Journal of Innovation in Psychology, Education and Didactics
ISSN: 2247-4579, E-ISSN: 2392-7127
http://www.jiped.ub.ro/
Covered in: EBSCO, CEEOL, ProQuest, ERIH PLUS, DOAJ, Scipio,
International Innovative Journal Impact Factor, CiteFactor,
EuroPub database, Open Academic Journals Index, ResearchBib,
Universal Impact Factor
2025, Volume 29, Issue 1, pages: 119-132, doi:10.29081/JIPED.2025.29.1.09

Using Stepped Wedge in Educational Research: Methodological Considerations

Oana ONCIU [1]*, Flavia PRISACARU [2]

Received: 03 June 2025/ Accepted: 29 August 2025/ Published: 02 September 2025

Abstract

The stepped-wedge clustered randomized design (SW-CRT) is an increasingly recognized research design, as it is particularly valuable when ethical or logistical constraints preclude conventional randomization. However, due to several limitations, it is underutilized in educational research. The study involved 80 children (ages 5 and 9), balanced by gender, distributed across 12 educational clusters. Data were collected across 7 waves during a 6-week implementation period. The classroom-based intervention comprised structured, age-appropriate metacognitive activities designed to enhance self-awareness of learning processes. Mixed-effects linear models accounted for repeated measures and hierarchical data. Findings indicated statistically robust immediate gains and sustained improvement in metacognitive knowledge across time. No moderation effects emerged for age or gender, while baseline-dependent effects suggested greater responsiveness among initially low-performing children. These outcomes highlight the potential of targeted interventions to reduce early metacognitive disparities. Methodologically, the SW-CRT offered a compelling balance between causal inference and ecological validity, though limitations—such as coordination demands, contamination risk, and confounding with time—warrant careful design calibration. Based on implementation insights, the paper proposes specific design adaptations and reporting standards to advance the rigor and applicability of SW-CRT in high-stakes educational contexts.

Keywords: assessment; early-childhood; metacognition; metacognitive development; stepped-wedge

How to cite: Onciu, O., & Prisacaru, F. (2025). Using Stepped Wedge in Educational Research: Methodological Considerations. *Journal of Innovation in Psychology, Education and Didactics*, *29*(1), 119-132. doi:10.29081/JIPED.2025.29.1.09

[1] PhD Student, "Alexandru Ioan Cuza" University of Iași, Romania, E-mail: onciu.oana@uaic.ro
[2] Master Student, "Alexandru Ioan Cuza" University of Iași, Romania. E-mail: flavia.prisacaru@student.uaic.ro
* Corresponding author

1. Introduction

The stepped-wedge cluster randomized trial (SW-CRT) is an increasingly adopted experimental design, particularly suited for evaluating complex interventions in real-world contexts. In this design, intact groups (e.g., schools, classrooms, institutions) are randomized to cross over from control to intervention at different, pre-specified time points, until all groups have received the intervention. Initially, all clusters function as controls; then, at each successive step, one or more clusters are randomly assigned to begin the intervention. Data are collected at each wave across all groups, allowing for both within-cluster and between-cluster comparisons over time. This structure offers a powerful alternative to the traditional parallel-group randomized controlled trial (RCT), by enabling full treatment coverage while preserving causal inference.

The empirical beginnings of the stepwise cluster-randomized trial (SW-CRT) design occurred in the field of public health research (Hooper, 2021; Hughes et al., 2024; Hussey, 2007). In such fields, ethical and logistical constraints often preclude traditional randomized controlled trials. Initially, the concept of staggered or phased implementation emerged informally in the 1970s and 1980s. In early use, groups of participants received the intervention at different times, not out of methodological novelty but out of practical necessity. However, the design gained academic legitimacy relatively recently with the work of Hussey and Hughes (2007). Their work coherently articulated a statistical procedure for designing and analysing stepwise trials. Thus, this contribution provided researchers not only with the rationale for choosing this type of design but also with the tools necessary to model the complexity of variation within and between clusters over time. After this, SW-CRT studies have become increasingly visible in both the methodological literature and applied research. Further, building on these foundations, other researchers have refined the design, providing practical guidance on its use and clarifying its advantages in situations where simultaneous randomization is not possible (Hemming, 2015). These research extensions highlight the ethical and pragmatic appeal of the design, especially in contexts where interventions are expected to do more good than harm and therefore cannot reasonably be retained by any group in the long term.

In recent years, research trends have placed SW-CRT as a good option for fields such as education or sciences that deal with implementation and evaluation (Nevins, 2024). This may be precisely because in the context of these fields, SW-CRT is extremely useful for its ability to reflect real-world conditions in contrast to clinical trials (Bronfenbrenner, 1977). Moreover, by integrating the intervention into existing programs and institutional constraints, this design thus facilitates stronger external validity, without compromising methodological rigor. However, this design also has significant disadvantages and challenges. A first example is the complexity of coordination required; This increases with the number of clusters and waves, and the risk of contamination (e.g., spillover effects between intervention and control groups) can threaten internal validity. In addition, the management of time points must be approached carefully, as fluctuations over time could be attributed to incidental factors not directly linked to the intervention (Tong, 2025). To sum up, these limitations require careful planning, transparent reporting, and sophisticated analytical strategies.

