

Construct and inference alignment in mathematical problem-solving research

Rogen A. Doronila

doronila.rogen@nddu.edu.ph

Notre Dame of Dadiangas University

Research Article

Keywords: Mathematical problem solving, Preservice mathematics teachers, Construct validity, Metacognition, Secondary documentary analysis, Assessment

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11339567/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

Research on mathematical problem solving often connects psychological dispositions to mathematical performance, yet the meaning of these connections depends on how constructs are measured and interpreted. This methodological study examines construct and inference alignment through secondary documentary analysis of a historical doctoral study involving 118 preservice mathematics teachers in the Philippines. Analysis of the methods, results, and concluding chapters maps theoretical labels, operational measures, reported relations, and substantive claims. Inventory-based findings are compared with process summaries from six participants and two published solution episodes. Three interpretive transitions require additional evidence: from dispositions to enacted mathematical cognition, from reported metacognitive awareness to task regulation, and from personal problem-solving performance to teaching competence. The reported reduced-model association between need for cognition and test performance was .864, whereas the conditional association between metacognitive awareness and performance was $-.164$. Neither coefficient directly represents observed reasoning or regulation. Process summaries encompass 42 nested participant-by-task opportunities. Contrasting episodes show how a correct result can emerge through pattern search or through revision of a spatial representation, supplying information absent from inventory scores. An evidence-chain framework makes model-dependent associations a separate evidentiary level, locating changes from measured dispositions to mechanisms and educational uses. A worked mathematical task illustrates how correctness, justification, regulation, and pedagogical response support different judgments. The conclusions are bounded by the published documentary record and do not constitute new model estimation or validation of the framework.

Introduction

Mathematical problem solving requires more than producing a correct answer. A solver must interpret constraints, coordinate representations, select relationships, and judge whether a solution is mathematically adequate. Psychological dispositions also matter to the questions researchers ask about this activity: beliefs about knowledge, interest in demanding thinking, expectations of success, and awareness of strategies may shape engagement (Schukajlow et al., [2023](#); Voica et al., [2020](#)). However, a disposition, an enacted process, and a mathematical product are different objects of evidence. A model connecting their scores becomes educationally informative only when its interpretation preserves these differences.

For preservice mathematics teachers, the distinction has an additional consequence. Solving a problem oneself does not directly establish an ability to interpret a student's reasoning or organize instruction around that reasoning. Research on teacher knowledge for problem solving treats the knowledge and skills involved in teaching as a substantive research domain (Depaepe et al., [2020](#); Lit et al., [2026](#); Weyers et al., [2024](#)). Assessments that name a broad competence while observing only a narrow response risk directing feedback or instructional decisions toward a phenomenon that has not actually been assessed.

Task design can make these differences visible. Yao et al. ([2021](#)) found that success in performing fraction division did not ensure that preservice teachers exhibited conceptual understanding through graphical representation or problem posing. This result concerns a particular content domain, but it demonstrates why the evidence elicited by an assessment matters to the meaning of performance. Similarly, Böswald and Schukajlow ([2023](#)) distinguished preservice teachers' own task value and self-efficacy from their judgments of students' motivation. A response about one's own experience and a judgment about a student are related without being equivalent.

These distinctions are harder to inspect when a statistical diagram uses broad labels such as cognition and metacognition. The labels may identify theoretical interests accurately while obscuring the narrower phenomena represented by the indicators (Lawson & Robins, [2021](#); Schukajlow et al., [2023](#)). A strong coefficient then encourages a broad interpretation even though its evidentiary referent remains an inventory score. Likewise, a negative conditional coefficient may be interpreted as evidence against an educationally valued process although the measure records reported awareness rather than that process in action.

Contemporary evidence makes this a current measurement problem rather than only a historical reporting issue (Flake, [2021](#); Ing et al., [2024](#)). Ing et al. ([2024](#)) reviewed 192 measures of elementary students identified across 1,582 articles and found response-process evidence for only 10% of the measures. That finding concerns their elementary-education corpus, not the prevalence of problems in preservice teacher research, but it establishes that evidence about what respondents actually do remains a substantive concern. Torregrosa and Albarracín ([2026](#)) examined metacognitive processes alongside mathematical skills during sixth-grade pattern-based problem solving, while Lit et al. ([2026](#)) distinguished multiple dimensions of knowledge and skills for teaching problem solving. These studies make the need to connect score meaning, observed activity, and educational use concrete. The present case examines a specific point in that connection: a statistical relation can be reported accurately while its explanation changes the phenomenon represented by the measure. The analysis does not classify these contemporary studies as examples of overreach.

The present study examines this problem through the historical 3P model on problem solving among preservice mathematics teachers (Author, [2012](#)). The dissertation uses 3P as a short name for Philosophical and Psychological Constructs of Problem Solving Performance (Author, [2012](#), p. 169). Its design combines inventories, a mathematics test, structural equation modeling, and individual problem-solving sessions. This combination provides an informative methodological case: the model represented cognition with a need-for-cognition inventory, whereas a separate process component observed mathematical activity. The same archive therefore contains evidence that can expose the consequences of assigning related measures a shared conceptual label.

The purpose is to determine what the reported evidence supports and where interpretation requires an additional warrant. The study is framed as methodological research using secondary documentary evidence. It does not present the historical models as newly estimated or independently replicated, and it

does not treat a persuasive illustration as validation of a new assessment instrument. Its contribution is an explicit account of the transitions between construct definition, measurement, statistical relation, and educational inference in mathematical problem-solving research.

Three research questions guide the analysis. First, how were cognition, metacognition, and problem-solving performance represented in the reported 3P design? Second, what interpretive limits become visible when the inventory-based results are compared with the reported mathematical process evidence? Third, what consequences follow for assessment and research concerning preservice mathematics teachers? The questions address a methodological concern that extends beyond the Philippine cohort: how researchers preserve the meaning of evidence as they move from psychological measurement to claims about mathematical learning and teaching.

Theoretical and analytical framework

Score interpretations and evidentiary warrants

Argument-based validation provides the foundation for this analysis. Kane (2013) locates validation in the interpretations and uses of scores and the assumptions required to move from responses to claims. An instrument cannot supply an unrestricted license to use its score for every conclusion associated with its construct label. Flake and Fried (2020) similarly emphasize transparency about measurement decisions so that readers can evaluate the resulting conclusions. These perspectives make the operational definition central to the interpretation of a statistical result (Flake, 2021; Flake & Fried, 2020; Kane, 2013).

For mathematical problem solving, an operational definition must specify more than an instrument's name. It must identify the response or activity recorded, the task or context in which it occurs, the scoring rule, and the unit of analysis. Correctness on a multiple-choice test, an explanation of why a conjecture holds, a self-report about checking, and a recorded revision can each contribute evidence. Their relationship must be investigated rather than assumed from common terminology (Baumanns & Rott, 2023; Craig et al., 2020; Vorhölter, 2025).

