

African Journal of Education and Practice (AJEP)

**Guided Instruction and Observed Numeracy Mastery in Kenyan Lower-Primary
Classrooms**

Reuben Gitonga Mutegi (PhD)

Guided Instruction and Observed Numeracy Mastery in Kenyan Lower-Primary Classrooms



¹*Reuben Gitonga Mutegi (PhD)

Department of Educational Management Policy and Curriculum Studies, University of Nairobi

Article History

Received 17th August 2026

Received in Revised Form 23th September 2026

Accepted 20th October 2026



How to cite in APA format:

Mutegi, R. (2026). Guided Instruction and Observed Numeracy Mastery in Kenyan Lower-Primary Classrooms. *African Journal of Education and Practice*, 12(5), 14–35.
<https://doi.org/10.47604/ajep.4019>

Abstract

Purpose: Kenya's Competency-Based Curriculum obliges lower-primary teachers to document their planning in detail while adopting learner-centered pedagogies, however, whether either obligation translates into what learners do in numeracy lessons remains poorly evidenced.

Methodology: This study analyzed 31 structured lesson observations across 26 public and low-cost private lower-primary schools in three Kenyan counties. A three-item index of observed learner numeracy mastery served as the outcome, combining strategic flexibility in producing correct answers, generalization of concepts beyond the textbook, and an assessor-verified check of learners' exercise books. Forty-one (41) binary practice indicators were screened on this outcome with Benjamini-Hochberg correction and then subjected to multivariable regression, exact categorical analysis, cluster-robust estimation and leave-one-observer-out resampling in exploratory analysis. Four practices were corrected.

Findings: Guided instruction was the most measured, with a massive difference between the average mastery of 97.1% observed and 54.2% unobserved ($r = .68$, $d = 2.03$, $q = .001$). The association remained statistically significant after teaching experience, class enrollment and county were controlled ($b = 42.31$, 95 % CI [24.99, 59.63]) and in all the leave-one-observer-out replications, using observer-clustered standard errors. There was no positive association between mastery and documentary preparation, and this effect was reversed after statistically controlling for enacted pedagogy.

Unique Contribution to Theory, Practice and Policy: The results cast doubt on the notion that inspection documents of planning reflect quality of teaching. The results are provided as a hypothesis only, due to the small sample and exploratory screening.

Keywords: Guided Instructions, Foundational Numeracy, Classroom Observation, Competency-Based Curriculum, Kenya

©2026 by the Authors. This Article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>)

INTRODUCTION

For the first time in the history of Africa, there is a way to gauge the continent's basic learning problem, and it has not gotten better in 30 years. Around 1 in every 7 children in primary education on the continent reaches minimum proficiency in reading and math, and this drops to 1 in 9 when an incomplete primary school education is factored in (UNESCO 2025). Kenya has adopted one of the most sweeping curricula changes in the region. The Competency Based Curriculum (CBC) which was introduced in 2017 and later re-branded as Competency Based Education, restructures mathematics teaching at the lower primary levels to focus on identifiable competencies instead of coverage of the syllabus and requires a new approach for teachers that many were never trained in (Kenya Institute of Curriculum Development [KICD], 2023).

The reform has two demands; the link between the two is seldom explored. The first is documentary. School-level supervision and external quality assurance most easily looks at teachers' artefacts, which should include schemes of work, lesson plans, lesson notes and records of work covered, all aligned to the curriculum design. The second is pedagogical. Teachers will be encouraged to shift practices from transmitting to learner-centred, using concrete materials, be differentiated, and be formatively assessed. The second demand has been a persistent challenge in Kenya and comparative African studies, and this has been attributed to different factors such as poor retooling, lack of resources, administrative burden among others (Mutai et al., 2024; Nketsia & Opoku, 2024).

What has not been established is whether either demand predicts what learners accomplish in a lesson. The distinction matters practically. If documentary compliance indexes instructional quality, then the substantial supervisory apparatus built around it is well targeted. Teachers note that apparatus consumes teacher time without returning learning, and the professional development that would return learning is something else entirely. Answering the question requires observational data in which practices and learner performance are recorded in the same lesson, which is precisely what most system-level monitoring in Kenya does not produce.

The question also has a particular urgency in lower primary. It is in the first three years of schooling that the numeracy foundations on which later mathematics depends are either laid or lost, and it is in these grades that the CBC's pedagogical demands depart most sharply from the practice teachers inherited. Supervisory attention in Kenyan primary schools is finite and, in many schools, is exercised by head teachers carrying substantial teaching loads of their own. What that attention is directed towards is therefore not a neutral administrative choice but a decision about where the limited capacity of the system to influence classroom practice is spent. Evidence on which observable features of a lesson carry information about learning, and which do not, bears directly on that decision.

Statement of the Problem

Kenya's competency-based reform rests on an assumption that has not been tested: that the documentary and pedagogical demands it places on lower-primary teachers serve the same purpose, namely learners who can demonstrably use mathematics. The supervisory system operates as though the assumption holds. School-level supervision and external quality assurance inspect what can most readily be verified after the fact, such as schemes of work, lesson plans and records of work, so that documentary completeness functions in practice as the system's working proxy for instructional quality. However, regional evidence, increasingly locates the binding constraint on foundational numeracy inside the lesson, in what teachers do

while learners are working, rather than in curricular provision or formal preparation (Lynch et al., 2025; UNESCO, 2024, 2025). If documentary compliance and enacted pedagogy are only weakly related, or are related to learning in different ways, then a considerable share of the system's limited supervisory capacity is directed at indicators that carry little information about whether learners are acquiring early numeracy.

This possibility cannot at present be evaluated, for three reasons. First, the observation instruments most widely used in the region record teacher practice without recording what learners produce in the same lesson, so the link between a practice and learning must be inferred from separate assessments administered at a different time and level of aggregation (World Bank, 2024). Second, studies of CBC implementation in Kenya and the wider region have documented teacher preparedness, documentation burden and difficulty with learner-centred pedagogy, but have not examined whether any of these is associated with learner performance (Mutai et al., 2024; Nketsia & Opoku, 2024).

Third, the pedagogical contrast in the reform discourse is that of teacher-centred versus learner-centred instruction, leaving a detailed look at what instructional moves are made to support learner thinking during a task largely unexamined. The question that this study seeks to answer is thus why there is a lack of evidence for lessons-level that can inform on the observable practices in the classroom, for better or worse, that correlate with what the lower primary child demonstrates in numeracy. In Kenya's lower primary grades, supervision, professional development and monitoring is organized around indicators whose relationship to learning is presumed without evidence.

This study employs observational data of the type needed. Thirty-one lower primary numeracy lessons across three contrasting counties in Kenya were observed using a structured protocol that captured both teacher practices and learner performance in the same observation window of the programme assessment. The analysis investigates which of forty-one observable practices most effectively explains the difference between lessons where learners showed mastery of numeracy and those where they did not, and whether the preparation of a document contributes to this.

