Raz-Plus ESSA Tier 2 Efficacy Study

Classroom Usage of Raz-Plus in Third Grade Associated With Accelerated Reading Growth

Technical Research Report

Megan Conrad, Ph.D. and Parnika Bhatia, Ph.D. | ExploreK12 | July 2026

Table of Contents

Executive Summary

The Challenge of Reading Proficiency by the End of Third Grade: Research indicates that reading proficiency by the end of third grade is a predictor of long-term academic success. However, national proficiency rates suggest that students need significantly more support to meet these goals: across the US, only 31% of fourth graders fully meet reading proficiency standards. Elementary educators often feel unequipped to address the wide range of student support needs within their classrooms. While educators recognize the need for Science of Reading (SoR)- aligned instruction, fewer than 25% are highly satisfied with the literacy curricula available to them, highlighting a gap between evidence-based practice and classroom resources.

The Solution and Study Design: Raz-Plus® from ExploreK12TM is a comprehensive K-5 supplemental literacy solution that bridges this resource gap through foundational skills instruction, decodable texts, high-quality teacher instructional materials, and engaging student lessons for differentiated and independent practice. To evaluate its real-world impact, a quasi-experimental design (QED) study was conducted to meet Every Student Succeeds Act (ESSA) Tier 2 Moderate evidence standards. The study compared the reading proficiency scores and growth of 82 third-grade classrooms with regular, teacher-directed implementation of Raz-Plus to 82 matched comparison classrooms with no Raz-Plus usage from August to December 2025, for a total sample of 3,229 diverse students across 15 different US states.

Key Impact Findings: After just 16 weeks of implementation, Raz-Plus delivered statistically significant improvements in reading proficiency.

- **Accelerated Growth:** Classrooms using Raz-Plus achieved an average growth of **7.41 RIT points**, compared to 6.05 points for classrooms that did not use Raz-Plus.
- **Weeks of Extra Learning:** The **1.36** RIT growth advantage translates to roughly **4.4 extra weeks of reading instruction** within a single semester.
- **Exceeding National Benchmarks:** The Classrooms using Raz-Plus achieved **1.27 times the anticipated national norm growth**.
- **Closing the Gap in High-Poverty Schools:** The largest gains were seen in under-resourced schools (where >75% of students qualify for Free and Reduced-Price Lunch). In these environments, Raz-Plus classrooms achieved +**7.7 points of growth**, significantly outperforming the +5.4 points of the comparison group and vastly exceeding the national average.
- **Effective Upward Differentiation:** Students performing in the Tier 1 (25th–90th percentile) and Gifted (>90th percentile) ranges demonstrated significantly higher growth when using Raz-Plus, proving its ability to stretch on-grade-level and advanced readers.

Conclusion for District Leaders: A strong core curriculum alone is rarely sufficient to address the wide range of proficiency levels within a single elementary classroom. This study demonstrates that regular, teacher-directed implementation of Raz-Plus provides the differentiated, SoR-aligned supplementation needed to accelerate reading growth. Notably, its outsized impact in high-poverty schools indicates that Raz-Plus is an effective, scalable lever for driving educational equity and ensuring more students safely cross the third-grade reading threshold.

Introduction

Developing well-rounded readers who have mastered foundational skills and are set to continue growing in language and comprehension remains a persistent challenge for elementary educators. Over the past 30 years, fourth-grade reading proficiency in the United States has remained low and largely unchanged; in 2024, only 31% of fourth-grade students met proficiency standards, only a marginal increase from 29% in 1992 (NCES, 2024). The challenge is even more acute for teachers of non-white and economically disadvantaged students, as well as children with diverse learning needs, who are two to three times more likely to fall below grade-level standards (NCES, 2017).

Catching students early is critical, as reading proficiency by the end of third grade is a well-documented predictor of long-term academic success, including eighth-grade reading proficiency, high school graduation, and college attendance (Goldhaber et al., 2020; Hernandez, 2011; Lesnick et al., 2010). This is largely due to the instructional shift from "learning to read" to "reading to learn" (Chall, 1983) that occurs between third and fourth grade. Students who enter fourth grade lacking foundational reading skills struggle to comprehend text, directly impeding their ability to master grade-level content across all subject areas (Lesnick et al, 2010; Annie E. Casey Foundation, 2010).

To mitigate these long-term academic risks, teachers must be equipped with High-Quality Instructional Materials (HQIM) — research-based resources that provide rigorous, grade-level content aligned with national and state standards — that are rooted in the Science of Reading (SoR). Instructional programs aligned to the SoR utilize a structured literacy approach, focusing on the explicit, systematic, and sequential teaching of foundational reading skills and an ongoing commitment to increasing language proficiency, vocabulary, and reading comprehension. However, a standardized core curriculum alone is rarely sufficient for teachers to effectively address and support the wide variance of proficiency levels within a single classroom. Teachers require flexible, SoR-aligned supplemental resources to deliver targeted differentiation before the critical shift to "reading to learn."

However, a stark resource gap persists between understanding evidence-based best practices and available classroom resources. Despite state-level SoR mandates, a recent survey found **that fewer than 25% of early elementary teachers are highly satisfied with the literacy curricula available to them** (The Harris Poll, 2025). Similarly, while 90% of New York educators endorse SoR practices, only half report having access to a fully vetted, SoR-aligned curriculum (Science of Reading Center, 2025).

Raz-Plus addresses this gap as a comprehensive K-5 supplemental literacy solution that aligns with SoR principles. Raz-Plus resources integrate into existing frameworks to both *complement* broad core instruction and *supplement* the curriculum by filling specific foundational gaps. It equips educators with:

- **Decodable texts and targeted vocabulary tools** to reinforce decoding skills and meaning-making.
- **Engaging books and explicit lessons** tied to research-backed strategies that deepen comprehension and fluency.
- **A game-based student portal** that drives independent practice and student engagement.
- **Built-in professional development** and ongoing webinars to provide the structural support teachers need to execute differentiated instruction effectively.