Nonetheless, when designed and implemented with methodological rigor, the stepped-wedge trial offers a compelling solution for ethically and logistically constrained intervention research. Its adoption in education is expected to grow as the field continues to move toward more ecological-contextually grounded and implementation-sensitive approaches to causal inference.

1.1. Strengths and Rationale

One of the primary strengths of the stepped-wedge cluster randomized trial (SW-CRT) lies in its capacity to reconcile internal validity—through randomization and temporal control—with ecological validity, by embedding the intervention within authentic educational settings. Namely, ecological validity refers to the extent to which research findings can be generalized to, or are representative of, real-life settings and situations. It concerns whether the conditions under which a study is conducted and the behaviors observed reflect the natural environment in which those behaviors typically occur (Shadish, 2002). Unlike conventional experimental designs that risk artificiality by isolating interventions from their natural contexts, the stepped-wedge framework enables the examination of intervention effects as they unfold organically within real classrooms and schools. Epistemologically, the design aligns with a pragmatic-constructivist paradigm, which values not only causal inference but also the generation of contextually situated and practically relevant knowledge. By capturing change over time and across diverse educational environments, the SW-CRT allows researchers to address not only whether an intervention works, but also *how*, *for whom*, and *under what conditions*—a central concern in contemporary educational research and policy (Biesta, 2007)

Another particularly compelling rationale for employing the SW-CRT in educational contexts is its congruence with ethical imperatives and institutional constraints. Unlike traditional parallel-group randomized controlled trials (RCTs), which may permanently withhold interventions from some participants, the stepped-wedge model ensures that all clusters eventually receive the intervention. This is especially pertinent in education, where fairness, access, and equity are not only ethical concerns for researchers but also salient issues for educators, families, and policymakers. Presumably, developmental educational interventions generally aim to confer cognitive, emotional, or developmental benefits. Hence, applying interventions directly in the educational environment of students makes the exclusion of some of them not only problematic but deeply questionable from an ethical point of view.

In addition, it is worth mentioning another advantageous functional aspect of the SW-CRT design: namely, the fact that it contributes to a collaborative research culture. The participatory ethos of the design can strengthen the legitimacy and sustainability of educational research initiatives, especially when schools are involved not only as data providers for researchers but as active partners in an approach oriented towards mutual benefits and institutional development. In essence, the stepped wedge model marks a transition in educational research: from rigid, top-down experiments to more context-sensitive and socially responsive methodologies. It recognizes that schools are not laboratories, but dynamic communities with complex rhythms and responsibilities.

Continuing, from an operational perspective, we can say that the stepped wedge design (SW-CRT) offers notable advantages in terms of institutional feasibility. Specifically, in many educational systems, practical limitations — such as curricular calendars, facilitator availability, or infrastructure preparation — make it impossible to implement the intervention simultaneously in all locations. These facts have a direct impact on the adjustment of the design, especially in terms of extending the time dedicated to research. In contrast, the phased rollout, specific to this type of design, allows for the escalation of these constraints. Thus, far from being a limitation, this sequential structure can be leveraged to support adaptive implementation and real-time adjustments. Overall, the ethical sensitivity, operational adaptability, and epistemological relevance of this design transform it into a valuable tool for promoting educational research that aligns with the values and realities of 21st-century education.

1.2. Methodological and Analytical Challenges

Although the stepped wedge design (SW-CRT) offers multiple advantages, it is not without methodological and logistical challenges—some of which stem from the very features that make it attractive. Managing these limitations requires rigorous planning, strong interdisciplinary collaboration, and the use of advanced analytical strategies.

First of the central difficulties is the temporal complexity of the design. Unlike classic parallel group experiments, in which the transition from control to intervention occurs in a single stage, SW-CRT involves progressive implementation in waves across multiple clusters. Implementing the intervention in stages requires that some units gradually enter the program while others remain in the control phase, which requires careful management of timing throughout the study. In the educational context, this planned sequence is further complicated by the rigidity of the school calendar, fixed assessment periods, institutional holidays, staffing variations, and possible disruptions along the way, such as absences or suspensions. Any lag in the delivery of the intervention in a unit can affect the coherence of the overall implementation, jeopardizing both the comparability between groups and the internal validity of the results. To prevent such disruptions, it is necessary to develop detailed implementation plans, create alternative scenarios, and maintain constant communication with all participating units. Infrastructure that allows centralized coordination—such as implementation monitoring software—can reduce some vulnerabilities, but it also involves significant resources in terms of staff, training, and logistics, which is often an obstacle in the educational context.