The analysis is guided by two research questions. Which observable classroom practices are most highly correlated with observed mastery of numeracy skills in the same lesson? RQ2: Does doing the preparation of the document relate with the observed learner numeracy mastery when enacted practice is controlled?

Three contributions follow. Empirically, the study compares the strength of the association between each observed practice and learner mastery and determines that guided instruction is associated with learner mastery the most. Conceptually, it provides evidence that documentary compliance and pedagogical enactment are distinct constructs in CBC classroom and have different behaviours in relation to learning. Methodologically, it illustrates that it is possible to defensibly analyse a small number of purposively sampled observations using multiplicity correction, exact inference and resampling, rather than over-interpret the results of a standard large-sample analysis.

Foundational Numeracy and the Teaching Quality Problem

Numeracy problems are shifting towards the inside of the classroom and not to access or enrolment in Africa. The UNESCO (2024) analyses that underpinned the 2024 Spotlight report on foundational learning in Africa found that the materials, guides and knowledge available to

teachers do not match their ambition for the classroom. The 2025 Spotlight took this argument further to instructional leadership, concluding that the role of school leaders in shaping learning is primarily as clarifiers of objectives and pedagogy, rather than as administrators of learning (UNESCO 2025). Monitoring instruments have been developed along the same lines. Classroom practice is now a system indicator for reporting, alongside curriculum and assessment policy, in the Global Foundational Learning Action Tracker (UNICEF, 2025).

This reorientation gave a lot of emphasis to the measurement of teaching. The scores in large-scale observation systems (such as the World Bank's Teach instrument, the most widely used measurement) are similar in low- and middle-income countries, and systems have been created to describe national instructional profiles (World Bank, 2024). For present purposes their limitation is that they do not measure the learner's production in the same lesson as they measure the teacher's production, and the connection between a practice score and the learner's production must be made out of the lesson, usually using independent assessments. Research that concurrently studies practices and learner performance continues to be rare in the region.

The available evidence base in Kenya confirms the same. Low levels of foundational attainment before the introduction of the CBC were documented in the citizen-led household assessments: before the introduction of the CBC, around three out of ten of Class 3 learners could do the work of Class 2 (Uwezo, 2016). A structured response, with access to pupil resources, teachers' guides and mathematics training for teachers (with a limit on hours), was tested in the Primary Math and Reading (PRIMR) Initiative, with modest but statistically significant gains in mathematics and smaller gains in reading (Piper et al., 2016).

The national Tusome programme then introduced the structured-pedagogy model to government systems, and their evaluation found a direct relevance here: visits for classroom observations were increased and more focused on teaching quality, but there was not a full use of the classroom observation data to inform provision of support (Piper et al., 2018). The CBC departs from these programmes in an important respect. While the CBC provides lesson-level guides for teachers, these are less structured and are supplemented by a variety of classroom-based resources, PRIMR and Tusome provided teachers with lesson-level guides and accompanying support in the classroom. Whether that change has maintained the focus of instruction of its predecessors is an empirical question that is addressed by the present study.

Guided Instruction and the Structure of Pedagogical Support

Within the framework of gradual release of responsibility, guided instruction has a well-defined place. It refers to the point in between "*teacher modelling*" and "*independent practice*" where the teacher asks questions, gives hints, cues and explanations, and corrects the learner's mistakes and misunderstandings, but does not give them the answer outright (Frey, 2025). These phases are not sequentially ordered, but are deployed, as needed, depending on what learners are doing, making it different from didactic transmission and unassisted discovery.

The case for structured instructional support in math is empirical and has been re-established time and again. Explicit and systematic instruction has been shown to have large effects on mathematics achievement in the synthesis of explicit and systematic instruction (Peltier, 2024), with effects most pronounced among students at the lower end of the performance distribution. Evidence also points to professional development in math and science having an impact in similar ways, albeit from a different perspective: Professional development that results in changes in teachers' knowledge also results in changes in their classrooms, and these changes in classrooms are the ones that relate to changes in student performance (Lynch et al., 2025).

These literatures imply that what the teacher does to organize the learner's thinking in a task is the variable, not what the teacher knows or has written down in advance.

Guided instruction should be distinguished from guided discovery, with which it is often conflated. Guided discovery front-loads exploration, positioning the teacher as the designer of a problem space that learner's traverse. Guided instruction front-loads support, positioning the teacher as a responsive presence during the traverse. Both appear in the observation protocol used here and separating them empirically is one purpose of the present analysis.

Documentary Compliance in Competency-Based Reform

Competency-based reforms across Africa have generated a substantial documentation burden. In Kenya, early grade teachers have been found to be inadequately prepared for the CBC, with training needs identified in pedagogy and assessment as well as in the preparation of the required teaching documents (Momanyi & Rop, 2019), and reviews of the rollout point to shortages of human and material resources and a mismatch between curriculum intentions and classroom practice (Akala, 2021). Studies of teacher preparedness in Kenya have found that possession of the required documents is widespread while confidence in the associated pedagogies is not (Mutai et al., 2024).

The theoretical potential this creates is the decoupling of the formal structures from the what is happening, as developed in the theoretical framework below: the ceremonial use of formal structures that meet external inspection, but day-to-day practice remains largely the same. According to Decoupling, there is a specific empirical pattern: high compliance on documentary indicators, weak association between those indicators and learner outcomes, and stronger association between those indicators that cannot be produced before the lesson, and learner outcomes. This pattern can be tested from the current data.

Theoretical Framework

Two complementary theories inform the study. Sociocultural learning theory, specified through the construct of scaffolding, explains why enacted and responsive practice should be associated with learner mastery. Institutional theory, specified through the concept of decoupling, explains why documentary compliance need not be. Together they account for both sides of the contrast the study examines: which features of teaching are associated with learner performance in the lesson, and why the most inspectable features of teaching may not be among them.

Sociocultural Theory and Scaffolding

Vygotsky's (1978) sociocultural theory holds that higher cognitive functions develop first through social interaction and are subsequently internalised by the learner. Its central construct, the *zone of proximal development*, denotes the distance between what a learner can accomplish independently and what the same learner can accomplish with the assistance of a more knowledgeable other. Learning occurs within that zone, and instruction is productive to the extent that it operates there. Wood et al. (1976) introduced the term *scaffolding* to describe the form such assistance takes: the tutor recruits the learner's interest, reduces the degrees of freedom in the task, keeps the learner oriented to the goal, marks critical features, controls frustration and demonstrates solutions, withdrawing support as competence grows.

Subsequent reviews identify contingency, fading and the transfer of responsibility as the defining characteristics of scaffolding, and emphasise that contingency, the calibration of support to the learner's moment-to-moment understanding, is what distinguishes scaffolding from assistance in general (van de Pol et al., 2010). The gradual release of responsibility model

translated these principles into an instructional sequence in which guided practice mediates between teacher modelling and independent work (Frey, 2025; Pearson & Gallagher, 1983).

