

A Culturally Grounded Data Literacy Assessment for Middle School Students: Development and Psychometric Evaluation

Fitry Wahyuni¹, Mukhtar¹, Elmanani^{1*}

¹Universitas Negeri Medan

Corresponding Author: elmanani_simamora@unimed.ac.id

HOW TO CITE:

Fitry Wahyuni, Mukhtar, Elmanani (2026). A Culturally Grounded Data Literacy Assessment for Middle School Students: Development and Psychometric Evaluation. International Journal of Special Education, 41(15s), 628-637

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

Data literacy is increasingly important in middle school mathematics, yet many assessments capture only isolated skills and rely on contexts far removed from students' everyday experiences. This study developed and psychometrically evaluated a culturally grounded data literacy assessment for Grade 8 students. The instrument contained 30 four-option multiple-choice items covering six dimensions: data knowledge; data management and preparation; analysis and interpretation; critical evaluation and reflection; communication and presentation; and ethical-situational reasoning. Six experts reviewed the instrument, and 128 students from two junior secondary schools in Medan, Indonesia, completed it. Evidence was examined using Aiken's V, classical item analysis, Cronbach's alpha, McDonald's omega, and theory-guided exploratory factor analysis based on tetrachoric correlations. Mean Aiken's V ranged from 0.919 to 0.924, item difficulty from 0.438 to 0.617, and point-biserial discrimination from 0.259 to 0.528. Fifteen items were retained, 10 were flagged for review, and five were prioritized for revision. Alpha ranged from 0.628 to 0.768, while omega ranged from 0.770 to 0.843. Data knowledge, analysis and interpretation, and critical evaluation and reflection showed the strongest psychometric performance. Communication and presentation and ethical-situational reasoning require refinement. Overall, the instrument supports low-stakes formative diagnosis, although broader or high-stakes use requires further independent empirical validation.

Keywords: Data Literacy Assessment, Culturally Grounded Assessment, Formative Diagnosis, Middle School, Psychometric Evaluation

1. Introduction

Data literacy has become an essential competence for students in contemporary mathematics education because learners are increasingly expected to read, organize, interpret, evaluate, and communicate information from data-rich situations. In middle school, this

competence is not limited to performing statistical procedures; it also involves making sense of tables, graphs, variability, data sources, and contextual claims. Recent reviews show that data literacy and statistical literacy in K–12 education have been conceptualized through multiple dimensions, including data

interpretation, data-based reasoning, communication, and critical evaluation (Cui et al., 2023; Friedrich et al., 2024; Witte et al., 2025). These dimensions are particularly relevant for Grade 8 students, who are transitioning from basic representation of data toward more reflective and evidence-based reasoning.

Despite this growing attention, assessment of data literacy in K–12 settings remains an emerging and uneven field. Existing instruments vary in construct definition, item format, target population, and the strength of validity and reliability evidence reported (Cui et al., 2023). Recent work has begun to address the need for K–12 data literacy competence assessment, yet many instruments are still broad, self-report oriented, or insufficiently connected to classroom diagnostic use (Donate-Beby et al., 2025; Duygulu et al., 2025). For middle school mathematics, there is still a need for student-facing assessment tasks that can capture multiple components of data literacy while providing interpretable evidence for formative diagnosis rather than merely producing a total score.

Another limitation in current assessment practice is that data literacy tasks are often presented in decontextualized forms, even though students' interpretation of data is strongly shaped by the familiarity, authenticity, and meaning of the context. Studies on authentic data and data preparation suggest that students need opportunities to work with data situations that preserve the connection among data, chance, variability, and real-world meaning (English & Watson, 2018; Wilkerson et al., 2022). In culturally grounded mathematics education, local contexts can serve as entry points for sense-making, provided that cultural elements are used to support interpretation rather than to test hidden cultural knowledge. Ethnomathematics-based learning has been shown to support mathematical literacy, and culturally rooted statistical tasks can help students connect statistical reasoning with meaningful social and local situations (Pratama & Yelken, 2024; Ramadhani et al., 2025). In this study, cultural grounding is therefore positioned as a fairness-by-design strategy: items are designed to be

meaningful and accessible, but the instrument is not claimed to be culturally bias-free without further response-process and invariance evidence.

