores from the baseline covariates listed above. Matching itself will use nearest-neighbor matching (either 1:1 or 1:k), with a caliper set at 0.2 standard deviations of the logit of the propensity score to prevent poor matches.
- **Balance Checks:** Before and after matching, the team will compute standardized mean differences and variance ratios on all covariates to confirm that the groups are genuinely equivalent.
- **Outcome Analysis:** For continuous outcomes like GPA, test scores, and attendance rates, we'll use analysis of covariance (ANCOVA) or multilevel regression to compare post-program outcomes between the matched groups, with pretest scores and any lingering covariate imbalances included as controls. Effect sizes (Cohen's d) will be reported alongside significance tests so that readers can judge practical importance, not just statistical significance.
- **Difference-in-Differences:** A DiD regression model estimates the treatment effect through the interaction between group assignment and time. Covariates are included in the model as well.
- **Dose-Response Analysis:** Within the treatment group only, regression models will test whether total hours of participation predict outcomes after controlling for baseline characteristics.
- **Subgroup Analyses:** We'll run moderator analyses to see whether the program works differently for different kinds of students, broken down by grade level, baseline achievement, free/reduced lunch status, race/ethnicity, and so on.
- **Sensitivity Analysis:** Rosenbaum bounds will tell us how large an unmeasured confounder would have to be to nullify the results. That serves as a useful reality check on the strength of the findings.

Qualitative Analysis

For the qualitative data, the team will use thematic analysis following the approach laid out by Braun and Clarke (2006). Two researchers will code the transcripts independently,

then sit down together to reconcile their codes and pull out the themes that keep coming up. Those themes will be organized around the evaluation questions. Where it's feasible, member checking (going back to participants to see if the interpretations ring true) will be used to validate the findings. Open-ended survey responses get the same thematic treatment.

Integration of Quantitative and Qualitative Findings

The quantitative and qualitative strands will come together through a convergent parallel mixed-methods design (Creswell & Plano Clark, 2017). The numbers tell us how big the program's effects are (or aren't); the qualitative findings help explain the why behind those numbers, specifically what made the program work in some areas and what fell flat in others. In the final report, joint displays will put the quantitative and qualitative findings next to each other so that readers can see how the two lines of evidence converge or diverge.

Matrix of Summative Questions and Data Collection Methods

Evaluation Question	Data Source(s)	Method	Analysis
1. Academic achievement impact	GPA, MAP scores, course grades	PSM + ANCOVA / DiD	Regression, effect sizes
2. Attendance and chronic absenteeism	SIS attendance records	PSM + ANCOVA / DiD	Regression, chi-square
3. Behavior and engagement	Disciplinary records, engagement survey	PSM + ANCOVA; Pre/post survey	Regression, paired t-tests
4. Stakeholder perceptions	Interviews, focus groups, surveys	Qualitative; Descriptive stats	Thematic analysis, frequencies
5. Implementation fidelity and outcomes link	Fidelity scores, dosage, outcomes	Correlation / regression	Multilevel modeling

Participant Sample

The treatment group will be made up of students who enroll in the program and actually attend enough to constitute a meaningful dose, tentatively at least 30 days or 60 hours over the course of the school year. Setting a minimum threshold matters because including students who came twice and never returned would water down the analysis. The comparison group comes from students in the same schools and grade levels who didn't participate. Propensity score matching narrows this pool down to the non-participants who most closely resemble the participants on baseline characteristics.

On the qualitative side, purposive sampling will ensure we hear from students and parents with a range of experiences, not just the most enthusiastic families. That means including high-dosage and low-dosage participants, students from different grade levels and school sites, and so on. Every program site will have at least one staff member interviewed.

Once enrollment numbers firm up, a formal power analysis will be needed to confirm that the sample is large enough to detect an effect size of practical significance, something in the neighborhood of Cohen's $d = 0.20$, with 80% power at $\alpha = 0.05$. Based on early projections of 200 to 300 participants, a 1:1 matched design should give us enough statistical power for the main outcome analyses, but this will need to be verified once actual numbers are in hand.

