The Good Behavior Game and Classroom Disruptions in Grade 4 Mathematics

An Action Research

[Copy for blind review: the researchers' names, positions, and school are removed. The final copy lists them here.]

Conducted June to September 2026

Abstract

This action research tested whether the Good Behavior Game, in which teams of learners earn a reward by keeping their marks for broken rules low, would reduce disruptions in a Grade 4 Mathematics class of 38 learners in a public elementary school. An ABAB reversal design alternated phases without the game (A1, 7 sessions), with it (B1, 9), without it (A2, 7), and with it again (B2, 9) from July 10 to August 27, 2026. Two observers counted disruptions in four 10-minute intervals of each period; the second counted 12 of the 32 sessions, and agreement was excellent, ICC(A,1) = .96, 95% CI [.93, .98]. Disruptions averaged 27.6 and 19.3 per 40 minutes without the game and 7.8 and 8.4 with it, and changed within days of each phase change. Nonoverlap of all pairs was .97, .95, and .94 at the three changes, and Tau combined across them was .91, $p < .001$. A negative binomial model with generalized estimating equations, which allowed for overdispersed counts and for correlated intervals within a period, estimated that the game cut the rate of disruptions by 65%, incidence rate ratio = 0.35, 95% CI [0.26, 0.48], $p < .001$. A t test on the interval counts would have treated correlated intervals as independent; a simulation showed it would call a useless game effective about twice as often as it should. Most learners liked the game and wanted to play it again. The teacher will continue the game in Mathematics.

Keywords: Good Behavior Game, positive discipline, single-case design, negative binomial regression, generalized estimating equations

Acknowledgment

The researchers thank the Grade 4 teacher who counted disruptions in all 32 periods during her vacant period; the Master Teacher in Mathematics who counted the second set; the school head, who approved the study and adjusted the observers' schedules; the Schools Division Research Committee, whose reviewers asked for the observers' practice sessions before the baseline; the parents, who gave their consent at the June general assembly; and the learners of the section, who played the game and told us what they thought of it. The study was funded by the Basic Education Research Fund (BERF) for 2026.

The Good Behavior Game and Classroom Disruptions in Grade 4 Mathematics

Context and Rationale

DepEd expects teachers to manage learner behavior without punishment. The Child Protection Policy defines positive and non-violent discipline as "a holistic, constructive and pro-active approach to teaching that helps children develop appropriate thinking and behavior in the short and long-term and fosters self-discipline" (Department of Education [DepEd], 2012, Sec. 3[P]), and DepEd Order No. 006, s. 2026 directs schools to "strengthen positive discipline, avoiding corporal or punitive actions" (DepEd, 2026a, Sec. IX.D.3). The Philippine Professional Standards for Teachers ask a teacher in the Proficient career stage to "manage learner behavior constructively by applying positive and non-violent discipline to ensure learning-focused environments" (Indicator 2.6.2; DepEd, 2017b). The Classroom Observation Tool rates that strand as its Indicator 12; under the multi-year performance management guidelines it was rated in teachers' second classroom observation of School Year 2025–2026 and is not scheduled in School Years 2026–2027 or 2027–2028 (DepEd, 2025).

Our Grade 4 section needed such an approach in Mathematics. In June 2026, the class adviser kept a tally of disruptions for five periods: learners leaving their seats, calling out, and making noise while the class worked. She counted 15 to 40 a period, enough that explanations had to be repeated and seatwork ran out of time. DepEd's rules on learner discipline treat "disruptive behavior and/or pranks in class" as a first-level offense, answered first by a

written reprimand and a notice to the parents (DepEd, 2026a, Annex I). A reprimand answers one learner's act after it happens; the adviser wanted a way to make the whole class's ordinary behavior calmer before acts happened.

The Good Behavior Game is such a way. Barrish et al. (1969) divided a fourth-grade class into two teams "to play a game" in which each out-of-seat or talking-out act earned a mark against the learner's team, and a team that kept its marks low won privileges; they applied it "first during math period and then during reading period" in a design with reversals. In the five decades since, a review of 22 studies found "moderate to large effects" on challenging behaviors that were "immediate," and noted that "correct application of reward procedures" mattered (Flower et al., 2014). A meta-analysis of 21 single-case studies with 1,580 students found a Tau-U of .82 across 137 phase contrasts and judged the game "most effective in reducing disruptive and off-task behaviors" (Bowman-Perrott et al., 2016). A trial that played the game in first- and second-grade classrooms followed the learners into young adulthood and found fewer drug and alcohol disorders, less regular smoking, and less antisocial personality disorder among the boys, especially those who had been aggressive and disruptive in first grade (Kellam et al., 2008).

