Linear Number Board Games and the Number Sense of Kindergarten Learners

An Action Research

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

Conducted August to November 2025

Abstract

This action research tested whether a linear number board game, played in small groups during Circle Time 2, would improve the number sense of Kindergarten learners. One teacher's two Kindergarten sessions in a public school took part in School Year 2025–2026. A coin toss gave the morning session the number game for four weeks while the afternoon session played the same board with colors; for the next four weeks the roles switched. Each child had eight 12-minute sittings with each game. An assessor who did not know which session played the number game gave every child the Number Task one at a time in September, October, and November: naming numerals, choosing the bigger of two numbers, and giving a set of bottle caps (0 to 25), and placing numbers on a 0-to-10 line. A second scorer who sat in on 36 assessments agreed with the assessor on 98.8% of items, ICC(A,1) = .992 for the totals. With the September score held constant, the morning session scored 2.68 points higher in October, 95% CI [1.02, 4.34], $F(1, 41) = 10.58$, $p = .002$, Hedges' $g = 0.52$. When the afternoon session played the number game, it scored 3.20 points higher than the morning session in November, 95% CI [1.48, 4.93], $p < .001$, $g = 0.62$, and the morning session kept its gain. Number line errors fell in each session while it played. The game is recommended for Kindergarten classes, with a check of how long the gain lasts.

Keywords: kindergarten, number sense, number board game, waitlist design, one-to-one assessment

Acknowledgment

The researchers thank the school head, who approved the study, tossed the coin, and kept the Kindergarten schedule unchanged for the eleven weeks; the Grade 1 teacher who assessed every child three times without knowing which session played which game, and the Grade 2 teacher who scored beside the assessor; the school librarian, who lent the reading corner and witnessed each child's assent; the four Grade 1 learners and their parents, who allowed the assessors to practice the task with them; and the 46 Kindergarten children and their parents. The study was funded by the researchers.

Linear Number Board Games and the Number Sense of Kindergarten Learners

Context and Rationale

Filipino children start school behind in mathematics, and the gap shows early. In the Trends in International Mathematics and Science Study (TIMSS) 2019, Philippine Grade 4 learners averaged 297 in mathematics, the lowest of the 58 countries, and 19% reached the Low International Benchmark, which the international median country reached with 92% (Mullis et al., 2020, Exhibits 1.1 and 1.8). The early years matter for what follows: across six longitudinal studies, "early math skills have the greatest predictive power" for later achievement, "followed by reading and then attention skills" (Duncan et al., 2007, p. 1428).

Our Kindergarten learners showed the same need in September of School Year 2025–2026. On the pretest of this study, the 44 children analyzed averaged 13.52 and 12.67 of 25 points in the two sessions; 14 of them scored 5 or fewer of 10 when asked which of two numbers is bigger, which guessing alone gives. On a line from 0 to 10, the children placed 2 near the middle: the median estimate for 2 was 3.9 in the morning session and 4.2 in the afternoon session. The MATATAG Kindergarten curriculum asks learners to "match numerals to a set of concrete objects around me" and to "compare quantities using one to one correspondence to determine which has more, less, or equal," and expects them "to develop thinking skills, and insights into patterns of mathematics, concepts of numbers" through play (Department of Education [DepEd], 2023, Theme I, Competencies 16 and 17; Cognitive Development).

One activity has strong evidence for building this knowledge in young children from low-income homes: a board game whose squares are numbered 1 to 10 in a row. In a randomized study in 10 Head Start centers, preschoolers who played such a game in four sessions of 15 to 20 minutes over two weeks improved in numerical magnitude comparison, number line estimation, counting and numeral identification, and "the gains remained 9 weeks later"; classmates who played the same game with colored squares "did not improve on any measure" (Ramani & Siegler, 2008, pp. 375, 378). The game also worked as a small-group activity in Head Start classrooms, including when it was led by "a paraprofessional from the children's classroom" (Ramani et al., 2012, p. 661). A meta-analysis of 18 studies from preschool to Grade 2 found a smaller average effect on mathematics outcomes than the first trials, g = 0.21, 95% CI [0.08, 0.34], with wide variation between studies (Nelson et al., 2025).

Whether the game helps five-year-olds in a public Kindergarten class, taught by their own teacher in the Blocks of Time, is a question for our own classroom. It addresses the Teaching and Learning theme of the Basic Education Research Agenda (DepEd, 2016).

Innovation, Intervention, and Strategy

The innovation was a linear number board game played in small groups during Circle Time 2, the 40-minute block of the Kindergarten day for "engaging activities" that integrate mathematics with language, science, music, and the arts (DepEd, 2024, Table 4). The game took one of the block's activity stations; the other children worked at the block's other stations, as

on any day. Nothing else in the two sessions changed: the teacher, the themes and competencies of the quarter, the songs, stories, and counting routines of Meeting Time and Circle Time 1, and the time.

The Number Board Game

The game was the linear number board game of Ramani and Siegler (2008): a board with 10 squares in one row, numbered 1 to 10 from left to right, with "Start" before the first square, and a spinner with a "1" half and a "2" half. On each turn, a child spun, said the number spun, and moved the token one square at a time, saying the number on each square: a child on 3 who spun 2 said "4, 5" (Ramani & Siegler, 2008, p. 379). When a child named a square wrongly or counted "1, 2" instead of the squares' numbers, the teacher said the numbers and the child moved again, saying them, as the original protocol does (p. 379). The first child to reach 10 won, and the group played again.

Groups of four children, or three where a session's number did not divide by four, played with the teacher for 12 minutes, about four games, at the station. Each child sat at the game table twice a week for four weeks: eight sittings, about 96 minutes of play. Two or three groups played each day, so the station fit within the 40 minutes of Circle Time 2.

