

Evaluation Plan for a District After-School Program

February 10, 2026

Everett Stuckey

Introduction

This evaluation plan lays out a structured approach for assessing the district's new after-school program, covering both implementation and effectiveness. Two major phases organize the work. The formative evaluation runs during early implementation; its job is to monitor fidelity and catch problems before they become entrenched. The summative evaluation comes after a full program year and tackles the harder question: did the program actually produce measurable changes in student outcomes? Both phases address the same core topics, including purposes, stakeholders, evaluation questions, design and methods, data collection and analysis, logistics, and budget.

Random assignment is not viable here. Ethical constraints rule out withholding services from eligible students, and the voluntary nature of the program makes randomization impractical. The plan therefore relies on a quasi-experimental framework. Specifically, the summative component will use a nonequivalent comparison group design paired with propensity score matching (PSM), which approximates the rigor of a randomized controlled trial without requiring anyone to be turned away. This allows for stronger causal inference while respecting the fact that students choose whether or not to participate.

A mixed-methods approach undergirds the entire evaluation. Quantitative measures will capture academic performance, attendance patterns, and behavioral indicators. On the qualitative side, interviews, focus groups, and open-ended survey items will document the lived experiences of students, families, and staff. Using both kinds of evidence together, and triangulating across sources, strengthens the validity of the conclusions and provides a more complete picture of program quality and impact than either approach could offer on its own.

Background

The district is preparing to launch a new after-school program designed to provide academic support, enrichment, and a structured environment for students outside of regular school hours. The ambitions go beyond childcare. District leadership wants to see improved achievement, stronger engagement, fewer risky behaviors, and dependable after-school options for working families. Day-to-day programming will combine homework help and tutoring with STEM enrichment, arts, physical fitness, and social-emotional learning (SEL).

The program will span multiple grade levels and school buildings. Any student can enroll, but the district plans targeted outreach to those flagged as at-risk based on grades, attendance, or behavior. That voluntary structure is what creates the evaluation's central methodological challenge: students who choose to participate may be systematically

different from those who do not. Accounting for that selection bias is the single most important design problem 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 called dosage) matters a great deal. So do instructional quality, how well programming lines up with the school-day curriculum, and the strength of relationships between staff and students (Durlak et al., 2010; Lauer et al., 2006). This evaluation will examine these implementation factors alongside student outcome data to build a comprehensive picture of both program quality and impact.

Evaluation Standards

This evaluation will be guided by the following standards and frameworks:

Program Quality Standards

- Joint Committee on Standards for Educational Evaluation: Program Evaluation Standards (Utility, Feasibility, Propriety, Accuracy, Accountability)
- Collaborative for Academic, Social, and Emotional Learning (CASEL) framework for SEL programming
- Missouri Department of Elementary and Secondary Education (DESE) standards for extended learning programs
- District curriculum standards and student learning outcomes

Research Design Standards

- What Works Clearinghouse (WWC) Standards for quasi-experimental designs
- American Evaluation Association Guiding Principles for Evaluators
- Institute of Education Sciences (IES) guidelines for evidence levels

Formative Evaluation Plan

Purposes

The formative evaluation is designed to monitor and improve the after-school program during its initial rollout. The focus will be on program fidelity, quality of implementation, participant engagement, and early indicators of progress. Rather than waiting until the end of the year, formative findings will feed directly into real-time adjustments to programming, staffing, scheduling, and content delivery. The primary deliverable is a mid-year report with actionable recommendations.

Stakeholders

The primary client of the evaluation is the district administration that commissioned this plan. Primary stakeholders include the following groups:

- **District Administration:** Decision-makers responsible for program funding, continuation, and expansion.
- **Program Coordinators and Staff:** The people on the ground every day, running sessions and dealing with whatever comes up. Nobody knows what is working and what is not better than they do.
- **Building Principals:** School leaders who oversee logistics, space, and coordination between the school day and after-school hours.

Classroom teachers can spot changes that nobody else would. If a student starts turning in homework more often or seems more focused, teachers notice first. Students are the direct beneficiaries of the program, and their engagement and growth are the whole reason for doing this evaluation. Parents and guardians round out the primary group; their buy-in can make or break attendance and retention. Beyond these, 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

The formative evaluation will inform a range of practical decisions: whether scheduling or staffing ratios need adjusting, which students need targeted support for low attendance or engagement, what professional development program staff might need, how to communicate with families about enrollment and retention, and where after-school content does not line up with the school-day curriculum.

Evaluation Questions

Four questions drive the formative evaluation. First, is the program actually being run the way it was designed? That means checking whether planned activities are happening consistently at every site, whether staffing levels and qualifications meet program standards, and whether the schedule has held steady.