The theory yields three expectations for the present study. First, practices enacted during learner task engagement, and responsive to what learners are doing, should be more strongly associated with mastery than practices fixed before the lesson begins. Guided instruction, as coded in the observation protocol, is a close operational analogue of contingent scaffolding: the teacher prompts, cues and questions in response to learner errors without supplying the answer. Second, the contrast of consequence is between contingent and non-contingent instruction rather than between teacher-directed and learner-directed instruction, since scaffolding is teacher-directed yet centred on the learner's reasoning.

Third, the theory predicts that scaffolding is contingent, which suggests that there will be reciprocal influences. Support is calibrated to learner performance, so learners who are successful determine the kind of support received and the kind of support determines their success. Sociocultural theory thus predicts the simultaneity that can only be addressed in a single lesson; the hypotheses are therefore put forward as associations, rather than effects.

Institutional Theory and Decoupling

The institutional theory helps to understand why organizations create formal structures that are irrelevant to the primary activity of the organization. Meyer and Rowan (1977) proposed that the institutions with strong expectations are able to gain their legitimacy through ceremonial behaviour; by decoupling the formal aspects of the institution from the use of its technical services, it shields itself from the review of the technical services. The school is a typical example.

In a loosely coupled system, as Weick (1976) described educational organisations, administrative structures and classroom instruction are only loosely connected in such a way that change in one does not necessarily trickle down into the other. Later studies differentiate policy–practice decoupling, that is, when formal requirements are not followed; and means–ends decoupling, that is, when the requirements are followed but have little connection to the ends they are meant to achieve (Bromley & Powell, 2012).

Kenya's competency-based reform offers circumstances conducive to the possibility of means–ends decoupling. Planning documents are externally driven, easily observed and clearly linked to being accountable, while responsive classroom practice is largely invisible and difficult to observe regularly. The form of the documentation requirements for the reform may then become disconnected from the instruction.

The present study overlooks a teacher's intention, but an empirical pattern emerges from the process of decoupling: the indicators in the documentation should be broadly satisfied, internally coherent, demonstrate a weak or no correlation with learner mastery, and be outperformed by indicators of practice that cannot be prepared prior to the lesson.

Integrating the Two Perspectives

Both theories exist at different levels and are complementary to one another. The instructional mechanism that is linked to the learning, as outlined by soc-cultural theory, is the mechanism which the organisation may be monitoring, and institutional theory is addressing the reason. By integration of the two notions, they foretell an association continuum that grows closer to learner mastery the closer to learner task engagement that it moves, starting at the documentary

artefacts and physical environment end and ending at the contingent moves end. This conceptual framework is then operationalised across five practice domains as follows:

Conceptual Framework and Hypotheses

The study identifies five areas of classroom input, which are arranged sequentially from the closest to the furthest from the moment of learning. Documentary preparation includes documents created prior to the lesson which can be shown during the lesson. Learning environment refers to the conditions of the physical setting and displays that are ongoing throughout the lessons. Pedagogical content enactment is the teacher's handling of the mathematical content in the lesson, which involves the representation, sequencing and language of the mathematics. The pedagogical strategy involves the moves that a teacher chooses and implements in the lesson. Classroom management includes items of order, transitions and time use. What the learners demonstrably did with the mathematics during the observation window results in outcome – observed learner numeracy mastery.

The conceptual framework predicts ordering based on the theoretical framework. Domains that follow the learning moment, which are the domains related to the learning action that the teacher takes, should be more strongly related to the mastery domain than the domain that is fixed beforehand. Three hypotheses follow. There is a positive correlation between pedagogical strategy and pedagogical content enactment and the observed mastery of learners' numeracy. R1: After controlling for enacted practice, there is no positive correlation between observed learner numeracy mastery and documentary preparation. When individual practices are considered, those involving responsive support during learner's task engagement contribute more variance in mastery than do prior specification or provision. All three hypotheses involve concurrent association, as predictors and outcome are observed in the same lesson, and as scaffolding is responsive to learner performance, none points to a direction of influence.

METHODOLOGY

Study Design

The study adopted a cross-sectional observational study design with an embedded qualitative strand. Within each lesson, both quantitative indicators and open-ended observer commentary were recorded simultaneously; the qualitative data were used to interpret the quantitative patterns instead of verifying them. The design includes statements about associations and about the relative ordering of predictors. It does not permit causal inference, and no causal language is employed when reporting results.

Setting and Sampling

Observations took place across three counties, which cover the socio-economic and ecological spectrum in Kenya; Nairobi, the most densely provisioned, Nyandarua, a high-potential agricultural county in the central highlands, and Isiolo, a high aridity/semi-aridity county in the north where service provision historically has been weak. A purposive sampling approach was used to select schools from those who had participated in the STEM programme assessment, and one lesson in numeracy, at the lower primary level, was observed in each school (unless this was not possible due to scheduling). A total of 31 lessons were included in the final sample, which involved 26 schools.

Lessons covered the Numbers, Measurement and Geometry strands of the lower-primary mathematics curriculum design, with sub-strands including addition with regrouping, multiplication, number value, telling time, mass, capacity, length and properties of shapes. Two

lessons in the sample addressed pre-numeracy and psychomotor competencies in a pre-primary setting and were retained because the observation protocol and outcome indicators applied without modification.

Instruments

The Classroom Observation Protocol was developed for the STEM programme assessment and organised into five sections: teacher preparation, learning environment, teaching and learning process, classroom management, and final reflections. Each indicator was recorded dichotomously as present or absent, and each was paired with a free-text field in which the observer recorded the evidence supporting the judgement. Two further fields captured recommendations for lesson delivery and for teacher support. Observations were entered on mobile devices with automatic geolocation and timestamp capture.

The one feature that is important for present purposes is the free-text substantiation feature of the instrument. Since observers had to document what they observed, the binary codes can then be crosschecked with the narrative evidence, and a code like guided instruction can be crosschecked for consistency of interpretation across observers. This audit is reflected in the qualitative substantiation part of the Results.

Measures

Outcome: Observed Learner Numeracy Mastery

The outcome index contained the three protocol indicators that were not related to teacher behaviour or classroom conditions, but recorded learner performance: whether learners were consistently answering correctly with various strategies; whether learners were able to generalise the concepts to real-life situations that were not from the textbook; and an assessor-administered check of three to five learners' exercise books to determine whether most learners were correctly answering the questions. The third indicator is a work-sample verification which gives the index an anchor that is a representation independent of the general impression the observer has of the lesson. The three items were scored 1 for present and 0 for absent, averaged, and expressed as a percentage. Inter-item phi-coefficients ranged from .24 to .52, corrected item-total correlations from .42 to .64, and internal consistency was $\alpha = .60$. For a three-item dichotomous index this is acceptable though not high, and results are reported at both index and item level.