The Color Board Game

The afternoon session played the same board in the same block, with the same teacher, groups, and minutes, but its squares had colors and no numbers, and its spinner showed the board's five colors; a child moved to the next square of the color spun and named each color passed. Ramani and

Siegler (2008) used the same comparison, which gives both groups a game, an adult's attention, and turn-taking, so that a difference between them is due to the numbers on the board.

The Order of the Sessions

A coin toss after the pretest decided which session played the number game first: the morning session, from September 8 to October 3, 2025 (Phase 1). In Phase 2, from October 13 to November 14, 2025, with the mid-school year wellness break of October 27 to 31 between its two halves, the afternoon session played the number game and the morning session the color game. No game was played in the three assessment weeks.

Strategy and Evaluation Plan

The research coordinator observed six sittings of each game in each phase with a six-item checklist of the game's planned components (Appendix D). Three criteria for success were set in the proposal: (a) in each session's number-game phase, at least 80% of the children attended at least 7 of the 8 sittings, and the observed sittings showed at least 80% of the planned components; (b) scoring that agreed well enough to analyze: Cohen's kappa of at least .80 on each part of the task and an ICC(A,1) for the totals whose lower 95% confidence bound reached .90, both on the assessments scored by two people; and (c) in each phase, a difference in favor of the session playing the number game, significant at the .05 level with the score before the phase held constant.

Action Research Questions

Main question: Does playing a linear number board game improve the number sense of Kindergarten learners?

1. What were the Number Task scores and number line errors of the two sessions in September, October, and November?
2. How closely did the assessor and the second scorer agree?
3. After Phase 1, is there a significant difference between the two sessions' October scores when their September scores are held constant?
4. After Phase 2, is there a significant difference between the two sessions' November scores when their October scores are held constant, and did the morning session keep its gain?
5. Did the number line errors of each session fall in the weeks it played the number game?

Three null hypotheses were tested at the .05 level of significance:

$H_0$1: There is no significant difference between the adjusted October scores of the morning session and the afternoon session.

$H_0$2: There is no significant difference between the adjusted November scores of the afternoon session and the morning session.

$H_0$3: There is no significant difference between the two sessions' adjusted number line errors after either phase.

Action Research Methods

Research Design

The study used a switching-replications design, a quasi-experimental design in which two groups receive the same intervention one after the other,

so that each serves as the other's comparison in turn (Shadish et al., 2002, p. 146). It suits a Kindergarten teacher with two sessions: the session that waits is not denied the intervention, only given it later, and the effect can be seen twice, once in each session. Children could not be assigned one by one, because each Kindergarten session is taught as a class, so a coin toss assigned the sessions: the school head tossed it on September 5, 2025, after the last pretest, with the research coordinator as witness. With one session on each side, the toss decided which session went first but could not make them equal. The pretest showed how similar they were, the analysis of covariance adjusted for the difference that remained, and the second phase tested the game again with the roles switched, so that the session's time of day could not explain both results.

Participants and Other Sources of Data and Information

The participants were the children of the first researcher's two Kindergarten sessions in a public elementary school, School Year 2025–2026: a morning session that met from 6:45 to 10:15 a.m. and an afternoon session that met from 12:15 to 3:45 p.m., each following DepEd's sample Blocks of Time for 180 minutes of class (DepEd, 2024). All 49 enrolled children were invited; no sampling was done. Table 1 shows how many were analyzed.

Table 1

Children Enrolled, Excluded, and Analyzed, by Session

Children	Morning session	Afternoon session	Total
Enrolled	25	24	49
No parental consent	1	2	3
Did not want to do the task in September (two tries)	0	1	1

Children	Morning session	Afternoon session	Total
Transferred to another school before the October assessment	1	0	1
Analyzed in Phase 1 (September and October)	23	21	44
Absent for the November assessment and its make-up day	1	0	1
Analyzed in Phase 2 (October and November)	22	21	43

Note. Every child played the games with their session as a class activity; consent and assent decided only whose answers were recorded. Nothing was recorded about a child without a parent's consent.

The children analyzed were 15 girls and 8 boys in the morning session and 11 girls and 10 boys in the afternoon session. On September 1, 2025, they were 56 to 73 months old (morning M = 63.3, SD = 3.9; afternoon M = 62.3, SD = 3.5); 9 were still four years old, as the age rule of School Year 2025–2026 allows: five years old by October 31, or by December 31 with a year in a child development center or the Philippine Early Childhood Development (ECD) Checklist (DepEd, 2025a). Two children, one in each session, were admitted on a certificate from a child development center. Before Kindergarten, 14 children of the morning session and 12 of the afternoon session had attended a child development center. The medium of instruction was Filipino, with the regional language as an auxiliary medium (DepEd, 2025b).

Sample Size

Because whole sessions took part, the number of children was fixed. A sensitivity analysis for an analysis of covariance with one covariate, as G*Power computes it (Faul et al., 2007), showed that the 46 children with consent give a power of .80 to detect an effect of f = 0.42 at α = .05, a difference

of 0.85 standard deviations of the adjusted score. With a correlation of .70 between the pretest and the posttest, assumed in the proposal, that is d = 0.60 in standard deviations of the posttest, against d = 0.84 for a comparison of posttests without the pretest. The pooled within-session correlation found in Phase 1, .85, gave the 44 children analyzed a detectable effect of d = 0.46.

Other Sources

A Grade 1 teacher of the school, who had taught Kindergarten before and taught neither session, gave every assessment. A Grade 2 teacher scored beside the assessor on the assessments drawn for a second score. The school's research coordinator kept the list that linked codes to children, drew the second-score assessments, and observed the game sittings. The school librarian witnessed each child's assent.

