

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
“Navigating Global Transformation in an Uncertain World: Interdisciplinary
Perspective on Economics, Business, Education, Law, Agriculture, Food
Technology, Politics and Society” July 30th, 2026

USING THE CERTAINTY RESPONSE INDEX TO SCREEN FOR POSSIBLE MISCONCEPTIONS IN YEAR TWO 2D AND 3D SHAPE LEARNING

Hartini Ismail^{1*}, Mohd Azarul Mohd Mokhtar¹, Oktiana Handini²

¹ Department of Educational Studies, Faculty of Human Development, Universiti
Pendidikan Sultan Idris, Perak, Malaysia

² Universitas Slamet Riyadi, Surakarta, Indonesia

*Corresponding Author: hartini.ismail@fpm.upsi.edu.my

ABSTRACT

Correctness alone does not distinguish an uncertain geometry response from a confidently held error. This descriptive pilot examined the Certainty Response Index (CRI) in 2D and 3D shape learning among 30 purposively selected Year Two students at one Malaysian primary school. The 10-item instrument paired an answer with a confidence rating on a five-point scale; ratings of 1–3 were classified as low and 4–5 as high. Data collection extended over one month through one-hour sessions held twice weekly. The student-level summary classified 10 students (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. These aggregate profiles support differentiated follow-up assessment, but they do not establish item-level misconception prevalence or diagnostic accuracy. Formal instrument validation, test–retest reliability assessment, and independent corroboration of possible misconceptions were not conducted. The rule for deriving one profile per student from multiple items is also unspecified. High-confidence incorrect profiles are therefore treated as screening flags requiring reasoning evidence. Within these limits, CRI can organise further assessment of early shape learning; it cannot independently confirm the conceptual source of an error.

Keywords: Certainty Response Index, diagnostic assessment, geometry learning, misconception screening, response confidence

INTRODUCTION

Early geometry supports later mathematical learning because students must coordinate visual perception, spatial language, classification, and reasoning about properties. Spatial thinking is associated with mathematics performance, and experiences with dynamic or three-dimensional representations can strengthen how students interpret geometric relationships (Gilligan-Lee et al., 2022; Ng et al., 2020; Oughton et al., 2024; Tam & Chan, 2022). Young learners may nevertheless identify shapes through familiar prototypes or surface appearance rather than properties that remain unchanged across

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
“Navigating Global Transformation in an Uncertain World: Interdisciplinary Perspective on Economics, Business, Education, Law, Agriculture, Food Technology, Politics and Society” July 30th, 2026

orientation and representation. Their judgments can 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 test score provides limited information about the source of such errors. An incorrect response may reflect uncertainty, insufficient knowledge, or a strongly held but inaccurate interpretation. A correct response may likewise represent stable understanding or an uncertain selection. When assessment records correctness alone, these possibilities receive the same score even though they may require different follow-up questions and instructional responses. Diagnostic assessment therefore benefits from information that helps teachers decide which responses need confirmation, foundational support, or deeper conceptual investigation.

The Certainty Response Index (CRI) combines response accuracy with students' reported certainty (Hasan et al., 1999). Research on multi-tier diagnostic instruments indicates that confidence information can sharpen the interpretation of correct and incorrect responses and reduce some false inferences associated with correctness-only testing (Caleon & Subramaniam, 2010; Kaltakci Gurel et al., 2015; Pesman & Eryilmaz, 2010). Confidence is not a direct measure of conceptual structure, however, and it should not be treated as a substitute for reasoning evidence. In primary assessment, confidence in an answer may also differ from confidence in the reason supporting that answer (Yang, 2022).

This study examines the interpretation of confidence information in selected Year Two 2D and 3D shape content. Two questions guide the study: (1) What aggregate response-confidence profiles are identified in this pilot? (2) What follow-up assessment implications and interpretive limitations arise from these profiles? The contribution is a bounded examination of screening and reporting, rather than a validation of diagnostic effectiveness.

METHODS

Design and participants

A quantitative descriptive pilot involved 30 Year Two students purposively selected from one Malaysian primary school. The assessment focused on identifying and interpreting 2D and 3D shapes. Data collection extended over one month, with approximately one-hour sessions conducted twice weekly. This schedule describes the overall data-collection period; the duration of each administration of the 10-item test is not specified. The small, single-school group supports bounded exploratory description; no population prevalence or causal effect is inferred. The purposive selection criteria, participant ages, and an a priori sample-size rationale are not specified.