Predictors: Five Practice Domains

Forty-one (41) teacher-practice indicators were assigned a priori to five domains based on the protocol's own structure and the conceptual framework set out above. Documentary preparation (nine indicators) covers schemes of work, lesson plans, lesson notes, records of work, digital planning tools, recent numeracy continuing professional development (CPD), socio-emotional learning in planning documents, curriculum design availability and alignment, and supervisory feedback. Learning environment (six indicators) covers wall displays, seating, display of learners' work, resource accessibility, timetable display and socio-emotional learning displays. Pedagogical content enactment (twelve indicators) covers lesson introduction, sequencing, mathematical language, representational progression, feedback and materials. Pedagogical strategy (ten indicators) covers questioning, modelling, monitoring, targeted feedback, experimentation, peer teaching, guided discovery and guided instruction. Classroom management (five indicators) covers control, work checking, transitions, time use and orderliness.

Domain scores were computed as the percentage of constituent indicators recorded as present. Internal consistency varied considerably. Documentary preparation cohered well ($\alpha = .81$), pedagogical content enactment and pedagogical strategy moderately ($\alpha = .59$ and $.50$), while learning environment ($\alpha = .26$) and classroom management ($\alpha = .13$) did not. The latter two are therefore treated as formative indices, in which the constituent indicators define rather than reflect the construct, and no reflective-scale interpretation is placed on them. Item-level analysis is reported alongside domain-level analysis throughout for this reason.

Covariates

The following three covariates were available: teacher years of teaching experience, total class enrolment and county. The gender of the teacher was noted but not used as no male teachers were observed, 30 of which were female.

Data Collection

Observations were made in June-July 2025 by 13 trained observers. The observers were trained education specialists who were purposefully recruited for the assessment of the programme and explained the protocol beforehand. There was uneven workload with one observer completing 8 lessons and seven observers completing 1 lesson each. This distribution gives rise to a possible concentration of judgement in observer, which is not assumed away but handled analytically.

Analytic Strategy

There were five stages of analysis. Construct-level tests of H_1 and H_2 were done through correlating domain scores with the outcome index. Second, we performed point-biserial correlation and Cohen's d for each of the 41 practice indicators against the outcome, correcting the false discovery rate using Benjamini-Hochberg across the entire screening family. Lessons with fewer than three in each of either group were not screened because estimates would be unstable. Third, multivariable ordinary least squares models were modeled both at domain level and for the leading indicator, with or without the covariates.

Fourth, the result was dichotomised at the maximum score of full mastery and analysed using Fisher's exact test because the distribution of the index does not meet the normality assumptions. Fifth, three robustness procedures were used: observer-clustered standard errors, bias-corrected bootstrap confidence intervals (using 10,000 resamples), and leave-one-observer-out re-estimation. The screening stage is explorative and is reported as such. There are 41 tests for 31 observations, and a serious risk of false discovery arises, and the main defense is the false discovery rate correction.

The entire screening table appears in appendix A for reader reference so one can judge the problem of multiplicity instead of just its surviving indicators. The models of multiple variables are deliberately limited. The models are not used to test confirmatively whether the screening result would change if there were other plausible confounders, and no model contains more than 5 predictors with 31 observations. The p values and confidence intervals in the models including guided instruction are descriptive and may be inflated due to selection for modelling being guided instruction. The inferential power of the categorical analysis and the robustness procedures is the primary. The analyses were performed in Python version 3.11 with the library, pandas, statsmodels and SciPy.

Ethical Considerations

Observations were conducted with the authorisation of county education authorities and with the consent of head teachers and observed teachers. No learner-identifying information was collected, and learners' exercise books were inspected for task completion without recording names. Teacher and school identifiers have been removed from the analytic dataset. The observation was formative in purpose, and results were returned to schools as recommendations rather than as evaluative judgements of individual teachers.

RESULTS

Sample Characteristics

Table 1 summarises the sample. Class enrolment averaged 48.5 learners (SD = 22.8, range 10 to 84), with an approximately gender split (boys M = 25.5; girls M = 23.0). Teachers averaged 11.4 years of experience (SD = 8.1, Mdn = 10, range 2 to 37). The observed teaching workforce was almost entirely female.

Table 1: Sample Characteristics (N = 31 Lessons in 26 Schools)

Characteristic	Value
Counties	Nairobi (n = 12), Nyandarua (n = 10), Isiolo (n = 9)
Schools	26
Observers	13
Class enrolment, M (SD)	48.5 (22.8)
Enrolment range	10 to 84
Boys enrolled, M	25.5
Girls enrolled, M	23.0
Teaching experience in years, M (SD)	11.4 (8.1)
Teaching experience, Mdn (range)	10 (2 to 37)
Female teachers	30 of 31
Strands observed	Numbers, Measurement, Geometry

The Outcome: Distribution and Reliability

The average level of learner numeracy mastery observed was 86.0%, (SD = 28.3). There is a serious negative skewness in the distribution (skew = -2.09): 23 out of 31 lessons with a score at the ceiling, five with a score of 66.7%, one with 33.3%, and two with zero. The individual item endorsement ranged from 25 to 27 out of 30 valid observations (Table 2).

There are two implications of this distribution to the analysis. It helps to reduce variance and so may underestimate correlations, so correlations reported below will be conservative. It also contradicts the "normality" assumption of the OLS model, hence the categorical analysis presented later is the main inferential result, while the regression models are treated as auxiliary analysis.

Table 2: Outcome Index: Item Statistics and Internal Consistency

Items	Endorsed	Corrected item-total r
Learners consistently produce correct answers using multiple strategies	25 / 30	.42
Learners generalize concepts to real-life situations beyond the textbook	26 / 30	.64
Books check: most learners complete the task correctly	27 / 30	.51
Index (3 items)	M = 86.02, SD = 28.25	$\alpha = .60$

Domain-Level Associations

Table 3 reports domain scores, internal consistency and zero-order correlations with mastery. The pattern supports H_1 and H_2 . Pedagogical strategy correlated most strongly with mastery ($r = .54$, $p = .002$), followed by pedagogical content enactment ($r = .42$, $p = .018$). Learning environment and classroom management showed no significant association. Documentary preparation showed effectively none ($r = .03$, $p = .889$), despite being the most internally consistent of the five domains.

The domains are not independent. Documentary preparation correlated with pedagogical content enactment ($r = .52$) and pedagogical strategy ($r = .43$), indicating that teachers who document more also tend to enact more. The absence of any bivariate association between documentation and mastery is therefore not an artefact of documentation being confined to weak teachers.

Table 3: Practice Domains: Descriptive Statistics, Reliability and Correlation with Observed Learner Numeracy Mastery

Domain	k	α	M	SD	r with mastery	p
Pedagogical strategy	10	.50	72.59	20.51	.540	.002
Pedagogical content enactment	12	.59	82.55	18.07	.423	.018
Learning environment	6	.26	76.88	19.09	.205	.269
Classroom management	5	.13	91.29	12.78	.129	.491
Documentary preparation	9	.81	60.33	28.57	.026	.889

Note. k = number of indicators. Learning environment and classroom management are treated as formative indices; their alpha coefficients are reported for transparency and should not be interpreted as evidence of construct reliability.

