

Multiple methods in quadratic-equation teaching: Same-lesson correspondence and discordance between coded material opportunity and observed student activity across eight sites

Jonas A. Mandalunes^{1,2}

jonas.mandalunes24414@oed.com.ph

AMA Computer University

Research Article

Keywords: quadratic equations, multiple solution methods, instructional materials, classroom observation, international comparative education

Posted Date: August 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10682079/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Instructional materials can make multiple solution methods available, but they do not establish what students do in the classroom. We examined same-lesson correspondence and bidirectional discordance between coded material opportunity and observed multiple-approach student activity in public OECD Global Teaching InSights data on quadratic-equation teaching. Following the Technical Report, we linked materials to videos by site, school, teacher, instrument, and exact date across eight participating sites. The sample comprised 1,270 lesson-days from 662 teachers; 608 teachers contributed two lessons. We estimated four correspondence states and a logistic model with site and filmed-occasion fixed effects and teacher-clustered uncertainty. Coded material opportunity occurred in 38.9% of lessons and observed activity in 35.5%. Overall, 36.1% were discordant: 19.8% showed opportunity without observed activity and 16.4% showed activity without coded opportunity. Model-standardized observed activity was 47.5% with coded opportunity versus 27.7% without it, a risk difference of 19.8 percentage points (teacher-cluster bootstrap 95% confidence interval 14.5–25.1); the corresponding odds ratio was 2.51 (95% confidence interval 1.95–3.22). All eight site-specific point estimates were positive, and the association remained positive in within-teacher and measurement, content, clustering, observation-coverage, and site-omission analyses. Because video raters could consult materials, the non-equivalent indicators were not fully information-independent; estimates are associational, fixed-site, and not nationally representative. In these lessons, material coding documented mathematical opportunity but did not by itself recover observed student activity, so it should not be treated as interchangeable with direct classroom evidence.

1 Introduction

Students can solve a quadratic equation by factorising, completing the square, using a quadratic formula, reasoning from a graph, or combining these approaches. Access to more than one method matters because mathematical competence includes not only executing a procedure but also recognising alternatives, comparing their properties, and selecting a suitable method (Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2008). A recent systematic review of 38 empirical studies on quadratic equations likewise concluded that flexible knowledge requires connecting multiple solution strategies with conceptual understanding and that instructional evidence remains limited and context-sensitive (Gustafsson, 2026). Classroom work with multiple strategies can support procedural flexibility, yet the benefit depends on the tasks selected, the presentation of the strategies, and the discussion through which similarities and differences become visible (Durkin et al., 2017; Silver et al., 2005). Multiple-method work is therefore both mathematical and didactical: it concerns what ways of solving are made available and what students are observed doing with them.

Instructional materials are one place where such opportunities become visible. Lesson plans, worksheets, slides, textbook pages, and assigned tasks can invite or require more than one method. They do not, however, provide a transcript of classroom activity. A teacher can select, adapt, omit, reorder, or supplement a resource, and students can respond in ways not specified in the submitted materials. Research on teachers' use of curriculum resources accordingly treats materials and classroom teaching as related through use and interpretation rather than as equivalent layers (Ball & Cohen, 1996; Gueudet & Trouche, 2009; Remillard, 2005). The distinction also echoes task research: a task as represented in curricular or material form and the activity that unfolds around it are analytically separable (Henningsen & Stein, 1997; Stein et al., 1996).

This distinction creates a measurement problem for mathematics education. Materials are often more scalable to collect and audit than video, but the inference from a material indicator to observed student activity is rarely exact. A positive relation would show that the material measure carries information about classroom mathematical activity. Discordance would show the limits of treating one evidence layer as a substitute for the other. Both directions matter. A material set may display a multiple-method opportunity without observed multiple-approach student activity in the

sampled video, while students may use multiple approaches when the collected materials do not display that opportunity. Neither state alone establishes implementation failure or improvisation. Sequencing, uncollected resources, interaction, measurement error, and the non-equivalence of the two codes can all contribute.

The OECD Global Teaching InSights study offers an unusual opportunity to examine this relation within the same quadratic-equation lesson. It collected teaching materials and videos under a common protocol in eight participating sites and coded both sources using related, but not identical, multiple-method indicators (OECD, 2020a, 2021). Prior work already covers important parts of this terrain. In the German Global Teaching InSights sample, Herbert and Schweig (2021) validated a broader artefact-based potential-for-cognitive-activation scale against a broader video-based scale at the lesson level and reported a moderate association; that analysis did not isolate the focal multiple-method pair or estimate the two discordant states. Ingram and Smith (2021) described observed multiple approaches across all eight sites and related them to solution-method content coded from lesson artefacts, but did not estimate the exact same-date relation between the focal material-opportunity and student-activity codes. Other Global Teaching InSights analyses compare observer, teacher, and student reports (Cheng et al., 2023) or examine the cross-system validity of student perceptions (Herbert et al., 2022). Against this cumulative background, our contribution is the exact-date, focal-item analysis of correspondence and both directions of discordance across eight fixed participating sites. We make no claim that an unpublished or unindexed analysis cannot exist.

The study asks:

1. How often do coded material opportunity and observed multiple-approach student activity occur together, separately, or not at all in the sampled lessons?
2. How strongly is coded material opportunity associated with observed activity after accounting for the participating site and filmed occasion?
3. How stable is the estimated association across participating sites and under alternative specifications addressing repeated lessons, lesson content, observation coverage, rater rules, and classification thresholds?

The contribution is cumulative and measurement-focused. Substantively, the study tests an assumption built into resource-based evidence: that a coded opportunity for multiple methods can stand in for students' observed engagement with multiple approaches. Analytically, the four-state representation keeps both directions of noncorrespondence visible rather than reducing the relation to one coefficient. We do not estimate a causal effect of materials, trace a particular resource into classroom activity, evaluate fidelity, measure student learning, test measurement invariance, or compare nationally representative populations.