Those studies come from other school systems. Whether the game calms a Filipino Grade 4 class of 38 in a 45-minute Mathematics period, with rewards a public school can afford, is a question for our own classroom. This study answers it with a design that shows the game's effect three times: when it starts, when it stops, and when it starts again. It addresses the Teaching and Learning theme of the Basic Education Research Agenda (DepEd, 2016).

Innovation, Intervention, and Strategy

The innovation was the Good Behavior Game, played during the Mathematics period. Everything else about the period stayed the same: the teacher, the competencies of the week, the time, and the seating.

The Rules and the Teams

Four rules were posted in English and Filipino at the front of the room: stay in your seat unless the teacher allows you to leave it; raise your hand to speak; work quietly; and keep your hands and feet to yourself. The adviser divided the class into four teams of nine or ten learners, mixing learners who often broke the rules with those who rarely did, and let each team choose a name.

How the Game Was Played

At the start of each period, the teacher said, "We are playing the game today," and reviewed the rules in one minute. Each time a learner broke a rule, she put a mark under the learner's team on a corner of the board and named the rule, calmly and without naming or scolding the learner. At the end of the period, every team with four marks or fewer won; teams did not compete with one another, and all four could win. A winning team put a sticker on the class chart and lined up first for recess. On Friday, every team that had won on at least three days that week chose the class's energizer for the next Monday. The rewards cost only the stickers.

Safeguards

A learner whose marks repeatedly cost the team was spoken to privately, never in front of the class, and the teacher stopped any teasing at

once. No learner was moved to a team of one, and no privilege a learner already had was taken away.

Strategy and Evaluation Plan

The design called for the game to start, stop, and start again, so that its effect could be seen three times. A phase was changed only after at least five sessions and once the counts had steadied, by the researchers' reading of the graph. Two observers, trained before the baseline, counted the disruptions; the teacher's own marks were part of the game and were not the study's data.

Action Research Questions

The study answered five questions:

1. What were the disruptions per 40 minutes and per 10-minute interval in each phase, and what does visual analysis of the four phases show?
2. Is there a significant difference between the rates of disruption with and without the game, taking overdispersion and the correlation of intervals within a period into account?
3. How large and how consistent was the change at each of the three phase changes?
4. How closely did the two observers agree?
5. How did the learners rate the game?

One null hypothesis was tested at the .05 level of significance:

H_0: There is no significant difference between the rates of disruption in Mathematics periods with and without the Good Behavior Game.

Action Research Methods

Research Design

The study used an ABAB reversal design, a single-case design in which the class serves as its own comparison (Horner et al., 2005; Kratochwill et al., 2013). A baseline without the game (A1) was followed by the game (B1), its withdrawal (A2), and its return (B2), giving three chances to see the game's effect at three points in time, the number the single-case standards require (Kratochwill et al., 2010). The design met the What Works Clearinghouse's current standards for data points: the first baseline had at least six sessions, and at least two phases of each condition had five or more (What Works Clearinghouse [WWC], 2022).

Participants and Other Sources of Data and Information

The participants were the 38 learners of one Grade 4 section of a public elementary school, 20 boys and 18 girls, in Term 1 of School Year 2026–2027, and their class adviser, who taught them Mathematics for 45 minutes a day (DepEd, 2024). Attendance in the 32 observed periods ranged from 34 to 38 learners, with a mean of 35.8. The observers counted the class's disruptions, not any one learner's; no learner's behavior was recorded by name.

Data Gathering Methods

Counting Disruptions

A disruption was any of three acts: leaving the seat without the teacher's permission; speaking without being called on, loudly enough for the observer to hear; or making noise or contact that interrupts others, such as tapping, throwing, or poking. An act counted once; the same learner's act

counted again only after it had stopped for at least three seconds. The observers sat at the back of the room and recorded each disruption on a sheet with four boxes, one for each 10-minute interval of the first 40 minutes of the period, timed by a phone set to vibrate every 10 minutes.

The primary observer, a Grade 4 teacher, counted all 32 sessions during her vacant period. The second observer, the Master Teacher in Mathematics, counted 12 sessions, three in each phase, chosen in advance and not told to the primary observer. Before the baseline, the two practiced in two Mathematics periods of another section until their interval counts agreed within one disruption in at least 80% of the intervals.

The Teacher's Marks and the Learners' Ratings

During the game, the adviser's marks for each team were copied from the board after each period. On August 27, 2026, after the last session, the 36 learners present rated three statements by circling one of three faces (Yes, A little, No): "I liked playing the game," "The game helped me listen and work in Math," and "I want to play the game again."