The present study is situated in a HIEVS learning context, in which inquiry, visual statistical representation, and local cultural contexts are used to support students' engagement with data. However, this article does not evaluate the effectiveness of HIEVS as an instructional model. HIEVS is used only as the classroom context in which the assessment was developed and administered. From a measurement perspective, developing a culturally grounded data literacy assessment requires documentation of intended score use, construct domain, test specifications, item review, empirical item analysis, reliability, and evidence for score interpretation as part of a validation argument (AERA, APA, & NCME, 2014; Kane, 2013; Linn, 2006). Therefore, this study aimed to develop and evaluate a culturally grounded data literacy assessment for Grade 8 middle school students. The instrument consists of multiple-choice items representing six dimensions: data knowledge, data management and preparation, analysis and interpretation, critical evaluation and reflection, communication and presentation, and ethical-situational reasoning. The psychometric evaluation collected several forms of evidence, including expert-based content validity, item difficulty and discrimination, internal consistency reliability, and exploratory evidence of internal structure. Thus, the central contribution of this article lies in the development and psychometric evaluation of the instrument, not in testing a pretest–posttest learning intervention.

This article contributes to the literature in three ways. First, it proposes a culturally grounded data literacy assessment specifically designed for middle school students rather than adapting a general data literacy or teacher-oriented scale. Second, it integrates cultural context, authentic data situations, and fairness-by-design considerations into the item-development process. Third, it provides classroom-based psychometric evidence to support formative diagnostic use while maintaining a conservative

validity claim. The instrument is intended to help teachers identify students' strengths and weaknesses across data literacy dimensions, but broader use requires additional validation, including cognitive interviews, categorical confirmatory factor analysis, measurement invariance, differential item functioning, and item response theory analysis.

2. Methodology

2.1 Design

This study used an instrument-development and psychometric-evaluation design. The main purpose was not to test the effectiveness of a learning model, but to develop and examine the quality of a culturally grounded data literacy assessment for Grade 8 students. The study was therefore organized around the development of the assessment blueprint, expert review of item content and cultural appropriateness, classroom-aligned administration, and psychometric evaluation of student responses.

The design was primarily quantitative because the main evidence was obtained from expert ratings and students' dichotomous item responses. However, qualitative information from expert comments, cultural-fairness review, and item-revision notes was also used to refine the instrument before and after administration. In this sense, the study followed a development-oriented validation logic: evidence was collected to support the intended formative diagnostic interpretation of the assessment scores, not to make high-stakes decisions about students.

The validation framework was guided by the view that validity refers to the appropriateness of score interpretation and use rather than to a single statistical coefficient. Accordingly, the study combined content-related evidence, item-function evidence, reliability evidence, and exploratory evidence of internal structure. Content validity was examined through expert judgment using Aiken's V, while empirical item performance was evaluated using classical test theory, internal consistency estimates, and theory-guided exploratory factor analysis. This layered design allowed the study to examine whether the instrument was sufficiently coherent

for classroom formative diagnosis and to identify which items or dimensions required further refinement.

2.2 Sample

The participants were 128 Grade 8 students from two junior secondary schools in Medan, Indonesia. The implementation involved four Grade 8 classes. The two schools represented different classroom contexts, including urban and suburban settings and different levels of access to digital learning support. This variation was useful because the assessment was designed to function in a culturally grounded and hybrid inquiry-based learning environment, while still being administered as a supervised classroom assessment.

Students were included in the analytic dataset when they participated in the classroom administration of the instrument and produced valid responses for the 30 assessment items. The final response matrix consisted of 128 rows representing students and 30 columns representing the implemented multiple-choice items. Before analysis, the dataset was audited to check missing responses, invalid response codes, and duplicate student identifiers. The audit indicated that the dataset was suitable for item analysis and internal-structure exploration.

Because the participants came from two schools and four classes, the sample was treated as an implementation sample rather than a nationally representative sample. The results were therefore interpreted as initial psychometric evidence for formative diagnostic use in middle school classrooms. They were not used to justify high-stakes classification, cross-school ranking, or broad generalization across all Grade 8 students.