2 Conceptual framing

2.1 Multiple methods as mathematical opportunity and activity

Knowing more than one way to solve an equation creates the possibility of comparing procedures and choosing strategically. In experimental work on linear equations, comparing solution methods supported procedural knowledge and flexibility (Rittle-Johnson & Star, 2007), while direct instruction in multiple strategies and prompts to discover alternatives each improved flexible problem solving (Star & Rittle-Johnson, 2008). Reviews of classroom-based comparison research emphasise that multiplicity alone is insufficient. Teachers must decide which methods and problems to juxtapose, when comparison is productive, and how to focus discussion on similarities, differences, constraints, and efficiencies (Durkin et al., 2017). Classroom studies likewise document the practical demands of eliciting and discussing multiple solutions (Silver et al., 2005).

For quadratic equations, the mathematical alternatives are especially visible. Factorising, completing the square, applying a formula, and identifying roots graphically foreground different structures and representations. Their availability depends partly on curriculum organisation and lesson content. The Global Teaching InSights report documented substantial differences in which approaches were included and how quadratic-equation units were sequenced across the participating sites (OECD, 2020a). This makes the focal topic a useful setting for studying multiple methods, but it also prevents a simple interpretation of absence: a lesson centred on introducing one procedure may reasonably show one approach, whereas another lesson may invite comparison.

We therefore distinguish an opportunity coded from collected materials from student activity coded in the video. The first indicates that a submitted material set permitted or required multiple mathematical methods under the study rubric. The second indicates that students used multiple approaches within at least one rated segment. Both are instructionally meaningful. They are not the same construct, and the analysis does not require them to be interchangeable.

For example, a worksheet may permit $x^2 - 5x + 6 = 0$ to be solved by factorising or by the quadratic formula, while classroom work publicly develops only factorising. Conversely, students may compare graphical roots with symbolic factorisation during discussion even when the submitted worksheet specifies one procedure. These hypothetical cases show why either discordant state is mathematically plausible without identifying its cause.

2.2 Resources, tasks, and classroom activity

Curriculum-use scholarship offers a strong reason not to infer classroom activity directly from a resource. Ball and Cohen (1996) proposed that curriculum materials can support teacher learning and instructional change, while recognising that materials operate through teachers' interpretation and use. Remillard's (2005) review similarly conceptualised curriculum use as a participatory relationship between teacher and resource rather than mechanical delivery. The documentational approach to didactics further distinguishes available resources from the documents teachers develop through schemes of use (Gueudet & Trouche, 2009). Across these perspectives, a material carries possibilities and constraints, but its instructional meaning is not exhausted by what appears on the page.

Research on mathematical tasks makes a related distinction. The features of a task can set up possibilities for reasoning, multiple representations, or solution strategies, yet the cognitive activity sustained during a lesson depends on classroom conditions and interaction (Henningesen & Stein, 1997; Stein et al., 1996). The present study does not observe a single task's trajectory from document to enactment: the released material score aggregates a submitted daily set, and the video code records student activity in segments. Still, the resource–activity distinction provides a defensible interpretation of their relation. Correspondence indicates empirical coupling between two layers; discordance signals that an audit of one layer does not deterministically recover the other.

The four-state representation is therefore more informative than a correlation alone. “Both” and “neither” are concordant states. “Material opportunity without observed activity” and “observed activity without coded material opportunity” are discordant states. We do not label either discordant state a deficit. The substantive question is whether discordance is sufficiently uncommon to treat the layers as substitutes, or sufficiently prevalent to require separate measurement.

The four states distinguish two didactical questions: what the submitted resource set makes mathematically possible and what students are observed doing. We did not fix in advance a threshold below which discordance would be negligible, and no such conventional threshold exists; we therefore report the observed prevalence of each discordant state and treat its consequences as dependent on the intended use. The higher the required accuracy of lesson-level classification, the more consequential a given discordance rate becomes. On this reading, resource-based opportunity

and classroom mathematical activity are better modelled as linked but separate instructional layers than as one layer standing in for the other.

2.3 Contextual recurrence without national comparison

The participating sites differ in curriculum, sampling, lesson content, coding reliability, and the prevalence of the focal indicators (OECD, 2020a, 2021). The study was not designed to produce a representative ranking of countries, and Germany used a convenience sample. We treat site as a fixed context and ask whether the relation is directionally recurring under one protocol. A positive estimate in every site would not establish measurement invariance or universal generality, but it would show that the pooled relation is not created by a single site. Conversely, heterogeneous magnitudes would be expected when baselines and measurement properties differ. Site estimates and leave-one-site-out analyses are thus boundary diagnostics, not league tables.

3 Method

3.1 Data source and study context

We conducted a secondary analysis of the public-use Global Teaching InSights files (OECD, 2020a, 2021). The original study recorded two lessons from a quadratic-equation unit for each participating teacher and collected lesson materials, questionnaires, teacher logs, and student assessments. The present analysis used only the artefact-rating and component-video files. The eight fixed participating sites were Chile, Colombia, England, Germany, Japan, Madrid, Mexico, and Shanghai. Site labels follow the public release; they should not be interpreted as nationally representative populations.

The public archive was acquired through a scripted, hash-checked workflow. The analysis used GTI-Artefact-Data.csv and GTI-VidComp-Data.csv from the public-use data files (OECD, 2020b). The archived bundle and both source files were verified against recorded SHA-256 checksums before analysis. No released respondent identifiers are reported in this article.

3.2 Lesson linkage and analytic sample

The unit of analysis was a videotaped lesson-day. A documentation conflict required explicit resolution. Table 14.10 of the Technical Report identifies TAA as the material set for the first video (TVA) and TAC as the material set for the second video (TVB). Page 9 of the public User Guide assigns the latter instrument labels differently. We followed the Technical Report mapping because it was strongly supported by the released dates: 639 TAA–TVA and 633 TAC–TVB records matched on the full exact-date lesson key, whereas the conflicting TAB–TVB implication produced only 15 exact-date matches.