Ethical Considerations

The Schools Division Research Committee approved the proposal before the study began; its reference number is withheld for blind review. The school head gave written permission, parents gave written consent at the June 2026 general assembly, and the learners gave their assent. The game is a form of positive discipline: it takes no privilege away and names no learner (DepEd, 2012). The study recorded only the class's counts, kept the sheets with the researchers, and reported no individual learner. Withdrawing the game in A2

held back an approach that seemed to work; the phase was kept to seven sessions, and the game resumed in B2 and continued after the study. The researchers declared their use of an artificial intelligence (AI) tool in Appendix E (DepEd, 2026b).

Data Analysis

Each phase was described by the mean, median, standard deviation, and range of its session totals, the least-squares trend of those totals, and the mean and variance of its interval counts. The four-phase graph was read for the six features of visual analysis: level, trend, variability, immediacy of the effect, overlap, and consistency across similar phases (Kratochwill et al., 2010). Immediacy compared the means of the last three sessions of one phase and the first three of the next. The What Works Clearinghouse's checks for bias were applied: a nonoverlap of .85 or less between the last three baseline sessions and the earlier ones shows a minimal baseline trend, and a nonoverlap of .85 or less between the withdrawal phase and the first baseline shows that the behavior returned (WWC, 2022).

Overlap at each phase change was measured with the nonoverlap of all pairs (NAP), the share of all pairs of one session before and one after the change in which the later session went the way the game predicts, ties counting half (Parker & Vannest, 2009), and with Tau, the same pairs' improvements minus deteriorations over all pairs (Parker et al., 2011). Tau-U, which also subtracts the earlier phase's own trend, was computed for comparison; following Parker et al. (2011), as Brossart et al. (2018) report, the trend correction was to be applied only where the earlier phase's trend

reached a Kendall's tau of .40. The three phase changes' Tau values were combined with inverse-variance weights.

For the second question, the interval counts were analyzed with count models in the statsmodels library for Python (Seabold & Perktold, 2010). A Poisson model's ratio of the Pearson chi-square to its degrees of freedom tested for overdispersion, and a negative binomial model estimated the dispersion parameter, α. The main analysis was a negative binomial regression fitted with generalized estimating equations (GEE; Liang & Zeger, 1986), with each session as a cluster, a first-order autoregressive (AR[1]) working correlation among its four intervals, and robust standard errors, with the game, the interval's position in the period, and the second half of the study (A2 and B2) as predictors. The exponentiated coefficient of the game is the incidence rate ratio (IRR), the factor by which the game multiplied the rate of disruptions. Three checks followed: an exchangeable working correlation; a linear trend across sessions in place of the second-half term; and, because the sessions themselves may run alike from one day to the next (Shadish & Sullivan, 2011), the lag-1 autocorrelation of the session residuals with the Ljung-Box test and a negative binomial model of the session totals with Newey-West standard errors, which allow for that autocorrelation, though with 32 sessions only in part (Newey & West, 1987). A second GEE with the four phases as factors estimated the IRR at each phase change. For contrast, Welch's t test was run on the interval counts, and a simulation of 1,000 classes for each of five levels of day-to-day autocorrelation, with no game effect at all, measured how often each analysis would report one at the .05 level.

The observers' agreement was measured on the 48 intervals both counted with the intraclass correlation from a two-way mixed-effects model, for absolute agreement of a single rater, interpreted by its 95% confidence interval (Koo & Li, 2016); it equals the two-way random-effects estimate, which pingouin labels ICC(A,1) (Vallat, 2018), and with the percentages single-case studies report: total-count agreement (the smaller session total over the larger), mean count-per-interval agreement, and exact agreement. The learners' ratings were counted.

All analyses were run in Python 3.12 with pandas, statsmodels, and pingouin, free software that a school can use without buying a license; the notebook and the Excel data set reproduce every result (Appendix A).

Discussion of Results and Reflection

Implementation

The class played the game in all 18 sessions of B1 and B2. The Mathematics period did not meet on July 28, the day of Term 1's second teacher-made summative test (DepEd, 2026c); on August 6, when there were no classes; or on August 21, Ninoy Aquino Day (Proclamation No. 1006, 2025). That left 32 sessions from July 10 to August 27; the Term 1 Examination began on August 28 (DepEd, 2026c). All four teams won in 15 of the 18 game sessions. The adviser's marks came to 74.0% of the disruptions the primary observer counted in the same sessions, $r = .93$: she marked most disruptions while teaching, and the observers, who did nothing else, saw more.

Question 1: Disruptions in Each Phase

Without the game, the class averaged 27.6 disruptions in 40 minutes in A1 and 19.3 in A2; with it, 7.8 in B1 and 8.4 in B2 (Table 1). The game phases were also steadier: their sessions ranged from 1 to 14 disruptions, against 10 to 44 without the game.