Equally, the SW-CRT is considerably more complex than a conventional RCT when it comes to data analysis. For example, estimating effects requires multilevel models that account for the hierarchy of data, repeated measures, and potential time-related confounds. Even more, standard models typically include random intercepts for cluster and time, fixed effects for intervention and time, and often interaction terms to assess variations in the treatment effect (Copas, 2015). When outcomes follow nonlinear trajectories, exhibit autocorrelation, or time-varying responses, more sophisticated specifications—such as random slopes, autoregressive structures, or nonparametric smoothing methods—are required. These requirements involve a considerable learning effort, especially for educational researchers who do not have advanced training in hierarchical or longitudinal modelling. At the same time, Bayesian approaches, although useful in small samples or complex contexts, in the case of the steed wedge, add additional difficulty, both statistically and computationally. Similarly, estimating sample size and statistical power is also more complicated (Kristunas, 2019).

In contrast to traditional designs, power in SW-CRT depends on an extensive series of parameters: the number of stages, the number of clusters, cluster-period combinations, intra-cluster correlation coefficient (ICC), variance of results, dropout rate, and possible temporal trends. Unequal cluster sizes or variation in intervention effects over time can considerably reduce the power of the analysis. Although tools such as Shiny CRT Calculator, the R package swCRTdesign, or the steppedwedge module in Stata can support the design optimization process, their effective use requires advanced methodological knowledge. Incorrect application can lead to underpowered studies or erroneous inferences (Martin, 2016).

Given all this, interdisciplinary collaboration becomes essential in this type of study. Not only for the analysis of quantitative data. Moreover, statisticians and methodological experts should be involved from the early design phase, not only in the post hoc analysis stage, but as active partners in the development of the research protocol (Mdege, 2011). Such partnerships help to avoid model specification errors, validate inferences, and generate relevant and practically applicable conclusions.

1.3. SW-CRT in Educational Contexts

Extensively validated in clinical research and public health, the stepped wedge experimental design (SW-CRT) finds its applicability in an educational setting only after a substantial reconceptualization. More precisely, in education, the application of this model cannot be dissociated from the particularities of the institutional environment, its own epistemological frames of reference, and daily operational realities. These conditions decisively influence not only the implementation of interventions, but also the way in which they are analysed and supported from an ethical perspective. Therefore, methodological transfer requires a careful adjustment to the educational specifics, going beyond the logic of direct transposition from other fields.

First, educational interventions are inherently socially embedded. Educational innovations are deeply interconnected with curriculum, school culture, and classroom dynamics. As well as teachers' beliefs and institutional norms. As a result, the fidelity of implementation can vary greatly between settings, and outcomes often depend on local, legal, and cultural interpretation and adaptation (Bronfenbrenner, 1977).

Secondly, another essential element that differentiates educational research from clinical research is the contextual nature of the school environment. In education, the effects of an intervention cannot be understood in isolation, as they are strongly influenced by a wide range of external factors – from the socioeconomic conditions of the beneficiaries and the educational policies in force, to the governance style adopted at the institutional level. In comparison, clinical environments often have standardized protocols that reduce contextual variability. In contrast, schools operate in more varied and unstable settings, which makes the same intervention produce different results depending on the context (Tong, 2025). In this sense, the stepped wedge design is obliged to respond to this complexity not only through analytical tools capable of capturing contextual variability, but also through implementation strategies adaptable to local realities.

Furthermore, another major challenge in educational research is related to the structural constraints specific to this field. Fixed school calendars, standardized assessment periods, holidays, and limits on the availability of educational staff establish rigid frameworks that significantly condition when and how an intervention can be put into practice (Copas, 2015). In many cases, these institutional realities make it impossible to apply the intervention simultaneously in all the units involved. The stepped wedge design becomes, in this context, a particularly valuable methodological option, as it allows for staggered implementation, in line with the temporal dynamics specific to schools, without compromising the experimental control necessary to validate the results.

Additionally, another difficulty specific to educational research lies in the complexity of evaluating outcomes. Essential dimensions of learning, such as motivation, engagement, or academic and cognitive progress, are dynamic processes developing over time. Moreover, these phenomena are influenced by numerous contextual factors, often impossible to control or quantify directly. For this reason, current research trends say that simple point-in-time measurement is not enough. Repeated evaluations are needed to capture both the immediate effects of the intervention and its long-term impact. These considerations imply a carefully calibrated analytical approach that is sensitive to change over time.

Finally, educational research is increasingly shaped by participatory and constructivist epistemologies. Traditional experimental approaches often assume a top-down logic that prioritizes control over context (Shadish, 2002). However, contemporary scholars emphasize that schools should be dynamic, relational systems where change emerges through collaboration and reflection (Biesta, 2007). The SW-CRT is compatible with this orientation. Rather than treating variability as methodological noise, the design positions it as a source of insight into how interventions function across diverse educational landscapes.

To conclude, we can say that the transposition of a research model established in fields such as medicine or public health to the educational context cannot be achieved by direct adoption and application. The process must involve careful adaptation to the particularities of the educational sector — including its normative values, institutional constraints, and specific internal dynamics. When these dimensions are rigorously integrated into the methodological process, the stepped wedge design (SW-CRT) becomes a pertinent and efficient analytical tool for investigating educational interventions.