For each candidate pair, we linked records on site, school, teacher, and exact date and required a one-to-one lesson-level match after rater aggregation. The 1,272 exact-date links included two material sets with no valid focal rating, leaving 1,270 complete lesson records: 638 first-video and 632 second-video lessons. These records came from 662 teachers in 612 schools. Six hundred eight teachers contributed two complete lessons and 54 contributed one.

3.3 Measures

Coded material opportunity

The material indicator was derived from MUL_MATH_METHOD_RATE. The artefact rubric evaluates the extent to which students are asked to use more than one mathematical method on a task or related set of tasks and to understand relationships among methods (OECD, 2021, Annex C). Rubric level 1 indicates that one method is specified or implied;

level 2 that different methods may be used; and level 3 that multiple methods must be used or compared. We retained only released rubric values 1, 2, and 3, averaged all valid ratings for the lesson's material set, and classified a lesson as displaying coded material opportunity when the mean was at least 2.

The collected set could include lesson plans, slides, textbook pages, worksheets, assignments, homework, quizzes, and other submitted resources. Teachers did not annotate exactly how every item was used. The indicator therefore describes a coded opportunity in the submitted materials, not teacher intention, student receipt, or use of a particular task.

Observed multiple-approach student activity

The video indicator was derived from CCE2MA ratings for up to eight approximately 16-minute segments. Under the video rubric, level 1 indicates one approach or no evidence of multiple approaches; level 2 brief use of a second approach by at least one student; level 3 a second approach in some depth; and level 4 generally two approaches, or more than two in some depth (OECD, 2021, Annex A). We retained only rubric values 1 through 4. Within each segment, we averaged valid rater scores. We then used the maximum segment mean for the lesson and classified observed activity as present when that maximum was at least 2. This aggregation respects the rubric requirement that multiple approaches occur within a segment rather than being accumulated across separate parts of the lesson.

The video code focuses on students' use of approaches, whereas the material code concerns an opportunity or requirement in a submitted set. Video raters were permitted to consult lesson materials. Accordingly, the indicators are related but non-equivalent and not fully information-independent; the analysis is not an independent criterion-validation study.

Lesson content and recording diagnostics

The locked sensitivity covariates described whether the same material set included four quadratic-equation method contents: completing the square, factorising, using the quadratic formula, and finding roots. Further diagnostics used the number of valid video segments and mean recorded lesson duration. These covariates test whether the association persists under reasonable content and observation-coverage adjustments; they do not eliminate confounding.

3.4 Statistical analysis

First, we estimated the count and percentage of four mutually exclusive lesson states: neither indicator, observed activity only, material opportunity only, and both indicators. We also reported the marginal prevalence of each indicator and the combined proportion of discordant lessons.

Second, we fitted an unweighted logistic regression predicting observed activity from coded material opportunity, with fixed indicators for the eight sites and the first versus second filmed occasion. We treated the participating sites as fixed contexts rather than a random sample of countries. Most teachers contributed two records, so the primary covariance was a teacher-clustered CR1 sandwich with 662 clusters. From the fitted model, we standardized predicted probabilities under opportunity present and opportunity absent over the observed site-by-occasion distribution. We calculated the risk difference and used 5,000 deterministic teacher-cluster bootstrap resamples within site for its confidence interval. All tests were two-sided; effect sizes and intervals, rather than a significance threshold, guide interpretation.

The target of inference requires explicit statement because the analysis is unweighted. The estimand is the conditional association among the realized 1,270 complete, exact-date lesson-days contributed by 662 teachers in the eight

participating site samples. Each complete lesson-day receives equal weight, so the four-state percentages describe that realized set of records, and the standardized risk difference averages fitted probabilities over its observed site-by-occasion distribution. Sites and teachers therefore contribute in proportion to the analysable lesson-days they supplied. These quantities are not national parameters, cross-national contrasts, equally weighted international averages, or population-average teacher or lesson parameters.

Two features of the public release govern that choice. First, the artefact and video-component rating files contain no lesson-level, teacher-level, or school-level sampling weight, and no stratum, primary sampling unit, or replicate-weight variable; we inspected the released variable inventories of both files and found none. No released weight therefore identifies a population parameter for this lesson-level estimand, and we did not construct one, because a self-invented weight would imply a sampling frame the release does not document. Second, the original sampling was not designed to support national inference for these records. Participation was voluntary, Germany used a convenience sample, and the Technical Report documents that random selection within participating schools did not operate in every context (OECD, 2021). Site fixed effects prevent between-site differences in baseline prevalence from driving the pooled coefficient, but they do not create representativeness. Accordingly we report an explicitly bounded, design-consistent association with teacher-clustered uncertainty rather than a weighted population estimate, and we do not apply plausible-value or replicate-weight machinery, which would be appropriate for latent achievement estimands rather than for the lesson observation classifications analysed here.

Third, we inspected crude site-specific odds ratios and tested the seven site-by-opportunity interaction terms jointly using a cluster-robust Wald test. Because most teachers contributed two lesson-days, site-specific uncertainty was also estimated in a way that respects that dependence rather than treating lessons as independent. For each site we report a 5,000-replicate teacher-cluster bootstrap percentile interval as the primary interval, because resampling whole teachers preserves the one-lesson or two-lesson cluster structure and does not rely on a normal reference with only 50 to 95 teacher clusters per site. As an analytic check we also refitted each site-specific crude model with a teacher-clustered CR1 sandwich covariance, the same estimator used for the pooled model. The conventional independent-cell interval from the site two-by-two table is reported in the supplement for comparison only; it is not used for inference. With only eight sites, the interaction test has limited power and cannot establish equality of slopes. We also fitted a conditional logistic model to teachers contributing both lessons. This model controls all stable teacher and site attributes but not changing lesson-level confounding.