Table 1

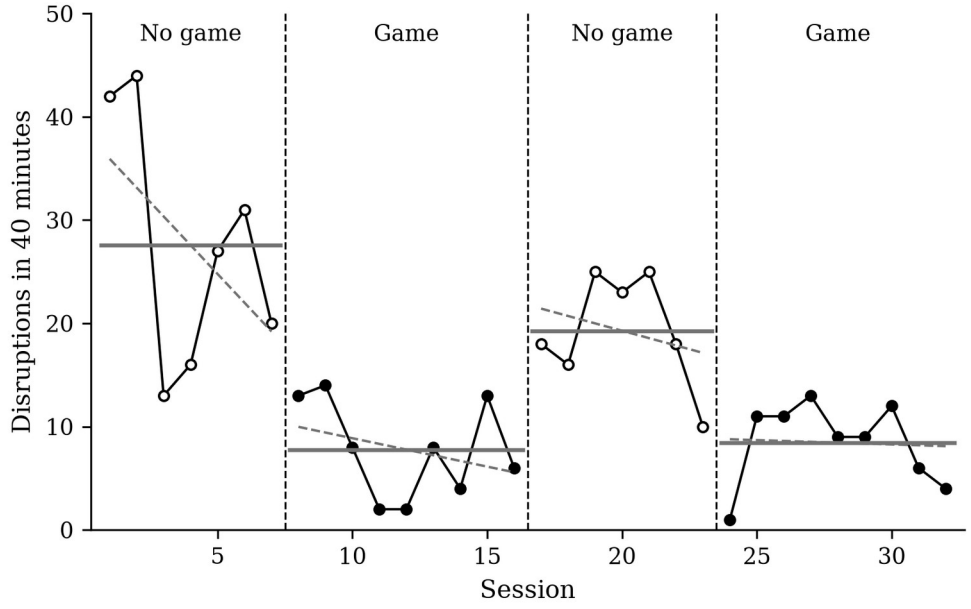
Disruptions in 40 Minutes and per 10-Minute Interval, by Phase

Phase	Sessions	*M*	*SD*	*Mdn*	Range	Trend per session	Interval *M*	Variance / mean
A1, no game	7	27.6	12.2	27.0	13–44	−2.79	6.89	3.27
B1, game	9	7.8	4.7	8.0	2–14	−0.55	1.94	1.76
A2, no game	7	19.3	5.5	18.0	10–25	−0.71	4.82	3.00
B2, game	9	8.4	4.0	9.0	1–13	−0.08	2.11	1.46

Note. Counts by the primary observer. A1 ran from July 10 to 20, B1 from July 21 to August 3, A2 from August 4 to 13, and B2 from August 14 to 27, 2026. The trend is the least-squares slope of the session totals within the phase. The variance-to-mean ratio is that of the interval counts; above 1, they spread more than a Poisson count.

Figure 1

Disruptions in 40 Minutes in the Four Phases



Note. Filled markers show sessions with the game. Horizontal lines show each phase's mean; dashed lines show its trend. Vertical dashed lines mark the phase changes.

The graph shows the game's effect three times (Figure 1). The level fell when the game began, rose when it stopped, and fell again when it resumed. The change was immediate: the means of the last three sessions before each change and the first three after it went from 26.0 to 11.7, from 7.7 to 19.7, and from 17.7 to 7.7. Both no-game phases drifted downward, A1 by 2.8 disruptions a session, a slope pulled down by two loud first days, but their trend lines ended at 19.2 and 17.1, far above the game phases' levels. Few sessions overlapped: 3 of the 9 B1 sessions fell within the range of A1, 1 of the 7 A2 sessions within the range of B1, and 4 of the 9 B2 sessions within the range of A2. The two game phases nearly matched; the second no-game phase rose but stayed below the first, a sign that part of the calmer routine stayed after the game stopped. The What Works Clearinghouse's checks were

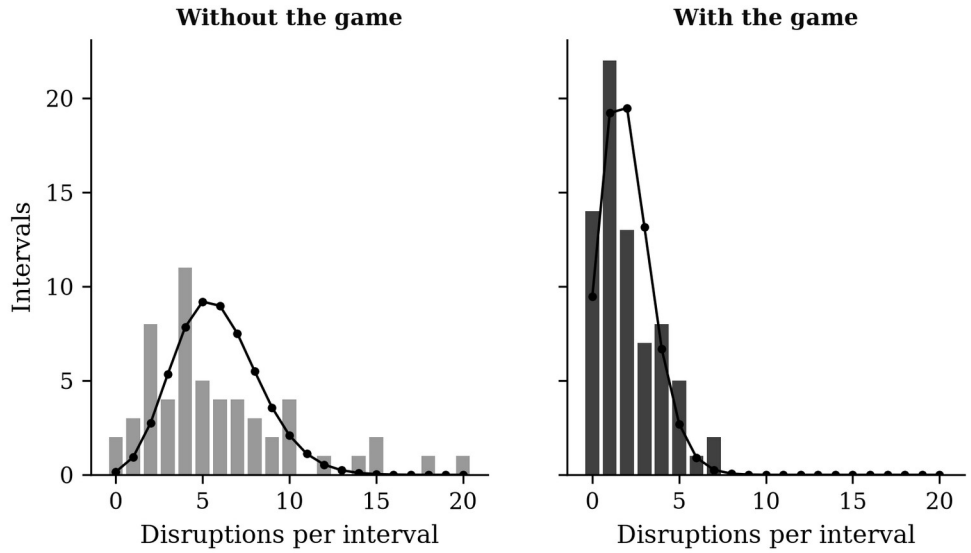
met: the last three A1 sessions against its earlier ones gave a nonoverlap of .50, showing a minimal baseline trend, and A2 against A1 gave .70, showing that the behavior returned when the game stopped.