The evidence-chain framework in Table 1 makes four levels jointly inspectable. The first identifies the construct being named. The second identifies the observation or response representing it. The third describes the relation estimated among those measures. The fourth identifies the substantive interpretation or educational use. Alignment means that the meaning established at the measurement level is retained or explicitly justified when a broader claim is made.

Table 1 *Four levels of the evidence chain for mathematical problem solving*

Level	Question	Illustration	Required warrant
Construct	What phenomenon is named?	Mathematical cognition or metacognition	Define the phenomenon independently of its instrument.
Measure	What response or activity is recorded?	Disposition inventory, task solution, or checking episode	Specify the response, scoring, context, and unit.
Statistical relation	What relationship is estimated?	Conditional relation among inventory and test scores	Retain the operational referents and design conditions.
Inference	What conclusion or use is proposed?	Claim about reasoning, regulation, or teaching	Identify the additional evidence needed for a broader claim.

Note. The framework is the present methodological synthesis; its usefulness has not been empirically validated.

Argument-based validation already permits examination of the assumptions connecting responses, scores, and proposed uses (Kane, [2013](#)). The evidence chain organizes that examination for explanations involving several measures and a statistical model; it does not introduce a new validity inference. Its separate statistical level requires the analyst to state the measured predictor, measured outcome, conditioning variables, and specification before assessing the explanation attached to a coefficient. This makes two different problems distinguishable: a change in the phenomenon being explained and a change in the statistical evidence for a relation between the same measures. Table 2 compares the approaches. The added analytical value claimed here is the explicit location of these transitions within the case, not superiority over argument-based validation.

Table 2 *Comparison with argument-based validity reasoning*

Analytical focus	Argument-based validity approach	Added emphasis in the evidence chain
Score meaning	Evaluate support for the proposed interpretation of responses and scores.	Retain the response, task, scoring, and analytical unit beside each construct label.
Relations among measures	Evaluate the assumptions supporting a claim based on scores.	Identify the exact measures, model conditioning, and specification represented by a coefficient.
Explanatory and educational claims	Require evidence for the proposed interpretation or use.	Locate where a score association is recast as an enacted process, mechanism, or teaching outcome.

Note. The comparison is the present synthesis of Kane ([2013](#)). The separate statistical level is an analytical emphasis within validity reasoning, not a replacement for an interpretation and use argument.

The separate level supports three checks. First, it detects a change of referent: a disposition factor becomes a claim about observed reasoning even though its indicators have not changed. Second, it detects an explanatory change: a conditional path is treated as a mechanism despite alternative specifications or untested temporal assumptions. Third, it detects a change of outcome: a relation involving personal task scores becomes a recommendation about teaching effectiveness without a teaching observation. None of these checks establishes that the original relation is false. Each identifies the additional evidence required for the stronger claim.

Distinguishing dispositions and mathematical activity

Need for cognition concerns a disposition toward effortful thinking. Mathematical cognition, as used here, concerns the reasoning represented in interpreting a problem, constructing relationships, choosing representations, and justifying conclusions. A disposition may be associated with such activity, but evidence about the first is not a direct observation of the second. The distinction permits a useful inventory score to retain its legitimate role while preventing its label from absorbing unobserved mathematical processes (Dang et al., [2020](#); Lawson & Robins, [2021](#)).

Metacognitive awareness and enacted regulation also require separate treatment. Awareness inventories capture what participants report about their knowledge or regulatory tendencies (Craig et al., [2020](#); Dang et al., [2020](#); Zepeda & Nokes-Malach, [2023](#)). Task-based evidence can locate planning, monitoring, evaluation, and revision in the mathematical activity itself. Baumanns and Rott ([2023](#)) identified metacognitive behavior through task-based interviews in problem posing; Vorhölter ([2025](#)) analyzed its timing and purposes within group modelling processes. These studies concern different tasks and populations, but they illustrate the explanatory information supplied by observing regulation in context (Baumanns & Rott, [2023](#); Torregrosa & Albarracín, [2026](#); Vorhölter, [2025](#)).

Epistemological and motivational measures require the same discipline. Rott ([2021](#)) examined epistemological beliefs in relation to justification of knowledge, while Voica et al. ([2020](#)) investigated motivation and self-efficacy in problem solving and posing. Such work supports studying relationships among beliefs, engagement, and mathematical activity without collapsing their evidence into an undifferentiated psychological explanation (Rott, [2021](#); Schukajlow et al., [2023](#); Voica et al., [2020](#)). Contemporary literature is used here to interpret the historical case, not to retrospectively confirm its coefficients.

Three analytical categories organize the examination. Construct substitution denotes an extension from the operationally represented phenomenon to a different or broader phenomenon without sufficient evidence. Design overreach denotes an extension from the reported design to an unestablished temporal or causal relation. Transfer overreach denotes an extension from a candidate's personal task performance to instructional competence without directly examining teaching. These categories are interpretive tools, not prevalence measures or a validated coding scale.

Method

Design and documentary corpus

The study uses secondary documentary analysis of Author's (2012) doctoral dissertation, available through ProQuest. The corpus comprises Chapter III on methodology (printed pp. 59–82), Chapter IV on results and interpretation (pp. 83–183), and Chapter V on conclusions and recommendations (pp. 184–194). These chapters contain instrument descriptions, sampling statements, numerical summaries, full and reduced structural models, process-coding summaries, two extended solution transcripts, and educational implications. The study analyzes this published research record without accessing the underlying participant-level dataset or conducting a new round of process coding on original recordings.

Historical values were checked against the dissertation chapters and retained as reported except for explicitly identified percentage recalculations. Source locations use the dissertation's printed pagination, not the page position in a separately uploaded chapter file. The principal locators are the instrument descriptions and reliability table (pp. 67–74, Table 7); descriptive summaries (Tables 18, 23, 27, 31–34); process transcripts (pp. 158–161); process summaries (pp. 162, 166–167, Tables 47–49); and structural-model fit and paths (pp. 172, 174, 177–178, Tables 50–54). Article tables are renumbered independently. These locators make the documentary comparisons traceable but do not permit reproduction of the original models without the underlying data and project files.

The case was selected because its design brings inventory measures and mathematical process observations into the same research account. This is an information-oriented methodological selection. The two extended solution transcripts printed in Chapter IV were both retained for contrastive examination; they were not selected from the original recordings on the basis of the present argument. No claim is made that the case represents the frequency of measurement problems across mathematics education research. The contemporary publications cited in the theoretical framing were selected for conceptual relevance and checked against publisher records; they do not constitute a systematic review corpus.