2.3 Instruments

The main instrument was a culturally grounded data literacy assessment consisting of 30 four-option multiple-choice items. Each item was scored dichotomously, with 1 assigned to a correct response and 0 assigned to an incorrect response. The items were distributed across six dimensions of data literacy: data knowledge, data management and preparation, analysis and

interpretation, critical evaluation and reflection, communication and presentation, and ethical-situational reasoning. Each dimension was represented by five items to maintain balance across the blueprint.

The first dimension, data knowledge, focused on students' understanding of variables, data types, simple representations, and basic statistical terms appropriate for Grade 8. The second dimension, data management and preparation, assessed students' ability to select relevant data, organize data, check units, recognize incomplete information, and prepare data for analysis. The third dimension, analysis and interpretation, measured students' ability to read patterns, compare groups, interpret graphs, and connect numerical results with contextual questions. The fourth dimension, critical evaluation and reflection, focused on evaluating claims, judging the sufficiency of data, identifying possible bias, and recognizing limitations of interpretation. The fifth dimension, communication and presentation, assessed students' ability to choose suitable representations and communicate data-based conclusions. The sixth dimension, ethical-situational reasoning, measured students' responses to situational problems involving privacy, responsible use of data, careful claims, and openness to revised conclusions.

Cultural grounding was embedded through item stimuli that used familiar and locally meaningful situations, such as household water consumption, traditional market data, school activities, community information, and privacy issues in the use of data. These contexts were used as bridges for meaning-making, not as hidden cultural knowledge required to answer the items. All necessary information was provided in the item stimulus so that students could solve the problem by applying data literacy skills rather than by relying on prior knowledge of a specific cultural practice.

Two supporting instruments were also used during development. The first was an expert-review form that asked reviewers to rate the items and blueprint using ordinal scales. The reviewed aspects included content relevance, clarity,

construct representation, cultural grounding, alignment with the hybrid inquiry learning context, and appropriateness for Grade 8 students. The second was a cultural-fairness checklist used to examine whether item language, contexts, labels, units, visual elements, and answer options could unintentionally advantage or disadvantage particular groups. The checklist was used as a design safeguard, not as evidence that the instrument was empirically free from cultural bias.

2.4 Procedure

The study was conducted in five phases. First, the research team mapped the construct, defined six dimensions, developed operational indicators, and prepared the assessment blueprint. Second, 30 multiple-choice items were written with contextual stimuli, one correct answer, and three distractors reflecting common misconceptions in data interpretation, graph reading, claim evaluation, and communication. Third, six experts reviewed the instrument for relevance, clarity, construct representation, cultural grounding, inquiry alignment, and Grade 8 appropriateness; their ratings were used for content-validity analysis and their comments guided item revision. Fourth, the revised instrument was administered in a supervised classroom setting after students participated in a hybrid inquiry-based learning sequence. Fifth, student responses were coded, audited, and analyzed for content validity, item functioning, reliability, internal structure, and item-decision mapping.

2.5 Data Analysis

Data analysis was conducted in several stages. First, expert-rating data were analyzed using Aiken's *V* to estimate the degree of agreement among experts regarding item relevance and other content-validity aspects. Aiken's *V* was selected because the expert judgments were collected using ordinal ratings and the coefficient is appropriate for summarizing expert-based evidence of content validity (Aiken, 1985). The coefficient was calculated for each review aspect, and confidence intervals were used to describe the uncertainty around the estimates. The results

were interpreted as content-validity evidence, not as complete evidence of score validity.

Second, student responses were analyzed using classical test theory. Item difficulty was calculated as the proportion of students who answered each item correctly. Item discrimination was estimated using the point-biserial correlation between each dichotomously scored item and the total score. These indices were used to examine item functioning and to support retain, review, and revise decisions for each item (Thorndike & Thorndike-Christ, 2010).

Third, internal consistency was summarized using Cronbach's alpha (α) and McDonald's omega (ω) (Hayes & Coutts, 2020). Alpha was reported because it is widely used in educational measurement and provides a familiar index of internal consistency. Omega was also reported because it is less restrictive than alpha when item contributions are not fully equivalent. The two coefficients were interpreted together with item statistics and dimensional structure, especially because each dimension contained only five items.