Sensitivity analyses changed one feature at a time: clustering by school; bootstrapping schools within site; adjusting for the four content indicators; adjusting for valid segment count and duration; requiring activity in at least half of valid segments; analysing each filmed occasion separately; classifying both indicators by any rater, all raters, or a higher aggregate threshold; restricting sites by published reliability diagnostics; excluding Germany; and omitting each site in turn. Rater agreement was summarised for lessons with exactly two valid raters using percent agreement and Fleiss's kappa.

3.5 Reproducibility and computational checks

The analyses were developed after access to the public data and were not preregistered. Before preparation of this manuscript, the executable workflow, estimands, sensitivity suite, result ledger, and verification checks were frozen. The scripted workflow validates archive and file hashes, applies exact rubric whitelists, audits the record linkage, constructs the paired lesson file, estimates all models, and generates machine-readable tables and figures. A second program, written separately from the analysis script, rereads the source files using different aggregation code and a different numerical optimizer; it passed all locked checks. The reported results were also reproduced by a separate implementation in R, which recovered the primary coefficient, clustered standard error, standardized probabilities, and

within-teacher conditional model. Both cross-checks were implemented by the author rather than by an independent third party, so they establish computational reproducibility rather than external replication. A clean rerun with 5,000 teacher and 5,000 school bootstrap replicates reproduced all tables and figures byte for byte, apart from the generated timestamp in the ledger. The complete workflow was additionally re-executed from the freshly downloaded public archive, whose two source files matched their recorded SHA-256 checksums, and this reproduced every derived output byte for byte.

The reported results were produced with CPython 3.14.3 on macOS (arm64) using numpy 2.4.4, pandas 3.0.2, and scipy 1.17.1, the versions pinned in the supplied environment specification. All resampling uses fixed integer seeds, and per-site bootstrap seeds are derived deterministically so that results do not depend on process-level hash randomisation.

OpenAI ChatGPT 5.6 Sol was used to assist with drafting and code review. It was not used to generate data, results, or citations. All reported analyses were produced by archived deterministic scripts and reproduced as described above, and every citation was checked by the author against the cited source. The author reviewed and takes responsibility for the study design, source use, analyses, interpretations, citations, and manuscript text.

4 Results

4.1 Sample and four lesson states

Table 1 describes the sample by participating site. Counts ranged from 96 lessons in Germany to 176 each in Japan and Mexico. Marginal prevalence varied substantially: coded material opportunity ranged from 23.4% in England to 67.7% in Germany, and observed activity ranged from 19.9% in Mexico to 53.3% in Shanghai. These descriptive differences are one reason the pooled analysis includes fixed site indicators and does not interpret site coefficients as national comparisons.

Table 1
Analytic sample and indicator prevalence by participating site

Site	Lessons	Teachers	Schools	Coded material opportunity, n (%)	Observed activity, n (%)
Chile	156	90	90	60 (38.5)	40 (25.6)
Colombia	162	83	83	50 (30.9)	36 (22.2)
England	167	85	78	39 (23.4)	70 (41.9)
Germany	96	50	38	65 (67.7)	49 (51.0)
Japan	176	89	73	87 (49.4)	86 (48.9)
Madrid	168	85	70	51 (30.4)	45 (26.8)
Mexico	176	95	95	62 (35.2)	35 (19.9)
Shanghai	169	85	85	80 (47.3)	90 (53.3)
Total	1,270	662	612	494 (38.9)	451 (35.5)

The four lesson states appear in Table 2 and Fig. 1. Neither indicator was present in 568 lessons (44.7%), and both were present in 243 (19.1%). The two discordant states together accounted for 459 lessons (36.1%): 208 lessons (16.4%) displayed observed activity without coded material opportunity, while 251 (19.8%) displayed material opportunity without observed activity. The raw proportion showing observed activity was 49.2% among lessons with

coded opportunity and 26.8% among lessons without it. These raw percentages should not be confused with the model-standardized probabilities below.

Table 2
Same-lesson material-opportunity and observed-activity states

Coded material opportunity	No observed activity	Observed activity	Total
Absent	568 (44.7%)	208 (16.4%)	776 (61.1%)
Present	251 (19.8%)	243 (19.1%)	494 (38.9%)
Total	819 (64.5%)	451 (35.5%)	1,270 (100.0%)
Note. Cell percentages use all 1,270 complete exact-date lessons as the denominator.			

4.2 Primary association

Lessons with coded material opportunity had a higher model-standardized probability of observed multiple-approach activity than lessons without it: 47.5% versus 27.7%. The standardized risk difference was 19.8 percentage points (teacher-cluster bootstrap 95% confidence interval [CI] 14.5–25.1). In the site- and occasion-fixed logistic model, the corresponding odds ratio (OR) was 2.51 (teacher-cluster-robust 95% CI 1.95–3.22, $p < .001$).

All eight crude site estimates were positive, ranging from OR 1.29 in Mexico to 3.55 in Japan (Fig. 2). The cluster-robust interaction test did not detect heterogeneity, Wald $\chi^2(7) = 5.54$, $p = .594$. This low-powered result is not evidence that the site associations are equal or that the measures are invariant. Using teacher-cluster bootstrap intervals, Mexico was the only site whose interval included 1; the intervals for the other seven sites excluded 1. The teacher-clustered CR1 intervals agreed with that classification, as did the superseded independent-cell intervals reported in the supplement. These are pointwise intervals without adjustment for eight comparisons, and the lower limits for Colombia, England, and Germany fell only just above 1 (1.004, 1.04, and 1.05 respectively). The site pattern should therefore be read as a directional recurrence diagnostic, not as seven separately established site associations.