Indicator Screening

Thirty-seven of the forty-one indicators met the minimum cell-size criterion for screening. Four survived false discovery rate correction at $q < .05$, and all four belong to the enacted-practice domains (Table 4). Guided instruction was the strongest by a considerable margin. Lessons in which it was observed averaged 97.1% mastery against 54.2% where it was not, a difference of 42.9% and a standardized effect of $d = 2.03$.

Three further indicators fell just outside the corrected threshold and are noted for completeness: display of socio-emotional learning skills ($q = .093$), experimentation as a teaching strategy ($q = .093$) and use of concrete or pictorial representations ($q = .115$). None of the nine documentary preparation indicators approached significance, and four carried negative point estimates. Participation in numeracy CPD in the preceding year showed the largest negative coefficient among documentary indicators ($r = -.26$, $q = .591$), a result discussed in Table 4.

Table 4: Practice Indicators Surviving False Discovery Rate Correction ($q < .05$)

Practice indicator	Present	M	M	r	d	p	q
		present	absent				
Guided instruction used	23 / 31	97.10	54.17	.676	2.03	< .001	.001
Lesson stimulates learner interest	24 / 31	94.44	57.14	.561	1.57	.001	.019
Guided discovery used	22 / 31	95.45	62.96	.531	1.33	.002	.026
Learner responses reinforced	27 / 31	91.36	50.00	.499	1.66	.004	.040

Note. Screening covered 37 indicators meeting the minimum cell-size criterion. q = Benjamini-Hochberg adjusted p value. Full screening results appear in Appendix A.

Multivariable Models

Table 5 reports four regression models. In line with the analytic strategy, these models are exploratory and supplement, rather than replace the screening and categorical results. Model 1 enters all five domain scores and explains 42.2% of variance in mastery, $F(5, 25) = 3.65$, $p = .013$. Two coefficients reach significance. Pedagogical strategy is positive ($b = 0.73$, 95% CI [0.17, 1.30], $p = .013$). Documentary preparation is negative ($b = -0.39$, 95% CI [-0.77, -0.01], $p = .045$). The negative sign is the first indication that documentation is not merely uninformative about mastery but is associated with lower mastery once enacted practice is held constant.

Model 2 enters guided instruction alone and explains slightly more variance than all five domains combined ($R^2 = .457$, adjusted $R^2 = .438$), $F(1, 29) = 24.41$, $p < .001$. Lessons using guided instruction scored 42.93% higher on mastery (95% CI [25.16, 60.71]). Model 3 adds teaching experience, class enrolment and county fixed effects, and explains 60.2% of variance, $F(5, 24) = 7.25$, $p < .001$. The guided instruction coefficient is essentially unchanged ($b = 42.31$, 95% CI [24.99, 59.63]), indicating that the association is not confounded by teacher seniority, class size or geography. Enrolment carries a small positive coefficient ($b = 0.35$, $p = .040$), discussed in the limitations. Teaching experience is unrelated to mastery ($b = 0.21$, $p = .648$), as is county ($F = 0.84$, $p = .440$ in one-way analysis of variance).

Model 4 tests whether guided instruction is simply a proxy for stronger content enactment. Entering both, guided instruction retains a large coefficient ($b = 38.26$, $p < .001$) while pedagogical content enactment does not reach significance ($b = 0.30$, $p = .204$). Cohen's f^2 for the increment attributable to guided instruction is 0.60, conventionally a large effect.

Table 5: Regression Models Predicting Observed Learner Numeracy Mastery

Predictor	M1 b [95% CI]	M2 b [95% CI]	M3 b [95% CI]	M4 b [95% CI]
Guided instruction	—	42.93** [25.16, 60.71]	42.31** [24.99, 59.63]	38.26** [19.20, 57.33]
Pedagogical strategy	0.73* [0.17, 1.30]	—	—	—
Pedagogical content enactment	0.45 [-0.20, 1.10]	—	—	0.30 [-0.17, 0.77]
Learning environment	0.20 [-0.31, 0.71]	—	—	—
Classroom management	-0.07 [-0.82, 0.69]	—	—	—
Documentary preparation	-0.39* [-0.77, -0.01]	—	—	—
Teaching experience	—	—	0.21 [-0.74, 1.17]	—
Class enrolment	—	—	0.35* [0.02, 0.68]	—
County fixed effects	No	No	Yes	No
n	31	31	30	31
R ²	.422	.457	.602	.488
Adjusted R ²	.306	.438	.519	.451
F	3.65*	24.41**	7.25**	13.34**

Note. Unstandardised coefficients with 95% confidence intervals. * $p < .05$, ** $p < .001$.

Categorical Analysis

Because the outcome distribution is ceiling-bound, the primary inferential test dichotomises mastery at the ceiling and uses exact inference. Of 23 lessons using guided instruction, 21 attained full mastery. Of eight (8) lessons not using it, two did. The association is strong and exact (OR = 31.50, Fisher's exact $p < .001$; $\phi = .66$). The Haldane-corrected 95% confidence interval for the odds ratio is wide, [4.44, 223.71], reflecting the small cell counts; the risk ratio is a more stable summary at 3.65.

Table 6: Guided Instruction and Attainment of Full Observed Mastery

Items	Full mastery	Below full mastery	Total
Guided instruction observed	21	2	23
Guided instruction not observed	2	6	8
Total	23	8	31

Note. Fisher's exact $p < .001$; OR = 31.50, 95% CI [4.44, 223.71]; risk ratio = 3.65; $\phi = .66$.

Robustness

Three checks were applied. First, standard errors were clustered by observers to account for the uneven and potentially correlated judgement across the thirteen observers. The guided instruction coefficient remains significant with substantially inflated standard errors ($b = 42.93$, $SE = 15.75$, $t = 2.73$, $p = .006$). Second, bootstrap resampling with 10,000 replications produced

a correlation of .676 with a 95% interval of [.406, .878] and a mean difference of 42.9% with an interval of [16.2, 70.0]. Neither interval includes zero.

Third, the model was re-estimating thirteen times, each omitting all observations from one observer. The guided instruction coefficient ranged from 23.8% to 56.3%, and the association remained significant in every replication (all $p \leq .002$). The lower bound arises when the observer contributing three lessons with atypical coding is removed, and the upper bound when the observer contributing eight lessons is removed. The association is therefore not generated by any single observer's coding tendency. Uptake of guided instruction did not differ substantially by county (Isiolo 66.7%, Nairobi 83.3 %, Nyandarua 70.0%), and mastery did not differ significantly by county, so the association is not a geographical artefact.