2. Methodology

2.1. The MKIT Pilot Trial – A Stepped-Wedge Cluster Randomized Evaluation

The present study investigated the effectiveness of MKIT (Metacognitive Knowledge Intervention for Thinking) — an educational intervention aimed to stimulate the development of metacognitive knowledge in children aged 5 to 9 years. The choice of a stepped-wedge experimental design was determined by the constraints related to the limited time available for implementation. Moreover, the progressive development of the intervention among the participating clusters allowed for an efficient allocation of resources, guaranteeing full access to the program for all participants and facilitating detailed monitoring of metacognitive transformations. Specifically, quantitative analyses aimed at the persistence of effects over time, differences in receptivity according to age, as well as variations in progress in relation to the initial performance level. The chosen design fully allowed for a rigorous and scientifically valid evaluation, while maintaining contextual relevance for the educational environment, operational feasibility, and compliance with ethical standards.

2.2. Participants

The study involved N = 80 children, included in 12 educational clusters (C1–C12). Subjects were recruited from 1 kindergarten and 1 primary school in a single geographical region of Romania. The sample included 40 five-year-olds (50%) and 40 nine-year-olds (50%), with 37 boys (46.3%) and 43 girls (53.8%). All participants were monolingual Romanian speakers with no reported developmental delays, based on parent and teacher screening. As Table 1 shows, participants were recruited via institutional partnerships and randomly assigned to clusters, which were then allocated to staggered intervention phases based on a predefined SW-CRT schedule. The sample was diverse in socio-educational background and designed to allow comparisons across age, gender, and baseline performance. Power analyses were not applicable due to the pilot nature of the study; however, the stepped-wedge design maximized inferential potential through repeated within- and between-cluster comparisons.

Table 1. *Participant distribution by cluster, age group, and gender*

Cluster	Age group	Size (N)	Boys	Girls	Note
C1	5 yrs	7	4	3	random 57 % boys
C2	5 yrs	7	3	4	random 43 % boys
C3	5 yrs	7	4	3	random 57 % boys
C4	5 yrs	7	3	4	random 43 % boys
C5	5 yrs	6	2	4	random 33 % boys
C6	5 yrs	6	4	2	random 67 % boys
C7	9 yrs	7	3	4	random 43 % boys
C8	9 yrs	7	3	4	random 43 % boys
C9	9 yrs	7	4	3	random 57 % boys
C10	9 yrs	7	2	5	random 29 % boys
C11	9 yrs	6	3	3	random 50 % boys
C12	9 yrs	6	2	4	random 33 % boys
Total		80	37	43	satisfies study requirements

2.3. Design and procedure

The metacognitive intervention was implemented using a stepped-wedge design over a period of 6 consecutive weeks within naturalistic classroom settings. A total of 12 educational clusters (C1–C12), stratified by age (children aged 5 and 9), participated in the study. The overall sample included 80 children, and the intervention was gradually introduced to two new clusters per week, following a predefined rotation schedule. This design allowed for controlled, staggered exposure while maintaining ecological validity.

Each week, children engaged in four structured metacognitive activity sessions delivered from Monday through Thursday. The sessions incorporated four recurring task formats designed to strengthen metacognitive knowledge: illustrated narrative comprehension, semi-structured cognitive games, writing and drawing activities, and strategic reasoning exercises. Fridays were reserved for data collection, during which individual interviews were conducted with each child in quiet, distraction-free settings. This weekly structure ensured a clear separation between instructional and evaluative components, minimizing performance contamination and social desirability bias.

The intervention was delivered by the principal investigator in a group-based classroom format, following a standardized implementation and observation protocol to ensure high procedural fidelity and consistency across all clusters, as shown in Figure 1.