4.3 Dependence, content, and measurement sensitivities

Table 3 summarises the principal sensitivity analyses. School-clustered inference was virtually identical to the teacher-clustered primary result, and a 5,000-replicate within-site school bootstrap gave a standardized risk-difference interval of 14.6–25.1 percentage points. The within-teacher conditional model yielded OR 2.20 (95% CI 1.47–3.29, $p < .001$). The conditional estimate was based on 608 teachers with two complete lessons; 249 had different observed-activity outcomes across the two lessons, and 112 were informative about a change in both the predictor and outcome. This model removes stable teacher and site main effects but not time-varying lesson confounding.

Adjustment for four exact-day quadratic-method content indicators reduced the estimate modestly to OR 2.29 (95% CI 1.78–2.96) and a standardized risk difference of 17.6 percentage points. Adjustment for valid video-segment count, and then for both segment count and recorded duration, also preserved the association. A stricter outcome requiring positive activity in at least half of valid segments classified 17.1% of lessons as positive and yielded OR 2.50 (95% CI 1.83–3.41), with a smaller standardized risk difference of 12.8 percentage points.

Results were also positive when each filmed occasion was analysed separately, under any-rater and all-raters classifications, at an aggregate threshold of 2.5, after excluding Germany, in the reliability-restricted analysis, and in every leave-one-site-out model. The leave-one-site-out ORs ranged from 2.32 to 2.71.

Table 3
Primary model and selected sensitivity analyses

Analysis	N	Odds ratio (95% CI)	Standardized risk difference, percentage points (95% CI)
Primary: site and occasion fixed, teacher-clustered	1,270	2.51 (1.95–3.22)	19.8 (14.5–25.1)¶
School-cluster covariance	1,270	2.51 (1.95–3.22)	19.8 (14.4–25.1)¶
Within-teacher conditional logistic model	249 teachers¶	2.20 (1.47–3.29)	Not applicable
Four quadratic-method content indicators	1,270	2.29 (1.78–2.96)	17.6 (12.2–23.0)¶
Valid segment count	1,270	2.44 (1.90–3.14)	18.9 (13.6–24.2)¶
Segment count and recorded duration	1,270	2.42 (1.88–3.11)	18.7 (13.4–24.0)¶
Activity in at least half of valid segments	1,270	2.50 (1.83–3.41)	12.8 (8.4–17.3)¶
<p>Note. CI = confidence interval. ¶ Within-site teacher-cluster bootstrap interval (5,000 replicates). ¶ Delta-method interval; the separate school-cluster bootstrap interval for the primary risk difference was 14.6–25.1 percentage points. ¶ Of 608 teachers with two complete lessons, 249 had discordant observed-activity outcomes and entered the conditional likelihood; 112 also varied on coded material opportunity.</p>			

4.4 Rater agreement

Among 1,246 material sets with exactly two valid raters, binary agreement was 71.8% and Fleiss's κ was .436. Among 1,269 videos with exactly two valid raters, agreement was 69.9% and κ was .356. Agreement varied by site; the Madrid video ratings were especially weak (58.9%, κ = .023). Threshold, rater-rule, reliability-restriction, and site-omission analyses preserved a positive association, but they do not remove measurement error. The agreement results therefore limit the precision and interpretation of both lesson indicators.

5 Discussion

5.1 Association and discordance are simultaneous findings

Coded material opportunity and observed multiple-approach student activity were positively related in the same sampled lesson. After accounting for site and filmed occasion, the standardized probability of observed activity was about 20 percentage points higher when coded opportunity was present. The direction was positive in all eight sites and persisted within teachers and across a broad set of dependence, measurement, content, duration, and omission diagnostics.

This association should not eclipse the second result: more than one third of lessons were discordant. Coded material opportunity was present without observed activity in 19.8% of lessons, and observed activity appeared without coded opportunity in 16.4%. Thus the material indicator was informative but not interchangeable with the video indicator. A material-only audit would miss a substantial portion of observed multiple-approach activity and would also classify as opportunities many lessons in which the sampled video did not show such activity.

The four-state result adds information that a pooled odds ratio cannot. It keeps both forms of mismatch visible and prevents the positive association from being interpreted as deterministic transfer. The states should not be moralised. “Opportunity without observed activity” may reflect sequencing, selective use, a brief event missed by the sampling/coding, or an opportunity that students did not take up. “Observed activity without coded opportunity” may reflect oral interaction, teacher or student contributions, uncollected resources, or limitations of the material code. Because the released data do not link a specific task to a specific observed approach, these explanations remain hypotheses.

5.2 Implications for studying curriculum resources

The findings are consistent with resource-use theories that position teaching between what materials make possible and what classrooms produce (Ball & Cohen, 1996; Gueudet & Trouche, 2009; Remillard, 2005). A submitted material set is neither inert nor self-executing. The positive relation shows that the coded material indicator carries information about observed activity; the bidirectional discordance shows why visible material features cannot be equated with classroom mathematical activity.

This distinction matters methodologically. Material coding can support large or historical audits where observation is impossible, but its estimand should be named precisely: a coded opportunity in collected resources. When the research question concerns students’ mathematical activity, a material proxy requires empirical justification and an explicit error model. Where feasible, studies should sample materials and observation on the same date, represent both discordant states, and avoid reporting only a correlation or mean difference. Future research could trace particular tasks prospectively, record teacher annotations of intended and actual use, and link specific methods in a material to specific student approaches in interaction.

The problem generalises beyond artefact coding to the wider question of which scalable evidence sources can stand in for costly direct observation in large classroom studies. Global Teaching InSights was itself designed partly to trial video- and artefact-based measurement procedures at international scale (OECD, 2020a, 2021), and subsequent work in the same data has tested whether automated multimodal scoring can approximate human teaching ratings (Fütterer et al., 2026). Our result contributes a quantity that such substitution decisions require: not only whether two sources agree on average, but how often each direction of disagreement occurs. A source that is positively associated with observation can still misclassify a substantial minority of lessons in both directions, and the tolerable rate depends on the intended use. Reporting both discordance directions, rather than a single agreement coefficient, is therefore a modest but practical reporting recommendation for studies that propose one measurement layer as a substitute for another.