Question 2: The Rate of Disruption With and Without the Game

The counts called for a count model. The Poisson model's dispersion ratio was 2.26, and the negative binomial model fitted far better, $\alpha = 0.31$, likelihood-ratio $x^2(1) = 53.26$, $p < .001$, AIC 576.8 against 628.1. Without the game, the interval counts' variance was three times their mean (Table 1; Figure 2).

Figure 2

Disruptions per 10-Minute Interval With and Without the Game



Note. Bars show the number of intervals at each count; points show what a Poisson count with the same mean would give.

Table 2

Negative Binomial GEE of Disruptions per 10-Minute Interval

Term	*b*	*SE*	*z*	*p*	IRR	95% CI of IRR
Intercept	1.691	0.185	9.13	< .001	5.427	[3.775, 7.802]

Term	*b*	*SE*	*z*	*p*	IRR	95% CI of IRR
Game (vs. no game)	−1.043	0.156	−6.70	< .001	0.353	[0.260, 0.478]
Interval (each 10 minutes later)	0.094	0.067	1.40	.160	1.099	[0.963, 1.253]
Second half (A2, B2)	−0.160	0.152	−1.05	.293	0.853	[0.633, 1.148]

Note. 128 intervals in 32 sessions (clusters). Negative binomial with α = 0.312; AR(1) working correlation within a session, estimated at .21; robust standard errors. IRR = incidence rate ratio.

Taking the dispersion and the correlation of a period's intervals into account, the game multiplied the rate of disruptions by 0.35, 95% CI [0.26, 0.48], $z = -6.70$, $p < .001$ (Table 2): a 65% reduction. The null hypothesis is rejected. The intervals of one period were correlated, AR(1) = .21, and disruptions did not rise significantly later in the period. The estimate held under every check (Table 3). The session residuals showed some day-to-day autocorrelation, lag-1 $r = .28$, Ljung-Box $Q(1) = 2.40$, $p = .122$, and the model of the session totals with Newey-West standard errors, which allows for it, gave the same IRR, 0.35.

Table 3

The Game's Incidence Rate Ratio Under Other Analyses

Analysis	IRR	95% CI	*p*
GEE, AR(1) within a session (main)	0.353	[0.260, 0.478]	< .001
GEE, exchangeable correlation	0.350	[0.256, 0.478]	< .001
GEE with a linear trend over sessions	0.393	[0.292, 0.529]	< .001
Session totals, Newey-West errors (2 lags)	0.350	[0.253, 0.483]	< .001
A1 to B1 (game begins)	0.285	[0.179, 0.452]	< .001
B1 to A2 (game stops)	2.464	[1.649, 3.682]	< .001
A2 to B2 (game resumes)	0.440	[0.309, 0.627]	< .001

Note. The last three rows come from a GEE with the four phases as factors; an IRR above 1 is a rise in disruptions.

A *t* test gives a smaller standard error than these analyses because it treats each interval as independent. Welch's *t* test on the interval counts gave a difference of −3.83 disruptions per interval, 95% CI [−5.07, −2.59], $t(69.3) =$

−6.14, $p < .001$; its conclusion here is the same, but its method is not sound for these data. The four intervals of one period are not independent, the counts are skewed, and their variance grows with their mean. In the simulation of classes in which the game did nothing, the *t* test on the intervals reported an effect in 9.4% of classes with no day-to-day autocorrelation and in 15.6% at an autocorrelation of .4, against the 5% it promises; the GEE did so in 6.2% and 12.0%. Autocorrelation between days inflated every analysis that treats days as independent, which is why the study rests on three replications and visual analysis, not on one *p* value.