The current study provides a rigorous evaluation of the real-world impact of Raz-Plus on student reading achievement. Guided by Every Student Succeeds Act (ESSA) Tier 2 (Moderate) evidence standards, this quasi-experimental design (QED) study compared reading growth over a 16-week period (August to December 2025). Using propensity score matching to ensure baseline equivalence, the study analyzed NWEA MAP Reading Growth (RIT) scores for 82 third-grade classrooms with Raz-Plus usage, and compared them with a matched comparison group of 82 third-grade classrooms drawn from across 15 states.

Specifically, the study explores two primary research questions:

1. Does classroom use of Raz-Plus improve third-grade students' reading proficiency compared to matched classrooms without use over 16 weeks?
2. Do the impacts vary across different student subgroups, including academically at-risk students and students in under-resourced schools?

Study Design and Methods

Study Design

This study examined retrospective reading assessment data using a quasi-experimental design (QED) with a matched-comparison group to examine the relationship between Raz-Plus use and student reading growth. Treatment and comparison classrooms were selected based on the availability of fall 2025 and winter 2025 NWEA MAP Growth Reading scores and classroom-level implementation data. Classroom-level performance on the beginning-of-the-year reading assessment was used to create baseline-matched groups (see Appendix A for full details of the analytic approach used to create matched groups). After matching, the study design meets *WWC standards, with reservations* for baseline equivalency.

Raz-Plus: Usage and Implementation

The Raz-Plus solution provides a range of differentiated instructional resources, including decodable texts, vocabulary development tools, and systematically designed lessons. For third-grade students, the solution supports both guided and independent learning through an interactive student portal. Typical usage includes accessing a digital library of grade-appropriate texts, completing associated comprehension quizzes, and engaging in self-paced practice activities that reinforce foundational literacy skills and build independent reading volume.

Treatment classrooms were selected based on evidence of high and consistent implementation of Raz-Plus. To ensure a focus on active, teacher-driven engagement rather than optional or inconsistent access, we restricted the treatment group to classrooms where 90% or more of the students demonstrated frequent usage throughout the study period. The comparison group included classrooms where usage was minimal across the majority (90%+) of the students. Implementation within the treatment group was teacher-directed and guided by local instructional needs rather than a researcher-defined protocol.

NWEA MAP Growth: Assessing Reading Skills and Comprehension Growth

The outcome measure of student reading achievement and growth in the current study is the NWEA MAP® Growth™ Reading assessment. MAP Growth assessments are adaptive, interim assessments used to track student achievement and growth before end-of-year summative tests, to inform instructional planning, and to evaluate the efficacy of interventions. The reading assessment for independent readers (grades 2-12) is designed to measure achievement in reading comprehension and vocabulary skills. Tests are delivered online and typically administered at three time points in the school year: Fall (around week 4), Winter (around week 20), and Spring (around week 32).

Performance is presented as **RIT scores** (Rasch Unit), an equal-interval scale independent of grade level, and percentile ranking based on comparison to nationwide performance for that grade and subject area. Growth is calculated as RIT score difference between tests. Since exact test dates are set by individual schools and districts, and significant variation was observed, we calculated **Fall-to-Winter growth** using a standardized 16-week growth score for each student:

$$\text{Standardized Growth} = \left(\frac{\text{Winter RIT - Fall RIT}}{\text{Instructional Weeks}}\right) \times 16 \text{ weeks}$$

Study Sample

After matching, the sample comprised 164 third-grade classrooms across 15 U.S. states. Demographic information was not available at the student level, so school-level data from the National Center for Education Statistics (NCES) was used to estimate the sample student characteristics. When school-level data were available, we derived student-level demographic estimates by applying the school's demographic percentages to the number of study participants within that school. This proportional assignment method enabled aggregation of estimated student-level characteristics across the study sample.

Table 1.
Estimated Demographic Characteristics of Treatment and Comparison Groups

	Treatment Students		Comparison Students	
Ethnicity	**Number of Students**	**%**	**Number of Students**	**%**
Asian	227	16%	93	7%
Black	148	10%	176	13%
Hispanic	331	23%	344	26%
White	654	45%	609	47%
Multi-racial/Other Background	101	7%	81	6%
Low Socioeconomic Status	543	37%	652	50%
Students Missing School-Level Data	106	7%	360	22%

Within the treatment group, classrooms averaged four logins per week to the Raz-Plus platform between fall and winter. Each week, classrooms typically completed an average of seven activities, including independent reading, comprehension quizzes, and foundational skills activities. Usage was confirmed to be low (fewer than 2 total logins per student on average across all 16 weeks) for classrooms in the comparison group. Across groups, the classroom's average fall percentile was 56, confirming that the groups represented the average norm-reference for their national peer group. Average baseline (fall) RIT scores between treatment and comparison groups differed by .04 points and satisfied WWC standards for baseline equivalence (< .05).

Table 2.
Raz-Plus Usage and Baseline Performance by Treatment Group

	Treatment Classrooms (n = 82)	Comparison Classrooms (n = 82)
	M (SD)	M (SD)
Days Usage of Raz-Plus	64.1 (58.7)	1.7 (6.8)
# Activities Completed	115.2 (105.4)	0.7 (3.1)
Baseline (Fall) RIT Score	188 (7.16)	187 (7.04)
Baseline Percentile	56.1 (12.3)	55.5 (11.9)

Analytic Approach

Statistical tests were used to determine whether students in classrooms implementing Raz-Plus demonstrated significantly greater reading growth than matched students in comparison classrooms.