Fourth, the internal structure of the instrument was explored using exploratory factor analysis (EFA). Because the items were scored dichotomously, the EFA was based on a tetrachoric correlation matrix rather than a Pearson correlation matrix. Factors were extracted using minimum residual extraction and rotated using oblique rotation because the six dimensions of data literacy were theoretically expected to be related. The six-factor solution was retained as a theory-guided exploratory model based on the blueprint, the interpretability of the factor pattern, and the diagnostic purpose of the instrument. Factor loadings, cross-loadings, and communalities were examined to identify items that aligned well with their intended dimensions and items requiring revision. For reporting purposes, the exploratory factor solution was summarized for each

dimension using its dominant factor (F^*), the range of standardized factor loadings ($|\lambda|$), and the mean item communality ($\bar{h}_D^2$) (Marôco, 2024; Rogers, 2022).

Finally, all evidence was integrated to support item-level and dimension-level interpretation. Items with acceptable difficulty, adequate discrimination, meaningful factor loading, and clear alignment with the blueprint were retained. Items with weak discrimination, low communality, ambiguous factor behavior, or potential cultural-fairness concerns were flagged for review or revision. The total score and dimension scores were interpreted for formative diagnosis only. Further validation through cognitive interviews, categorical confirmatory factor analysis, measurement invariance, differential item functioning, and item response theory is recommended before the instrument is used for broader comparison or high-stakes assessment.

3. Results

3.1 Content Validity Evidence

The expert review provided content-validity evidence for the instrument. Six aspects were evaluated: relevance, clarity, representation, cultural grounding, alignment with hybrid inquiry, and appropriateness for Grade 8 students. Following the use of Aiken's V for ordinal expert ratings, the analysis summarized the degree of agreement among experts regarding the adequacy of the assessment blueprint and item content (Aiken, 1985). As shown in Table 1, the Aiken's V results and confidence intervals indicated consistently strong expert agreement across the reviewed aspects. These findings indicate that the instrument had sufficient content-validity evidence for classroom implementation. However, additional empirical evidence from student responses was still needed before the scores could be interpreted for broader or higher-stakes purposes.

Table 1. Content validity evidence based on expert review

Aspect	Mean Aiken's V	Lower 95% CI	Upper 95% CI	Items with $V \geq 0.75$
Relevance	0.921	0.824	1	26
Clarity	0.919	0.822	1	26
Representation	0.924	0.829	1	28
Cultural grounding	0.921	0.824	1	27
Hybrid inquiry alignment	0.921	0.824	1	27
Grade 8 appropriateness	0.922	0.826	1	27

3.2 Descriptive Score Profile

The assessment was administered to 128 Grade 8 students from two junior secondary schools and four classes. As shown in Table 2, the final instrument consisted of 30 four-option multiple-choice items, with a theoretical total-score range of 0–30. The total score had a Mean (M) of 15.81 and a Standard Deviation (SD) of 5.67. The median total score was 16, and the Interquartile Range (IQR) was 8, indicating adequate score variability for psychometric analysis. Across the six dimensions, each consisting of five items, the

mean scores ranged from 2.32 to 2.88 out of 5. The highest mean score was observed for D3 Analysis and interpretation (M = 2.88, SD = 1.73), followed by D4 Critical evaluation and reflection (M = 2.85, SD = 1.68). The lowest mean score was found for D1 Data knowledge (M = 2.32, SD = 1.80). These descriptive results suggest that the instrument did not produce ceiling effects and that student performance varied across the intended data literacy dimensions.

Table 2. Descriptive statistics for total and dimensional scores

Score variable	Number of items	Theoretical range	N	Mean	SD	Median	IQR
Total score	30	0–30	128	15.81	5.67	16	8
D1 Data knowledge	5	0–5	128	2.32	1.80	2	3
D2 Data management and preparation	5	0–5	128	2.50	1.66	2	3
D3 Analysis and interpretation	5	0–5	128	2.88	1.73	3	3
D4 Critical evaluation and reflection	5	0–5	128	2.85	1.68	3	2
D5 Communication and presentation	5	0–5	128	2.69	1.63	3	2.5
D6 Ethical-situational reasoning	5	0–5	128	2.58	1.59	2	3

3.3 Item Difficulty, Discrimination, and Item Decisions

Item performance was examined using the item difficulty index (p) and the point-biserial discrimination coefficient (r_{pb}). In Table 3, SR refers to the student-response item code, p

represents the proportion of students who answered an item correctly, and r_{pb} represents the correlation between each item score and the total score. As shown in Table 3, the item difficulty values ranged from 0.438 to 0.617, indicating that all 30 items were within the operational difficulty range of 0.30–0.70. The

point-biserial discrimination values ranged from 0.259 to 0.528. Based on the combined item statistics and post-implementation item-decision rules, 15 items were retained, 10 items were

flagged for review, and five items were prioritized for revision. The five priority revision items were SR13, SR21, SR24, SR27, and SR28.