Logistics

Summative data collection spans the full school year. At the front end, the team will pull baseline administrative data (prior-year GPA, attendance, test scores, disciplinary records) from the SIS right at the start of the year. The student engagement survey goes out in September as a pretest and again in May as a posttest, and both treatment and comparison students take it. Program attendance logs accumulate continuously. Interviews and focus groups happen in May, once the program year is wrapping up. The final pull of administrative outcome data comes in June, after the school year closes.

Team Responsibilities

Task	Responsible Party	Timeline
Baseline data extraction and PSM	Lead Researcher	September–October
Survey administration (pre/post)	Research Assistant	September & May
Program observation and fidelity monitoring	Research Team	Ongoing
Interviews, focus groups, and qualitative analysis	Qualitative Lead	May–June

The final summative report goes to district administration by August, early enough that the findings can actually shape decisions about the next program year. The report itself will include an executive summary, detailed findings organized around each evaluation question, a frank discussion of limitations, and concrete recommendations for moving forward.

Budget

Item	Estimated Cost
Personnel (research team, ~250 hours @ $35/hr)	$8,750

Survey platform (Qualtrics or equivalent)	$500 (or in-kind)
Statistical software (e.g., R/Stata; may be in-kind)	$500
Transcription services for interviews/focus groups	$1,200
Incentives (student focus groups, parent interviews)	$800
Printing and materials	$250
Data management and security	$500
Contingency	$500

Total Estimated Summative Evaluation Budget: $13,000

Total Combined Evaluation Budget (Formative + Summative): $18,000

Worth noting: these numbers could come down if the district is able to contribute in-kind resources, things like existing Qualtrics or survey platform licenses, statistical software the research office already owns, and personnel time that's already budgeted for program evaluation work.

Limitations and Threats to Validity

No evaluation design is perfect, and a quasi-experimental approach, even one that uses propensity score matching, comes with real limitations. Being upfront about them is important:

- **Selection Bias from Unobserved Confounders:** PSM only accounts for the variables we can observe and measure. Things like a student's innate motivation, how involved their parents are at home, or who their friends are. None of that gets captured in the matching. Sensitivity analyses (Rosenbaum bounds) will help us quantify just how big an unmeasured confounder would have to be to flip the findings, but the limitation itself doesn't go away.
- **Attrition:** Some students will inevitably drop out of the program partway through the year, and that dropout isn't random. It could introduce bias. To handle this, intent-to-treat (ITT) analysis will be the primary approach, meaning all enrolled students stay in the analysis regardless of how much they actually attended. A supplementary treatment-on-the-treated (TOT) analysis will focus on students who met the dosage threshold.
- **Generalizability:** These findings will be specific to this district, its students, and its particular context. They shouldn't be assumed to transfer to other districts or settings without replication.
- **Contamination:** Students in the comparison group aren't sitting at home doing nothing. Some of them will be in other after-school activities, tutoring, or enrichment programs. That outside participation could shrink the apparent difference between the two groups. Where feasible, we'll collect data on what comparison students are doing after school to at least document the extent of this issue.
- **Implementation Variability:** The program won't look exactly the same at every school site. Differences in staffing, space, leadership, and local context will produce variation in implementation quality. Multilevel models that account for site-level clustering will help us deal with this analytically, but the variation itself is something to keep in mind when interpreting results.

Appendix A: Recommended Data Collection Instruments

Student Engagement Survey (Pre/Post)

Both groups take this survey at the start and end of the school year. It uses a 5-point Likert scale. We deliberately kept it short so kids do not zone out halfway through. A few of the items get at belonging and safety: do students feel like they fit in at school, do they feel safe there, is there at least one adult they could go to if something went sideways. Other items are more academic. How much effort are they putting in? Do they actually pay attention? Are they getting homework done? Then there are a couple about motivation and whether school feels like it matters for their future.

Parent/Guardian Satisfaction Survey

This survey goes home to parents of students in the program at the end of the school year. It is mostly closed-ended ratings, but we threw in a couple open-ended questions too, because sometimes the most useful feedback is whatever parents decide to write in their own words.