Data Gathering Methods

The Number Task

The Number Task was given one child at a time in about 8 to 10 minutes. Its items follow the tasks of Ramani and Siegler (2008, p. 379), shortened for a Kindergarten class, and the two curriculum competencies above (DepEd, 2023):

1. Part A, Naming Numerals (10 points): the numerals 1 to 10 on cards, in a fixed mixed order; "Anong numero ito?" ("What number is this?").
2. Part B, Which Is More? (10 points): ten pairs of numerals from 1 to 9 on cards, after two practice pairs with feedback; "Alin ang mas malaking numero?" ("Which is the bigger number?"). The bigger number was on

the left in five pairs and on the right in five, and guessing gives about 5.

3. Part C, Give Me (5 points): from a plate of 10 bottle caps, the child put 4, 2, 7, 5, and 9 caps into a cup, one request at a time; a point for exactly the number asked.
4. Part D, the Number Line (measured, not scored right or wrong): six sheets, each with a line 200 mm long marked 0 and 10 and a number above it (2, 7, 4, 9, 6, 3); "Kung dito ang 0 at dito ang 10, saan ang ___?" ("If 0 goes here and 10 goes here, where does ___ go?").

The task was given in the weeks of September 1 to 5, 2025, October 6 to 10, 2025, and November 17 to 21, 2025. Parts A to C gave the Number Task score, 0 to 25, the primary outcome, named in the proposal. For the number line, each mark was measured in millimeters from the 0 end of the 200 mm line, the estimate was the measurement divided by 20, and the error was the percent absolute error, the distance between the estimate and the number divided by the length of the scale (Siegler & Booth, 2004, p. 432); a child who marked 5 at the point for 9 has an error of 40% (Ramani & Siegler, 2008, p. 382). The child's number line error was the mean over the six numbers. The number line error was the secondary outcome. The same items were given at the three assessments, without feedback, as in Ramani and Siegler (2008). The 25 scored items were internally consistent, KR-20 (Kuder & Richardson, 1937) = .80 in September, .84 in October, and .84 in November.

Giving the Task Fairly and Safely

The assessor gave the task in the reading corner of the school library, next to the Kindergarten rooms, which both sessions visited each week. The assessor sat beside the child at a low table with the door open and the librarian in the room. The assessor took children during Arrival Time and Indoor/Outdoor Play, never during Circle Time 2, so that no child missed a game sitting.

Before each assessment, the assessor asked the child in Filipino whether they wanted to play a number game and said that they could stop at any time; the regional language was used when a child needed it. A child who said no, shook their head, or would not come was not assessed that day, and was asked once more on another day of the week. A child who asked to stop or turned away twice in a row was asked once whether to go on or rest, and a child who wanted to stop finished the remaining parts on another day of the same week. A child who cried was comforted and taken back to the teacher at once. Answers in any language counted, and in Parts B and D a pointed answer counted. The assessor gave no sign of right or wrong, said the same words of thanks to every child after each part, and gave every child a sticker at the end.

Blinding and the Second Score

The assessor and the second scorer were not told the result of the coin toss or which session played which game, and the scoring sheets carried only a three-digit code and the date. Both sessions played a board game with the teacher, so a child's mention of "the game" did not reveal the session's game.

Blinding matters even with a second scorer: "good interobserver agreement does not prevent observer bias" (Hróbjartsson et al., 2013, p. E209), and in clinical trials with subjective measurement scale outcomes, nonblinded assessors exaggerated the pooled effect size by 68% (p. E201). After the October and November assessments, the assessor and the second scorer were each asked which session had played the number game in the weeks before.

The research coordinator drew six children of each session for a second score at each assessment, 36 assessments in all, from the children assessed on the first try. The second scorer sat beside the assessor, scored every item on a separate sheet without comparing, and measured each number line mark separately. The analysis used the assessor's scores. Agreement on the items was computed for each part as the percentage of items scored alike and as Cohen's (1960) kappa, with a 95% confidence interval from 2,000 bootstrap resamples of the assessments. Agreement on the totals was computed as the intraclass correlation from a two-way model for absolute agreement and a single scorer, ICC(A,1) (McGraw & Wong, 1996), judged against Koo and Li's (2016) guidelines: below .50 poor, .50 to .75 moderate, .75 to .90 good, and above .90 excellent.

Training the Assessors

The assessor and the second scorer were trained together for three hours on August 28, 2025: the script, the assent and stopping rules, the scoring rules, and the measuring of marks. They practiced on each other, then with four Grade 1 learners of the assessor's class whose parents agreed; those children's answers were not kept.

Implementation Records

The teacher kept a log of each sitting: the date, the group, the children present, and the games played. The research coordinator observed six sittings of each game in each phase with a checklist of six components: for the number game, the 1-to-10 board and the 1-and-2 spinner; each child saying the numbers on the squares while moving; a child who miscounted counting again from the square; games played to 10; about 12 minutes; and four children or fewer.

Storage

The scoring sheets, the number line sheets, and the consent forms were kept in a locked cabinet in the school. The list linking codes to children stayed with the research coordinator, apart from the data. The data set holds codes and each child's age in months; birth dates stayed in the school's own records. No photograph or recording of a child was made.