For primary analysis, we conducted a Hierarchical Linear Model (HLM) to account for the nested structure of the data, where classrooms are grouped within schools. The analytic sample for the HLM consisted of 164 matched third-grade classrooms, split equally between the treatment and comparison groups, nested within 97 schools. Classroom Fall-to-Winter RIT growth was the outcome of interest. Because the testing window differed across usage groups, the treatment effect was standardized post hoc to a 16-week window. This corresponds to NWEA's default fall-to-winter testing interval and serves as the basis for national growth norms. A subsequent HLM was conducted that controlled for specific school-level demographics. Data for school-level demographics were sourced from the National Center for Education Statistics (NCES). Effect sizes were calculated using Hedges' g.

To determine if the efficacy of Raz-Plus implementation varied across student populations, we conducted exploratory subgroup analyses based on initial academic risk and school-level socioeconomic status. Academic risk was categorized by baseline fall NWEA MAP performance: lower-performing students (25th percentile and below, typically qualifying for Tier 2/3 intervention), Tier 1 students (25th to 90th percentile), and higher-performing students (90th percentile and above).

Additionally, analyses were segmented by school-level socioeconomic status, using the National Center for Education Statistics (NCES) definition of high-poverty schools as those with more than 75% of students eligible for Free and Reduced-Price Lunch (FRPL). Students at high-poverty schools are statistically more likely to struggle with reading comprehension and slower growth due to complex, systemic stressors — such as housing instability and community trauma — while simultaneously attending underfunded schools that may struggle to retain certified, experienced educators. By examining these subgroups, we aimed to determine if Raz-Plus provides consistent, meaningful support across these diverse instructional contexts.

Results

Raz-Plus Impact on Growth (Primary Analysis)

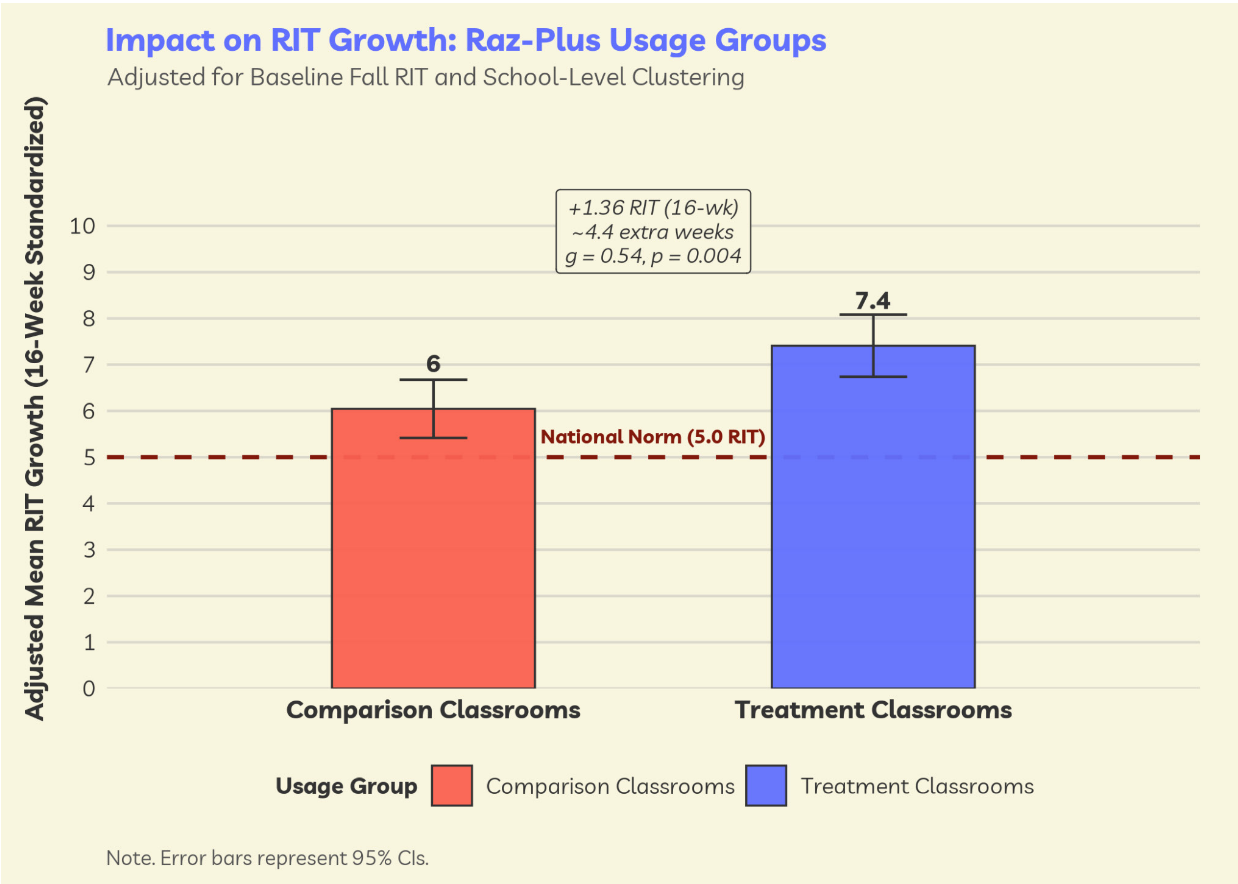
After accounting for school-level clustering and controlling for baseline fall RIT scores, Raz-Plus usage had a statistically significant impact on reading growth. Classrooms in the treatment group achieved an adjusted mean of 7.41 RIT points of growth over a standardized 16-week window, compared to 6.05 RIT points in the comparison group — a difference of 1.36 RIT points in favor of the treatment group (beta = 1.36, SE = 0.46, t(97.01) = 2.94, p = 0.004, Hedges' g = 0.54). These results demonstrate that consistent engagement with Raz-Plus solution reliably accelerates reading growth, even when accounting for the powerful role that school-level environments play in student outcomes.

The +1.36 RIT growth advantage seen in the treatment group over the matched comparison group represents an instructional gain comparable to 4.4 extra weeks of reading instruction, based on a standardized 16-week testing window. Furthermore, given that the national NWEA-MAP norm for third-grade progress during this Fall-to-Winter interval is 5.0 RIT points, treatment classrooms outpaced typical national growth expectations by 27%.