5.3 Implications for multiple-method work with quadratic equations

The study does not test whether multiple-method activity improved learning. It does, however, show that the opportunity–activity distinction is empirically consequential in a topic where strategy flexibility matters. Research on comparison indicates that students benefit not merely from the co-presence of methods but from carefully designed opportunities to notice, explain, and evaluate relations among them (Durkin et al., 2017; Rittle-Johnson & Star, 2007). A worksheet that permits several methods may not organise comparison; conversely, a classroom discussion may generate alternatives beyond the written task. The moderate association and substantial discordance are compatible with both possibilities.

For researchers and designers, the implication is to specify the instructional layer being claimed. Material design can ask whether alternatives are available, required, juxtaposed, or compared. Observation can ask whether students

propose, use, explain, or evaluate alternatives. These are related stages, not interchangeable labels. The distinction is particularly useful for quadratic equations because methods vary in applicability and in the structures they foreground. A future design could code method identity, problem type, and comparison discourse, allowing analysis of whether a particular material invitation is taken up and how its mathematical purpose is developed.

5.4 What the eight-site design adds

The contextual value of the eight sites is recurrence under a common data-collection and coding protocol, not a country ranking. Baseline prevalence differed widely, but every crude site estimate was positive and every leave-one-site-out analysis retained a positive pooled relation. This makes it less likely that the headline association is an artefact of one site's prevalence pattern. At the same time, eight contexts provide little power for interaction tests, site reliabilities differed, and the samples are not equally representative. We therefore do not infer equality, invariance, or global generality from the non-significant interaction test.

The most defensible international claim is bounded: among these sampled quadratic-equation lessons, a positive relation recurred across varied participating sites while discordance remained material. Replication with other topics, resource systems, and observation instruments is needed before generalising the magnitude.

5.5 Limitations

First, the design is observational. Materials were not assigned, and teachers may select or design them in response to students, content, and lesson goals. Fixed-site, content-adjusted, and within-teacher analyses cannot remove changing lesson-level confounding. The estimates are associations, not effects.

Second, the two measures are non-equivalent. The material rubric describes an opportunity or requirement across a submitted set; the video rubric describes students' activity in a segment. A mismatch can therefore be substantively correct rather than measurement error. Video raters could consult the materials, so the association is also not based on fully independent judgments.

Third, the material set does not establish which document or task was used. No student work was collected, and unsubmitted or oral resources may have shaped activity. The data cannot identify the mechanism of correspondence.

Fourth, both indicators are fallible. Overall binary agreement was moderate, site agreement varied, and Madrid's video kappa was very low. The sensitivity analyses show that the positive direction is not tied to a single threshold or rater rule, but they do not establish measurement invariance or correct attenuation.

Fifth, only two lessons were sampled for most teachers. These records should not be read as a stable teacher trait or complete representation of the quadratic-equation unit. The within-teacher model is useful for stable confounding but rests on a limited window of lesson variation.

Sixth, the site samples do not support national estimates. Participation was voluntary, Germany was a convenience sample, random selection within participating schools did not operate in every context, and the eight fixed sites are not a probability sample of educational systems (OECD, 2021). The public artefact and video-component rating files also contain no sampling weight, stratum, or replicate-weight variable, so no weighted population quantity can be recovered for this lesson-level estimand. Site-specific estimates are diagnostics for this dataset, and the site intervals are pointwise rather than simultaneous.

Seventh, the record-linkage decision resolves a conflict in the public documentation by combining Technical Report definitions with exact released dates. The strong date evidence supports the selected mapping, but the conflict itself remains a data-release limitation that future users should audit.

Finally, the study concerns one mathematical topic and one pair of related codes. Replication is needed before extending the correspondence pattern to other domains or resource features.

6 Conclusion

Across 1,270 exact-date quadratic-equation lessons, the two evidence layers were related but far from interchangeable. The model-standardized probability of observed multiple-approach activity was 19.8 percentage points higher in lessons with coded material opportunity (95% confidence interval 14.5–25.1), and all eight site-specific point estimates were positive. Yet 36.1% of lessons were discordant: 19.8% showed material opportunity without observed activity and 16.4% showed observed activity without coded opportunity. These patterns do not identify a causal effect, implementation failure, or the pathway from a particular resource to student action. They establish a narrower measurement point with a content-specific implication for mathematics education. In these sampled lessons, the material indicator documented coded opportunity for multiple methods but did not by itself recover observed student activity, so a study that needs to characterise students' mathematical activity should not treat this artefact indicator as interchangeable with classroom observation unless the correspondence has been established for the intended use. That conclusion is bounded by one mathematical topic, one pair of non-equivalent codes, and eight non-representative site samples; it does not establish that no artefact measure could ever support calibrated probabilistic claims about classroom activity.

Declarations

Funding

This research received no external funding.

Competing Interests

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics Approval

This secondary analysis used publicly released, de-identified data and involved no new data collection, recruitment, intervention, or contact with participants. Ethics approval was therefore not applicable to this secondary analysis.

Consent to Participate

Not applicable to this secondary analysis. No new participants were recruited or contacted; consent for the original Global Teaching InSights data collection was administered under the OECD study protocol.

Consent for Publication

Not applicable; no identifiable participant information is reported.