Table 3. Item difficulty, discrimination, and post-implementation decisions

Item	p	r_{pb}	Decision	Item	p	r_{pb}	Decision
SR01	0.461	0.428	Retain	SR16	0.609	0.328	Review
SR02	0.445	0.497	Retain	SR17	0.609	0.365	Review
SR03	0.484	0.528	Retain	SR18	0.547	0.462	Retain
SR04	0.484	0.417	Retain	SR19	0.555	0.34	Review
SR05	0.445	0.428	Retain	SR20	0.531	0.371	Review
SR06	0.539	0.424	Retain	SR21	0.57	0.259	Revise
SR07	0.438	0.336	Review	SR22	0.539	0.344	Review
SR08	0.516	0.38	Retain	SR23	0.547	0.353	Review
SR09	0.562	0.411	Retain	SR24	0.523	0.262	Revise
SR10	0.445	0.439	Retain	SR25	0.508	0.366	Review
SR11	0.578	0.409	Retain	SR26	0.508	0.297	Review
SR12	0.516	0.466	Retain	SR27	0.500	0.282	Revise
SR13	0.586	0.278	Revise	SR28	0.477	0.267	Revise
SR14	0.617	0.389	Retain	SR29	0.586	0.334	Review
SR15	0.578	0.431	Retain	SR30	0.508	0.477	Retain

3.4 Reliability and Internal Structure

Internal consistency and exploratory internal-structure evidence were examined after the item-level analysis. As shown in Table 4, Cronbach's alpha ranged from 0.628 to 0.768, while McDonald's omega ranged from 0.770 to 0.843 across the six dimensions. D1, D3, and D4 showed relatively stronger reliability evidence,

whereas D5 and D6 produced lower reliability estimates and required closer review. The theory-guided exploratory factor analysis showed that each intended dimension was mainly associated with one dominant factor, although several items showed lower communality or weaker loading patterns.

Table 4. Reliability and internal structure summary

Dimension	Items	α	ω	F*	$ \lambda $	$\bar{h}_D^2$	Main decision
D1 Data knowledge	SR01–SR05	0.768	0.843	F1	0.471–0.774	0.416	Retain
D2 Data management and preparation	SR06–SR10	0.688	0.8	F6	0.465–0.614	0.317	Retain with review of weak items
D3 Analysis and interpretation	SR11–SR15	0.735	0.826	F2	0.470–0.716	0.424	Retain; revise SR13
D4 Critical evaluation and reflection	SR16–SR20	0.705	0.809	F4	0.444–0.703	0.381	Retain with review
D5 Communication and presentation	SR21–SR25	0.662	0.787	F3	0.393–0.713	0.354	Review and refine
D6 Ethical-situational reasoning	SR26–SR30	0.628	0.77	F5	0.230–0.882	0.351	Priority refinement

Note. F1–F6 denote the six factors extracted from the theory-guided EFA. Factor numbering reflects the statistical extraction and rotation results and does not indicate rank, importance, or correspondence with the numerical order of the theoretical dimensions.

The reliability and internal-structure results were consistent with the item-decision pattern in Table 3. D1, D2, D3, and D4 generally showed sufficient evidence for formative diagnostic interpretation, while D5 and D6 required further refinement. In particular, the communication and ethical-situational reasoning dimensions should be strengthened through revision of item wording, distractor quality, and situational clarity before broader use of the instrument.**4.**

4. Discussion

This study developed and evaluated a culturally grounded data literacy assessment for Grade 8 students by combining expert review, classical item analysis, reliability estimation, and exploratory factor analysis. The findings support the use of the instrument as a low-stakes formative diagnostic tool, but they do not justify broader or high-stakes score interpretations. This position is consistent with the view that validity depends on the intended interpretation and use of assessment results (AERA, APA, & NCME, 2014; Kane, 2013).