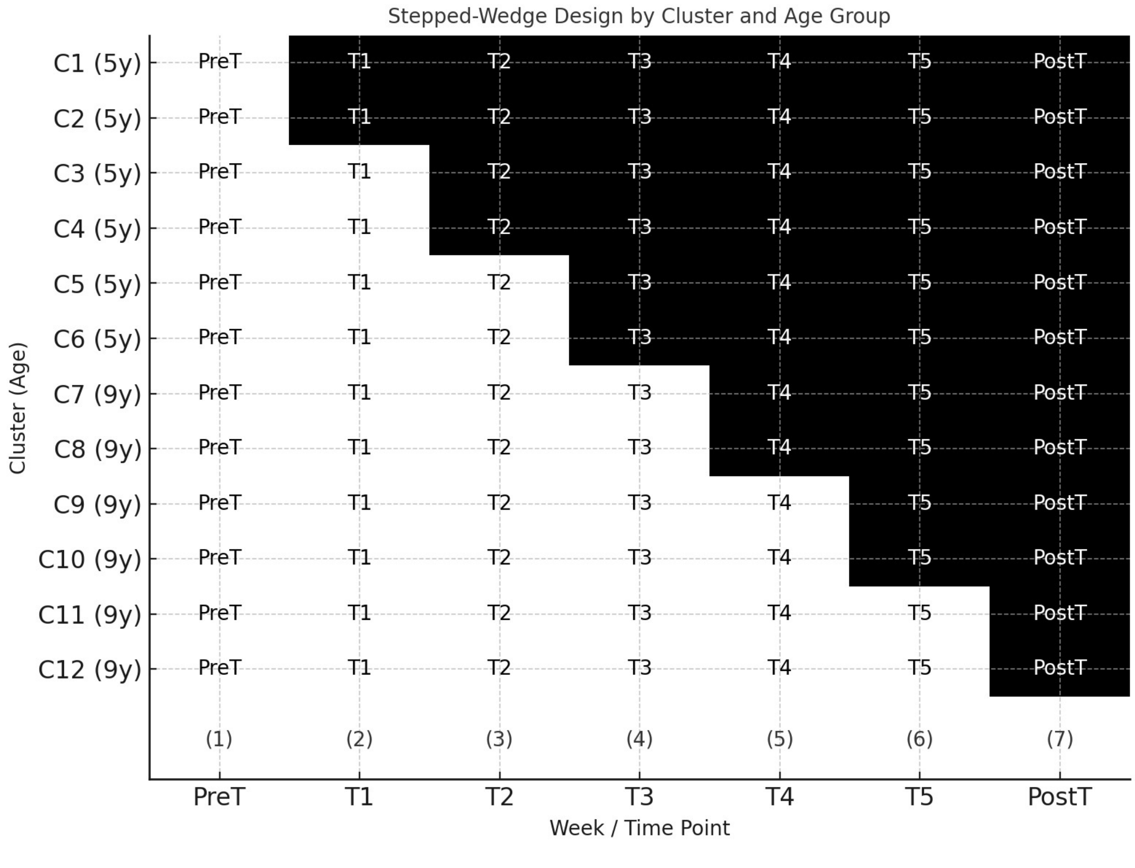

Figure 1. *Stepped-Wedge trial design with cluster crosspoints schedule and data collection timeline*

Data were collected individually across seven assessment waves, each employing a uniform administration procedure and lasting approximately 17 to 20 minutes per child. This repeated-measures design allowed for the systematic and reliable monitoring of developmental changes in metacognitive knowledge over the course of the intervention.

2.4. Measures and analytical framework

Building upon the standardized implementation and consistent data collection procedures described above, the present section outlines the measures used to assess metacognitive knowledge, followed by the analytical framework employed to evaluate intervention effects. Hence, given the longitudinal structure of the study—comprising seven repeated measurement waves and a nested data hierarchy (children nested within educational clusters)—a linear mixed-effects modelling (LMM) approach was employed.

This method was selected to appropriately model both within-individual change and between-individual variability across time, while accounting for the dependency structure of the data. The analytical strategy was guided by a set of theoretically grounded hypotheses, each addressing distinct dimensions of the intervention's effects:

H1: It was hypothesized that the MKIT intervention would produce a significant increase in children's metacognitive knowledge immediately following its implementation, relative to pre-interventional levels. The expected effect size was in the small-to-moderate range, consistent with findings from prior intervention research.

H2: The effectiveness of the intervention was expected to vary by age group, with older children (9 years) predicted to show greater gains than younger children (5 years), reflecting age-related differences in metacognitive responsiveness.

H3: The intervention's effects were hypothesized to be sustained or amplified over time. A positive linear post-intervention trend was anticipated, alongside potential non-linear growth components (e.g., quadratic terms) to capture curvilinear developmental patterns. Random slopes for time were included to account for inter-individual variability in growth trajectories.

H4: Finally, a compensatory effect was expected, whereby children with lower baseline metacognitive scores would benefit disproportionately more from the intervention than their higher-performing peers. This effect was hypothesized to manifest as a negative interaction between baseline score and intervention phase, with a predefined criterion of $\beta \leq -0.10$.

To test these hypotheses, a series of linear mixed-effects models (LMMs) was estimated, incorporating the following fixed effects: intervention status, time, age group, gender, and baseline performance. Random intercepts were specified for participants and clusters to account for the nested structure of the data.

The main effect of intervention status was examined to evaluate H1, while interaction terms were included to assess moderation by age (H2), time (H3), and baseline level (H4). This analytical framework ensured alignment with the longitudinal, hierarchical, and unbalanced nature of the dataset, providing robust estimation of both average and individual developmental patterns over time.

3. Results

Immediate Effects

As shown in Table 2, descriptive statistics indicated that scores increased in post-intervention from $M = 10.19$ ($SD = 2.96$) at pre-test to $M = 20.68$ ($SD = 1.89$) at post-test, with a mean difference of –10.49 points (SD = 3.16, SE = .13). The effect size was large, d = 1.94, indicating a substantial educational impact. The findings are detailed in Table 3, and revealed a strong and highly significant intervention effect, $t(559) = -78.51$, $p < .001$, suggesting that treatment condition accounted for a substantial proportion of the variance in post-intervention outcomes. Aligned with Hypothesis 1, children demonstrated higher levels of metacognitive knowledge following participation in the intervention.