Second, are students showing up and staying engaged? We need to look at daily and weekly attendance, average dosage in hours per week, and whether students seem genuinely checked in during activities or just physically present.

Third, what are the people involved actually saying? Students, parents, and staff each have a different vantage point on what is working and what is not. Their feedback matters.

Fourth, are there any early signs that the program is headed in the right direction? Homework completion going up would be one indicator. So would shifts in school-day attendance or early movement on Star Reading and Star Math benchmark scores.

Participant Sample

Formative data will be collected from all students enrolled in the after-school program. Additionally, a convenience sample of program staff (all staff members) and a stratified random sample of parents/guardians will be surveyed. Program observations will be conducted at each participating school site. The research team will aim for a parent survey response rate of at least 30% to ensure adequate representation.

Methods

Data collection for the formative phase draws 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 give program staff a place to record weekly notes about what is going well, what is not, and any adjustments they have made.

Logistics

Formative work starts 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 that no one building feels singled out. Each month, the research team will meet with program coordinators to go over preliminary findings and talk through what might need to change.

By semester's end, 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
Incentives for parent survey completion	\$300
Contingency	\$250

Total Estimated Formative Evaluation Budget: \$4,050

Summative Evaluation Plan

Purposes

The summative evaluation takes on a different question: did the program actually work? Using a matched comparison group design with propensity score matching, this phase will assess whether the program achieved its intended goals, specifically improving academic achievement as measured by Star Reading and Star Math, increasing school-day attendance, reducing chronic absenteeism, and strengthening student engagement and behavior. Comparing outcomes between participants (treatment group) and a matched control group of similar non-participants will allow the evaluation to estimate the program's causal impact.

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 are the ones making 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 the summative findings. The most basic one: should the program continue, expand, or get shut down? Beyond that, decision-makers will want to know whether there is a statistically meaningful impact on academics and attendance, 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 improve academic achievement? The evaluation will compare GPA gains, Star Reading and Star Math growth (scaled scores and student growth percentiles), and MAP scores between participants and matched non-participants. A dose-response analysis will also test whether more hours in the program tracked with larger academic gains.

Second, did the program improve school-day attendance and reduce chronic absenteeism? Attendance rates from the SIS will be compared between the two groups, along with the proportion of students missing 10% or more of school days.

Third, did behavior and engagement improve? Fewer disciplinary referrals and higher self-reported school connectedness among participants would both count as evidence here.

Fourth, how do the people closest to the program perceive its value? Students, parents, and staff will be asked what worked, what fell short, and what mattered most.

Fifth, was the program implemented with fidelity? Sites with higher fidelity scores should, in theory, produce better outcomes, and the evaluation will test whether that relationship holds.

Research Design: Quasi-Experimental Approach

Because random assignment of students to treatment (program participation) and control (non-participation) conditions is not feasible, this evaluation will employ a quasi-experimental nonequivalent comparison group design with propensity score matching (PSM) to construct a matched control group. The primary outcome measures for the treatment-control comparison will be (a) Star Reading scaled scores and student growth percentiles (SGPs), (b) Star Math scaled scores and SGPs, and (c) school-day attendance rates and chronic absenteeism indicators.

By matching each program participant to one or more non-participating students with similar baseline characteristics, the design creates a credible counterfactual: what would have happened to program participants had they not enrolled in the after-school program. This is the most rigorous design available when randomization is not possible and meets the What Works Clearinghouse (WWC) standards for quasi-experimental designs with reservations, provided that baseline equivalence between groups is established.

Propensity Score Matching (PSM)

The fundamental concern with a voluntary program is selection bias: students who choose to show up may be different in important ways from those who do not. Propensity score matching (PSM) is the tool this evaluation will use to address that. In short, a propensity score captures the likelihood that a given student would participate in the program, based on observed baseline characteristics. By matching each participant to a non-participant with a similar propensity score, the design creates two groups that look much more alike on the variables we can measure, which makes any comparison between them far more credible. The matched groups will then be compared on three primary outcome domains: Star Reading performance, Star Math performance, and school-day attendance.

Covariates used to estimate propensity scores 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

After matching, balance diagnostics will be conducted to verify that the treatment and comparison groups are comparable on all observed covariates. Standardized mean differences of less than 0.25 on each covariate will be the threshold for acceptable balance, consistent with WWC guidelines. If balance is not achieved through one-to-one nearest-neighbor matching, alternative approaches such as caliper matching, stratification on propensity scores, or inverse probability of treatment weighting (IPTW) will be considered.