Reported historical participants and measures

The original analytical sample comprised 118 preservice mathematics teachers, including 55 men and 63 women, from ten public and private teacher education institutions in Region XII, southern Philippines (Author, 2012, p. 83). The results chapter describes purposive selection, whereas the concluding summary describes random selection (p. 185). The more detailed purposive account is followed provisionally. The discrepancy remains unresolved because a sampling frame and selection record are unavailable. Six participants were selected purposively for individual problem-solving sessions.

Table 3 maps the reported constructs and their measures. Epistemological beliefs, curiosity, motivation, metacognitive awareness, and need for cognition were represented through adapted inventories. The

mathematical outcome was a 20-item multiple-choice nonroutine problem-solving test, reduced from 27 items after expert review. The separate process component used seven problems and a four-level rubric for understanding, analysis, exploration, application, representation, calculation, and review or extension.

Table 3 *Historical constructs and operational roles in the 3P study.*

Domain	Variables/indicators	Instrument and role
Epistemic profile	Rationalism; empiricism; metaphorism	Psycho-Epistemological Profile (PEP), 78 items; descriptive classification, not an SEM factor.
Epistemology	Certainty; structure; source; personal control; general control; speed	Epistemological Beliefs Inventory (EBI), 60 items; six SEM indicators.
Curiosity	Epistemic; perceptual; exploration; absorption	Curiosity Inventory; 35 items in the third trial and 30 in final data collection, as retrospectively clarified by the author; four SEM indicators.
Motivation	Goal orientation; task value; self-efficacy	Motivation for Learning Inventory, 30 items; three SEM indicators.
Cognition	Need for cognition; separately, seven observed problem-solving processes	Need for Cognition Inventory, 20 items and one SEM indicator; process rubric used with six cases.
Metacognition	Knowledge about cognition; regulation; separately, process-session self-assessment	Metacognitive Awareness Inventory, 35 items; two SEM indicators; 20-item Self-Assessment Questionnaire in sessions.
Performance	Correct answers on 20-item nonroutine test; seven selected tasks in process sessions	Problem Solving Test; single observed SEM outcome.

Note. SEM = structural equation modeling. Cognition in the SEM is represented by need for cognition; the observed mathematical processes come from six separate process cases. The Curiosity Inventory comprised 35 items in the third trial and 30 items in final data collection; the author retrospectively clarified this transition.

Source: Author ([2012](#)), pp. 68–74 and the model description on pp. 168–171

Instrument development reportedly involved expert review and three trials: 68 preservice teachers at a national teacher education institution in October 2010, 66 at a teacher education institution in November 2010, and 307 in Region XII in January–February 2011. The reported third-trial alpha coefficients ranged from .793 for need for cognition to .946 for the psycho-epistemological profile. These coefficients refer to trial scores; they do not establish score reliability in the analytical sample or the validity of a later use. The Curiosity Inventory contained 35 items in the third trial (Author, [2012](#), p. 68, Table 7) and was subsequently reduced to 30 items for final data collection, as retrospectively clarified by the author.

Although the dissertation reports both lengths, it does not explicitly describe this transition in the methods account (pp. 67–69). The third-trial reliability coefficient for curiosity ($\alpha = .939$) pertains to the 35-item version and cannot be assumed to apply to the final 30-item version. The reasons for deleting the five items and their allocation across subscales remain undocumented in the material examined.

The historical process sessions reportedly included audio and video recording, self-assessment, playback, retrospective interviews, and a focus group discussion. The account reports Cohen's $\kappa = .72$, $p = .015$, but does not specify the agreement unit or pairwise aggregation. That value is treated as a historical report, not an independently established reliability estimate for the present documentary analysis. The original analytical sequence included descriptive summaries, regressions with backward elimination in selected specifications, exploratory product terms, and maximum-likelihood structural equation models fitted in Analysis of Moment Structures (AMOS) to the same 118 participants.

Analytical procedure and checks

The present analysis proceeded by mapping each focal evidentiary chain across the four levels in Table 1. Instrument descriptions supplied the operational meaning; sample and task descriptions supplied the analytical context; reported tables supplied the relations; and the broader interpretations supplied the claims to be assessed. The main focal chains concerned cognition and performance, metacognition and performance, global model fit, and transfer to teaching. The process summaries were examined as a comparison source because they represented mathematical activity through a different mode of observation.

Three kinds of issue were kept distinct. Arithmetic discrepancies concern a recoverable inconsistency between a count and its reported percentage. Documentary inconsistencies concern incompatible descriptions of a procedure or instrument. Inferential mismatches concern claims requiring evidence beyond the described measure or design. Keeping these categories separate prevents an unresolved reporting discrepancy from being presented as proof that a substantive relation is false.

Table 8 pairs the construct description on p. 185, Conclusions 3 and 6 on pp. 192–193, and Recommendations 1–2 on p. 193 with their evidentiary referents. Classification depends on the proposition asserted. A definition broader than its operational measure indicates a construct–measure discrepancy; it does not by itself demonstrate that every associated coefficient was interpreted as that broader process. An explanation asserting an unobserved process or mechanism is a documented explanatory extension. A statement of general empirical validity exceeds in-sample fit if it claims processes or populations beyond those examined. A conditional recommendation remains a proposal unless the source asserts that its educational effect was demonstrated. These rules separate demonstrated extensions, definitional discrepancies, and untested applications.

Arithmetic checking was limited to published counts. The test-score groups contain 45, 37, and 36 participants; dividing by 118 gives 38.1%, 31.4%, and 30.5%, rounded to one decimal place. Process

counts were examined with the denominator of six participants working on seven tasks, or 42 participant-by-task opportunities. No significance tests were recomputed. Reported values printed as $p = .000$ were expressed as $p < .001$. No participant-level models, missing-data analyses, sensitivity tests, confidence intervals, or indirect-effect intervals were generated.

Each comparison identifies the additional observation needed for the stronger claim. The two published episodes were read for representation, conjecture, checking, and final answer, and their mathematical relationships were worked through. This reading concerns visible transitions, without new numerical coding or reconstruction of unrecorded thoughts. Scrutiny of Table 8 can proceed independently of the original dataset: a reader can retrieve each cited passage with its surrounding paragraph, identify the proposition asserted, compare it with the instrument and design description, and judge whether the narrower interpretation retains everything supported. A classification should be revised if surrounding text explicitly limits the claim to score relations, if the proposed mechanism was measured and linked to the outcome, or if an identified teaching outcome supports the transfer. Disagreement should retain the source passage, competing interpretations, and reason for revision. No independent second-reader review or agreement estimate has been completed; the explicit rules and locators support such examination without substituting for it.