Table 3.
Impact Results by Treatment Group

Outcome	Treatment Group		Comparison Group		Pooled SD	Difference	SE	p	g
	M	*SD*	*M*	*SD*					
RIT Score (16 wk)	7.41	2.24	6.05	2.79	2.53	1.36	0.46	.004	0.54

Figure 1:
Impact Results by Treatment Group

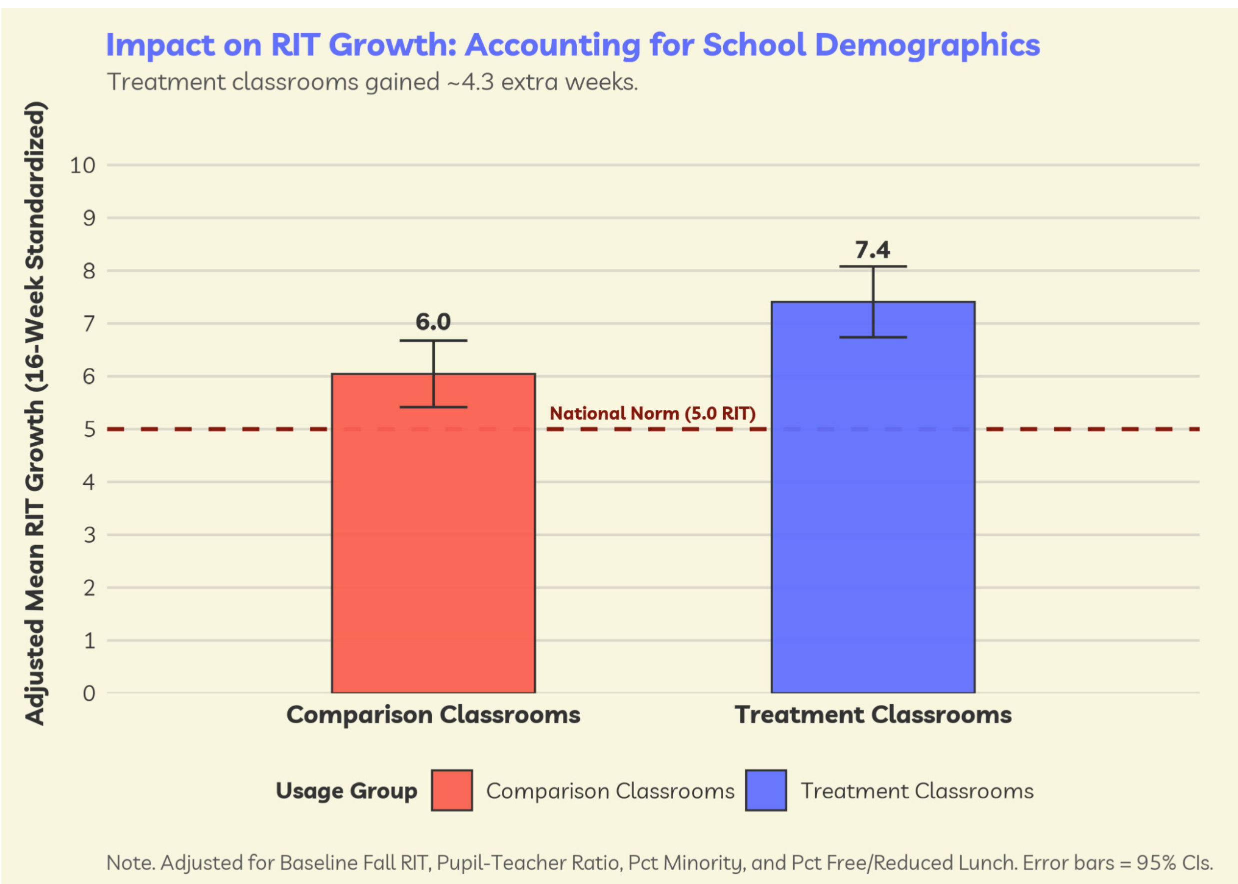


Follow-Up Analyses:
The HLM above, incorporating school-level random effects and classroom-mean fall RIT score as a covariate, noted a substantial clustering of RIT growth outcomes within schools (ICC= 0.374). By examining the specific school-level demographic characteristics — sourced from the National Center for Education Statistics — we wanted to investigate specific differences in school composition that might be driving the observed differences in RIT growth between the treatment and comparison, rather than the Raz-Plus solution itself. The analytic sample for which school-level data were available consisted of 141 third grade classrooms across 83 schools (Low usage: n = 64; Consistent usage: n = 77), averaging 1.7 classrooms per school (range: 1–6). The groups were equivalent on fall reading scores (classroom-mean fall RIT Hedges' g = –0.127), meeting the WWC threshold for equivalence with statistical adjustment.

Demographic balance checks revealed an imbalance: schools in the Treatment group enrolled a substantially lower percentage of students qualifying for free or reduced-price lunch (M = 36.6% vs. 50.6%; g = –0.565), exceeding the WWC non-equivalence threshold and indicating that treatment classrooms were, on average, located in less economically disadvantaged schools. Smaller imbalances were observed across the remaining school-level variables, pupil-to-teacher ratio (g = –0.198), percentage minority (g = 0.154), and percentage female (g = –0.071), all of which fell within the WWC range requiring statistical adjustment but not non-equivalence. All variables were retained as covariates in this model.