Table 2. *Descriptive statistics and paired samples t-Tests for McKI scores*

Variable	M	SD	SE
McKI Pre-intervention	10.19	2.96	.125
McKI Pre-intervention	20.68	1.89	.080

Table 3. *Paired samples t-Tests results*

Pair	Mean Difference	SD	SE	t	df	p
McKI Pre-intervention McKI Pre-intervention	-10.487	3.161	.134	-78.510	559	.000

Note: A paired-samples t-test revealed a statistically significant increase in McKI scores after the intervention, *t*(559) = –78.51, p < .001. The negative mean difference indicates higher post-intervention scores. The effect size was large (Cohen's d = 1.94), suggesting substantial educational impact.

Age Effects

Hypothesis 2 predicted that older participants (age 9) would exhibit greater metacognitive gains than their younger peers (age 5). This assumption was not supported by the data. As Table 4 shows, no significant differences emerged between the two age groups, *F*(1, 556) = 0.011, *p* = .917, indicating that the intervention was equally effective across developmental stages, highlighting its robustness across age levels.

Age × Intervention interaction

To address Hypothesis 3, which posited that the impact of the intervention might vary as a function of age, interaction effects were examined using mixed-effects modelling. Thus, Age × Intervention interaction did not reach statistical significance, *F*(1, 556) = 1.620, *p* = .204. As illustrated, results demonstrate that children of different ages showed comparable improvements, and age did not significantly moderate the efficacy of the intervention.

Table 4. *Tests of between-subjects effects for age, intervention, and their interaction*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	50.464	3	16.821	171.809	.000
Intercept	846.158	1	846.158	8642.555	.000
Age	0.001	1	0.001	0.011	.917
Intervention Status	50.266	1	50.266	513.406	.000
Age × Intervention	0.159	1	0.159	1.620	.204
Error	54.436	556	0.098	—	—
Total	1154.917	560	—	—	—
Corrected Total	104.899	559	—	—	—

Note: A significant main effect was found for intervention status. No significant main effect of age or Age × Intervention interaction was observed.

Group × Time Interaction

Hypothesis 4 suggested that the benefits of the intervention would be retained or even amplified over the course of the study. Longitudinal analyses were conducted across seven measurement waves showed a progressive and sustained improvement in MK scores within the intervention group, increasing from *M* = 0.93 (*SD* = 0.27) at Wave 1 to *M* = 1.78 (*SD* = 0.19) at Wave 7. In contrast, participants in the control group exhibited only minimal gains over the same period, with clearly divergent performance trajectories between groups. To further examine the persistence of effects, a linear regression analysis was conducted using baseline MKI scores as a predictor of outcomes. Results confirmed a significant predictive relationship, $F(1, 558) = 26.614$, $p < .001$, with an R^2 of .046 (adjusted $R^2 = .044$), and a standard error of the estimate of 1.856. The regression model illustrated in Table 5 confirms that initial metacognitive performance was a significant, though modest, predictor of final scores. Specifically, MKI initial score had a positive effect, $\beta = 0.137$ (SE = 0.026), standardized $\beta = .213$, $t = 5.159$, $p < .001$, while the intercept was $\beta = 19.284$ (*SE* = 0.281), $t = 68.676$, $p < .001$.

Table 5. *Linear regression predicting post MKI scores from pre scores*

Component	Value
ANOVA	
Regression	$F(1, 558) = 26.614$, $p < .001$
R^2	.046
Adjusted R^2	.044
Std. Error	1.856
Regression Coefficients	
Intercept	*$\beta = 19.284$, $SE = 0.281$, $t = 68.676$, $p < .001$*
MKI Initial Score	*$\beta = 0.137$, $SE = 0.026$, $\beta = .213$, $t = 5.159$, $p < .001$*

Note: Dependent variable: McKI Final Mean (Wave 7). The regression equation is: Y = 19.284 + 0.137 × X, where X is the initial McKI score. The model explained 4.6% of the variance in post-intervention metacognitive knowledge scores, indicating a small but statistically significant predictive effect, consistent with durable learning gains over time.

Baseline-Dependent Effects

In line with Hypothesis 5, it was hypothesized that children with lower baseline metacognitive scores would experience greater relative benefits from the intervention. This pattern was confirmed by a one-way ANOVA, conducted using four baseline performance groups. As Table 6 shows, results revealed a statistically significant effect of baseline metacognitive knowledge on gain scores, $F(3, 556) = 250.97$, $p < .001$, indicating that the level of improvement varied systematically by initial performance level.

Table 6. *One-Way ANOVA comparing metacognitive gains across baseline performance groups*

Baseline Group	*n*	Mean Gain	SD
Low	56	–14.50	3.07
Medium	266	–11.92	2.11
High	161	–8.61	1.64
Very High	77	–6.55	1.79

Note: Gain scores represent the difference between post- and pre-intervention MKI scores ($\Delta MK = MK\_post - MK\_pre$). A one-way ANOVA revealed a significant effect of baseline group, F(3, 556) = 250.97, p < .001. Bonferroni post hoc comparisons indicated that all between-group differences were statistically significant (p < .001).