Results

Operational differences within the problem solving account

The first result is a difference between the phenomena represented under related labels. In the structural model, cognition was represented by a need-for-cognition inventory. In the process component, mathematical activity was represented by observed understanding, analysis, exploration, application, representation, calculation, and review. The inventory and process rubric therefore supplied different evidence. Their inclusion in one design did not make the structural-model cognition variable a measure of those seven observed processes.

The descriptive record reinforces the need to separate dispositions and products. The problem-solving test had $M = 9.680$, $SD = 2.558$, on a possible range of 0–20, with observed scores from 3 to 17. The original classifications placed 85.6% of participants in the high-motivation group and 74.6% in the high-metacognitive-awareness group. Table 4 retains the corresponding reported means and classifications. These observations describe aggregate score patterns; they do not identify which individuals combined high inventory scores with low performance.

Table 4 *Reported descriptive scores and selected classifications in the historical sample $N = 118$.*

Measure	M	SD	Selected classification
Epistemology, overall	3.53	1.36	35/118 (29.7%) classified sophisticated
Curiosity, overall	4.11	0.90	92/118 (78.0%) classified high
Motivation, overall	4.17	0.83	101/118 (85.6%) classified high
Need for cognition	4.12	0.96	Overall mean classified high
Metacognition, overall	4.00	0.88	88/118 (74.6%) classified high
Problem-solving test (0–20)	9.680	2.558	45 scored 11–17; 37 scored 9–10; 36 scored 3–8

Note. M = mean; SD = standard deviation. Inventory values retain the original scoring and classification rules. Test-score group proportions recalculated from counts are 38.1%, 31.4%, and 30.5%. Descriptive means on different scales are not directly comparable.

Source: Author (2012), Tables 18, 23, 27, and 31–34 pp. 97, 103, 109, 115, 117, and 119–120

The aggregate contrast cannot establish miscalibration or estimate agreement between inventory and test scores. The measures use different scales and classification rules, and the necessary individual cross-tabulations are unavailable. Its methodological significance is more limited: a high self-report classification does not itself warrant a high-performance classification. The same constraint applies to the reported epistemic profiles: 101 participants were classified predominantly rationalist, 17 predominantly empiricist, and none predominantly metaphorist. These categories reflect the original scoring system, not enduring types of mathematical thinker.

Conditional associations and model dependence

The largest reported reduced-model path connected need for cognition to problem-solving test performance, $\beta = .864$, $p < .001$. Its operational referents were a self-report disposition and correctness on selected mathematics items. The coefficient does not identify the representations, justifications, or revisions used by participants. Interpreting it as a direct estimate of enacted mathematical cognition would change the phenomenon represented by the predictor.

The metacognitive-awareness path to performance was negative in the reduced model, $\beta = -.164$, $p = .016$. This finding concerns an association conditional on the model's other relations. It does not compare the effectiveness of observed planning or checking episodes. A negative coefficient could motivate further study of perceived awareness, task difficulty, overlapping predictors, or omitted prior attainment, but the reported record cannot establish which explanation is correct. In particular, it does not show that metacognitive regulation impaired mathematical problem solving.

Model comparison supplies a second limit. Table 5 presents the reported fit statistics, defining the comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA),

and goodness-of-fit index (GFI). Both specifications had favorable reported CFI and RMSEA values. Nevertheless, the reduced model had lower CFI and TLI and higher RMSEA than the full model. Its GFI was .891, below the .90 criterion stated in the source account. Reduction consequently cannot be described as a uniform improvement in fit.

Table 5 *Reported full and reduced structural model fit.*

Fit statistic	Full model	Reduced model
χ^2 (degrees of freedom)	119.57 (106)	132.11 (112)
p	.173	.094
Comparative fit index	.987	.981
Tucker–Lewis index	.984	.977
Root mean square error of approximation	.033	.039
Goodness-of-fit index	.900	.891

Note. Reported estimates are retained without re-estimation. Reduction does not improve every index; both specifications were evaluated in the same historical sample.

Source: Author (2012), Tables 50 and 53 pp. 172 and 177 with degrees of freedom on pp. 170 and 176

More fundamentally, correspondence with the observed covariance structure does not determine whether a variable measures enacted mathematical activity. The strongest legitimate interpretation of global fit remains conditional on the chosen indicators and specification. The reported model was refined and evaluated in the same sample; its fit provides no fresh-data confirmation of the selected paths.

Table 6 retains all eight reduced-model paths reported in the source account, with operational labels made explicit. Paths from epistemology, motivation, and curiosity to need for cognition, and from motivation and curiosity to metacognitive awareness, connect inventory-based representations. Directional arrows reflect model specification rather than observed temporal ordering. In the absence of design assumptions and temporally informative evidence, those paths cannot establish that changing one disposition would produce the reported change in another score or in performance.

Table 6 *Reported standardized paths in the reduced structural model.*

Path in reduced SEM	β	Reported p
Epistemology → need-for-cognition factor	.343	.030
Motivation → need-for-cognition factor	.299	.003
Curiosity → need-for-cognition factor	.271	.008
Need for cognition → metacognitive awareness	.143	.050
Motivation → metacognitive awareness	.513	< .001
Curiosity → metacognitive awareness	.284	.002
Need for cognition → problem-solving test	.864	< .001
Metacognitive awareness → problem-solving test	-.164	.016

Note. β = standardized path coefficient. Operational labels distinguish need for cognition from enacted mathematical cognition. Values printed as $p = .000$ are reported as $p < .001$. The arrows indicate model specification, not established causal ordering.

Source: Author (2012), Table 54 p. 178

The full model illustrates dependence on specification. Its metacognitive-awareness coefficient was $-.163$ with $p = .112$; in the reduced model the coefficient was almost unchanged at $-.164$, but the reported p value was $.016$. The full model also contained a negative curiosity-to-performance path, $\beta = -.211$, $p = .011$, which was removed in the reduced specification. Changes in retained relations alter estimation and uncertainty. These results do not support describing statistical significance in the reduced model as a newly stronger substantive effect.

The reported regression summaries show a related conditional pattern. A performance model retaining need for cognition and regulation of cognition had $R^2 = .633$, with standardized coefficients $.863$ and $-.170$. A total-score model retaining need for cognition and overall metacognition had $R^2 = .630$, with coefficients $.859$ and $-.159$. These are separate specifications fitted to the same cohort, rather than independent replications. The reduced structural model reportedly explained approximately 63% of performance variance and 63.6% of metacognitive-awareness variance in sample. None of these values estimates instructional effectiveness or out-of-sample prediction.

Mathematical process evidence at a different analytical level

The process summaries represent six selected participants completing seven tasks each. Two of three raters had to identify a process for its occurrence to be counted. Analysis, application, and calculation were identified in all 42 opportunities; understanding in 39, representation in 34, exploration in 25, and

review or extension in 25. Table 7 presents the counts and mean rubric ratings. The opportunities are repeated observations nested within participants, not a sample of 42 independent individuals.