Ethical Considerations

The school head approved the study in writing on August 15, 2025. At the Kindergarten parents' meeting on August 20, 2025, the teacher explained the study in Filipino and English, showed the cards and the number line, and gave each parent a letter on the purpose, the games, the three short assessments, the right to refuse or withdraw without any effect on the child's schooling, and how the child's answers would be kept; 46 of 49 parents signed the consent form by August 27, 2025. Each child's assent was asked before every assessment in words a five-year-old understands, and a child's no, in words or in behavior, was respected: one child did not want to do the task on

either September try and was not asked again. DepEd's quality checklist for completed research asks for the "assent from minor participants and consent from their parents and/or legal guardians" (DepEd, 2022, Annex 2, Item 13), and the national ethical guidelines ask that a minor's assent be "properly documented and witnessed by a third party who has no unresolved conflict of interest" (Philippine Health Research Ethics Board Ad Hoc Committee for Updating the National Ethical Guidelines [PHREB Ad Hoc Committee], 2022, p. 93), which the librarian's initials on each sheet did. A child's dissent, in words or in behavior, is to be "upheld and respected" (p. 142). A child's code, answers, and age are personal information (Republic Act No. 10173, 2012, Sec. 3(g)), so the data were coded and the list of names kept apart. Because the afternoon session could benefit from the number game, it played the game in Phase 2, and the boards stayed in both rooms' mathematics corners after the study. The task's results were not used for the Kindergarten Progress Report or the ECD Checklist. The study followed the Research Management Guidelines (DepEd, 2017). The researchers' use of an artificial intelligence (AI) tool is declared in Appendix E (DepEd, 2026).

Data Analysis

For the first question, the scores and errors were summarized by session and assessment, and the sessions' September scores were compared with Welch's t test and as a standardized difference. For the second, the agreement statistics above were computed, and KR-20 gave the standard error of measurement, SEM = $SD \times \sqrt{(1 - \text{KR-20})}$, from which a child's change counted as reliable when it reached $1.96 \times \sqrt{2} \times \text{SEM}$, the reliable change

index (Jacobson & Truax, 1991, p. 14). Jacobson and Truax used test–retest reliability; KR-20 was used because the children learned between assessments.

For the third and fourth questions, an analysis of covariance (ANCOVA) compared the sessions on the score after each phase with the score before it as the covariate: the October score on the September score and the session for Phase 1, and the November score on the October score and the session for Phase 2. The adjusted difference is reported with its 95% confidence interval, F, p, and partial η^2, and as Hedges' g, the adjusted difference divided by the pooled standard deviation of the unadjusted outcome with the small-sample correction (What Works Clearinghouse [WWC], 2022). The WWC judges two groups' pretests equivalent when they differ by 0.05 standard deviations or less, and requires a difference between 0.05 and 0.25 to be adjusted for in the analysis (WWC, 2022, p. 53). The ANCOVA's assumptions were checked: equal slopes of posttest on pretest in the two sessions (the session × pretest interaction), normal residuals (Shapiro & Wilk, 1965), and equal residual variances (Levene's test, median-centered). Four checks followed for Phase 1: the ANCOVA with age in months as a second covariate, with heteroskedasticity-consistent (HC3) standard errors, and without the children who attended fewer than 6 of the 8 sittings, and a Mann–Whitney U test of the gains, which uses only their ranks. For the morning session, the change from September to November was tested with a paired t test, and the two sessions' November scores were compared with the September score as the

covariate. For the fifth question, the same ANCOVAs were run on the number line error.

The analyses were run in Python 3.12 with pandas, statsmodels (Seabold & Perktold, 2010), SciPy, and scikit-learn, free software; the Excel data set computes as a formula every statistic Excel has a function for and labels Python's values for the rest (the bootstrap intervals, the interval of the ICC, Shapiro–Wilk, the HC3 standard errors, and the power analysis), and the notebook reproduces every result from the data (Appendix A).

Discussion of Results and Reflection

Implementation

In Phase 1, the morning session's children attended a mean of 7.3 of the 8 number-game sittings (range 5–8), and 19 of 23 (82.6%) attended at least 7; the afternoon session's attended 7.5 color-game sittings. In Phase 2, the afternoon session's children attended 7.7 number-game sittings (range 6 –8), and 20 of 21 (95.2%) attended at least 7. In the 12 observed sittings of the number game, 70 of the 72 planned components were seen (97.2%), and none of the 24 observed sittings of either game showed fewer than 5 of its 6. The first criterion for success was met.

Question 1: The Scores at the Three Assessments

The sessions began close together, and each rose most in the four weeks it played the number game (Table 2 and Figure 1). In September, the morning session's mean was 0.86 points higher, $t(41.84) = 0.58$, $p = .568$, a difference of 0.17 standard deviations, which the ANCOVA adjusted for. In Phase 1, the morning session gained 3.61 points and the afternoon session, waiting, 1.05.

In Phase 2, the afternoon session gained 3.29 points, and the morning session's mean changed by −0.41.

Table 2

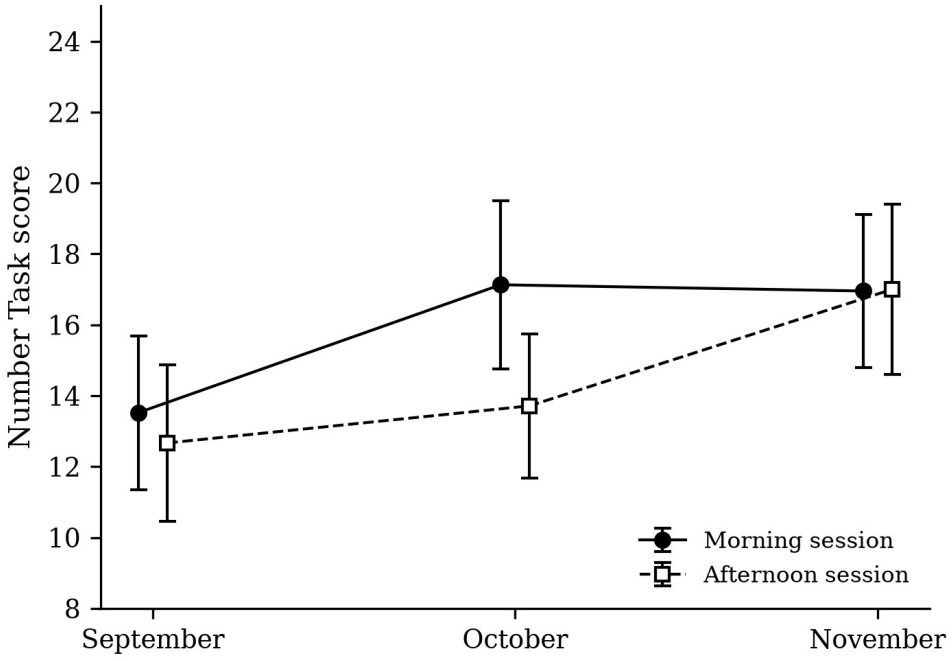
Number Task Scores and Number Line Errors by Session and Assessment