Results showed a clear descending gradient in gains across groups: children in the Low baseline group exhibited the largest improvements (*M* = –14.50, *SD* = 3.07), followed by the Medium (*M* = –11.92, *SD* = 2.11), High (*M* = –8.61, *SD* = 1.64), and Very High (*M* = –6.55, *SD* = 1.79) groups. Post hoc Bonferroni comparisons confirmed that all pairwise differences between groups were statistically significant (*p* < .001). These results indicate that the intervention was particularly advantageous for initially lower-performing children, thereby supporting a threshold or equity-enhancing effect.

4. Discussions

The stepped-wedge cluster randomized trial (SW-CRT) design offered a strategic and ethical advantage by enabling all participating clusters to eventually receive the intervention. This facilitated a robust examination of causal relationships while addressing equity concerns often encountered in educational research. By distributing the intervention sequentially across clusters, the design permitted both within- and between-group comparisons over time, strengthening the internal validity of the findings and ensuring that no group was permanently denied access to the potentially beneficial program.

One of the most salient contributions of this research was the demonstration that children can reliably engage with a structured metacognitive intervention and achieve significant, lasting improvements within a short period of time. The results were consistent with Hypothesis 1; children exhibited a marked improvement in metacognitive knowledge following intervention exposure. The SW-CRT design allowed for the analysis of these gains as a function of timing and sequence, rather than relying solely on pre–post comparisons. This enhanced the internal validity of the findings by helping to isolate the effect of the intervention from maturational or external time-related influences. The lack of significant differences between age groups (Hypothesis 2) and the absence of a meaningful Age × Intervention interaction (Hypothesis 3) suggest that the MKIT intervention was developmentally robust. Because the SW-CRT staggered entry points across clusters, age distributions were effectively balanced over time, further reducing confounding and increasing the confidence with which we can attribute gains to the intervention rather than to natural developmental change. The design also facilitated the assessment of longitudinal growth trajectories. Aligned with Hypothesis 4, the intervention's effects were not only immediate but also cumulative, as evidenced by sustained gains across successive waves. The SW-CRT structure, with repeated assessments of each participant, enabled the detection of progressive learning curves and supported a fine-grained examination of change over time—an advantage not afforded by traditional parallel-group designs. Moreover, conducting the trial as an SW-CRT within authentic classroom settings enhanced the study's ecological validity, reinforcing the relevance of the findings for everyday educational practice. Simultaneously, the structure permitted precise statistical control through the inclusion of random effects at both the individual and cluster levels.

However, even though the current findings highlight the efficacy and feasibility of the MKIT intervention, the SW-CRT is not without limitations. The temporal correlation between intervention exposure and time complicates interpretation unless properly modelled, and carryover effects cannot be ruled out if the intervention induces lasting classroom-level changes. Additionally, the assumption that the intervention effect is immediate and stable across clusters may fail to sustain statistical significance in other contexts. Future work would be well advised to consider more flexible designs for analysing delayed or decreasing effects and to include post-intervention follow-up. several avenues warrant exploration in future research. Simultaneously, one important next step is to evaluate the long-term impact and transferability of metacognitive gains to broader academic outcomes, such as problem-solving, reading comprehension, or self-regulated learning. Future trials should consider incorporating delayed follow-up assessments (e.g., 3–6 months post-intervention) to examine maintenance of effects, particularly in naturalistic settings without ongoing support. Another promising direction is to investigate the optimal dosage and timing of the intervention. Although the current implementation followed a 6-week format, it remains to be seen whether shorter or more intensive formats might yield comparable or enhanced outcomes.

To conclude, researchers employing Stepped-Wedge Cluster Randomized Trials (SW-CRT) in education should remain attentive to several methodological aspects regarding both the design and implementation of the study:

(*1*) Timing and temporal confounding: In SW-CRT designs, time and treatment are inherently correlated. Researchers should use carefully specified time-fixed effects or spline models to isolate intervention effects from secular trends (Salway, 2025).

(*2*) Readiness of clusters: clusters may vary in terms of implementation readiness, and staggered roll-out schedules must consider practical constraints, such as teacher or facilitator training, local calendar conflicts, or administrative capacity (Kristunas, 2019).

(*3*) Measurement fatigue and reactivity: repeated measurements across waves may induce fatigue or test-retest effects. Ensuring that assessments remain engaging and non-intrusive is critical, particularly when working with young children (Bronfenbrenner, 1977).

(*4*) Mixed-method integration: embedding qualitative components (e.g., teacher feedback, classroom observations) within the SW-CRT structure can yield richer insights into contextual factors that affect intervention success or failure, thus enhancing interpretation and generalizability (Hemming, 2015).