Supplementary Design Elements

To further strengthen causal inference, the evaluation will incorporate the following supplementary design elements:

- **Pretest-Posttest Design:** Baseline (pretest) measures of all academic and behavioral outcomes will be collected before the program begins. Comparing pre-to-post change between treatment and matched comparison groups strengthens the ability to attribute observed differences to the program rather than preexisting differences.
- **Dose-Response Analysis:** Within the treatment group, the relationship between the amount of program participation (total hours attended) and outcomes will be examined. A dose-response relationship, where more participation is associated with better outcomes, provides additional evidence supporting a causal interpretation.
- **Difference-in-Differences (DiD):** As an additional analytic strategy, a difference-in-differences approach will compare the change in outcomes from pre to post between treatment and comparison groups. This approach controls for any time-invariant unobserved differences between groups.
- **Sensitivity Analysis:** Rosenbaum bounds or similar sensitivity analyses will be conducted to assess how robust the findings are to potential unmeasured confounders. This provides transparency about the limits of causal inference in the absence of random assignment.

Data Collection

A comprehensive set of quantitative and qualitative data will be collected to answer the summative evaluation questions. The following table summarizes the recommended data types, sources, timing, and associated evaluation questions.

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 Student Information System (SIS). GPA, course grades, test scores, daily attendance, and disciplinary records are all collected as part of normal district operations, so pulling them adds no extra burden on students or staff. Prior-year values serve as pretests; current-year values become the posttests.

Program staff will track participation using a standardized sign-in/sign-out system, recording which activities each student attends and how many total hours they accumulate. Getting this dosage data right is essential for the dose-response analysis.

Both participants and comparison students will complete a validated engagement survey (such as the Student Engagement Instrument or a district-adapted version) at the start and end of the year, measuring school connectedness, cognitive engagement, and peer/teacher relationships.

Qualitative Data

Once the program year wraps up, the research team will conduct semi-structured interviews and focus groups with a purposive sample of students, parents, and staff. On the student side, four to six focus groups of six to eight students each, stratified by grade level and school site, will explore what the experience was actually like. Individual interviews with eight to ten parents and six to eight staff members will add depth. Every session will be recorded and transcribed.

Data Analysis Strategies

Quantitative Analysis

- **Propensity Score Estimation:** A logistic regression model will estimate propensity scores using the baseline covariates listed above (including prior-year Star Reading and Star Math scores). Matching will be performed using nearest-neighbor matching (1:1 or 1:k) with a caliper of 0.2 standard deviations of the logit of the propensity score.
- **Balance Checks:** Standardized mean differences and variance ratios will be computed for all covariates before and after matching to verify group equivalence.
- **Outcome Analysis:** For continuous outcomes (GPA, Star Reading and Star Math scaled scores and SGPs, MAP scores, attendance rates), analysis of covariance (ANCOVA) or multilevel regression models will compare post-program outcomes between matched treatment and comparison groups, controlling for pretest scores and any remaining covariate imbalances. Effect sizes (Cohen's d) will be reported.
- **Difference-in-Differences:** A DiD regression model will estimate the treatment effect as the interaction between group assignment (treatment vs. comparison) and time (pre vs. post), with covariates included.
- **Dose-Response Analysis:** Within the treatment group, regression models will examine the relationship between total hours of participation and outcomes, controlling for baseline characteristics.
- **Subgroup Analyses:** Moderator analyses will examine whether program effects differ by student subgroups (e.g., grade level, baseline achievement level, free/reduced lunch status, race/ethnicity).

- **Sensitivity Analysis:** Rosenbaum bounds will assess sensitivity to hidden bias from unmeasured confounders.

Qualitative Analysis

Interview and focus group transcripts will be analyzed using thematic analysis (Braun & Clarke, 2006). Two members of the research team will independently code transcripts, then meet to reconcile codes and identify emergent themes. Themes will be organized around the evaluation questions. Member checking will be used where feasible to validate interpretations. Open-ended survey responses will be analyzed using the same thematic approach.

Integration of Quantitative and Qualitative Findings

Findings from quantitative and qualitative analyses will be integrated using a convergent parallel mixed-methods design (Creswell & Plano Clark, 2017). Quantitative results will establish the magnitude of program effects, while qualitative findings will provide explanatory context, illuminating how and why the program did or did not produce its intended effects. Joint displays will be used in the final report to present aligned quantitative and qualitative findings side by side.

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 consists of students who enroll and attend enough to constitute a meaningful dose, tentatively at least 30 days or 60 hours over the school year. Setting a minimum threshold matters because students who came twice and never returned would dilute the analysis. The comparison group comes from students in the same schools and grade levels who did not participate; propensity score matching narrows this pool to those who most closely resemble participants on baseline characteristics.

