Beyond Correctness: Using the Certainty Response Index to Screen Primary Geometry Misconceptions

Hartini Ismail*, Mohd Azarul Mohd Mokhtar

Department of Educational Studies, Faculty of Human Development, Universiti Pendidikan Sultan Idris, Perak, Malaysia

ABSTRACT

Conventional achievement scores do not distinguish a confidently held error from uncertainty or a correct response from an unstable selection. This descriptive pilot examined whether the Certainty Response Index (CRI) could improve interpretation of responses to 10 two-dimensional and three-dimensional geometry items among 30 purposively selected Year Two pupils in one Malaysian primary school. For each item, pupils selected an answer and rated their confidence on a five-point scale (1 = lowest confidence; 5 = highest confidence). Scores of 1–3 were classified as low confidence and scores of 4–5 as high confidence. Data were collected over one month through one-hour sessions held twice weekly. The available summary dataset classified 10 pupils (33.3%) as correct-high confidence, 6 (20.0%) as correct-low confidence, 6 (20.0%) as incorrect-low confidence, and 8 (26.7%) as incorrect-high confidence. The latter profile was interpreted as a screening flag for a possible misconception rather than definitive evidence of its cognitive source. CRI adds practical diagnostic resolution to early geometry assessment, but should be combined with explanations, interviews, or multi-tier items. The study offers a cautious framework for selecting follow-up responses while recognising the limits of a small, single-school pilot and aggregate profile reporting.

Keywords: certainty response index; diagnostic assessment; geometry misconceptions; primary mathematics; response confidence; two-dimensional and three-dimensional shapes

INTRODUCTION

Early geometry requires learners to coordinate visual perception, spatial language, classification, and reasoning about properties. Spatial thinking is associated with mathematical performance, while experience with dynamic and three-dimensional representations can strengthen interpretations of geometric relationships (Gilligan-Lee et al., 2022; Ng et al., 2020; Oughton et al., 2024; Tam & Chan, 2022). Yet young learners often identify figures through familiar prototypes or surface appearance rather than properties that remain invariant across orientation and representation. A judgment may therefore change when a shape is rotated, embedded in a complex figure, or shown from an unfamiliar viewpoint (Fujita et al., 2017; Sinclair & Moss, 2012).

A total score reveals whether a response matches the key but not why it was chosen. An incorrect answer can reflect uncertainty, incomplete knowledge, or a stable but inaccurate interpretation. A correct answer can represent robust understanding or an uncertain selection. These cases require different instructional responses, but correctness-only assessment treats each pair identically.

Introduction (continued)

The Certainty Response Index combines answer accuracy with reported confidence (Hasan et al., 1999). Multi-tier diagnostic research shows that confidence information can refine the interpretation of alternative conceptions and reduce some false inferences produced by correctness-only tests (Caleon & Subramaniam, 2010; Kaltakci Gurel et al., 2015; Pesman & Eryilmaz, 2010). Confidence is not, however, a direct measure of conceptual structure. In children, confidence may also be shaped by scale wording, task familiarity, or social expectations; confidence in an answer may differ from confidence in the reason supporting it (Yang, 2022).

Compared with its established use in science education, CRI remains underused in lower-primary mathematics. Geometry is a particularly suitable test bed because visually plausible selections may conceal prototype dependence or weak property-based reasoning. A calibrated use of CRI may help a teacher decide which responses can be accepted provisionally, which require foundational support, and which require an explicit challenge to an existing interpretation.

Purpose and research questions

This pilot investigated the diagnostic contribution of confidence information to Year Two geometry assessment. It addressed two questions: (1) What response-confidence profiles emerge when geometry responses are classified using CRI? and (2) What instructional implications can cautiously be derived from those profiles? The contribution is methodological and practical: the study tests a low-burden way to add diagnostic resolution without claiming that confidence alone proves a misconception.

METHODOLOGY

Design and participants

A quantitative descriptive pilot design was used. Participants were 30 Year Two pupils purposively selected from one Malaysian primary school. The assessment focused on identifying and interpreting two-dimensional (2D) and three-dimensional (3D) shapes. Data collection extended over one month, with one-hour sessions conducted twice weekly. The small, single-school sample was suitable for piloting the response-confidence classification, but it was not intended to estimate the prevalence of misconceptions in the wider population.