Question 3: The Size and Consistency of the Change

At each phase change, the sessions after it went the way the game predicts in nearly every pair (Table 4). NAP was .97 when the game began, .95 when it stopped, and .94 when it resumed, and Tau combined across the three changes was .91, $SE = 0.17$, $z = 5.23$, $p < .001$; the combined test treats the three changes as independent, though they share B1 and A2, so it is a summary rather than a fourth test. None of the earlier phases' trends reached the .40 at which a correction is applied, so Tau is reported uncorrected. Tau-U is shown for comparison; at the B1 to A2 change it reached 1.02, beyond the bound of a correlation, because B1's slight downward trend, subtracted, added to the effect, a known weakness of the statistic (Tarlow, 2017).

Table 4

Nonoverlap at Each Phase Change

Phase change	Better	Worse	Ties	NAP	Tau	*p*	Trend of the earlier phase	Tau-U
A1 to B1 (fewer expected)	60	1	2	.968	.937	.002	.143	0.889

Phase change	Better	Worse	Ties	NAP	Tau	*p*	Trend of the earlier phase	Tau-U
B1 to A2 (more expected)	60	3	0	.952	.905	.003	−.194	1.016
A2 to B2 (fewer expected)	59	4	0	.937	.873	.004	.143	0.825

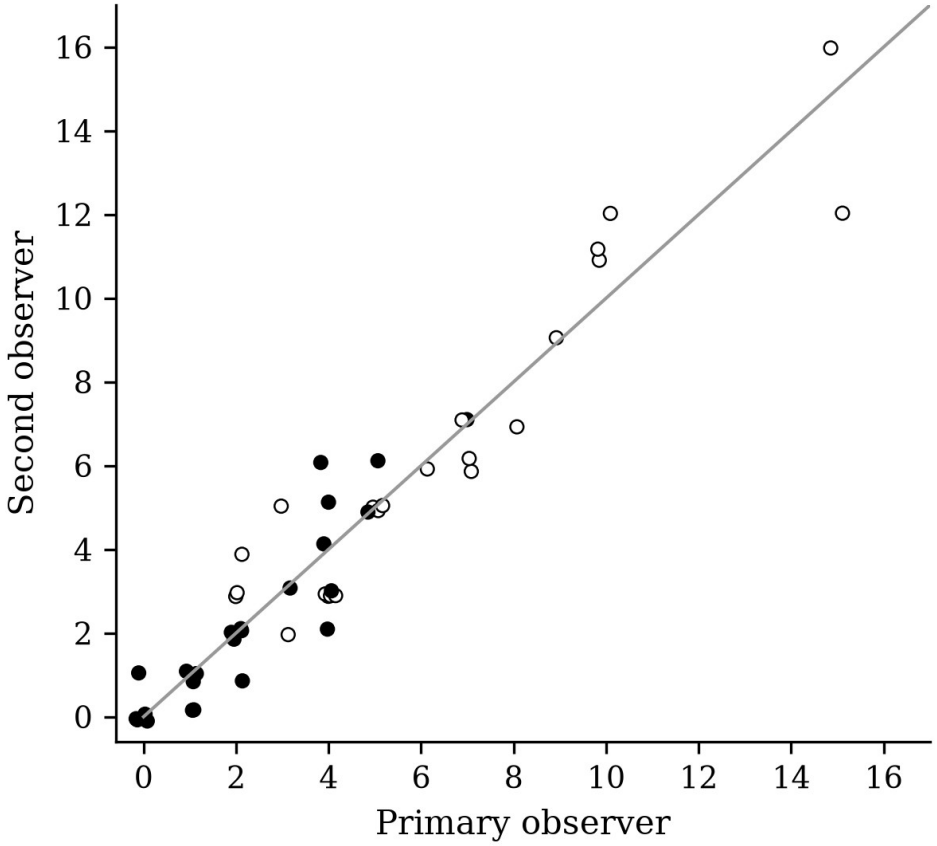
Note. Session totals; each change has 63 pairs (7 × 9). "Better" is a pair in which the later session went the way the game predicts. The trend is Kendall's tau of the earlier phase's sessions in time order, in the predicted direction. Tau-U subtracts that trend (Parker et al., 2011).

Question 4: Agreement Between the Observers

The two observers agreed closely (Figure 3). On the 48 intervals both counted, ICC(A,1) = .96, 95% CI [.93, .98], $F(47, 47) = 49.24$, $p < .001$: excellent agreement, since the whole interval lies above .90 (Koo & Li, 2016). Their session totals agreed by 94.1% on average and by at least 83.3% in every session; the mean count-per-interval agreement was 81.5%; and their counts of an interval were the same in 45.8% of intervals and within one disruption in 87.5%. Total-count and mean count-per-interval agreement both met the single-case standard of at least .80 for percentage agreement, collected in each phase and on 37.5% of the sessions (WWC, 2022).

Figure 3

Disruptions Counted by the Two Observers in the Same Intervals



Note. Each marker is one 10-minute interval both observers counted; filled markers are game sessions. Markers are spread slightly so that equal counts do not hide one another. The line marks perfect agreement.