The cultural contexts used in the items also supported the design rationale. Situations involving household water use, traditional markets, school activities, community information, and data privacy were intended to make the tasks meaningful without requiring hidden cultural knowledge. This approach is consistent with research emphasizing the value of authentic data contexts in supporting students' interpretation of variability, evidence, and real-world claims (English & Watson, 2018; Wilkerson et al., 2022). Nevertheless, strong expert ratings should not be interpreted as proof that the instrument is free from cultural bias. Cultural grounding functions here as a fairness-oriented design strategy that still requires evidence from student response processes, group comparisons, and measurement-invariance analysis.

The item-level results further support the instrument's potential classroom use. Item-difficulty values ranged from 0.438 to 0.617, indicating that the items were neither excessively easy nor excessively difficult for the

implementation sample. The total-score distribution also showed no strong floor or ceiling effect. Point-biserial discrimination coefficients ranged from 0.259 to 0.528, indicating positive but varied relationships between individual items and overall performance. These findings are consistent with classical test theory, in which difficulty and discrimination are used together to evaluate item functioning (Thorndike & Thorndike-Christ, 2010).

The classification of 15 items as retain, 10 as review, and five as revise shows the importance of combining statistical and substantive evidence. The five priority items—SR13, SR21, SR24, SR27, and SR28—had relatively weak discrimination and may require improvement in wording, distractor plausibility, contextual clarity, or cognitive demand. However, item decisions should not be based on a single cut-off. Some items may still be needed to preserve coverage of communication, ethical reasoning, or other dimensions that would otherwise be underrepresented. Therefore, the retain–review–revise procedure offers a more balanced approach than automatically removing items with weaker statistics.

The reliability and exploratory internal-structure results showed that D1 Data knowledge, D3 Analysis and interpretation, and D4 Critical evaluation and reflection had the strongest overall patterns. D2 Data management and preparation also showed usable psychometric evidence, although some items still need review. In contrast, D5 Communication and presentation and D6 Ethical-situational reasoning had lower alpha coefficients and less consistent loading and communality patterns. The higher omega values compared with alpha are reasonable when item contributions are not fully equivalent (Hayes & Coutts, 2020).

The weaker results for D5 and D6 may reflect both item-quality and item-format limitations. Communication and ethical reasoning are complex abilities that may not be fully represented by five dichotomously scored multiple-choice items. Selecting an appropriate

representation, explaining a conclusion, considering privacy, or revising a claim often involves judgment that cannot be captured completely by a single correct response. These dimensions may therefore benefit from clearer scenarios, stronger distractors, and complementary formats such as short constructed responses or situational judgment tasks. The factor results should also remain exploratory; the dominant-factor assignments summarize the observed pattern but do not confirm the proposed six-dimensional structure (Marôco, 2024; Rogers, 2022).

Practically, teachers may use the total score as a broad indicator of data literacy and the dimensional scores as diagnostic signals of specific learning needs. The item results can also help identify misconceptions and guide follow-up activities, worksheet revision, or targeted classroom discussion. Even so, the dimension scores, especially D5 and D6, should be interpreted cautiously because each dimension contains only five items and some show weaker reliability evidence. The instrument is therefore more appropriate as one source of formative information than as a stand-alone basis for consequential decisions.

Several limitations should be considered when interpreting the findings. First, the study involved 128 students from two schools and four classes; therefore, the generalizability of the results should be examined in larger and more diverse samples. Second, each dimension was represented by five items, which may partly explain the lower reliability estimates observed for some dimensions, particularly communication and presentation and ethical-situational reasoning. Third, the six-factor structure was examined through a theory-guided exploratory analysis and should therefore be confirmed using an independent sample. Finally, cultural fairness was evaluated through expert review and a design checklist; further evidence from cognitive interviews, measurement invariance, and differential item functioning is needed to examine how the items function across student groups.