Year Two shape learning provides a bounded setting for examining how answer accuracy and confidence can be interpreted together. Restricting the content to 2D and 3D shapes aligns the assessment focus with the stated research questions; it does not

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
“Navigating Global Transformation in an Uncertain World: Interdisciplinary Perspective on Economics, Business, Education, Law, Agriculture, Food Technology, Politics and Society” July 30th, 2026

justify extension to other geometry domains. The methodological role of the 30-student group is exploratory description, rather than diagnostic validation. The numerical sample size cannot establish that students understood the certainty scale. Recruitment justification and developmental suitability are therefore separate questions, and neither is resolved by calling the study a pilot.

Instrument

Students completed a 10-item CRI-enhanced answer-confidence instrument covering identification and interpretation of 2D and 3D shapes. For each item, they selected an answer and then rated certainty on a five-point scale, with 1 representing the lowest confidence and 5 the highest. Ratings of 1–3 were classified as low confidence and ratings of 4–5 as high confidence. Accuracy was scored against the answer key. The instrument comprised answer and confidence components without a separate reason tier; it is therefore not described as an answer-reason two-tier diagnostic test.

The instrument paired answer selection with confidence reporting. Formal instrument validation and reliability assessment were not conducted: no documented expert-rating validation, focus-group validation, test-retest assessment, or reliability coefficient is presented. There was no independent assessment of how students interpreted the items or confidence scale. Exact item wording, diagrams, and the answer key are not reproduced here, limiting scrutiny of ambiguity and content coverage. The instrument is therefore treated as an exploratory assessment tool, not a validated diagnostic measure.

Content validity, response-process evidence, and reliability address different questions. Content review examines whether tasks represent the intended learning content; response-process investigation examines how children understand the questions and certainty options; reliability assessment examines consistency. An apparently clear geometry item may still place unintended demands on reading, vocabulary, or spatial interpretation. Because these evaluations were not conducted in this study, confidence profiles cannot establish the instrument’s validity or stability.

Operational classification and analysis

Answer accuracy and confidence were cross-classified into four operational response profiles, as summarised in Table 1. At item level, the classification combines correct or incorrect scoring with low (1–3) or high (4–5) confidence. The available study summary assigns the 30 students to four mutually exclusive aggregate profiles. However, the rule for reducing multiple item responses to one profile per student is not specified. No majority, mean-score, dominant-profile, or selected-item rule is assumed. The aggregate categories are therefore described with the limitation that their derivation cannot be reproduced from the information presented.

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
“Navigating Global Transformation in an Uncertain World: Interdisciplinary Perspective on Economics, Business, Education, Law, Agriculture, Food Technology, Politics and Society” July 30th, 2026

The four profiles are defined by observable answer scoring and self-reported certainty. A misconception, by contrast, is an inaccurate conceptual interpretation that would require supporting reasoning evidence and, where relevant, consistency across equivalent tasks. Guessing refers to a response selected without sufficient supporting knowledge. A correct-low confidence profile is compatible with guessing but does not demonstrate it; an incorrect-high confidence profile is compatible with a misconception but does not demonstrate its conceptual basis. The profiles must therefore be interpreted as assessment signals rather than definitive cognitive labels.

Frequencies and percentages are reported using $N = 30$ students as the denominator. They are not interpreted as counts of item responses: 30 students answering 10 items would generate up to 300 answer-confidence pairs, but item-level distributions of those pairs are not presented. Item-specific distributions, missing-response patterns, and differences between 2D and 3D concepts cannot be reconstructed. No inferential test, sensitivity or specificity estimate, or independent diagnostic comparison is reported.

RESULTS AND DISCUSSION

The student-level summary contains all four profiles. Correct-high confidence was the largest category, with 10 students (33.3%); incorrect-high confidence comprised 8 students (26.7%). Correct-low confidence and incorrect-low confidence each comprised 6 students (20.0%). Table 1 defines the operational combinations, and Table 2 presents the aggregate distribution. These counts describe student-level categories in the available study summary rather than prevalence across the 10 items. The two low-confidence categories together comprise 12 students (40.0%), calculated from the category counts, and identify uncertainty within both correct and incorrect categories.

Variation among correct responses

In the study summary, 10 students (33.3%) were classified as correct-high confidence and 6 (20.0%) as correct-low confidence. Thus, the correct category did not uniformly coincide with strong reported certainty. The distinction supplies additional self-report information but does not establish stable property-based understanding. A correct-low profile may reflect partial knowledge, hesitation, or guessing; it cannot distinguish these explanations. Explanation and transfer prompts are proposed follow-up tasks; they were not administered in this pilot.