Score	Morning, Sept. *M* (*SD*)	Morning, Oct. *M* (*SD*)	Morning, Nov. *M* (*SD*)	Afternoon, Sept. *M* (*SD*)	Afternoon, Oct. *M* (*SD*)	Afternoon, Nov. *M* (*SD*)
Number Task (0–25)	13.52 (5.01)	17.13 (5.49)	16.95 (4.85)	12.67 (4.85)	13.71 (4.47)	17.00 (5.26)
A. Naming numerals (0–10)	5.04 (2.51)	6.57 (2.74)	6.45 (2.36)	4.67 (2.37)	4.81 (2.68)	6.38 (2.85)
B. Which is more (0–10)	6.39 (2.06)	7.70 (1.89)	7.77 (1.88)	6.24 (2.02)	6.57 (1.91)	7.71 (1.71)
C. Give me (0–5)	2.09 (1.41)	2.87 (1.55)	2.73 (1.35)	1.76 (1.26)	2.33 (1.39)	2.90 (1.48)
Number line error (%)	16.2 (3.2)	9.6 (3.1)	10.7 (5.2)	16.2 (4.4)	15.5 (4.9)	10.7 (3.3)

Note. September and October: morning $n = 23$, afternoon $n = 21$. November: morning $n = 22$, afternoon $n = 21$. The morning session played the number game between September and October, the afternoon session between October and November. Number line error: the mean percent absolute error over six numbers; lower is better.

Figure 1

Mean Number Task Scores of the Two Sessions at the Three Assessments



Note. Error bars are 95% confidence intervals. Phase 1 (September to October): the morning session played the number game. Phase 2 (October to November): the afternoon session played it.

In Phase 1, most of the difference between the sessions' gains came from naming numerals (1.52 against 0.14 points) and comparing numbers (1.30 against 0.33); on giving a set of bottle caps, the gains were closer (0.78 against 0.57). The parts were described, not tested one by one, as the proposal set. Few children reached the top of the task: 4 of the 43 children scored 24 or 25 in November.

Question 2: The Scorers' Agreement

The assessor and the second scorer agreed on 98.8% of the 900 items they both scored and disagreed on 11 (Table 3). Kappa was .96 for naming numerals, .99 for comparing numbers, and .98 for giving a set, and their totals agreed almost perfectly, ICC(A,1) = .992, 95% CI [.985, .996]: the same

total in 26 of 36 assessments and never more than 2 points apart. Most disagreements were on numerals said softly or in a mix of languages. The scorers' number line measurements differed by 0.5 mm on average and never by more than 2 mm. The second criterion for success was met. After the October assessment, the assessor guessed that the afternoon session had played the number game, which was wrong, and the second scorer could not tell; after November, the assessor could not tell and the second scorer guessed the morning session, also wrong.

Table 3

Agreement Between the Assessor and the Second Scorer

Part	Items scored by both	Scored alike (%)	Cohen's κ [95% CI]
A. Naming numerals	360	98.1	.96 [.93, .98]
B. Which is more	360	99.4	.99 [.97, 1.00]
C. Give me	180	98.9	.98 [.94, 1.00]
All 25 items	900	98.8	.97

Note. 36 assessments, 6 in each session at each of the three assessments, scored by both. The confidence intervals are from 2,000 bootstrap resamples of the assessments. Totals: ICC(A,1) = .992, 95% CI [.985, .996]. Number line: 216 marks measured by both, a mean difference of 0.5 mm.

The task's KR-20 of .80 in September gives a standard error of measurement of 2.17 points, with the September scores' standard deviation of 4.88. A single child's change therefore had to reach 6.01 points, that is, 7 points on a task scored in whole points, before it could be told apart from measurement error (Jacobson & Truax, 1991, p. 14). In Phase 1, 3 of the 23 children of the morning session gained that much, and 1 of the 21 of the afternoon session; in Phase 2, 1 of the 21 children of the afternoon session did. The task measured a session's change well and most single children's change only roughly.

Question 3: The Difference After Phase 1

With the September score held constant, the morning session's October score was 2.68 points higher than the afternoon session's, 95% CI [1.02, 4.34], $F(1, 41) = 10.58$, $p = .002$, partial $\eta^2 = .21$, Hedges' $g = 0.52$ (Table 4 and Figure 2). The adjusted means were 16.78 and 14.10 of 25. $H_0 1$ is rejected.