Question 5: The Learners' Ratings

Of the 36 learners present on August 27, 26 (72.2%) said yes, they liked playing the game, 8 said a little, and 2 said no. The same numbers said the game helped them listen and work in Mathematics, and 28 (77.8%) wanted to play it again (Table 5).

Table 5

The Learners' Ratings of the Game

Statement	Yes	A little	No	Yes (%)
I liked playing the game.	26	8	2	72.2
The game helped me listen and work in Math.	26	8	2	72.2
I want to play the game again.	28	4	4	77.8

Note. 36 learners present on August 27, 2026.

Reflection

The game worked in our class, and it worked each time it was used. Disruptions fell by about two thirds when it began, came back when it stopped, and fell again when it returned, within days each time. The effect is in line with the average across single-case studies of the game, a Tau-U of .82 (Bowman-Perrott et al., 2016), which found the game most effective for disruptive and off-task behavior, the kind counted here.

Two things the study cannot show. It counted the class, not each learner, so it cannot say whether the game helped the learners who disrupted most or only the many who disrupted a little; and it ran for seven weeks, so it cannot say whether the effect lasts once the stickers lose their novelty. The second no-game phase not returning to the first baseline's level suggests some of the routine stayed, but A2 was too short to tell.

The adviser's own marks, about three in four of the disruptions the observers saw, were enough to run the game. A teacher does not need an observer to play it; the observers were needed to know whether it worked. The analysis also taught a caution: a *t* test on the intervals would have given the

right answer here for the wrong reason, and in a class with a smaller effect it could have declared a useless method effective.

Action Plan

Table 6

Action Plan for School Year 2026–2027

Action	Purpose	Persons responsible	Time frame	Resources	Indicator of success
Continue the game in Mathematics and add it to Araling Panlipunan	Keep the calmer periods and test the game in a second subject	Class adviser	Term 2	Stickers, class chart	Game played in both subjects
Count disruptions in one period a week, with the teacher's marks as the record	Keep watch on whether the effect lasts	Class adviser	Terms 2 and 3	Tally sheet	Weekly count recorded
Share the game and the findings in a learning action cell session	Let other Grade 4 teachers try it	Class adviser, Master Teacher	October 2026	Results brief	Session held; two teachers try the game
Raise the criterion from four marks to three once the class wins most days	Keep the game a challenge	Class adviser	When all teams win on four of five days	None	Criterion changed
Study the learners who disrupt most, with individual counts	Learn whether the game helps them	Research team	2027 BERF call	BERF proposal	Proposal submitted

References

Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good behavior game: Effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis, 2*(2), 119–124. https://doi.org/10.1901/jaba.1969.2-119

Bowman-Perrott, L., Burke, M. D., Zaini, S., Zhang, N., & Vannest, K. (2016). Promoting positive behavior using the Good Behavior Game: A meta-analysis of single-case research. *Journal of Positive Behavior Interventions, 18*(3), 180–190. https://doi.org/10.1177/1098300715592355

Brossart, D. F., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall's Tau and Tau-U for single-case experimental designs. *Cogent Psychology, 5*(1), Article 1518687. https://doi.org/10.1080/23311908.2018.1518687

Department of Education. (2012). *DepEd child protection policy* (DepEd Order No. 40, s. 2012).

Department of Education. (2016). *Adoption of the Basic Education Research Agenda* (DepEd Order No. 39, s. 2016).

Department of Education. (2017a). *Research management guidelines* (DepEd Order No. 16, s. 2017).

Department of Education. (2017b). *National adoption and implementation of the Philippine Professional Standards for Teachers* (DepEd Order No. 42, s. 2017).

Department of Education. (2021). *Amendment to DepEd Order No. 016, s. 2017 (Research management guidelines)* (DepEd Order No. 026, s. 2021).

Department of Education. (2024). *Amendment to DepEd Order No. 010, s. 2024 (Policy guidelines on the implementation of the MATATAG Curriculum)* (DepEd Order No. 012, s. 2024).

Department of Education. (2025). *Guidelines on the multi-year performance management and evaluation system for teachers from school years 2025–2026 to 2027–2028* (DepEd Memorandum No. 089, s. 2025).

Department of Education. (2026a). *Guidelines on ensuring a safe and motivating learning environment* (DepEd Order No. 006, s. 2026).

Department of Education. (2026b). *Foundational guidelines on artificial intelligence (AI) in basic education* (DepEd Order No. 003, s. 2026).

Department of Education. (2026c). *Guidelines on the implementation of the three-term school calendar in basic education* (DepEd Order No. 009, s. 2026).