Table 7 *Reported process occurrences and rubric means across six participants and seven tasks.*

Observed process	Count of 42	Mean rubric rating (1–4)
Understanding	39	2.77
Analyzing	42	2.78
Exploring	25	2.69
Applying	42	3.06
Representing	34	3.05
Calculating	42	2.44
Reviewing and extending	25	2.60

Note. Counts refer to 42 nested participant-by-task opportunities. The original rating scale ranges from 1 (novice) to 4 (distinguished). Occurrence requires agreement by at least two of three raters; it is distinct from rated quality.

Source: Author ([2012](#)), Tables 47 and 48 pp. 162 and 166

Occurrence and quality convey different information. Calculation was counted in every opportunity, yet its mean rubric rating, 2.44, was the lowest reported process mean. Application also occurred in every opportunity and had the highest mean, 3.06. This comparison does not establish a statistical difference, but it shows why a count of process presence cannot substitute for its assessed adequacy. It also does not disclose whether calculation preceded representation, whether a review repaired an error, or how a process changed with task difficulty.

The process records were not integrated into the structural model at a comparable analytical level. They therefore cannot independently validate the interpretation that its cognition factor represented observed reasoning. The small, selected component shows what another form of evidence can capture; it does not estimate the disposition–process relationship across the 118 participants. Likewise, the reported process-session self-assessment means of 2.70–3.85 cannot be treated as observations of the timing or quality of regulation.

Contrasting published mathematical solution episodes

The first published episode concerns the number of line segments determined by 100 distinct points on a circle (Author, [2012](#), pp. 158–160). Participant A draws configurations, counts segments, tabulates small cases, and searches for a formula. A provisional count of 16 segments for six points is followed by

a table using the correct value 15. The episode then moves from examples and successive increments to the expression $(n/2)(n - 1)$, yielding 4,950 for 100 points. The visible mathematical activity includes coordinating drawings with a numerical pattern and revising a count; the final answer alone would conceal these transitions.

The transcript records a successful formula search but does not articulate a full combinatorial justification. Such a justification would explain that each unordered pair of distinct points determines exactly one segment, so counting ordered pairs gives $n(n - 1)$ and dividing by two removes double counting. This distinction concerns what the printed episode makes explicit. It does not establish that the participant lacked a justification that remained unspoken. Moreover, the researcher reminded the participant to think aloud and encouraged continuation. The episode is therefore evidence of reasoning within an elicited session, rather than an entirely unprompted account of persistence.

The second episode concerns a bookworm travelling from the first page of the first volume to the last page of the third volume of three books placed in order with their spines facing the observer (pp. 160–161). Participant B initially adds three book interiors and four covers, obtaining $5\frac{1}{2}$ inches. Re-reading the endpoint and orientation conditions leads to a revised drawing and a different representation of the route. The revised calculation includes one interior and four covers, producing $2\frac{1}{2}$ inches, given an interior thickness of $1\frac{1}{2}$ inches and cover thickness of $\frac{1}{4}$ inch.

Here the decisive revision concerns the situation represented mathematically. Correcting arithmetic within the original representation would have retained the wrong route. The printed sequence shows that re-examining spatial orientation changed the quantities included in the calculation. It thus supplies local evidence of a regulatory action with a mathematical consequence. It does not identify that action's general frequency, establish a causal effect across tasks, or explain a cohort-level negative awareness coefficient. Reminders to verbalize during pauses also limit what can be attributed to wholly spontaneous expression.

The contrast makes the evidence-chain problem concrete. Both episodes end with a correct numerical result, yet they expose different reasoning: movement from examples to a general expression in the first and reconstruction of a spatial model in the second. Process observation adds information about the relation between representations, errors, and repair. It does not retrospectively convert the need-for-cognition inventory into a direct measure of those processes. Nor does one successful revision justify a general claim about the beneficial effect of regulation. The episodes support narrower conclusions about the mathematical transitions visible on those occasions.

Evidentiary transitions requiring additional warrants

Taken together, the results identify three transitions. The first moves from a willingness or tendency to think effortfully to mathematical cognition enacted in a task. The second moves from reported metacognitive awareness to the consequences of actual regulation. The third moves from success on

personal mathematical tasks to competence in teaching those tasks. The reported measures support claims about their own responses, scores, or coded observations; broader claims require separately justified connections.

No directly observed teaching outcome appears in the reported record. Consequently, the case supports implications for what teacher educators might assess, but it does not demonstrate transfer to noticing, task selection, or classroom response. This absence is not evidence that transfer fails. It identifies the additional evidentiary level needed to investigate it. The author’s clarification resolves the 35-item trial and 30-item final inventory lengths. Remaining uncertainty concerns the deletion criteria, subscale allocation, and reliability of final-version scores; the sampling discrepancy also remains unresolved. These documentation limits do not establish that every reported association is incorrect.

Source claims and warranted interpretations

Table 8 distinguishes three kinds of conclusion. The cognition description documents a construct–measure discrepancy, while the monitoring-and-confidence explanation asserts a mechanism not tested by the reported score relation. The global-validity statement requires clarification of its scope: model fit supports a specified score structure, whereas a claim about the wider theory requires evidence about its constructs and mechanisms. The teaching recommendations remain conditional proposals. Their evidentiary gap is a proposed transfer to a separately observed outcome, not a demonstrated adverse effect or an established teaching benefit.

Table 8 *Source claims evidence and warranted interpretations*

Source claim and locator	Evidence available	Interpretive extension	Narrower interpretation warranted
Cognition concerns constructing a representation and selecting a strategy (p. 185, Summary).	Need-for-cognition inventory measures enjoyment of effortful thinking (pp. 70–71); reduced path to test score $\beta = .864$ (p. 178, Table 54).	Construct–measure discrepancy: the definition names enacted processes, but the model’s inventory factor records a disposition. The definition alone does not prove that every statistical claim asserted observed reasoning.	The model estimates a conditional relation between a thinking disposition and test performance; the six process cases are a separate source of activity evidence.
Better performers monitor less because they have ample self-confidence (p. 192, Conclusion 3).	Negative metacognition coefficient in the reported regression (p. 138, Table 40) and reduced SEM $\beta = -.164$ (p. 178, Table 54); selected process summaries (pp. 162, 166).	Explanatory overreach: the reported association does not test whether self-confidence produces less monitoring or whether monitoring lowers performance.	Scores show a conditional negative association. The confidence-and-monitoring explanation remains a hypothesis requiring linked process and confidence evidence.
The postulated 3P theory is empirically valid and generally acceptable (p. 193, Conclusion 6).	Fit indices and paths from full and reduced models fitted to the same 118 participants (pp. 172, 177–178, Tables 50, 53–54).	Scope of validation: in-sample fit supports the specified score model. A claim of wider theoretical validity requires evidence about the proposed constructs, mechanisms, and applicability.	The specified score models have reported in-sample fit and associations; their substantive mechanisms and applicability to other cohorts remain untested.
Epistemological experiences may support effective teaching; trainees will favor future students’ varied epistemologies (p. 193, Recommendations 1–2).	Inventories, personal problem-solving test, and six process cases; no directly observed teaching outcome in the examined record.	Proposed transfer: a conditional pedagogical recommendation and an anticipated student-related benefit require a separate teaching outcome.	The recommendations identify possibilities for teacher education. Their effectiveness and transfer to classroom responses require investigation.