Instrument and administration

Pupils completed a 10-item response-confidence instrument covering the identification and interpretation of 2D and 3D shapes. For each item, they first selected an answer and then rated their certainty on a five-point scale, where 1 represented the lowest confidence and 5 the highest. Confidence ratings of 1–3 were classified as low, whereas ratings of 4–5 were classified as high. The instrument contained an answer component and a confidence component but no separate reason tier; it is therefore described as CRI-enhanced rather than as a conventional answer-reason diagnostic test.

The geometry items were developed through discussion with experts and further examined in a focus group discussion (FGD). The review considered content relevance, age appropriateness, clarity of wording, and alignment with the targeted 2D and 3D geometry content. Administration followed the same sequence for all pupils during the one-month data-collection period. Each session lasted approximately one hour and sessions were held twice weekly. Accuracy was scored against the answer key. A test-retest procedure was used to examine stability; however, the numerical coefficient was not available in the retained summary dataset and is therefore not reported as quantified reliability evidence.

Operational classification and analysis

Accuracy and confidence were cross-classified into four operational profiles: correct-high confidence, correct-low confidence, incorrect-low confidence, and incorrect-high confidence. The retained summary dataset contained pupil-level totals for these four profiles. Frequencies and percentages were calculated using N = 30. Because item-by-item profile counts were not available in the retained dataset, no item-level prevalence or comparison across the 10 items was attempted. The profiles were treated as screening categories rather than definitive diagnoses of internal cognitive states, and no inferential test or population estimate was conducted.

Table 1. Operational CRI response-confidence classification

Accuracy	Confidence	Operational profile	Cautious interpretation and next step
Correct	High	Correct-high confidence	Evidence consistent with understanding; use explanation or transfer to confirm stability.
Correct	Low	Correct-low confidence	Uncertain correct response; guessing or unstable knowledge is possible but not confirmed.

Incorrect	Low	Incorrect-low confidence	Uncertain error; consider a knowledge gap, incomplete understanding, or task confusion.
Incorrect	High	Incorrect-high confidence	Flag for a possible misconception; elicit reasoning before assigning a conceptual label.

ANALYSIS AND RESULTS

The pupil-level summary produced all four response-confidence profiles (Table 2). Correct-high confidence was the largest profile, involving 10 pupils (33.3%). Eight pupils (26.7%) were classified as incorrect-high confidence, the group warranting the most immediate conceptual probing. Correct-low confidence and incorrect-low confidence each involved 6 pupils (20.0%). Thus, confidence differentiated pupils who would otherwise have been grouped only as correct or incorrect. These figures describe the available aggregate profiles and should not be interpreted as item-level prevalence across the 10 geometry questions.

Table 2. Pupil-level distribution across CRI response-confidence profiles

Accuracy	Confidence	Operational profile	n (%)
Correct	High (4–5)	Correct-high confidence	10 (33.3)
Correct	Low (1–3)	Correct-low confidence	6 (20.0)
Incorrect	Low (1–3)	Incorrect-low confidence	6 (20.0)
Incorrect	High (4–5)	Incorrect-high confidence	8 (26.7)
Total	—	—	30 (100.0)

Note. Confidence scores of 1–3 were classified as low and scores of 4–5 as high. The retained dataset supported pupil-level aggregate reporting only.

Correct-response profiles

A correct-high confidence response provided stronger preliminary evidence of stable understanding than correctness alone. Even so, a short explanation or transfer task remained useful because geometry answers may be based on familiar prototypes. A correct-low confidence response had a different implication. It could reflect guessing, weakly consolidated knowledge, difficulty reading the representation, or uncertainty about terminology. The profile did not prove any one explanation; it identified a response that needed confirmation.

Useful follow-up prompts include asking the pupil to name a defining property, identify a non-example, rotate the figure, or apply the idea to a less familiar representation. These probes test whether the correct choice generalises beyond the original item.

Incorrect-response profiles

The distinction between incorrect-low and incorrect-high confidence was more consequential. An incorrect-low confidence response indicated uncertainty and was consistent with a possible knowledge gap or incomplete understanding. It could initially be addressed through definitions, manipulatives, guided comparison, and visual examples.