Qualitative Substantiation

The free-text entries made by the observers enable the guided instruction code to be checked against the behaviour that it is applied to. If the code was marked as present, the descriptions that follow are consistent in describing scaffolding that is responsive to learners' work. One observer noted that the teacher led the learners on an individual basis and marked their work. Other examples included a teacher suggesting the right option to students in a question, and a teacher walking students through reading the time as quarter past two on a clock face. A fourth reported a series of steps where the students had to watch a demonstration of how to lace blocks and then were asked to make a necklace. The code was used fairly in a familiar type of behaviour.

The negative cases are also very instructive. When no one was guiding them, the observers noted that the transmission was in the absence of responsive support, which was mostly direct instruction and modelling, more demonstration and lecture, and teacher led. The lesson was largely rote learning, according to one observer. Their descriptions place the missing condition not in discovery but in didactic delivery: their contrast is between scaffolded and unscaffolded instruction, not between structured and unstructured lessons, which is important for interpretation.

The reinforcement indicator works in a similar fashion. Clapping, singing, praise and immediate marking were mentioned, but one observer recorded that correct responses were not heartily applauded but incorrect responses were rebuked, and the other that there was a lack of interest in individual responses, favouring, rather, chorus answering. The indicator reflects whether the learner is getting individualized response or if the classroom is affectively warm.

The broader commentary has two aspects that relate to interpretation. Learner-initiated questioning was infrequent, occurring in five of 30 lessons, and was suggested regularly by observers as a way to diversify questioning techniques instead of increasing the amount of documentation. One entry suggested that if a learner was not understanding a strategy, the teacher should switch to a different strategy, and not repeat the same one, which is a very precise description of the guided instruction deficit identified quantitatively. One observer commented that the curriculum designs used were from the 2017 edition and suggested ongoing professional development with classroom support, with a clear distinction between what is an artefact and the capability.

Discussion

Guided Instruction as the Operative Variable

The study establish that one observed and trainable teacher behaviour is more effective at distinguishing lessons that resulted in lower-primary learners achieving mastery in numeracy than any combination of the domains typically monitored. Guided instruction accounted for more variance than the five domain scores combined, was not sensitive to teacher seniority, class size, and county, and remained significant when pedagogical content enactment was controlled for.

The effect of $d = 2.03$ is much larger than the effects usually reported for guided instruction in the classroom and the study is not designed in a way that it can claim that the guided instruction has this effect in the population. There are two aspects of the design that overestimate the figure. The design has two characteristics that overestimate the figure. The consequence is looked at in the same observation window as the predictor; if the teacher is good at scaffolding, the student is likely to be successful, so part of the relation is definitional and not causal. Thus, ordering, rather than magnitude, is the defensible claim in the case of this protocol; guided instruction is the most tightly coupled with learner performance of the practices recorded in this protocol.

This is consistent with broader evidence. Research on professional development in mathematics shows that professional development effects on student achievement travel through the school's mathematics instruction and not through the teacher's knowledge (Lynch et al., 2025), and the explicit instruction literature says that the greatest returns come from responsive treatment of learner error in instruction (Peltier, 2024). Guided instruction, in the gradual release model, is specifically the responsiveness (Frey, 2025). This research builds upon this evidence in a Kenyan lower primary classroom and, crucially, using an assessment that is conducted in the classroom, not in an external exam.

The separation of guided instruction from guided discovery is a further contribution. Both managed to survive corrections but guided instruction was the more powerful and it prevailed in multivariable specification. The qualitative material indicates contextual explanations. The absent condition for guided instruction was didactic transmission, whereas the absent condition for guided discovery was closed questioning within an otherwise supported lesson. The discriminating feature is teacher responsiveness during learner work, not the degree of learner autonomy in problem selection.

The Planning and Enactment Gap

Documentary preparation showed no positive relationship with mastery at bivariate level and a negative relationship in multivariable specification. This is the study's most policy-consequential finding, and it requires careful statement. The negative coefficient does not mean that planning harms learning. It means that among teachers with comparable enacted practice, those with more complete documentation had somewhat lower observed mastery. Three interpretations are available.

The first is substitution. Documentation and enactment compete for a fixed quantity of teacher time and attention, and the marginal hour spent completing records is an hour not spent preparing to respond to learner thinking. Observers' commentary is consistent with this reading, repeatedly identifying method diversification rather than documentation as the improvement priority.

The second is decoupling. Where formal structures are adopted primarily to satisfy inspection, documentary completeness may become a signal of compliance orientation rather than of pedagogical investment. The internal consistency figures are suggestive: documentary preparation was by far the most coherent of the five domains ($\alpha = .81$), which is what one would expect of a set of practices performed as a bundle for a single external audience, whereas the enacted domains cohered weakly, as one would expect of practices selected situationally.

The third is measurement. The documentary indicators record presence, not quality or use. A scheme of work that is present but not consulted scores identically to one that structures the lesson. This interpretation implies that the finding is about the indicator rather than about planning, and it is the interpretation this study can least easily rule out. It is also, notably, an indictment of the indicator as a supervisory instrument, since presence is exactly what school-level supervision records.

The negative point estimate for recent numeracy CPD deserves separate mention. Lessons taught by teachers reporting CPD participation in the previous year showed lower mastery on average, though the difference did not survive correction. Selection is the most plausible explanation, in that teachers identified as needing support are those likely to be nominated for it, and the qualitative record supports this: one observer noted that a teacher recently transferred from upper primary had received only a single short CBC session. The quality and duration of the CPD itself also warrant consideration. Kenyan studies report that teachers remained inadequately prepared for the CBC after the training provided and call for further training that addresses pedagogy and assessment rather than documentation alone (Akala, 2021; Momanyi & Rop, 2019).

Where CPD takes the form of brief orientation sessions concerned largely with curriculum documents and assessment rubrics, rather than sustained work on pedagogical content knowledge, it would not be expected to change the responsive classroom moves that this study finds most closely associated with mastery, since professional development is associated with student achievement chiefly through changes in classroom instruction (Lynch et al., 2025). Because the protocol recorded only whether CPD had occurred, and not its content, duration or delivery model, selection and programme quality cannot be separated in these data. The finding should not be read as evidence that CPD is ineffective, but it does indicate that CPD participation, like documentary completeness, is a poor proxy for instructional quality.

Implications for Policy and Practice

The findings indicate a lack of balance in the use of documentary inspection and observation of enacted practice in supervision. In this sample, the documents which Kenyan lower-primary supervision can be relied on to inspect did not contain positive information on learner mastery, and a behaviour which can be observed in a few minutes of a lesson did. It would be relatively inexpensive to re-allocate a small portion of supervisory attention from record keeping to practical work, since the observation is already taking place.