This follow-up HLM, which included school-level demographic covariates alongside school random effects and baseline fall RIT score, revealed that ***treatment classrooms continued to demonstrate significantly greater reading growth than comparison classrooms (β = 1.47, SE = 0.52, t(70.64) = 2.80, p = .007).*** The presence of this effect after controlling for FRP, racial/ ethnic composition, and pupil-to-teacher ratio strengthens the inference that Raz-Plus solution is indeed effective. Baseline RIT scores remained a significant negative predictor (p = .043), consistent with regression toward the mean. School-level poverty (β = –0.039, p = .011) and minority (β = 0.025, p = .023) were associated with RIT growth, confirming that their inclusion in the model adjusted the estimates. Pupil-to-teacher ratio did not reach significance after accounting for the remaining covariates (p = .147). As a sensitivity check, an ANCOVA ignoring school clustering produced a nearly identical point estimate (β = 1.42, SE = 0.45, p = .002), indicating the finding is robust to modeling choice, though the HLM is a more rigorous, conservative approach, and produces appropriate standard errors. Importantly, post-hoc 16-week standardization of the effect also produced a meaningful and practical estimate of RIT growth. Consistent Raz-Plus usage classrooms grew 1.33 RIT points more than no-usage classrooms over the standard 16-week window. Relative to the NWEA third grade national norm of 5.0 RIT points, this represents an advantage equivalent to approximately 4.2 additional weeks of typical reading instruction (Figure 2).

Figure 2.
Impact Results, Accounting for School Demographics (NCES) and Clustering



Subgroup Analysis (Exploratory Analyses)

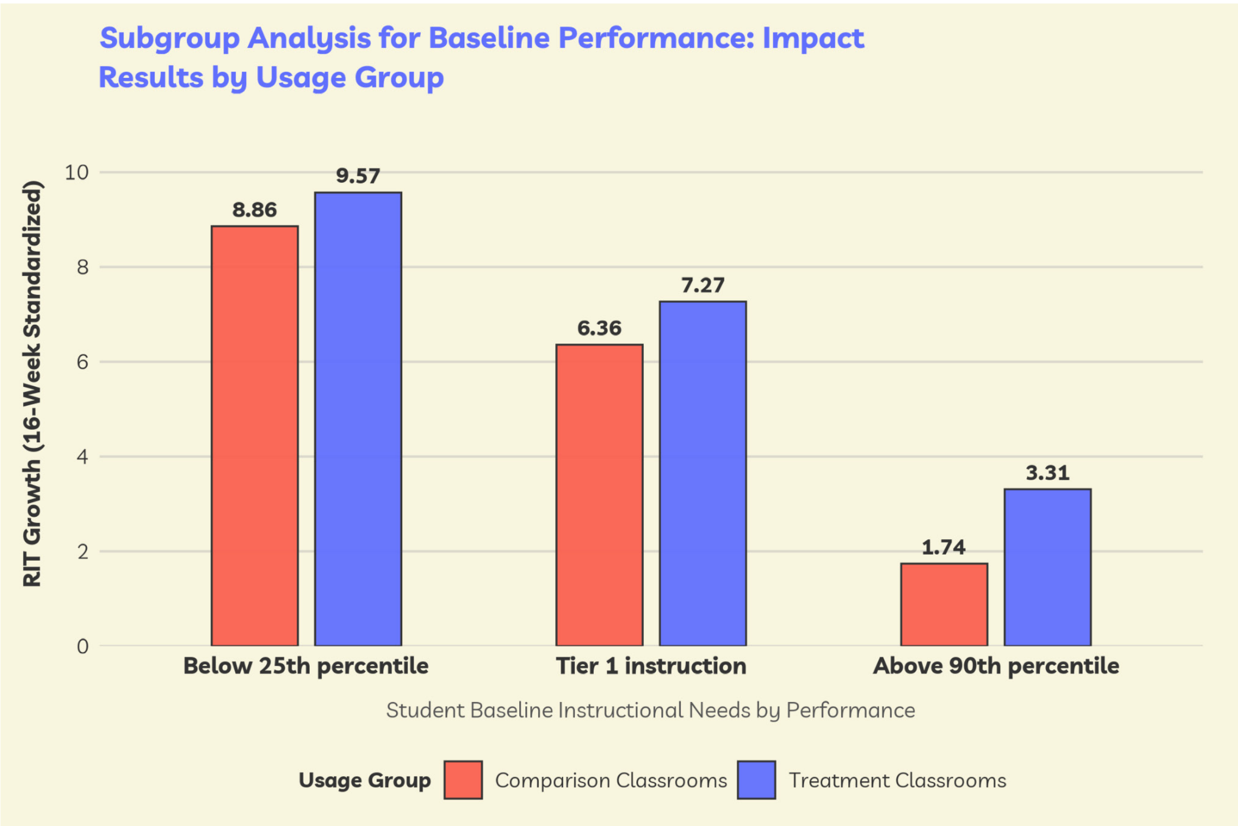
Exploratory analyses at the student level were also conducted to determine whether the impact of Raz-Plus varied by students' baseline fall RIT. Students from the matched classrooms were grouped into three performance ranges based on beginning-of-year percentile: below 25th percentile, 25th -90th percentile, and above 90th percentile. Students in the first group are typically placed in Tier 2 or Tier 3 intervention programs and may receive additional reading supports, such as individualized tutoring or remediation. Students in the second group typically receive only Tier 1 or core reading instruction. Students in the third group typically qualify for gifted enrichment or acceleration opportunities and need access to texts beyond the typical grade range.

When comparing Raz-Plus students to comparison students, students in the Tier 1 and high-performing groups who used Raz-Plus showed significantly higher growth than non-users. Raz-Plus users in the lowest-performing group also showed a larger increase than nonusers, but this difference did not reach statistical significance.