An incorrect-high confidence response represented a stronger screening signal. The pupil had selected an incorrect option while expressing confidence, so the teacher had a reason to investigate a stable alternative interpretation. However, the profile did not identify the rule, prototype, or spatial interpretation producing the answer. Confirmation required oral probing, written reasoning, or an answer-reason item.

Diagnostic value added by confidence

The value added by CRI can be stated as a decision problem. Correctness identifies success or error; confidence prioritises the next assessment action. High confidence strengthens provisional acceptance of a correct response and raises the priority of probing an incorrect one. Low confidence signals the need to stabilise or clarify knowledge regardless of whether the selected answer happened to be correct.

DISCUSSION

The findings support CRI as a practical screening layer for early geometry assessment. Its contribution is deliberately modest: it does not reconstruct a pupil's complete reasoning, but it separates responses that would be indistinguishable under ordinary scoring. This interpretation is consistent with the original CRI logic and with multi-tier diagnostic research showing that confidence can refine the interpretation of alternative conceptions (Caleon & Subramaniam, 2010; Hasan et al., 1999; Kaltakci Gurel et al., 2015).

Why correct answers still require interpretation

Correctness can be fragile in geometry. A pupil may identify a square in its canonical orientation but reject the same square after rotation, or recognise a solid by overall appearance while overlooking faces, edges, and vertices. These patterns are consistent with prototype-based rather than property-based classification (Fujita et al., 2017; Sinclair & Moss, 2012). Low confidence attached to a correct answer is therefore useful: it marks an opportunity to verify whether the idea transfers across orientation, size, context, and representation.

Why confident errors deserve priority

A high-confidence error is instructionally important because repetition of the same explanation may leave the existing interpretation intact. The teacher should first make the pupil's reasoning visible, then create cognitive contrast using non-examples, rotated figures, manipulatives, or linked 2D and 3D representations. Confidence helps prioritise this sequence, but the reasoning evidence determines the actual misconception and the response required.

A four-stage instructional cycle

The profiles suggest a four-stage cycle: diagnose the response-confidence combination; respond with an appropriate probe; rebuild the concept through representation and dialogue; and ask the pupil to reflect by explaining the revised idea. The cycle translates CRI information into classroom action. It is a proposed application framework rather than an intervention effect tested in this pilot.

Diagnose → Probe → Rebuild → Reflect

Interdisciplinary significance

The study also illustrates a broader data-analysis issue: a single observed outcome may combine qualitatively different latent states. Adding a confidence variable creates a simple two-dimensional classification that improves decision usefulness without requiring complex technology. This logic is relevant to assessment analytics, metacognition, educational measurement, and human decision research. At the same time, self-reported confidence is noisy data and must be triangulated rather than treated as ground truth.

Implications for teachers and instrument designers

For teachers, the profiles provide a manageable triage system. Correct-high responses can move to transfer; correct-low responses need confirmation; incorrect-low responses need clarification and foundational support; and incorrect-high responses need reasoning elicitation and conceptual challenge. For designers, the next step is to add an explicit reason tier and calibrate developmentally appropriate confidence labels. Item-level reporting should also separate concepts such as shape properties, orientation, classification, and 2D–3D relationships.

LIMITATIONS AND FUTURE RESEARCH

Five limitations constrain interpretation. First, the purposive sample comprised only 30 pupils from one school, so the findings cannot establish population prevalence or be generalised beyond comparable pilot contexts. Second, the retained dataset contained aggregate pupil-level profiles rather than profile frequencies for each of the 10 items; consequently, differences across geometry concepts, orientations, and representations could not be evaluated. Third, the absence of a separate reasoning tier or interview evidence means that an incorrect-high confidence profile cannot confirm the cognitive source of an error. Fourth, young pupils' confidence ratings may be influenced by language, task familiarity, and response style. Fifth, although a test-retest procedure was used, its numerical coefficient was unavailable for the present report, preventing quantified evaluation of score stability.

Future studies should retain and report profile frequencies for every item and geometry concept; document the composition of the expert panel and FGD process; quantify test-retest reliability with the coefficient and administration interval; and compare CRI classifications with oral explanations, drawings, manipulatives, or written reasons. A larger multisite sample would permit analysis by achievement level and task type. Intervention research should then test whether instruction matched to the four profiles produces greater conceptual change than correctness-only remediation.