Flower, A., McKenna, J. W., Bunuan, R. L., Muething, C. S., & Vega, R. (2014). Effects of the Good Behavior Game on challenging behaviors in school settings. *Review of Educational Research, 84*(4), 546–571. https://doi.org/10.3102/0034654314536781

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165–179. https://doi.org/10.1177/001440290507100203

Kellam, S. G., Brown, C. H., Poduska, J. M., Ialongo, N. S., Wang, W., Toyinbo, P., Petras, H., Ford, C., Windham, A., & Wilcox, H. C. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence, 95*(Suppl. 1), S5–S28. https://doi.org/10.1016/j.drugalcdep.2008.01.004

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703–708. https://doi.org/10.2307/1913610

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367. https://doi.org/10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*(2), 284–299. https://doi.org/10.1016/j.beth.2010.08.006

Proclamation No. 1006, s. 2025. (2025). https://lawphil.net/executive/proc/proc2025/proc_1006_2025.html

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 92–96). SciPy. https://doi.org/10.25080/Majora-92bf1922-011

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971–980. https://doi.org/10.3758/s13428-011-0111-y

Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification, 41*(4), 427–467. https://doi.org/10.1177/0145445516676750

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software, 3*(31), 1026. https://doi.org/10.21105/joss.01026

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education

Evaluation and Regional Assistance.

https://ies.ed.gov/ncee/wwc/Handbooks

Financial Report

The study was funded by a BERF grant of PhP 3,140.00 for school-based action research. Tranches apply only to proposals of PhP 5,000.00 and above, so the full amount was released on submission of the work plan and the signed memorandum of agreement (DepEd, 2017a); prepaid load is a communication expense the fund may pay for (DepEd, 2021). Python, its packages, and the school's computer cost nothing; the grant paid for no software, equipment, or statistician, which BERF does not cover.

Table 7

Statement of Expenses

Date (2026)	Particulars	Receipt No.	Category	Budget (PhP)	Actual (PhP)
July 2	Consent and assent forms and letters to parents (120 pages)	OR 0702	Reproduction	150.00	120.00
July 6	Observation sheets (48: 32 periods, the second observer's 12, and 4 for practice) and the learners' rating sheets (40)	OR 0706	Reproduction	100.00	90.00
July 7	Rules chart, team chart and markers	OR 1102	Supplies and materials	450.00	465.00
July 7	Reward stickers, two packs of 100	OR 1103	Supplies and materials	200.00	190.00
July to August	Prepaid load for coordination with the observers and parents, two months	OR 3410, 3462	Communication	600.00	600.00
September 21	Printing and binding of the final report (4 copies)	OR 0921	Printing and binding	1,400.00	1,435.00
September 25	Results brief for the teachers and the class's parents (120 copies)	OR 0925	Reproduction	240.00	240.00
	Total			3,140.00	3,140.00

Note. Receipt numbers are invented for this sample.

Appendices

Appendix A. Data Set and Notebook

The executive data set is the Excel workbook submitted with this report: the primary observer's and the second observer's counts for every interval; the session totals and the teacher's marks by team, with the teams that won, as formulas; each phase's level, spread, and trend; Welch's *t* test; NAP, Tau, and the baseline trend for each phase change, counted pair by pair; the observers' agreement, the ICC included, as formulas; the count models, Tau-U, and the simulation copied from Python; and the learners' ratings. The Jupyter notebook submitted with it reads the workbook and reproduces every result.

Appendix B. Letters, Consent, and Assent Forms

The school head's permission, the letter to parents, and the consent and assent forms are kept at the school and left out of the archived report.

Appendix C. The Game's Rules and Steps

Rules, posted in English and Filipino: (1) Stay in your seat unless the teacher allows you to leave it. (2) Raise your hand to speak. (3) Work quietly. (4) Keep your hands and feet to yourself. Steps: (1) Announce the game and review the rules in one minute. (2) Mark each broken rule under the team, naming the rule and not the learner. (3) At the end of the period, count each team's marks; every team with four or fewer wins. (4) Winning teams put a sticker on the class chart and line up first for recess. (5) On Friday, teams that won on at least three days choose Monday's energizer.

Appendix D. Observation Sheet and Definitions

1. The sheet has a header for the date, the observer, the session number, and the phase, and four boxes, one for each 10-minute interval, in which the observer makes a tally mark for each disruption. The definitions printed on the sheet: out of seat without permission; speaking without being called on, loudly enough for the observer to hear; noise or contact that interrupts others. An act counts once; the same learner's act counts again only after it has stopped for at least three seconds.

Appendix E. Declaration of AI Use

We, the researchers, declare that we used [name of the AI tool] only to check the grammar of the English text of this report, for the purpose of finalizing the research study. The prompt used was: "Check the grammar of this paragraph. Do not change its meaning or add content." We have reviewed, edited, and understood all the content, and we take full accountability for it. No names, counts, sheets, or other data from the study were entered into the AI tool.