Table 4.
Subgroup Analysis for Baseline Performance: Impact Results by Treatment Group

Student Baseline Performance	Treatment		Comparison		Results		
	N Students	M (SD)	N Students	M (SD)	Mean Difference (SE)	t-value	p-value
Below 25th Percentile	304	9.57 (10.70)	326	8.86 (9.65)	.72 (.81)	0.88	.378
25th - 90th Percentile	1093	7.27 (7.19)	1152	6.36 (7.38)	.91 (.31)	2.96	.003
Above 90th Percentile	169	3.31 (5.82)	185	1.74 (6.59)	1.57 (.66)	2.36	.019

Figure 2:
Subgroup Analysis for Baseline Performance: Impact Results by Treatment Group



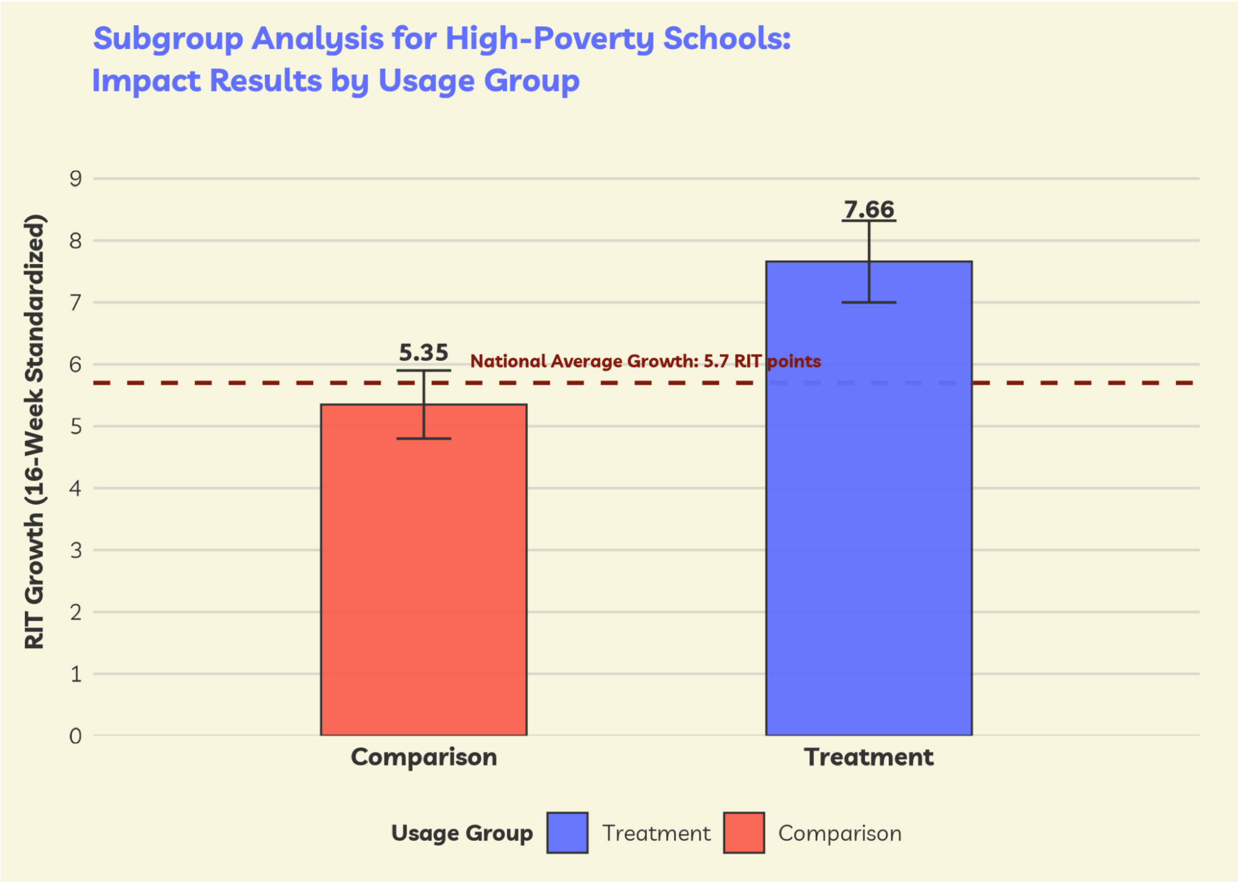
To assess the impact of Raz-Plus in high-poverty environments, exploratory analyses were performed on a subsample defined by the NCES as students attending schools where at least 75% of students qualify for Free and Reduced-Price Lunch (FRPL). This group included 180 Raz-Plus participants and 261 non-users.

The results within this high-poverty demographic were significant: Raz-Plus users achieved a growth of +7.7 points, outperforming the +5.4 points recorded by non-users. While the non-user group grew at a rate comparable to the national average for their baseline range (+5.7 RIT points), the Raz-Plus cohort was the only one to exceed that national benchmark. For students in under-resourced contexts who often face heightened academic risks, Raz-Plus was more effective at accelerating third-grade reading growth toward proficiency than core reading programs alone.

Table 5.
Subgroup Analysis for High-Poverty Schools: Impact Results by Treatment Group

Treatment		Comparison		Results		
N Students	M (SD)	N Students	M (SD)	Mean Difference (SE)	t-value	p-value
180	7.66 (8.36)	261	5.35 (7.99)	2.61 (.79)	2.93	.004

Figure 3:
Subgroup Analysis for High-Poverty Schools: Impact Results by Treatment Group



Discussion

The primary objective of this study was to evaluate the real-world efficacy of Raz-Plus as a supplemental literacy solution for third-grade students, a critical juncture in reading development. The findings offer compelling evidence that regular, active engagement with Raz-Plus significantly accelerates reading proficiency growth, surpassing gains from baseline-matched classrooms without Raz-Plus usage.

Amplifying Core Instruction

Classrooms using Raz-Plus achieved a +1.36 RIT advantage over their matched peers, a growth advantage of +27% over the national normative rate for the assessment period. In the context of the school year, this translates to approximately a month of additional reading instruction. This supports the hypothesis that flexible, SoR-aligned supplemental platforms like Raz-Plus are vital for complementing core curriculum and providing the volume and variety of targeted practice required to solidify decoding and meaning-making skills before the fourth-grade shift to "reading to learn."