CONCLUSION

"Correct" and "incorrect" are insufficient categories for interpreting Year Two geometry responses. In the available pupil-level summary, 10 pupils (33.3%) were classified as correct-high confidence, 6 (20.0%) as correct-low confidence, 6 (20.0%) as incorrect-low confidence, and 8 (26.7%) as incorrect-high confidence. By adding reported confidence, CRI exposed different instructional needs hidden within conventional correctness scores. Its strongest contribution is therefore as a screening and prioritisation tool: an incorrect-high confidence profile alerts the teacher to a possible misconception, whereas low-confidence profiles signal uncertainty that should be explored before firm conclusions are drawn.

CRI should supplement rather than replace content-rich tasks, explicit reasons, interviews, and teacher judgment. Used within these limits, it offers a low-burden method for making early geometry assessment more diagnostic and instructionally responsive. The pilot also demonstrates the value of combining outcome and confidence data in multidisciplinary assessment research.

DECLARATIONS

Ethics approval and consent: Permission to conduct the study was obtained from the participating school. Written informed consent was obtained from the pupils' parents or legal guardians, and assent was obtained from the pupils before their participation. Participation was voluntary, and confidentiality and anonymity were maintained throughout the study.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors declare no conflict of interest.

Data availability: The anonymised data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to ethical and institutional restrictions.

Author contributions: Hartini Ismail: Conceptualization, Methodology, Investigation, Formal Analysis, Writing – Original Draft, and Project Administration. Mohd Azarul Mohd Mokhtar: Validation, Formal Analysis, Writing – Review & Editing, and Visualization. Both authors reviewed and approved the final manuscript.

REFERENCES

1. Caleon, I. S., & Subramaniam, R. (2010). Do students know what they know and what they do not know? Using a four-tier diagnostic test to assess the nature of students’ alternative conceptions. Research in Science Education, 40(3), 313–337. https://doi.org/10.1007/s11165-009-9122-4
2. Fujita, T., Kondo, Y., Kumakura, H., & Kunimune, S. (2017). Students’ geometric thinking with cube representations: Assessment framework and empirical evidence. The Journal of Mathematical Behavior, 46, 96–111. https://doi.org/10.1016/j.jmathb.2017.03.003
3. Gilligan-Lee, K. A., Hawes, Z. C. K., & Mix, K. S. (2022). Spatial thinking as the missing piece in mathematics curricula. npj Science of Learning, 7, Article 10. https://doi.org/10.1038/s41539-022-00128-9
4. Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). Physics Education, 34(5), 294–299. https://doi.org/10.1088/0031-9120/34/5/304
5. Kaltakci Gurel, D., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students’ misconceptions in science. Eurasia Journal of Mathematics, Science and Technology Education, 11(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a
6. Ng, O.-L., Shi, L., & Ting, F. (2020). Exploring differences in primary students’ geometry learning outcomes in two technology-enhanced environments: Dynamic geometry and 3D printing. International Journal of STEM Education, 7, Article 50. https://doi.org/10.1186/s40594-020-00244-1
7. Oughton, R., Nichols, K., Bolden, D. S., Dixon-Jones, S., Fearn, S., Darwin, S., Mistry, M., Peyerimhoff, N., & Townsend, A. (2024). Developing “deep mathematical thinking” in geometry with 3- and 4-year-olds: A collaborative study between early years teachers and university-based mathematicians. Mathematical Thinking and Learning, 26(3), 306–325. https://doi.org/10.1080/10986065.2022.2119497
8. Pesman, H., & Eryilmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. The Journal of Educational Research, 103(3), 208–222. https://doi.org/10.1080/00220670903383002
9. Sinclair, N., & Moss, J. (2012). The more it changes, the more it becomes the same: The development of the routine of shape identification in a dynamic geometry environment. International Journal of Educational Research, 51–52, 28–44. https://doi.org/10.1016/j.ijer.2011.12.009
10. Tam, Y. P., & Chan, W. W. L. (2022). The differential relations between sub-domains of spatial abilities and mathematical performance in children. Contemporary Educational Psychology, 71, 102101. https://doi.org/10.1016/j.cedpsych.2022.102101
11. Yang, D.-C. (2022). Investigating the differences between confidence ratings in the answer and reason tiers in fourth graders via an online four-tier test. Studies in Educational Evaluation, 72, 101127. https://doi.org/10.1016/j.stueduc.2022.101127