Variation among incorrect responses

The study summary classified 6 students (20.0%) as incorrect-low confidence and 8 (26.7%) as incorrect-high confidence. The latter category warrants follow-up because it combines error with reported certainty. It does not establish that 26.7% of students held a confirmed geometry misconception. Carelessness, misunderstood instructions, unfamiliar drawings, or an inaccurate interpretation of the confidence options may

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
“Navigating Global Transformation in an Uncertain World: Interdisciplinary
Perspective on Economics, Business, Education, Law, Agriculture, Food
Technology, Politics and Society” July 30th, 2026

produce the same profile. Without reasoning evidence and the student-level aggregation rule, neither the conceptual source nor the stability of these errors can be established.

Table 1. Operational CRI response-confidence classification

Accuracy	Confidence	Operational profile	Interpretation
Correct	High	Correct high confidence	Correctness with reported certainty; understanding remains to be confirmed.
Correct	Low	Correct low confidence	Uncertain correctness; guessing is possible but unconfirmed.
Incorrect	Low	Incorrect low confidence	Uncertain error; inspect knowledge, wording, and representation.
Incorrect	High	Incorrect high confidence	Possible misconception flag; rule out other causes through follow-up.

Table 2. Reported student-level CRI profile distribution

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

Note. N = 30 students. Percentages use the student count as the denominator. The aggregation rule is unspecified, and item-level distributions are not presented; these percentages are not item-level misconception rates.

Interpretation and instructional implications

The reported profiles illustrate the information added when correctness is paired with certainty. This is a distinction in how responses are described, not evidence that CRI is more accurate than achievement testing. Establishing diagnostic accuracy would require a defined criterion for a confirmed misconception and comparison of CRI flags with an independent assessment. The original CRI approach and multi-tier diagnostic research provide a rationale for including confidence, but evidence from other settings does not validate this Year Two instrument automatically (Caleon & Subramaniam, 2010; Hasan et al., 1999; Kaltakci Gurel et al., 2015).

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
“Navigating Global Transformation in an Uncertain World: Interdisciplinary Perspective on Economics, Business, Education, Law, Agriculture, Food Technology, Politics and Society” July 30th, 2026

Geometry adds a representational limitation to confidence-based interpretation. A shape may look unfamiliar because of orientation, scale, perspective, or its resemblance to a taught example. Prototype-dependent recognition is a plausible explanation for an error, but it is not demonstrated in these students. Follow-up tasks should hold defining properties constant while varying appearance and should ask children to explain why examples belong together (Fujita et al., 2017; Sinclair & Moss, 2012). A consistent explanation across contrasting examples would be more informative than one incorrect identification.

A useful follow-up comparison would present the same defining shape properties in familiar and unfamiliar orientations. If the child changes the classification, the teacher should ask which feature prompted that change. The response might reveal prototype dependence, a misunderstood property, or difficulty interpreting the drawing. A single changed answer does not distinguish these explanations. The same caution applies to solid shapes: identifying a pictured object and explaining its geometric properties are different demands. These examples illustrate proposed probes, not errors documented in the present sample.

For incorrect responses, confidence can guide the next diagnostic step. An incorrect-low confidence response may be addressed initially through definitions, visual examples, manipulatives, and guided comparison. An incorrect-high confidence response calls for a different sequence: the teacher first elicits the student’s reasoning, then uses contrasting cases or multiple representations to test the existing interpretation. These are proposed instructional implications rather than intervention effects measured in this pilot.

The study framework presents this process as a four-stage cycle: diagnose the response-confidence profile, respond with an appropriate probe or activity, rebuild the concept through representation and dialogue, and ask the student to reflect by explaining the revised idea. This cycle is best understood as a practical application of the CRI framework, not as an outcome established by the present data.