Equity and Impact in Under-Resourced Environments

Perhaps the most notable finding emerged from the subgroup analysis of high-poverty schools (defined as >75% FRPL). Students in these environments face compounded systemic barriers that traditionally depress reading growth rates. However, Raz-Plus users in these schools not only outperformed similar non-user peers (+7.7 RIT vs +5.4 RIT) but were the only group in this demographic to exceed the national benchmark. This suggests that providing teachers in under-resourced schools with easily accessible, highly engaging, and differentiated resources can effectively mitigate some of the gaps that drive persistent achievement disparities.

Differentiated Support Across Proficiency Levels

The data also highlights how Raz-Plus functions across the proficiency spectrum. The significant growth observed in Tier 1 and high-performing (gifted) cohorts indicates that the platform's expansive digital library and self-paced activities successfully challenge on-level and advanced students.

Conversely, while students in the lowest-performing cohort (<25th percentile) experienced greater growth with Raz-Plus than without it, the difference did not reach statistical significance. This aligns with standard Multi-Tiered System of Supports (MTSS) frameworks: students with the most profound foundational deficits typically receive highly intensive, individualized, and teacher-led direct instruction. Raz-Plus serves as a highly effective Tier 1 supplement and practice tool, but educators should continue to pair it with intensive human-led interventions for the highest-risk readers.

Conclusion and Future Directions

Ultimately, these findings — meeting ESSA Tier 2 standards — confirm that integrating Raz-Plus into elementary literacy blocks yields significant, rapid improvements in reading achievement. Future research should expand on these findings by examining longitudinal impacts over a full academic year and by examining specific usage patterns (e.g., decodable texts vs. comprehension quizzes) to further optimize instructional implementation for varied learner profiles.

References

Chall, J. S. (1983). *Stages of reading development*. New York, NY: McGraw-Hill.

Feister, L. (2010). *Early Warning! Why Reading by the End of Third Grade Matters*. Annie E. Casey Foundation report. https://eric.ed.gov/?id=ED509795

Goldhaber, D., Wolf, M., & Daly, T. (2021). *Assessing the Accuracy of Elementary School Test Scores as Predictors of Students' High School Outcomes*. American Institutes for Research (AIR) CALDER Working Paper No. 235-0821-2 https://eric.ed.gov/?id=ED605735

The Harris Poll (2025). *Nemours Children's Reading BrightStart! Early Literacy Survey: A National Survey by Nemours Children's Health*. [Unpublished Report] https://www.nemours.org/content/dam/nemours/nemours-org/en/documents/NCRBS-harris-poll-executive-summary.pdf

Hernandez, D. J. (2011). *Double Jeopardy: How Third-Grade Reading Skills and Poverty Influence High School Graduation*. Annie E. Casey Foundation report. https://eric.ed.gov/?ID=ED518818

Lesnick, J., Goerge, R., Smithgall, C., & Gwynne, J. (2010). *Reading on Grade Level in Third Grade: How Is It Related to High School Performance and College Enrollment?* Annie E. Casey Foundation report. https://eric.ed.gov/?id=ED517805

National Center for Education Statistics. (2017). *The Nation's Report Card: NAEP Reading 2017 Assessment*. U.S. Department of Education, Institute of Education Sciences. https://www.nationsreportcard.gov/

National Center for Education Statistics. (2024). *The Nation's Report Card: NAEP Reading 2024 Assessment*. U.S. Department of Education, Institute of Education Sciences. https://www.nationsreportcard.gov/

Science of Reading Center (2025). *How is it Going? Insights from NYS Educators on the Implementation of the Science of Reading*. Science of Reading Center at SUNY New Paltz report. https://www.newpaltz.edu/media/science-of-reading/SUNY-New-Paltz-Science-of-Reading-Center-2025-New-York-State-Educator-Survey.pdf

APPENDIX A: Analytic Approach

Propensity Score Matching (PSM) for Matched Groups

To address the differences in baseline fall RIT (Hedge's g = 0.42) and baseline percentile (Hedge's g = 0.43) between treatment (N= 85) and comparison classrooms (N= 151), we employed a 1:1 nearest neighbor propensity score matching without replacement, implemented using the *MatchIt* package (Ho et al., 2011) in R. Propensity scores were estimated using logistic regression with baseline fall RIT scores as the covariate. Following Austin (2011), we used a caliper of 0.2 standard deviations of the logit of the propensity score. Essentially, to create matching groups, we estimated each classroom's probability of being in the treatment group based on its average baseline fall RIT score. We then matched each treatment classroom to a comparison classroom with the most similar probability. To ensure match quality, we discarded any pair in which the two classrooms were dissimilar, specifically when their scores differed by more than 0.2 standard deviations on the matching scale. Each comparison classroom was matched to only one treatment classroom, and the random seed was fixed (seed = 2026) to ensure reproducibility.

Baseline Equivalence and Matching Results

Baseline equivalence of the matched sample was assessed using Hedge's g value in accordance with the WWC standards. The primary equivalence check used baseline fall RIT scores, and the secondary check used baseline percentile as a robustness indicator. After matching, 82 classrooms were kept in each group (N = 164 classrooms total). The matched groups demonstrated strong baseline equivalence in RIT scores (g = 0.041, 95% CI [–0.265, 0.347]), meeting the WWC threshold for equivalence without statistical adjustment ($g \leq 0.05$). Baseline percentile also met the equivalence standard (g = 0.052, 95% CI [–0.255, 0.358]), though at the level requiring statistical adjustment ($0.05 < g \leq 0.25$).