Note. Source claims are paraphrased from Author (2012); page numbers use the dissertation’s printed pagination. The table identifies evidentiary limits, not proof that the proposed mechanisms or educational benefits are false. The 2012 source and the present analysis are distinct contributions.

Discussion

Preserving mathematical meaning across inferential levels

The case demonstrates a methodological problem with substantive consequences for mathematics education. A psychological predictor can be informative without representing the mathematical activity its interpretation names. The important question is whether the inferential transition is warranted. Reporting the large historical coefficient as an association between need-for-cognition and test scores preserves its operational meaning. Explaining it as a mechanism of mathematical reasoning requires observations capable of identifying that reasoning.

This distinction addresses a problem broader than the historical cohort. Mathematical products conceal variation in the pathways by which they are obtained. A correct answer might arise from a valid general argument, a restricted collection of examples, a remembered procedure, or successful guessing. Conversely, an incorrect answer might retain a productive representation or justification interrupted by an arithmetic mistake. A test score and a process account can therefore contribute different information even when both concern problem solving.

The separate statistical level yields three concrete distinctions. First, the .864 path remains a relation between need-for-cognition and test scores even when the surrounding construct description names representation and strategy selection. Relabeling the measured predictor preserves the estimate while exposing the observation needed for a process interpretation. Second, the metacognitive coefficients are nearly unchanged across specifications ($-.163$ and $-.164$), although their reported p values change from .112 to .016. The analytical issue is therefore altered statistical evidence under model reduction, not a larger estimated negative association. Without re-estimation, the analysis cannot identify the contribution of each deleted path to that change. Third, the proposed teaching implications change the outcome from personal test performance to instructional response; no coefficient in the reported model estimates that response. These distinctions lead to different evidentiary remedies: linked task observations, scrutiny of model selection and uncertainty, and an independently observed teaching outcome. A general warning about validity would not specify which remedy addresses which transition unless it undertook this same mapping.

The case demonstrates distinctions within a documented research account, not the effectiveness of the framework as an intervention. Its contribution is a compact procedure for preserving operational referents while examining statistical explanations and educational uses. Table 8 makes the judgments contestable through source locators, available evidence, alternative interpretations, and observations that would change the assessment. Ing et al. (2024) motivate the contemporary relevance of response-process evidence, but do not establish how often these particular transitions occur in preservice teacher research. Wider usefulness requires independent application to further studies and comparison with an explicit argument-based analysis.

Interpreting metacognition without turning associations into prescriptions

The negative historical metacognitive-awareness coefficient is particularly consequential because an overextended interpretation could discourage planning or checking. The present analysis provides no such prescription. Inventory responses may differ from enacted regulation (Craig et al., 2020; Dang et al., 2020; Zepeda & Nokes-Malach, 2023); regulation may also occur in response to difficulty or an emerging error. More visible checking can accompany harder work without causing poorer performance. These possibilities must be distinguished through appropriate observations rather than selected as explanations after a coefficient appears.

Task-based studies provide ways to investigate the issue. Baumanns and Rott (2023) identify regulatory behavior in problem posing, and Vorhölter (2025) attends to the timing and aims of metacognition in modelling. Ernst et al. (2025) separately investigated self-efficacy and self-assessment accuracy in mathematics student teachers. These contributions support distinguishing perceptions, judgments, and activity (Craig et al., 2020; Ernst et al., 2025; Zepeda & Nokes-Malach, 2023); they do not establish the cause of the 3P coefficient. A subsequent study would need to record the trigger, mathematical purpose, and consequence of regulation and relate these observations to task difficulty and prior attainment.

Causal interpretation requires explicit assumptions about relationships among variables. Statistical adjustment, a directional path, or an indirect-effect decomposition cannot establish those assumptions by itself (Grosz et al., 2020; Rohrer, 2018). For the historical design, explanatory pathways are hypotheses attached to cross-sectional score relations. The evidence-chain analysis clarifies the represented phenomena before any causal interpretation is attempted.

An assessment application for preservice mathematics teachers

The framework can inform assessment by making the intended judgment explicit. A teacher educator may wish to assess correctness, mathematical justification, purposeful regulation, or a response to student reasoning. These purposes require related but distinct observations (Cross Francis et al., 2022; Depaepe et al., 2020; Weyers et al., 2024). Table 9 presents a proposed evidence profile. It is a design illustration developed from the analysis, not an implemented intervention or a validated rubric.

Table 9 *Proposed evidence profile for assessment of preservice mathematics teachers*

Assessment focus	Evidence elicited	Warranted interpretation
Mathematical product	Dimensions, area, units, and domain in the solution.	Correctness for this task; not a complete profile of reasoning.
Mathematical reasoning	Explanation of the fixed-perimeter relation and why the maximum applies to all admissible cases.	Quality of justification, with examples distinguished from a general argument.
Enacted regulation	Recorded check or revision, its trigger, mathematical purpose, prompting status, and consequence.	Regulation observed on this occasion; not a general tendency inferred from frequency alone.
Pedagogical response	Interpretation of the fictional student's claim and a proposed question, representation, or counterexample.	Reasoning about teaching this idea; actual classroom responsiveness requires observation.
Reported disposition	A separately administered inventory or reflection on engagement and awareness.	Reported tendency or perception; interpreted alongside, not substituted for, task evidence.

Note. The profile is an illustrative design aid, not a validated rubric. Prompted and spontaneous actions must be distinguished. An unrecorded action is not necessarily an absent action.

Consider the task: A rectangle has perimeter 24 cm. Which dimensions give the greatest area? Explain why the answer applies to every admissible rectangle. With side lengths x and $12 - x$, the domain is $0 < x < 12$. The area satisfies $A = x(12 - x) = 36 - (x - 6)^2$. Since the squared term is nonnegative, the greatest area is 36 cm², attained at $x = 6$. A sound geometric argument is also acceptable. Numerical examples can support a conjecture, but a general argument establishes why untested admissible dimensions cannot exceed the proposed maximum.