Guided instruction is an unusually Malleable Object of Professional Development. It is not a philosophy, not just a way of working, it can be modelled and rehearsed, is not dependent on materials that schools in arid counties may not have. This sample showed that about one lesson was missing out of every four. The focus in these classrooms was on whether the instruction was scaffolded or didactic; then, rather than encouraging learner centeredness in abstraction, training should focus on what to do when the learner gets stuck.

The study recommends the addition of a few learner performance indicators for monitoring systems, along with teacher-practice indicators, in the region. The work-sample check employed here (with the assessor checking three to five exercise books as the lesson proceeds) is virtually free and provides a connection to something that the learners have produced. Systems that observe teacher practice can establish the link to learning internally, and systems that build on teacher practice must establish it externally.

Theoretical Implications

The results complicate a common framing in competency-based reform discourse, in which the relevant contrast is between teacher-centred and learner-centred pedagogy and progress consists of movement towards the latter. Guided instruction does not sit cleanly on that axis. The teacher is directive, and the direction is responsive to what the learner is doing. Lessons coded as lacking guided instruction were not insufficiently learner-centred in the usual sense; several featured group seating, songs and manipulatives while the teacher transmitted content without engaging individual reasoning.

This suggests that the productive distinction in these classrooms is between responsive and non-responsive instruction rather than between teacher-centred and learner-centred instruction. Reform messaging organised around the latter distinction may inadvertently direct teacher effort towards the visible markers of learner-centredness, including seating, display and group formation, which showed no relationship with mastery here, and away from the responsive behaviour that did.

Limitations

Sample size of 31 classroom observations were used in analyses reported but not sufficient for fine-grained modelling and 41 indicators were screened for use with this sample which invites false discovery. The study's defenses are the false discovery rate correction, and the reporting of the full screening table in Appendix A; they avert high risk in advance analysis. The multivariable models have the same limitation: If there are fewer than ~6 observations per predictor and the leading indicator is chosen in the screening stage, the coefficients and intervals reported in the multivariable model are descriptive not confirmatory. Results should be considered as hypothesis generating.

The variance is limited by a scope in the outcome distribution, which is not normal. The study focused on the categorical analysis, but the first has not been. The associations reported will probably be weakened. In replication it would be preferable to have a more finely graded outcome, such as an item-level count of the number of tasks completed correctly in the inspected books.

The predictor and outcome variables are measured by the same observer in the same window, raising the possibility of common-method variance, and part of the association is definitional. This can be alleviated by the above-mentioned qualitative audit, which demonstrates that the guided instruction code was used on a recognizable class of behaviour, and the work-sample indicator provides some degree of independence, but there is no way to fully separate the two. What would make the difference is an independent measure of learners' performance.

The schools are not representative of the lower-primary provision in Kenya, as sampling was done within a programme assessment. Inter-rater reliability statistics are not available because no lessons were double-coded; the consequence of this is addressed in the observer-clustered and leave-one-observer-out analyses but is not a substitute for the statistics. The instrument

included two protocol indicators that were repeated and one remarks field to which the label was not the correct one, indicating that it may need to be revised for future use.

Why class enrolment is positive is counterintuitive and likely a sampling effect. In this sample, larger classes were found to be in better resourced schools in urban settings, and it is because of this that the enrolment in the schools is probably a proxy for school resourcing. It should not be read as anything but a statement of fact. Lastly the design is cross sectional. There's no evidence here that guided instruction increases mastery. The connection may work either way around; where students are successful teachers may scaffold more easily, and vice versa.

Conclusion

In 31 lower-grade primary mathematics lessons in three counties in Kenya, one classroom technique was identified that set apart lessons in which students demonstrated learning from those in which they did not. The average for teachers who scaffolded learner understanding with prompts, cues and responsive questioning during learner work was 97% mastery; the average for teachers who did not scaffold was 54% mastery. The association was robust to teacher experience and class size, and under observer-clustered inference, and was consistent across all leave-one-observer-out replications. The documents indicate no positive information regarding learner mastery in the Kenyan supervision system where the implementation of supervision was controlled for. The documents do not contain any positive information concerning learner mastery in the Kenyan supervision system, when practice was controlled for.

Practical reading is simple. Of the 100 sample lessons, guided instruction occurs in approximately 25. The percentage of lessons where guided instruction occurs is sufficiently specific to be trainable, cost-effective to observe, and occurs in only about 25 of the 100 lessons in this sample. In relation to learning, these are none of the documentary completeness. The evidence in this study indicates that a slight increase in the efforts spent on supervisory work relative to professional development would result in a positive return.

The claim is still to be substantiated. The size of this study and the number of these indicators being studied, plus the way in which the outcomes and predictors are recorded by one observer, results in a hypothesis and not an estimate of the size of the study. The hypothesis is explicit and amenable to testing: responsive scaffolding during learner task engagement is the most likely classroom practice to have a strong claim on foundational numeracy outcomes in lower-primary classrooms in Kenya, while neither documentary compliance nor the visible indicators of learner-centred pedagogy is a viable proxy. For testing, a larger sample would be needed along with learner performance measured separately and preferably an experiment that involved the manipulation of guided instruction instead of its observation.

REFERENCES

- African Union Commission, & UNICEF. (2024). Report on the Conference on Foundational Literacy and Numeracy and the End Learning Poverty for All in Africa campaign. African Union.
- Akala, B. M. (2021). Revisiting education reform in Kenya: A case of Competency Based Curriculum (CBC). *Social Sciences & Humanities Open*, 3(1), Article 100107. <https://doi.org/10.1016/j.ssaho.2021.100107>
- Akongo, S. R., Ariapa, M., & Giacomazzi, M. (2025). What works in 21st century skills education in sub-Saharan Africa: A systematic review. *Frontiers in Psychology*, 16, Article 1619154. <https://doi.org/10.3389/fpsyg.2025.1619154>
- Bromley, P., & Powell, W. W. (2012). From smoke and mirrors to walking the talk: Decoupling in the contemporary world. *Academy of Management Annals*, 6(1), 483–530. <https://doi.org/10.5465/19416520.2012.684462>
- Frey, N. (2025, October). Revisiting the rules of gradual release of responsibility. *Educational Leadership*. ASCD.
- Kenya Institute of Curriculum Development. (2023). *Basic education curriculum framework*. KICD.
- Lynch, K., Gonzalez, K., Hill, H. C., & Merritt, R. (2025). A meta-analysis of the experimental evidence linking mathematics and science professional development interventions to teacher knowledge, classroom instruction, and student achievement. *AERA Open*, 11. <https://doi.org/10.1177/23328584251335302>
- Maeresera, J., Mutseekwa, C., & Muyengwa, B. (2024). Co-teaching as a strategy of implementing a competence-based curriculum: A case of two teacher training colleges in Rwanda. *African Journal of Teacher Education*, 13(2). <https://doi.org/10.21083/ajote.v13i2.7565>
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83(2), 340–363. <https://doi.org/10.1086/226550>
- Momanyi, J. M., & Rop, P. K. (2019). Teacher preparedness for the implementation of competency-based curriculum in Kenya: A survey of early grade primary school teachers' in Bomet East Sub-County. *The Cradle of Knowledge: African Journal of Educational and Social Science Research*, 7(1).
- Mutai, E. C., Chebet, E., & Oranga, J. (2024). Competency-based curriculum implementation and teacher preparedness in public schools in Moiben Sub-County, Uasin Gishu County, Kenya. *Journal of Research Innovation and Implications in Education*, 8(3), 356–365. <https://doi.org/10.59765/kzxwy84>
- Nketsia, W., & Opoku, M. P. (2024). Editorial: Transforming teacher education in sub-Saharan African countries towards the achievement of Sustainable Development Goal 4 by 2030. *Frontiers in Education*, 9, Article 1426730. <https://doi.org/10.3389/feduc.2024.1426730>