For the qualitative strand, purposive sampling will select students and parents representing a range of experiences: high-dosage and low-dosage participants, different grade levels, different school sites. Every program site will have at least one staff member interviewed.

A power analysis should be conducted once enrollment numbers are known to determine whether the sample size is sufficient to detect a minimum effect size of practical significance (e.g., Cohen's $d = 0.20$) with 80% power at $\alpha = 0.05$. Based on anticipated enrollment of 200-300 participants, a 1:1 matched design should provide adequate power for the primary outcome analyses.

Logistics

Summative data collection will span the full program year. Baseline administrative data (prior-year GPA, attendance, test scores, disciplinary records) will be extracted from the district SIS at the start of the school year. The student engagement survey will be administered in September (pretest) and May (posttest) to both treatment and comparison groups. Program attendance logs will be collected continuously throughout the year. End-of-year interviews and focus groups will be conducted in May. Administrative outcome data will be extracted after the close of the school year in June.

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

A final summative evaluation report will be delivered to district administration by August, allowing findings to inform decisions about the next program year. The report will include an executive summary, detailed findings organized by evaluation question, limitations, and recommendations.

Budget

Item	Estimated Cost
Personnel (research team, ~250 hours @ \$35/hr)	\$8,750
Incentives (student focus groups, parent interviews)	\$800
Contingency	\$500

Total Estimated Summative Evaluation Budget: \$10,050

Total Combined Evaluation Budget (Formative + Summative): \$14,050

Note: Costs may be reduced if the district can provide in-kind contributions such as existing survey software licenses, statistical software, and personnel time allocated from existing research office budgets.

Limitations and Threats to Validity

While the quasi-experimental design with propensity score matching represents the most rigorous approach available without random assignment, several limitations must be acknowledged:

- **Selection Bias from Unobserved Confounders:** PSM can only account for observed covariates. Unmeasured factors, such as student motivation, parental involvement, or peer influence, may still differ between groups. Sensitivity analyses (Rosenbaum bounds) will quantify how large such hidden bias would need to be to overturn findings.
- **Attrition:** Students may drop out of the program during the year, potentially introducing attrition bias. Intent-to-treat (ITT) analysis will be used as the primary analytic approach, supplemented by treatment-on-the-treated (TOT) analysis for students meeting the dosage threshold.
- **Generalizability:** Findings are specific to this district's context and may not generalize to other settings without replication.
- **Contamination:** Comparison group students may participate in other after-school activities or tutoring, which could dilute the estimated program effect. Data on comparison students' out-of-school activities will be collected where feasible.
- **Implementation Variability:** Differences in implementation quality across school sites may affect outcomes. Multilevel models that account for site-level clustering will be used to address this.

Appendix A: Recommended Data Collection Instruments

Student Engagement Survey (Pre/Post)

Both groups take this at the start and end of the school year. It uses a standard 5-point Likert scale. We kept it short on purpose so that students do not zone out halfway through. Some items ask about belonging and safety: does the student feel like they fit in, do they feel safe, is there at least one trusted adult in the building. Others focus on academics: effort, attention, homework habits. A couple get at motivation and whether school feels like it matters for the future.

Parent/Guardian Satisfaction Survey

Parents of program participants receive this at the end of the year. It is mostly closed-ended, but there are two open-ended write-in questions because sometimes the most telling feedback comes from what parents choose to say unprompted.

The closed-ended side covers satisfaction with the program overall, whether the child actually wants to go, academic benefit, safety, and likelihood of recommending it to another family. The write-in items ask what the program did best and what should change. One final item, formatted as yes/somewhat/no, asks whether having a child in the program has made the parent's own schedule 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 a yes or no.

Ask about safety and how comfortable they felt in the space. Then have them talk about the staff. Were the adults easy to talk to? Toward the end, ask whether the program changed how students think about school at all. Give each person the chance to name one thing they would change. Close by asking if they would recommend the program 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, semi-structured so that the interviewer has a set of topics but can follow up on whatever comes up naturally. Budget about 30 to 45 minutes per interview.

Start by asking staff to describe their role and a typical week. Then ask what the program is doing well and where it has been a struggle. Get them talking about student engagement: has it stayed steady all year or shifted? Have they noticed any academic changes, things like homework completion or attitude toward school?

Ask how well the after-school programming lines up with the regular school day. Find out whether there is training or professional development that would help them do their jobs better. Close by asking what they would change about the program if they could redesign part of it for next year.

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 student 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

The following is a consolidated summary of all data types recommended for collection in this evaluation:

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).