Candidates first solve independently and retain their working. They subsequently explain what convinced them that the maximum holds and identify any check or revision. The record distinguishes a spontaneous check from one elicited by a follow-up prompt. Finally, candidates respond to a fictional student who argues that making one side longer must increase area. A useful response addresses the fixed-perimeter condition, perhaps comparing a 2 cm by 10 cm rectangle with a 6 cm by 6 cm square. The comparison must be connected to the student's reasoning rather than presented as an unexplained counterexample.

This sequence yields separate evidence about a mathematical product, a justification, regulation, and pedagogical reasoning. A candidate might find the correct dimensions through a table while leaving the maximum unjustified. Another might correct an erroneous perimeter equation and develop a valid argument. These are hypothetical patterns: neither is a historical participant case. Their point is that the same final answer can support different feedback when the reasoning and revision are visible.

Feedback then follows the assessed evidence. An answer based on examples alone warrants work on generality. An argument with a unit error warrants attention to interpretation and checking. A valid

personal solution accompanied by a dismissive student response warrants attention to eliciting and using the student's thinking. A high awareness inventory score does not replace missing justification; a low confidence report does not invalidate an observed sound argument. Separate judgments preserve the instructional relevance of each observation.

The proposed profile should not be combined into a single readiness score without supporting evidence for its intended use. Familiar tasks may elicit little visible regulation, and silence does not establish that no monitoring occurred. A follow-up task with different constraints or content can investigate the scope of the observed reasoning. Actual classroom responsiveness requires a teaching observation beyond the hypothetical response (Cross Francis et al., [2022](#); Depaepe et al., [2020](#); Weyers et al., [2024](#)). Work on noticing struggle identifies a distinct need to interpret classroom evidence when deciding how to respond (Jarry-Shore & Richardson, [2025](#); Weyers et al., [2024](#)).

Research designs that connect dispositions with teaching

A subsequent study could collect disposition inventories before a task sequence, record mathematical reasoning and regulation during several tasks, assess the resulting products, and observe a later teaching episode. Variables would retain operational names such as self-reported need for cognition, observed strategy revision, justification quality, and response to student reasoning. This design would permit investigation of relationships among the evidence types rather than assuming that an inventory factor stands for all of them.

A process protocol should define the coding unit and its temporal boundaries. For a checking episode, it could record the trigger, mathematical object checked, prompting status, action taken, and consequence for the solution. Written work, recordings, and stimulated recall could be compared, while recognizing that retrospective explanations do not timestamp earlier thoughts. Pilot coding, independent checking, and an agreement measure suited to the coding scale would be necessary before treating process scores as sufficiently reproducible for the intended analysis.

Repeated task observations must retain their dependence structure. Episodes are nested within candidates and vary with content, difficulty, and prompting. A multilevel model or explicitly justified aggregation can represent this structure; adding task opportunities does not add independent participants. Sample planning should account for both candidates and observations per candidate, measurement precision, and model complexity. Fresh-data evaluation is especially important when selecting predictors or deleting paths using the original sample.

Three kinds of research question would then remain distinct. Does a disposition score predict justification quality after accounting for prior attainment? Does a particular revision repair a documented mathematical error? Does personal justification quality relate to a later instructional response? The first concerns an association among measures, the second a situated process, and the third transfer to a separately observed outcome. An intervention claim requires still further evidence concerning allocation,

comparison conditions, and common independently assessed outcomes. No single path coefficient answers all these questions.

Further evaluation should separate agreement about source interpretation from agreement about the framework's usefulness. Independent readers could first reconstruct the four focal claims from the cited passages, without seeing the present classifications, and then apply the stated rules. Comparing their judgments would reveal ambiguous claims or rules. A subsequent comparison using an argument-based analysis could examine whether the separate statistical level changes the evidentiary gaps identified, the specificity of the required observations, or the transparency of the resulting interpretations. Such work would test added practical value; it has not been conducted in this study.

Limitations

The principal limitation is dependence on the published documentary record. The focal numerical findings and procedural descriptions were checked against the original chapters, but underlying responses, statistical project files, and original recordings were unavailable. The numerical models cannot be reproduced, alternative specifications cannot be estimated, and the sampling discrepancy and the undocumented details of the final curiosity-inventory reduction cannot be resolved from the published record alone. The two published transcripts are a selected representation of the 42 sessions; the dissertation states that not every session was transcribed. They therefore permit local interpretation rather than a new comprehensive process analysis.

The historical study also has limits relevant to interpretation: conflicting sampling descriptions, an incompletely documented reduction of the curiosity inventory from 35 to 30 items, trial reliability estimates from other samples and, for curiosity, a longer instrument version, a short multiple-choice outcome, and six selected process cases. The case does not establish contemporary population parameters, the prevalence of inferential overreach, or the generality of its coefficients across institutions and cohorts. No confidence intervals or tests of differences between the process means are available from the reported record.

The present classifications have not received independent second-reader scrutiny, and the analyst conducted the original study. Explicit decision rules, source locators, and conditions for revising a classification make the judgments inspectable but do not establish their reproducibility. The literature was selected for conceptual relevance rather than systematically sampled. The case demonstrates particular construct–measure discrepancies and interpretive extensions; it does not establish their prevalence or validate the framework's usefulness across studies. The proposed assessment profile and research designs remain unevaluated.

Conclusion

Secondary documentary analysis of the 3P study identifies three evidentiary transitions requiring explicit warrants: from reported dispositions to enacted mathematical cognition, from awareness to regulation, and from personal problem solving to teaching. The historical coefficients retain meaning as relations among their operational measures. Process summaries and published solution episodes contribute another view of mathematical activity, but their presence cannot transform the inventory-based structural paths into estimates of observed reasoning or regulation.

For mathematics education research, the contribution is a framework for tracing these transitions through constructs, measurements, statistical relations, and educational claims. For teacher education, it can inform the collection of separate evidence and the provision of feedback concerning correctness, justification, regulation, and pedagogical response. The conclusions are methodological and bounded by the documentary record. Further research should directly connect these forms of evidence and investigate whether the framework improves the defensibility and usefulness of assessments and interpretations.

Declarations

Funding: No specific funding was received for this study or the preparation of this manuscript.

Competing interests: The author declares no financial or non-financial competing interests. The author's role as the researcher who conducted the original dissertation is disclosed in the manuscript.

Author contributions: Author conducted the original doctoral research. The present article develops a secondary documentary interpretation of that work.

Ethics approval and consent to participate: The present analysis recruited no participants, collected no new participant information, and did not access unpublished participant-level records. The examined chapters do not establish a specific original ethics-approval reference or complete consent procedure; no such approval is asserted here.

Consent for publication: Published solution episodes are paraphrased and identified using participant letters rather than the names appearing in the dissertation. No new participant images or unpublished records are included. The proposed student statement and assessment patterns are illustrative.