Reported confidence may be affected by children's developing capacity to monitor their own knowledge. Being familiar with a picture, wanting to please an adult, or interpreting “sure” as “I have seen this before” may lead to high ratings without secure understanding. These possibilities are interpretive risks, not observed behaviours established by the present data. Yang (2022) examined differences between answer and reason confidence among fourth graders; those findings support distinguishing the constructs but should not be treated as direct validation for younger Year Two students. Future administration should verify that children understand the options through neutral practice examples and should avoid equating confidence with correctness. Independent checks of students’ understanding of the confidence scale were not conducted in this pilot. Developmental suitability requires evidence that children distinguish “I am sure of my answer” from “my answer is correct” and “this picture is familiar”. Future protocols should use brief, neutral practice tasks, ask children to explain the certainty options in

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
“Navigating Global Transformation in an Uncertain World: Interdisciplinary Perspective on Economics, Business, Education, Law, Agriculture, Food Technology, Politics and Society” July 30th, 2026

their own words, and avoid feedback that makes confidence seem desirable. These comprehension checks were not conducted, leaving uncertainty about the meaning of individual ratings. Evidence from fourth graders in Yang (2022) cannot establish suitability for this Year Two group.

Limitations and future research

Several limitations constrain interpretation. The single-school purposive sample limits transfer beyond its setting, and selected 2D and 3D shape content cannot represent primary geometry generally. The available study summary contains student-level aggregate profiles without item distributions or a stated aggregation rule; differences across concepts and representations therefore cannot be assessed. The five-point scale and cutoff are specified, but scale comprehension and developmental suitability were not independently assessed. Formal expert-rating or focus-group validation and test-retest reliability assessment were not conducted. No reliability coefficient or quantified content-validity result is claimed. Interviews, a reason tier, concept maps, and independent teacher assessments were not conducted to corroborate the conceptual basis of flagged errors. Consequently, possible misconceptions cannot be distinguished conclusively from guessing, carelessness, question misunderstanding, or unfamiliar representations. Proposed instructional responses were not evaluated as interventions.

A stronger validation study would compare item-specific CRI flags with independently evaluated oral explanations or an answer-reason-confidence instrument. Concept maps and teacher assessments could provide supplementary evidence, but teacher judgments should follow explicit criteria rather than serve as an assumed gold standard. Sampling both flagged and unflagged responses would permit examination of false positives and false negatives. Larger and more varied samples would support tests of transfer across grades and schools. These are recommendations for future work, not procedures retrospectively attributed to this study.

Independent comparison should use explicit criteria for identifying an inaccurate conceptual explanation, rather than simply repeat the CRI label. Oral explanations could be evaluated without reference to the initial confidence category, and both flagged and unflagged responses should be examined. Agreement would provide evidence of convergence; disagreement would identify possible overconfidence, underconfidence, item ambiguity, or scoring problems. Interviews, concept maps, and teacher judgments are complementary sources, each with its own limitations. This design is proposed to address the reviewer’s validation concern and is not presented as completed triangulation.

CONCLUSION

The study summary differentiated 30 Year Two students into four aggregate CRI profiles: 10 correct-high confidence (33.3%), 6 correct-low confidence (20.0%), 6 incorrect-low confidence (20.0%), and 8 incorrect-high confidence (26.7%). The first

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
“Navigating Global Transformation in an Uncertain World: Interdisciplinary Perspective on Economics, Business, Education, Law, Agriculture, Food Technology, Politics and Society” July 30th, 2026

objective is addressed through this bounded description; the second through proposed profile-sensitive follow-up assessment. The 10-item instrument used a five-point confidence scale with 1–3 classified as low and 4–5 as high. Formal validation, reliability assessment, and independent diagnostic comparison were not conducted; the student-level aggregation rule also remains unspecified. CRI can prioritise questions for further assessment, but a high-confidence incorrect profile does not confirm a misconception and a low-confidence correct profile does not confirm guessing. These data support cautious screening rather than standalone diagnosis or population-level inference.

DECLARATIONS

Ethics and consent. Participating-school permission, written parental or guardian consent, and student assent were obtained. Participation was voluntary, and confidentiality and anonymity were maintained. No formal ethics committee approval reference is provided.

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

Competing interests. A competing-interest declaration covering all authors remains to be confirmed before submission.

Data availability. Anonymised supporting data may be requested from the corresponding author, subject to availability and ethical and institutional restrictions. The aggregate counts presented here do not substitute for item-level records.

REFERENCES

- 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>
- 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>
- 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>
- 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>

PROCEEDING

The 8th International Conference on Technology, Education, and Social Science
**“Navigating Global Transformation in an Uncertain World: Interdisciplinary
Perspective on Economics, Business, Education, Law, Agriculture, Food
Technology, Politics and Society” July 30th, 2026**

- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- Yang, D.-C. (2022). Investigating the differences between confidence ratings in the answer and reason tiers in fourth graders via online four-tier test. *Studies in Educational Evaluation*, 72, 101127. <https://doi.org/10.1016/j.stueduc.2022.101127>