Table 4

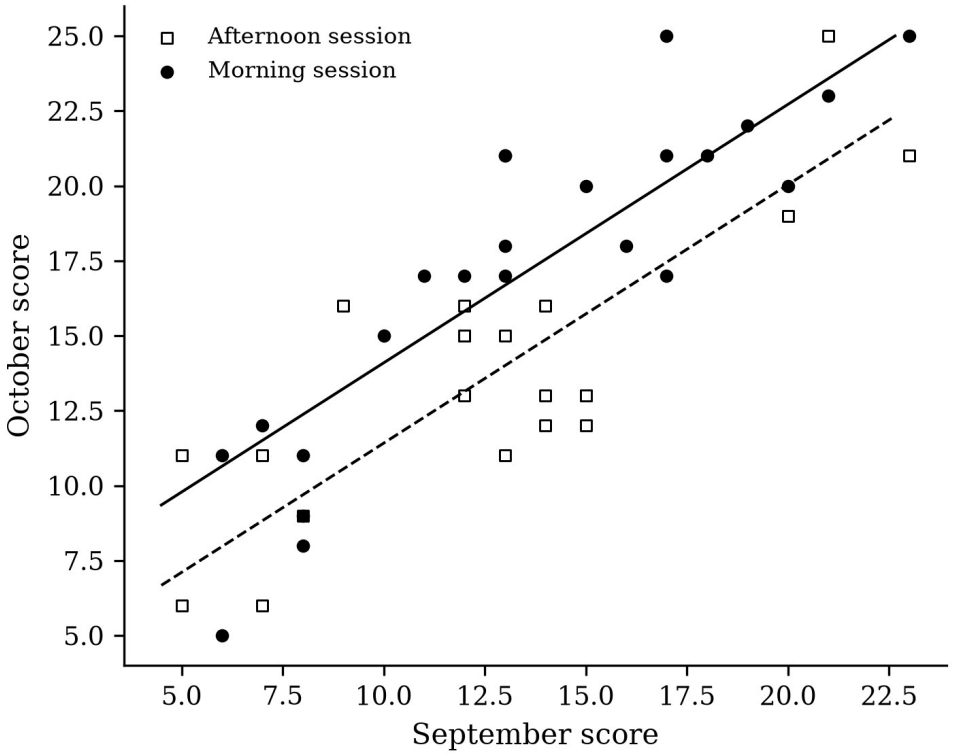
Analysis of Covariance of the October Score

Term	*b*	*SE*	*t*	*p*	95% CI
Intercept	2.79	1.23			[0.31, 5.28]
September score	0.86	0.09	10.14	< .001	[0.69, 1.03]
Session (morning vs. afternoon)	2.68	0.82	3.25	.002	[1.02, 4.34]

Note. $N = 44$; $df = 41$. $R^2 = .75$. Ordinary least squares. Adjusted means at the mean September score, 13.11: morning session 16.78, afternoon session 14.10.

Figure 2

October Score Against September Score, by Session, With the Fitted ANCOVA Lines



Note. Filled circles and the solid line: the morning session (number game). Open squares and the dashed line: the afternoon session (color game).

The ANCOVA's assumptions held. The slopes of the October score on the September score did not differ significantly between the sessions, $F(1, 40) = 1.68$, $p = .202$; the residuals showed no departure from normality, Shapiro–Wilk $W = .98$, $p = .684$; and their variances did not differ, Levene's $F(1, 42) = 0.40$, $p = .533$.

The checks agreed with the main result (Table 5). The difference was close to the same with age as a second covariate, with HC3 standard errors, and without the 3 children who attended fewer than 6 sittings. The Mann–Whitney test of the gains, which assumes only that they can be ranked, gave

the same conclusion, $U = 122.5$, $p = .005$, rank-biserial $r = .49$; the median gain was 4.0 points in the morning session and 1.0 in the afternoon session. The third criterion for success was met for Phase 1.

Table 5

The Phase 1 Difference Under Four Checks

Analysis	*n*	Difference (points)	95% CI	*p*
Main analysis	44	2.68	[1.02, 4.34]	.002
With age in months as a covariate	44	2.49	[0.84, 4.13]	.004
HC3 standard errors	44	2.68	[0.93, 4.43]	.004
Children with 6 or more of the 8 sittings	41	2.60	[0.82, 4.39]	.005

Note. Each row is the ANCOVA of the October score on the September score and the session. The Mann–Whitney test of the gains: $U = 122.5$, $p = .005$.

Question 4: The Difference After Phase 2

With the October score held constant, the afternoon session's November score was 3.20 points higher than the morning session's, 95% CI [1.48, 4.93], $F(1, 40) = 14.09$, $p < .001$, partial $\eta^2 = .26$, Hedges' $g = 0.62$; the adjusted means were 18.61 and 15.41. $H_0 2$ is rejected, and the third criterion for success was met for Phase 2. The slopes did not differ significantly, $F(1, 39) = 2.59$, $p = .115$. The residuals departed slightly from normality, $W = .95$, $p = .045$, and the Mann–Whitney test of the gains, which does not assume it, agreed, $U = 78.0$, $p < .001$, $r = .66$.

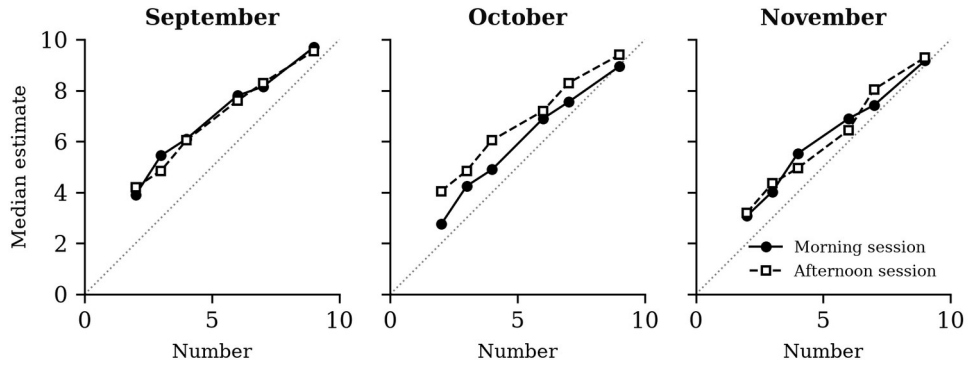
The morning session kept its gain six to seven weeks after its last sitting: its November score was 3.14 points above its September score, 95% CI [1.72, 4.55], $t(21) = 4.60$, $p < .001$. After both phases, the two sessions stood at about the same level: with the September score held constant, the November scores differed by −0.98 points, 95% CI [−2.97, 1.02], $p = .328$.