Hierarchical Linear Model (HLM)

RIT growth was analyzed using Hierarchical Linear Modeling (HLM), a statistical approach to account for classrooms nested within schools. Because classrooms in the same school share similar resources, leadership, and student populations, their outcomes cannot be treated as fully independent. Standard regression can produce artificially narrow standard errors and Type 1 errors. HLM accounts for school-level clustering explicitly, partitioning variance into the portion attributable to differences between schools and the portion attributable to differences between classrooms (specifically, teacher-led instruction) within schools. The analytic sample comprised 164 Grade 3 classrooms across 97 schools. The primary outcome was classroom-mean RIT score growth from Fall to Winter.

Model Development

Null model. Prior to introducing any predictors, an unconditional model was estimated to quantify the degree of school-level clustering. The resulting Intraclass Correlation Coefficient (ICC = 0.457) indicated that 45.7% of the total variance in reading growth was attributable to between-school differences, with the remaining 54.3% was at the classroom level, further justifying the need to use an HLM rather than standard ANCOVAs.

Classroom-level predictor. Classroom mean fall RIT score (baseline) was entered as a Level 1 covariate.

School-level demographic covariates. A subsequent model incorporated school-level demographic characteristics sourced from the National Center for Education Statistics (NCES, 2024–25). These covariates — percentage of students qualifying for free or reduced-price lunch, percentage of students from minoritized racial/ethnic groups, and pupil-to-teacher ratio — were included to ensure that observed differences in growth between usage groups were not confounded with broader school-level differences in school composition or resources.

Estimation. All models were estimated using Restricted Maximum Likelihood (REML) via the lme4 package (Bates et al., 2015) in R. Degrees of freedom and p-values were computed using Satterthwaite's approximation via the lmerTest package (Kuznetsova et al., 2017).

Post-Hoc 16-Week Standardization

The interval between Fall and Winter assessments varied across classrooms, ranging from 11.0 to 22.9 weeks (sample mean = 17.7 weeks, SD = 2.77). Because the testing window differed between usage groups, the treatment effect was standardized post hoc to a 16-week reference window. This window corresponds to NWEA's default fall-to-winter testing interval and is the basis of national growth norms. This approach was preferred over refitting the model on a ratio outcome (growth/weeks×16), as dividing by a classroom-varying quantity introduces heteroscedasticity that violates assumptions of the mixed-effects framework and inflates standard errors.

Effect Size Estimation

The magnitude of the difference between treatment and comparison classrooms was quantified using Hedges' g, a bias-corrected variant of Cohen's d that applies a small-sample correction factor. Hedges' g was computed for baseline equivalence checks on pretest variables, with WWC-standard thresholds applied: $|g| \le 0.05$ indicates practical equivalence requiring no adjustment; $0.05 < |g| \le 0.25$ indicates equivalence with statistical adjustment; $|g| > 0.25$ indicates a meaningful imbalance requiring covariate control. For the primary outcome, the treatment effect was additionally expressed in weeks of equivalent typical reading instruction, derived by dividing the 16-week standardized effect by NWEA's Grade 3 fall-to-winter weekly growth norm (5.0 RIT ÷ 16 weeks = 0.3125 RIT per week).

Table 1

Hierarchical Linear Model Predicting Third Grade Classroom Fall to Winter RIT Growth: Baseline Model (School Clustering + Fall RIT Covariate)

Parameter	Estimate (raw RIT)	*SE*	*df*	*t*	*p*	Estimate (16-wk RIT)	*SE* (16-wk)
Fixed Effects							
Intercept	15.851	6.065	151.30	2.61	.010		
Classroom Mean Fall RIT	−0.049	0.032	151.22	−1.51	.132		
Raz-Plus Usage (Consistent)	1.505	0.512	97.01	2.94	.004	1.360	0.463
Random Effects (Variance Components)							
School Intercept Variance	3.474					—	
Residual Variance	4.407					—	
ICC (School)	0.441					—	
Model Information							
N Classrooms / N Schools	164 / 97						
REML Criterion	788.6						

***Note.* Outcome is classroom mean RIT growth (fall to winter). Model estimated with REML via lme4; degrees of freedom and p values use Satterthwaite's approximation (lmerTest). The Raz-Plus reference category is No-Usage.**

Table 2

Hierarchical Linear Model Predicting Third Grade Classroom RIT Growth: Adjusted Model With School Demographic Covariates (School Clustering + Fall RIT Covariate)

Parameter	Estimate (raw RIT)	*SE*	*df*	*t*	p	Estimate (16-wk RIT)	*SE* (16-wk)
Fixed Effects							
Intercept	24.130	7.952	132.54	3.04	.003		
Classroom Mean Fall RIT	−0.082	0.040	130.73	−2.04	.043		
Raz-Plus Usage (Consistent)	1.466	0.524	70.64	2.80	.007	1.325	0.473
Pupil/Teacher Ratio	−0.125	0.085	75.10	−1.47	.147		
% Minority	0.025	0.011	61.99	2.33	.023		
% Free/Reduced Lunch	−0.039	0.015	95.87	−2.59	.011		
Random Effects (Variance Components)							
School Intercept Variance	2.334					—	
Residual Variance	3.909					—	
ICC (School)	0.374					—	
Model Information							
N Classrooms / N Schools	141 / 83						
REML Criterion	664.2						

***Note.* Outcome is classroom mean RIT growth (fall to winter). School demographic covariates were sourced from NCES (2024–25). Model estimated with REML via lme4; degrees of freedom and p values use Satterthwaite's approximation (lmerTest). The reference category is No-Usage. The analytic sample (N = 141) is smaller than in Table 1 (N = 164) because classrooms missing NCES covariate data were excluded.**

This meets WWC standards with Reservations for baseline equivalency, and Every Student Succeeds Act (ESSA) Tier 2 (Moderate) levels of evidence according to the U.S. Department of Education guidelines.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). *MatchIt: Nonparametric Preprocessing for Parametric Causal Inference.* Journal of Statistical Software, 42(8). https://doi.org/10.18637/jss.v042.i08

Austin, Peter C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." Multivariate Behavioral Research 46 (3): 399–424. https://doi.org/10.1080/00273171.2011.568786.