For parents, the survey covers the basics. Satisfaction with the program overall. Whether their kid actually wants to go. Academic progress. Safety. Would they tell another family about it. Then there are two write-in questions where parents can say in their own words what the program did well and what it should fix. The very last item just asks a yes, somewhat, or no about whether having their child in the program has made their own schedules any easier to manage.

Student Focus Group Protocol

Use this guide for end-of-year student focus groups. Keep each session between 45 and 60 minutes and cap groups at six to eight students. The whole point is for it to feel relaxed. Not like an interview.

Start by asking students to describe what a typical program day looked like. Then get into what they enjoyed most and, just as importantly, what they could have done without.

Shift the conversation toward academics. Did anyone feel like the program helped them with schoolwork? Push for specifics there, not just yes or no.

Ask about safety and how comfortable they felt in the space. Then have them talk about the staff. What kind of people were they? Were they easy to talk to?

Toward the end, ask whether the program changed how they think about school at all. Give each student the chance to name one thing they would change. Close by asking if they would recommend it to a friend and why, and leave room for anything else they want to bring up.

Program Staff Interview Protocol

These are one-on-one sit-downs with program staff. The format is semi-structured, which just means the interviewer has a set of questions but can follow up on whatever comes up naturally. Budget about 30 to 45 minutes per interview.

1. Give me the overview. What is your role, and what does a normal week look like for you?

2. In your opinion, what is the program doing really well right now?

3. Where has it been a struggle? What problems keep coming up?

4. Talk to me about student engagement. Has it stayed the same all year or has it shifted?

5. Have you seen students change at all academically? Things like turning in homework more, paying attention better, anything along those lines?

6. How well does the after-school programming line up with the regular school day, from what you have seen?

7. Is there training or professional development that would make your job easier?

8. If you could redesign part of the program for next year, where would you start?

Program Fidelity Observation Checklist

The research team fills this out during on-site visits. Each line gets a score from one to four, where one means the observer did not see it happen and four means it was fully in place. Some items are straightforward: did the session start on time, were there enough staff, did the activities match the curriculum. Others require more judgment. Are students actually engaged or just sitting there? Is the space organized? When a kid goes off-task, how do the adults handle it? The checklist also asks about enrichment programming and whether basic materials and supplies were available.

Appendix B: Summary of Recommended Data to Collect

Below is a consolidated list of every type of data this evaluation recommends collecting, pulled together in one place for easy reference:

Administrative / Existing Data (from District SIS)

- Student demographics (grade, gender, race/ethnicity, FRL status, EL status, IEP status)
- Prior-year and current-year GPA
- Prior-year and current-year standardized test scores (e.g., MAP)
- Prior-year and current-year daily attendance records
- Prior-year and current-year disciplinary referrals and suspensions
- Course grades by subject (quarterly)

Program-Generated Data

- Daily program attendance (sign-in/sign-out logs)
- Activities attended by each student
- Total hours of participation per student (dosage)
- Program fidelity observation scores
- Staff reflection logs

Survey Data

- Student engagement survey (pre and post, treatment and comparison groups)
- Parent/guardian satisfaction survey (post, treatment group)
- Student monthly engagement check-ins (formative)
- Staff surveys (formative)

Qualitative Data

- Student focus group transcripts
- Parent interview transcripts
- Program staff interview transcripts
- Open-ended survey responses
- Structured observation field notes

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

Creswell, J. W., & Plano Clark, V. L. (2017). Designing and conducting mixed methods research (3rd ed.). SAGE.

Durlak, J. A., Weissberg, R. P., & Pachan, M. (2010). A meta-analysis of after-school programs that seek to promote personal and social skills in children and adolescents. American Journal of Community Psychology, 45(3–4), 294–309.

Joint Committee on Standards for Educational Evaluation. (2011). The program evaluation standards (3rd ed.). SAGE.

Lauer, P. A., Akiba, M., Wilkerson, S. B., Apthorp, H. S., Snow, D., & Martin-Glenn, M. L. (2006). Out-of-school-time programs: A meta-analysis of effects for at-risk students. Review of Educational Research, 76(2), 275–313.

Rosenbaum, P. R. (2002). Observational studies (2nd ed.). Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2020). What Works Clearinghouse standards handbook (Version 4.1).