Question 5: The Number Line

Each session's number line error fell in the weeks it played the number game (Figure 3). After Phase 1, with the September error held constant, the morning session's error differed from the afternoon session's by −5.90 percentage points, 95% CI [−7.92, −3.87], $p < .001$; after Phase 2, with the October error held constant, the afternoon session's differed from the morning session's by −3.33 points, 95% CI [−6.37, −0.29], $p = .032$. $H_0 3$ is rejected for both phases. The children who had played the game placed small numbers closer to their places: the morning session's median estimate for 2 moved from 3.9 in September to 2.8 in October, while the afternoon session's was 4.2 in September and still 4.1 in October, before it played.

Figure 3

Median Number Line Estimates of the Two Sessions at the Three Assessments



Note. The dotted line marks exact estimates. Filled circles: the morning session. Open squares: the afternoon session.

Reflection

What We Found

The number board game raised the Number Task scores of each session in the four weeks it played, by about three points of 25 over the session that played the color game: 2.68 points in Phase 1 and 3.20 in Phase 2. The two results point the same way, from different sessions at different times of day, and the morning session's gain held six to seven weeks after its last sitting. The children also placed numbers on a line more accurately after playing.

How It Compares With the Research

Our effect, $g = 0.52$, is larger than the average of the meta-analysis of linear number board games, $g = 0.21$, but far less certain: with 44 children, its 95% confidence interval runs from 0.20 to 0.85, down to that average (Nelson et al., 2025). The Number Task also measured the skills the game practices, and the meta-analysis found larger effects on such measures than on broader ones, $g = 0.24$ against 0.13 (Nelson et al., 2025). Kindergarten studies in that meta-analysis averaged $g = 0.23$, with an interval that included zero, from only five studies (Nelson et al., 2025).

What Else Could Explain the Difference

The two sessions differ in more than the game: their times of day, the children's routines at home, and the classroom's heat in the afternoon. The switching design answers the largest of these doubts, because the afternoon session caught up when it played the game, so a morning advantage alone cannot explain both results. The teacher knew which game each session played and may have played the number game with more energy; the

observations checked the games' components but cannot measure enthusiasm. Children who are related or neighbors may have talked about the game across the sessions, which would make the difference smaller, not larger. The analysis treats the children as independent, though each session learned together; with one session on each side, the study cannot separate the game from everything else that differed between the two sessions in a given phase, and the replication in Phase 2 is the study's answer to that.

What a Waiting Session Shows

The afternoon session's score rose by 1.05 points in the five weeks it waited, 95% CI [−0.28, 2.38], from its own lessons, its routines, and growing older, the change a one-group design cannot separate from an intervention's effect (Shadish et al., 2002, p. 108). Had the study followed only the morning session from September to October, it would have credited all of that session's 3.61-point gain to the game. The comparison session shows how much a Kindergarten class grows in five weeks without it.

Limits

The task covered numbers to 10 and the skills the game practices; it does not show whether the game helps with counting beyond 10, with addition, or with Grade 1 mathematics. The study lasted eleven weeks, so it cannot show whether the gain lasts to the end of Kindergarten. A single child's change had to reach 7 points to exceed measurement error, so the results describe the sessions, not each child.

What We Would Change

We would give the Number Task again at the end of the school year, add a few items with numbers beyond 10, and, with another Kindergarten teacher's sessions, assign more sessions to each order.

Conclusion

The linear number board game, played in small groups for eight 12-minute sittings in Circle Time 2, improved the number sense of Kindergarten learners in each session that played it, with the color game as the comparison and an assessor who did not know which session played which. The boards remain in both Kindergarten rooms, and the school's Kindergarten teachers will try the game in the coming quarter.

Action Plan

Table 6

Action Plan

Objective	Activities	Persons involved	Timeline	Resources (PhP)	Success indicator
Keep the game in both sessions	The number board in each room's mathematics corner, played in Circle Time 2 once a week	First researcher	From November 24, 2025	0	Played weekly in both sessions
Find out whether the gain lasts	The Number Task again at the end of the school year, with the same blind assessor	Researchers, assessor	March 2026	250 (printing)	Both sessions above their September means
Share the game with the other Kindergarten teachers	A Learning Action Cell session: the game, the color comparison, and the one-to-one task	First researcher, Kindergarten teachers, school head	March 2026	400 (boards and spinners)	Each Kindergarten teacher plays the game with one group
Send the game home	A paper board and spinner for each child, with a one-page guide for parents in Filipino	First researcher, parents	March 2026	690 (printing)	Boards given to all children
Test the game across more classes	Proposal for a study of four Kindergarten sessions in two schools, the order of the sessions drawn by lot	Researchers, school research committee	Next call for proposals	Up to 150,000 (BERF, division level)	Proposal submitted and approved

References

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Department of Education. (2016). *Adoption of the Basic Education Research Agenda* (DepEd Order No. 39, s. 2016). https://www.deped.gov.ph/wp-content/uploads/2016/06/DO_s2016_039.pdf