Data Availability

This study analysed the publicly released OECD Global Teaching InSights public-use data files, cited in the reference list as OECD (2020b). The two source files used were GTI-Artifact-Data.csv (SHA-256 4fe251ad827eacacdb82e01f3f071a5531e1cf097dbe939d3765b180449912eb) and GTI-VidComp-Data.csv (SHA-256 7a5206b4361d3358dc0f1cbb41073eb7311262068d615f08b2cad1bced67e185). These checksums allow any user to confirm that they have obtained byte-identical source files. No new data were generated. Raw source records and the record-level derived lesson file are not redistributed with this article because reuse is subject to OECD terms; the supplementary material provides the scripted acquisition workflow that retrieves and hash-verifies the files, together with all derived aggregate outputs needed to check the reported results.

Code Availability

The acquisition, analysis, and independent-verification code, environment specification, codebook, aggregate results, and verification record are provided as Supplementary Material. The supplement does not redistribute OECD raw files or record-level derived data.

Author Contributions

J.A.M. conceived and designed the study; acquired, curated, and analysed the data; developed and verified the software; prepared the tables and figures; interpreted the findings; and wrote and revised the manuscript. J.A.M. approved the final manuscript and accepts responsibility for all aspects of the work.

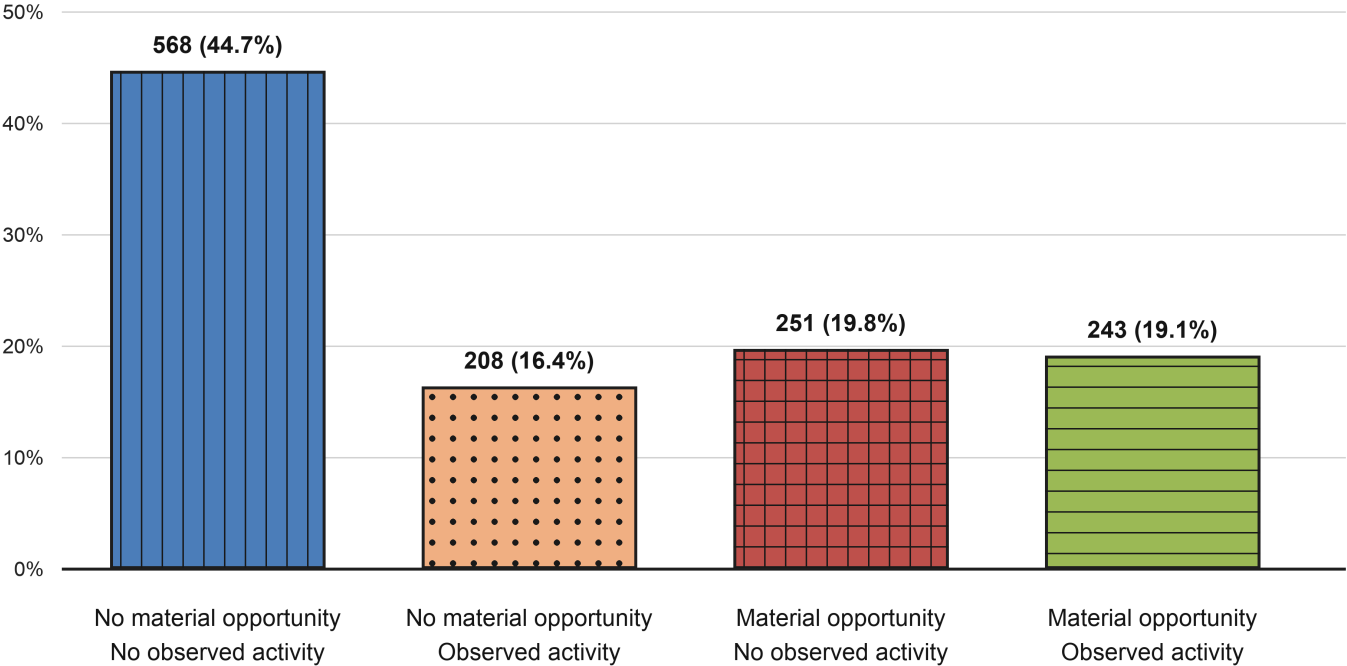
References

1. Ball, D. L., & Cohen, D. K. (1996). Reform by the book: What is—or might be—the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher*, 25(9), 6–8. <https://doi.org/10.3102/0013189X025009006>
2. Cheng, Q., Shen, J., & Zhang, S. (2023). Comparing perceived and observed instructional practices and their predictive power for student mathematics achievement: An analysis of Shanghai data from OECD Global Teaching InSights. *Asian Journal for Mathematics Education*, 2(4), 445–468. <https://doi.org/10.1177/27527263231210322>
3. Durkin, K., Star, J. R., & Rittle-Johnson, B. (2017). Using comparison of multiple strategies in the mathematics classroom: Lessons learned and next steps. *Zdm*, 49(4), 585–597. <https://doi.org/10.1007/s11858-017-0853-9>
4. Fütterer, T., Hou, R., Bühler, B., Bozkir, E., Bell, C., Kasneci, E., Gerjets, P., & Trautwein, U. (2026). Validating automated assessments of teaching effectiveness using multimodal data. *Learning and Instruction*, 101, 102264. <https://doi.org/10.1016/j.learninstruc.2025.102264>
5. Gueudet, G., & Trouche, L. (2009). Towards new documentation systems for mathematics teachers? *Educational Studies in Mathematics*, 71(3), 199–218. <https://doi.org/10.1007/s10649-008-9159-8>
6. Gustafsson, M. (2026). Empirical findings on the teaching and learning of quadratic equations: A systematic review. *Educational Studies in Mathematics*. Advance online publication. <https://doi.org/10.1007/s10649-026-10493-6>
7. Henningsen, M., & Stein, M. K. (1997). Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for Research in Mathematics Education*, 28(5), 524–549. <https://doi.org/10.2307/749690>