Data availability: The documentary source is Author's (2012) dissertation, available through ProQuest [identifying access details withheld for blind review]. The corpus and printed page and table locators are specified in Method and the table captions. Historical results are retained in Tables 3–7. The underlying participant-level dataset, original recordings, and AMOS project were unavailable to the present analysis. The published transcript episodes can be located on printed pp. 158–161.

Code availability: No new statistical model was estimated and no participant-analysis code accompanies this article. Arithmetic checks are described in Method.

Prior dissemination: This article draws on the author's 2012 doctoral dissertation, available through ProQuest. Identifying bibliographic and access details are withheld for blind review. Historical findings are distinguished from the present methodological interpretation.

Use of AI: ChatGPT (OpenAI) assisted with language editing, formatting, and citation-reference audit. It was not used to generate participant responses, statistical estimates, or analyses. Human verification was ensured by the author.

Researcher position

The author conducted the original doctoral study. This creates access to its conceptual history but also a potential investment in its interpretation. The analysis therefore retains unfavorable or ambiguous findings, distinguishes reported evidence from explanation, and treats the reduced model as exploratory. The author's retrospective clarification that the Curiosity Inventory was reduced from 35 items after the third trial to 30 items for final data collection is identified separately from the dissertation's documentary evidence. This clarification explains the two reported lengths but does not establish the deletion criteria or the reliability of final-version scores. Other missing sampling and statistical documentation is not reconstructed from the author's familiarity.

References

1. Author. (2012). Doctoral dissertation [identifying details withheld for blind review].
2. Baumanns, L., & Rott, B. (2023). Identifying metacognitive behavior in problem-posing processes: Development of a framework and a proof of concept. *International Journal of Science and Mathematics Education*, 21(5), 1381–1406. <https://doi.org/10.1007/s10763-022-10297-z>
3. Böswald, V., & Schukajlow, S. (2023). I value the problem, but I don't think my students will: Preservice teachers' judgments of task value and self-efficacy for modelling, word, and intramathematical problems. *ZDM – Mathematics Education*, 55(2), 331–344. <https://doi.org/10.1007/s11858-022-01412-z>
4. Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning*, 15(2), 155–213. <https://doi.org/10.1007/s11409-020-09222-y>
5. Cross Francis, D., Eker, A., Liu, J., Lloyd, K., & Bharaj, P. (2022). (Mis)alignment between noticing and instructional quality: The role of psychological and cognitive constructs. *Journal of Mathematics Teacher Education*, 25(5), 599–632. <https://doi.org/10.1007/s10857-021-09509-0>
6. Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, 24(4), 267–269. <https://doi.org/10.1016/j.tics.2020.01.007>
7. Depaepe, F., Verschaffel, L., & Star, J. (2020). Expertise in developing students' expertise in mathematics: Bridging teachers' professional knowledge and instructional quality. *ZDM –*

- Mathematics Education*, 52(2), 179–192. <https://doi.org/10.1007/s11858-020-01148-8>
8. Ernst, H. M., Prinz-Weiß, A., Wittwer, J., & Voss, T. (2025). Discrepancy between performance and feedback affects mathematics student teachers' self-efficacy but not their self-assessment accuracy. *Frontiers in Psychology*, 15, 1391093. <https://doi.org/10.3389/fpsyg.2024.1391093>
 9. Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist*, 56(2), 132–141. <https://doi.org/10.1080/00461520.2021.1898962>
 10. Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://doi.org/10.1177/2515245920952393>
 11. Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5), 1243–1255. <https://doi.org/10.1177/1745691620921521>
 12. Ing, M., Kosko, K. W., Jong, C., & Shih, J. C. (2024). Validity evidence of the use of quantitative measures of students in elementary mathematics education. *School Science and Mathematics*, 124(6), 411–423. <https://doi.org/10.1111/ssm.12660>
 13. Jarry-Shore, M., & Richardson, A. (2025). Noticing struggle during collaborative problem-solving in the middle-school mathematics classroom. *Journal of Mathematics Teacher Education*, 28(4), 953–978. <https://doi.org/10.1007/s10857-024-09638-2>
 14. Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
 15. Lawson, K. M., & Robins, R. W. (2021). Sibling constructs: What are they, why do they matter, and how should you handle them? *Personality and Social Psychology Review*, 25(4), 344–366. <https://doi.org/10.1177/10888683211047101>
 16. Lit, S., Kool, M., & Drijvers, P. (2026). Teacher knowledge and skills for mathematical problem solving: A systematic literature review. *Journal of Mathematics Teacher Education*. Advance online publication. <https://doi.org/10.1007/s10857-025-09737-8>
 17. Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27–42. <https://doi.org/10.1177/2515245917745629>
 18. Rott, B. (2021). Inductive and deductive justification of knowledge: Epistemological beliefs and critical thinking at the beginning of studying mathematics. *Educational Studies in Mathematics*, 106(1), 117–132. <https://doi.org/10.1007/s10649-020-10004-1>
 19. Schukajlow, S., Rakoczy, K., & Pekrun, R. (2023). Emotions and motivation in mathematics education: Where we are today and where we need to go. *ZDM – Mathematics Education*, 55(2), 249–267. <https://doi.org/10.1007/s11858-022-01463-2>
 20. Torregrosa, A., & Albarracín, L. (2026). Characterizing metacognitive processes and mathematical skills during problem solving by using a non-linear orientation base as a self-regulation tool.

- Educational Studies in Mathematics*, 121, 333–364. <https://doi.org/10.1007/s10649-025-10451-8>
21. Voica, C., Singer, F. M., & Stan, E. (2020). How are motivation and self-efficacy interacting in problem-solving and problem-posing? *Educational Studies in Mathematics*, 105(3), 487–517. <https://doi.org/10.1007/s10649-020-10005-0>
22. Vorhölter, K. (2025). Metacognitive behaviour when working on modelling problems in small groups. *ZDM – Mathematics Education*, 57(2–3), 503–517. <https://doi.org/10.1007/s11858-025-01661-8>
23. Weyers, J., König, J., Scheiner, T., Santagata, R., & Kaiser, G. (2024). Teacher noticing in mathematics education: A review of recent developments. *ZDM – Mathematics Education*, 56(2), 249–264. <https://doi.org/10.1007/s11858-023-01527-x>
24. Yao, Y., Hwang, S., & Cai, J. (2021). Preservice teachers' mathematical understanding exhibited in problem posing and problem solving. *ZDM – Mathematics Education*, 53(4), 937–949. <https://doi.org/10.1007/s11858-021-01277-8>
25. Zepeda, C. D., & Nokes-Malach, T. J. (2023). Assessing metacognitive regulation during problem solving: A comparison of three measures. *Journal of Intelligence*, 11(1), Article 16. <https://doi.org/10.3390/jintelligence11010016>