Department of Education. (2017). *Research management guidelines* (DepEd Order No. 16, s. 2017). https://www.deped.gov.ph/wp-content/uploads/2017/03/DO_s2017_016.pdf

Department of Education. (2022). *Adopting tools to improve quality management of completed research at the Department of Education* (DepEd Memorandum No. 028, s. 2022). https://www.deped.gov.ph/2022/04/05/april-5-2022-dm-028-s-2022-adopting-tools-to-improve-quality-management-of-completed-research-at-the-department-of-education/

Department of Education. (2023). *MATATAG K to 10 curriculum of the K to 12 program: Kindergarten*. https://www.deped.gov.ph/wp-content/uploads/6-FINAL-MATATAG-Kindergarten-CG-2023.pdf

Department of Education. (2024). *Policy guidelines on the implementation of the MATATAG curriculum* (DepEd Order No. 010, s. 2024). https://www.deped.gov.ph/wp-content/uploads/DO_s2024_010.pdf

Department of Education. (2025a). *Amendment to DepEd Order No. 47, s. 2016 (Omnibus Policy on Kindergarten Education) as amended by DepEd Order No. 020, s. 2018* (DepEd Order No. 015, s. 2025). https://www.deped.gov.ph/2025/06/04/june-4-2025-do-015-s-2025-amendment-to-deped-order-no-47-s-2016-omnibus-policy-on-kindergarten-education-as-amended-by-deped-order-no-020-s-2018/

Department of Education. (2025b). *Policy on the medium of instruction for Kindergarten to Grade 3 effective school year 2025–2026* (DepEd Order No. 020, s. 2025). https://www.deped.gov.ph/2025/07/03/july-3-2025-do-020-s-2025-policy-on-the-medium-of-instruction-for-kindergarten-to-grade-3-effective-school-year-2025-2026/

Department of Education. (2026). *Foundational guidelines on artificial intelligence in basic education* (DepEd Order No. 003, s. 2026). https://www.deped.gov.ph/2026/02/20/february-20-2026-do-003-s-2026-foundational-guidelines-on-artificial-intelligence-ai-in-basic-education/

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2013). Observer bias in randomized clinical trials with measurement scale outcomes: A systematic review of trials with both blinded and nonblinded assessors. *CMAJ, 185*(4), E201–E211. https://doi.org/10.1503/cmaj.120744

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

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

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2019/international-results/

Nelson, G., Boedeker, P., Devlin, B. L., Lesner, T., Drake, K. R., Sutherland, M., Wilke, E. J., Cook, M., & Schmitt, S. A. (2025). Investigating main effects and moderators of linear number board games: A meta-analytic review. *Review of Educational Research*. Advance online publication. https://doi.org/10.3102/00346543251383552

Philippine Health Research Ethics Board Ad Hoc Committee for Updating the National Ethical Guidelines. (2022). *National ethical guidelines for research involving human participants 2022*. Philippine Council for Health Research and Development. https://www.pchrd.dost.gov.ph/wp-content/uploads/2023/05/2022-NEGRIHP_Official-Gazatte_Ver-5-1.pdf

Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development, 79*(2), 375–394. https://doi.org/10.1111/j.1467-8624.2007.01131.x

Ramani, G. B., Siegler, R. S., & Hitti, A. (2012). Taking it to the classroom: Number board games as a small group learning activity. *Journal of Educational Psychology, 104*(3), 661–672. https://doi.org/10.1037/a0028995

Republic Act No. 10173. (2012). *Data Privacy Act of 2012*. https://lawphil.net/statutes/repacts/ra2012/ra_10173_2012.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). https://doi.org/10.25080/Majora-92bf1922-011

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

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75*(2), 428–444. https://doi.org/10.1111/j.1467-8624.2004.00684.x

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0* (WWC 2022008). U.S. Department of Education, Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf

Financial Report

Table 7

Statement of Expenses

Date	Particulars	Receipt no.	Amount (PhP)	Category
August 19, 2025	Consent letters in Filipino and English, 55 copies	OR 2214	110	Reproduction
August 26, 2025	Four number boards and four color boards, printed in color on A3 and laminated, with spinners	OR 2261	680	Supplies and materials
August 27, 2025	Numeral cards and pair cards, printed and laminated, two sets	OR 2263	360	Supplies and materials
August 29, 2025	Number line sheets (6 for each child at each assessment) and scoring sheets, 590 pages	OR 2290	590	Reproduction
August 29, 2025	Stickers, three sheets for each child	OR 0457	240	Supplies and materials
February 27, 2026	Printing and binding of the report, 3 copies	OR 2702	780	Reproduction
Total			**2,760**	

Note. Fund source: the researchers' own funds. No school funds or grants were used. Receipt numbers are invented for this sample. Bottle caps, plates, and cups came from the classroom.

Appendices

Appendix A. Data Set and Notebook

The Excel data set (every child's answers at the three assessments, the second scorer's scores, the sittings attended, every statistic Excel can compute as a formula, and Python's values for the rest, labeled) and the Jupyter notebook that reproduces every result from the data.

Appendix B. Consent Letter and Assent Script

Kept with the school's copy, with the signed forms, and left out of the archived report.

Appendix C. The Number Task

The assessor's guide (the setting, the assent and stopping rules, the script, the scoring rules), the scoring sheet, the numeral and pair cards, and the number line sheets.

Appendix D. The Board Games and the Fidelity Checklists

The number board, the color board, the two spinners, how to play, and the observer's six-item checklist for each game.

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, codes, scores, or other personal information were entered into the AI tool (adapted from Annex A of DepEd Order No. 003, s. 2026).