8. Herbert, B., Fischer, J., & Klieme, E. (2022). How valid are student perceptions of teaching quality across education systems? *Learning and Instruction*, 82, 101652. <https://doi.org/10.1016/j.learninstruc.2022.101652>
9. Herbert, B., & Schweig, J. (2021). Erfassung des Potenzials zur kognitiven Aktivierung über Unterrichtsmaterialien im Mathematikunterricht [Measuring the potential for cognitive activation via teaching materials in mathematics lessons]. *Zeitschrift für Erziehungswissenschaft*, 24(4), 955–983. <https://doi.org/10.1007/s11618-021-01020-9>
10. Ingram, J., & Smith, K. (2021). Connections made when teaching quadratic equations: Data from the OECD Teaching and Learning International Survey (TALIS) video study. *Proceedings of the British Society for Research into Learning Mathematics*, 41(1), 1–6. <https://bsrlm.org.uk/wp-content/uploads/2021/05/BSRLM-CP-41-1-09.pdf>
11. OECD. (2020a). Global Teaching InSights: A video study of teaching. *OECD Publishing*. <https://doi.org/10.1787/20d6f36b-en>
12. OECD (2020b). *Global Teaching InSights public-use data files* [Data set]. OECD Publishing. Stable archived copy dated 9 March 2021: https://web.archive.org/web/20210309131907id_/http://www.oecd.org/education/school/GTI%20Data%20csv.zip
13. OECD (2021). *Global Teaching InSights technical report*. <https://www.oecd.org/content/dam/oecd/en/about/projects/edu/global-teaching-insights/reports/technical-report/GTI%20Technical%20Report.pdf>
14. Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research*, 75(2), 211–246. <https://doi.org/10.3102/00346543075002211>
15. Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, 99(3), 561–574. <https://doi.org/10.1037/0022-0663.99.3.561>
16. Silver, E. A., Ghouseini, H., Gosen, D., Charalambous, C., & Strawhun, B. T. F. (2005). Moving from rhetoric to praxis: Issues faced by teachers in having students consider multiple solutions for problems in the mathematics classroom. *The Journal of Mathematical Behavior*, 24(3–4), 287–301. <https://doi.org/10.1016/j.jmathb.2005.09.009>
17. Star, J. R., & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and Instruction*, 18(6), 565–579. <https://doi.org/10.1016/j.learninstruc.2007.09.018>
18. Stein, M. K., Grover, B. W., & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. *American Educational Research Journal*, 33(2), 455–488. <https://doi.org/10.3102/00028312033002455>

Figures

Same-lesson material-opportunity and observed-activity states



N = 1,270 complete exact-date lessons; descriptive, unweighted site samples

Figure 1

Distribution of the four same-lesson material-opportunity and observed-activity states. Percentages use all 1,270 complete exact-date lessons as the denominator

Material opportunity–observed activity odds ratios by site

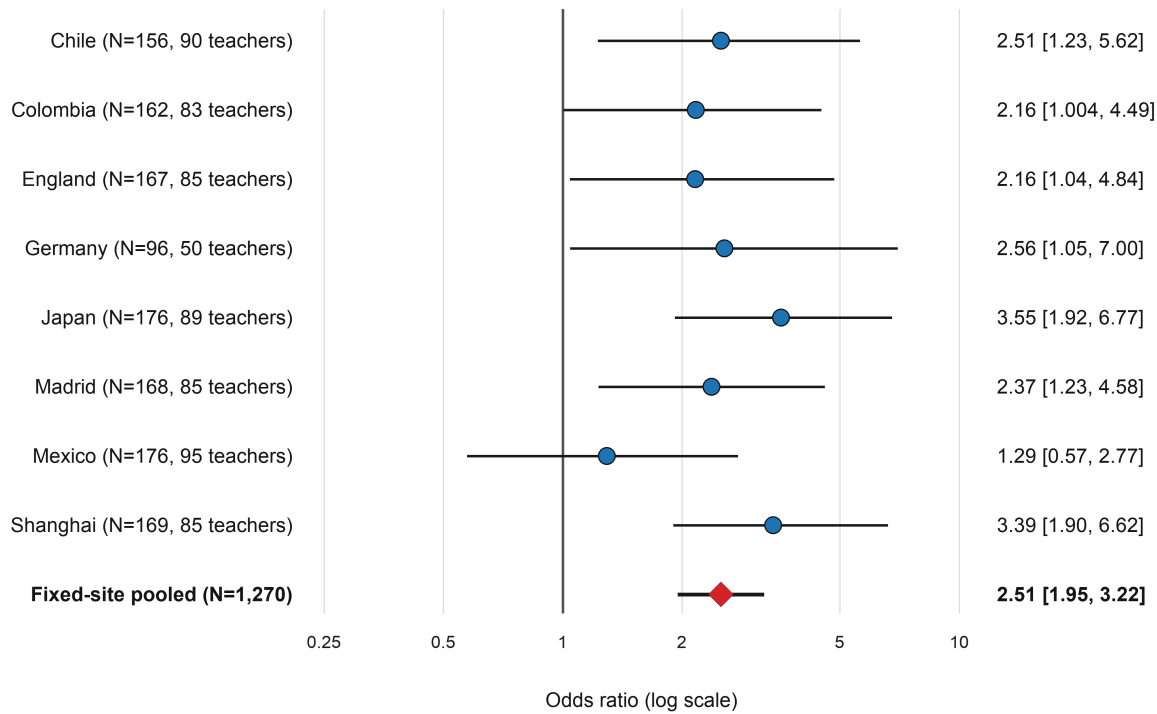


Figure 2

Crude same-lesson odds ratios by participating site and the fixed-site pooled estimate. Site intervals are 5,000-replicate teacher-cluster bootstrap percentile 95% intervals from site-specific models, so they respect the two lesson-days contributed by most teachers; teacher-clustered CR1 intervals give the same classification relative to 1 and are reported in the supplement. The pooled interval is teacher-cluster robust. Intervals are pointwise and are not adjusted for eight comparisons. Estimates are descriptive of the sampled sites, not national effects

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESMReproducibilitySupplement.zip](#)