In summary, the use of an SW-CRT design provided a methodologically rigorous and ethically sound framework for evaluating the MKIT intervention under real-world conditions. The findings demonstrate that the intervention not only produced significant and sustained improvements in metacognitive knowledge but did so in a way that was equitable, scalable, and developmentally inclusive. These results position the SW-CRT as a promising candidate for broader implementation in future implementations of this study, while also showcasing the general value of stepped-wedge designs in applied developmental research. The results reinforce the relevance of the SW-CRT design as a viable methodological solution for developmental interventions in multi-varied educational contexts.

Conclusions

In conclusion, the present study highlights the potential of SW-CRT designs in educational research, particularly when ethical, developmental, and contextual factors must be balanced with methodological rigor. The stepped-edge cluster randomized trial (SW-CRT) design offers an extremely valuable methodological architecture for contemporary educational research, as it allows for rigorous causal inference under ecological conditions, often impossible to achieve through traditional RCTs. Furthermore, from a statistical perspective, it allows for the coherent integration of hierarchical structures (e.g., students in classes) and longitudinal data by using linear mixed models (LMM) or generalized linear models (GLMM). This provides robust estimates of intervention effects by simultaneously controlling for inter- and intra-cluster variations. In addition, the design allows for testing developmental trajectories and differential effects through multiple interactions (e.g., time × treatment × baseline level), facilitating granular analysis of impact by demographic characteristics or baseline performance. Therefore, from a psychometric perspective, the repetitive structure of the measurements makes it possible to test metric invariance over time, estimate longitudinal reliability, and assess the sensitivity of the instruments used, thus providing an ideal framework for validating measurement scales in dynamic contexts. Pragmatically, SW-CRT optimizes resource use by maximizing statistical power in relatively small samples and allows for sequential implementation adapted to educational realities — such as school calendars, teacher training, or local infrastructure.

Overall, SW-CRT is not just an experimental design, but an integrative platform that supports ethical, equitable, and methodologically valid research, with a clear potential for transfer and scalability at the educational system level. Its widespread use can transform the paradigm of intervention evaluation from a technical approach to a collaborative and contextualized process, capable of generating solid evidence with direct relevance for public policies in education.

Additionally, future efforts aiming to design stepped-wedge trials are encouraged to integrate process-level measures, strengthen fidelity monitoring, and consider hybrid effectiveness-implementation models to maximize both impact and sustainability.

References

1. Biesta, G. J. (2007). Why "What Works" Won't Work: Evidence-Based Practice and the Democratic Deficit in Educational Research. *Educational Theory*, *57*(1), 1–22. doi:https://doi.org/10.1111/j.1741-5446.2006.00241.x
2. Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32*(7), 513–531. doi:https://doi.org/10.1037/0003-066X.32.7.513
3. Copas, A. L. (2015). Designing a stepped wedge trial: three main designs, carry-over effects and randomisation approaches. *Trials, 16*, Art. 352.
4. Hemming, K. H. (2015). The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *BMJ, 350*. doi:https://doi.org/10.1136/bmj.h391.
5. Hooper, R. (2021). Key concepts in clinical epidemiology: Stepped wedge trials. *Journal of Clinical Epidemiology, 137*, 159-162.
6. Hughes, J.P., Lee, W.Y., Troxel, A.B., & Heagerty, P.J. (2024). Sample Size Calculations for Stepped Wedge Designs with Treatment Effects that May Change with the Duration of Time under Intervention. *Prevention Science, 25* (Suppl 3), 348–355. https://doi.org/10.1007/s11121-023-01587-1
7. Hussey, M. A. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials, 28*(2), 182–191. doi:https://doi.org/10.1016/j.cct.2006.05.007.
8. Kristunas, C. A. (2019). Assessing the sample size requirements of stepped-wedge cluster randomised trials with varying cluster sizes. *BMC Medical Research Methodology*, https://doi.org/10.1186/s12874-019-0661-3.
9. Martin, J. T. (2016). Systematic review finds major deficiencies in sample size methodology and reporting for stepped-wedge cluster randomised trials. *BMJ, 6*(2). doi:https://doi.org/10.1136/bmjopen-2015-010166
10. Mdege, N. D.-S. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology, 64*(9), 936–948. doi:https://doi.org/10.1016/j.jclinepi.2010.12.003
11. Nevins, P. R.-P.-R. (2024). Adherence to key recommendations for design and analysis of stepped-wedge cluster randomized trials: A review of trials published 2016–2022. *Clinical Trials, 21*(2), 199–210. doi:https://doi.org/10.1177/17407745231208397
12. Salway, R. H.-S.-C. (2025). Designing stepped wedge trials to evaluate physical activity interventions in schools: methodological considerations. International Journal of Behavioral Nutrition and Physical Activity(22). doi:https://doi.org/10.1186/s12966-025-01720-z
13. Shadish, W. R. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

14. Tong, G. P. (2025). A review of current practice in the design and analysis of extremely small stepped-wedge cluster randomized trials. *Clinical Trials, 22*(1), 45–56. doi:https://doi.org/10.1177/17407745241276137.