- Pearson, P. D., & Gallagher, M. C. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology*, 8(3), 317–344. [https://doi.org/10.1016/0361-476X\(83\)90019-X](https://doi.org/10.1016/0361-476X(83)90019-X)
- Piper, B., Destefano, J., Kinyanjui, E. M., & Ong’ele, S. (2018). Scaling up successfully: Lessons from Kenya’s Tusome national literacy program. *Journal of Educational Change*, 19, 293–321. <https://doi.org/10.1007/s10833-018-9325-4>
- Piper, B., Ralaingita, W., Akach, L., & King, S. (2016). Improving procedural and conceptual mathematics outcomes: Evidence from a randomised controlled trial in Kenya. *Journal of Development Effectiveness*, 8(3), 404–422. <https://doi.org/10.1080/19439342.2016.1149502>
- Peltier, C. (2024). Explicit instruction: What students with disabilities need most. *Oklahoma Education Journal*, 2(3), 21–25.
- UNESCO. (2024). *Learning counts: Spotlight on basic education completion and foundational learning in Africa*. UNESCO Global Education Monitoring Report & Association for the Development of Education in Africa.
- UNESCO. (2025). 2025 Spotlight on basic education completion and foundational learning in Africa: Lead for foundational learning. UNESCO Global Education Monitoring Report.
- Uwezo. (2016). Are our children learning (2016)? Uwezo Kenya sixth learning assessment report. Twaweza East Africa.
- UNICEF. (2025). Foundations first: Insights from the Global Foundational Learning Action Tracker 2025. UNICEF.
- van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, 22(3), 271–296. <https://doi.org/10.1007/s10648-010-9127-6>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.
- Weick, K. E. (1976). Educational organizations have a loosely coupled systems. *Administrative Science Quarterly*, 21(1), 1–19. <https://doi.org/10.2307/2391875>
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>
- World Bank. (2024). Teacher practices in Indonesia: Results of the Teach Primary classroom observation study. World Bank.

Appendix A. Full Indicator Screening Results

All 37 practice indicators meet the minimum cell-size criterion, ordered by uncorrected p value. Indicators with fewer than three lessons in either category were excluded.

Table A1: Screening of All Practice Indicators against Observed Learner Numeracy Mastery

Indicator	n	Pres. %	M present	M absent	r	d	p	q
Guided instruction used	31	74.2	97.10	54.17	.676	2.03	< .001	.001
Lesson stimulates learner interest	31	77.4	94.44	57.14	.561	1.57	.001	.019
Guided discovery used	31	71.0	95.45	62.96	.531	1.33	.002	.026
Learner responses reinforced	31	87.1	91.36	50.00	.499	1.66	.004	.040
SEL skills displayed	31	80.6	92.00	61.11	.439	1.20	.013	.093
Experimentation used	30	70.0	93.65	66.67	.440	1.03	.015	.093
Concrete or pictorial representations	31	74.2	92.75	66.67	.411	1.00	.022	.115
Context-relevant realia used	30	76.7	91.30	71.43	.300	0.72	.107	.496
Lesson notes in use	29	75.9	89.39	71.43	.270	0.63	.157	.591
CPD in numeracy (past year)	30	36.7	78.79	92.98	-.258	-0.54	.169	.591
Feedback on learners' questions	29	79.3	89.86	72.22	.252	0.62	.188	.591
Learning resources accessible	31	74.2	89.86	75.00	.234	0.53	.205	.591
Variety of questions posed	31	90.3	88.10	66.67	.228	0.77	.217	.591
SEL captured in planning documents	30	60.0	90.74	77.78	.226	0.46	.230	.591
Concrete-semi-abstract-abstract progression	31	83.9	88.46	73.33	.200	0.54	.280	.591
Digital tools in planning	30	13.3	100.00	83.33	.201	0.58	.286	.591
Concepts linked to real life	29	65.5	91.23	80.00	.199	0.41	.301	.591
Peer teaching used	30	40.0	94.44	85.19	.195	0.39	.303	.591
Hints, prompts, targeted feedback	30	83.3	88.00	73.33	.194	0.51	.304	.591
Learner work monitored during task	30	90.0	90.12	77.78	.159	0.52	.402	.743
Multiple solution methods modelled	30	86.7	87.18	75.00	.147	0.42	.438	.771
Learners' work displayed	31	58.1	88.89	82.05	.121	0.24	.515	.850
Class timetable displayed	31	83.9	84.62	93.33	-.115	-0.31	.537	.850
Schemes of work in use	30	83.3	85.33	93.33	-.106	-0.28	.576	.850
Lesson plans in use	30	83.3	88.00	80.00	.106	0.28	.576	.850
Supervisory feedback present	31	67.7	84.13	90.00	-.099	-0.21	.597	.850
Links to prior knowledge	31	77.4	84.72	90.48	-.087	-0.20	.643	.871
Record of work covered	30	70.0	84.13	88.89	-.078	-0.16	.684	.871
New concepts modelled or explained	31	71.0	84.85	88.89	-.066	-0.14	.724	.871
Wall displays of learning materials	31	71.0	84.85	88.89	-.066	-0.14	.724	.871
Accurate responses to learners	30	86.7	88.46	83.33	.066	0.19	.730	.871
Additional time with selected learners	30	73.3	84.85	87.50	-.042	-0.09	.827	.916
Correct mathematical language	30	90.0	85.19	88.89	-.039	-0.13	.836	.916
Smooth activity transitions	31	90.3	85.71	88.89	-.034	-0.11	.857	.916
Curriculum design available and aligned	30	56.7	88.24	89.74	-.032	-0.06	.866	.916
Learner work checked in lesson	30	83.3	85.33	86.67	-.018	-0.05	.926	.952
Learning outcomes / KIQs stated	31	77.4	86.11	85.71	.006	0.01	.975	.975

Note. q = Benjamini-Hochberg adjusted p value across the family of 37 tests. Four indicators with insufficient variation were excluded: collaborative seating, class control maintained, effective use of time, and clean and organised learning space. SEL = socio-emotional learning; CPD = continuing professional development; KIQ = key inquiry question.