Future studies should therefore involve larger and more diverse samples, cognitive interviews, categorical confirmatory factor analysis, measurement invariance, differential item functioning, and item response theory. Further refinement is especially needed for the communication and ethical-situational reasoning dimensions. Finally, because the assessment was administered within a HIEVS learning context without a comparison group or pretest–posttest design, the findings cannot be used to evaluate the effectiveness of HIEVS. The instrument should currently be viewed as a formative diagnostic assessment supported by teacher judgment rather than as a high-stakes measure.**5.**

5. Conclusion

This study developed and evaluated a culturally grounded data literacy assessment for Grade 8 students. Expert review showed strong content-validity evidence, all items fell within the operational difficulty range, and all items showed positive point-biserial discrimination. Of the 30 items, 15 were retained, 10 were identified for review, and five—SR13, SR21, SR24, SR27, and SR28—were prioritized for revision. Reliability and exploratory internal-structure evidence supported the formative diagnostic use of the instrument, particularly for data knowledge, analysis and interpretation, and critical evaluation and reflection.

Communication and presentation as well as ethical-situational reasoning require further refinement because of their lower reliability and less consistent loading and communality patterns. The instrument should therefore be used for low-stakes formative diagnosis rather than high-stakes decisions. Further validation should involve larger and more diverse samples, cognitive interviews, categorical confirmatory factor analysis, measurement invariance, differential item functioning, item response theory, external-criterion evidence, and independent replication. Because the instrument was implemented within a HIEVS learning context without an experimental comparison, the findings should not be interpreted as evidence of HIEVS instructional effectiveness.

References

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Cui, Y., Chen, F., Lutsyk, A., Leighton, J. P., & Cutumisu, M. (2023). Data literacy assessments: A systematic literature review. *Assessment in Education: Principles, Policy & Practice*, 30(1), 76–96. <https://doi.org/10.1080/0969594X.2023.2182737>
- Donate-Bebby, B., García-Peñalvo, F. J., Amo-Filva, D., & Aguayo-Mauri, S. (2025). Filling the gap in K–12 data literacy competence assessment: Design and initial validation of a questionnaire. *Computers in Human Behavior Reports*, 17, Article 100583. <https://doi.org/10.1016/j.chbr.2024.100583>
- Duygulu, A., Doğan, S., & Yıldız, S. (2025). Data literacy at school: A scale development study. *Kuramsal Eğitim Bilim Dergisi*, 18(1), 106–130. <https://doi.org/10.30831/akukey.1472451>
- English, L. D., & Watson, J. M. (2018). Modelling with authentic data in sixth grade. *ZDM Mathematics Education*, 50(1–2), 103–115. <https://doi.org/10.1007/s11858-017-0896-y>
- Friedrich, A., Schreiter, S., Vogel, M., Becker-Genschow, S., Brünken, R., Kuhn, J., Lehmann, J., & Malone, S. (2024). What shapes statistical and data literacy research in K–12 STEM education? A systematic review of metrics and instructional strategies. *International Journal of STEM Education*, 11, Article 58. <https://doi.org/10.1186/s40594-024-00517-z>
- Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But... *Communication Methods and Measures*, 14(1), 1–24. <https://doi.org/10.1080/19312458.2020.1718629>
- 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>
- Linn, R. L. (2006). The standards for educational and psychological testing: Guidance in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 27–38). Lawrence Erlbaum Associates.
- Marôco, J. (2024). Factor analysis of ordinal items: Old questions, modern solutions? *Stats*, 7(3), 984–1001. <https://doi.org/10.3390/stats7030060>
- Pratama, R. A., & Yelken, T. Y. (2024). Effectiveness of ethnomathematics-based learning on students' mathematical literacy: A meta-analysis study. *Discover Education*, 3, Article 202. <https://doi.org/10.1007/s44217-024-00309-1>
- Ramadhani, R., Syahputra, E., & Simamora, E. (2025). Enhancing students' informal statistical reasoning skills through the development of a worksheet rooted in the Nias culture. *Mathematics Teaching Research Journal*, 17(2), 104–128.
- Rogers, P. (2022). Best practices for your exploratory factor analysis: A Factor tutorial. *Revista de Administração Contemporânea*, 26(6), Article e210085. <https://doi.org/10.1590/1982-7849rac2022210085.en>
- Thorndike, R. M., & Thorndike-Christ, T. M. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Pearson.
- Wilkerson, M. H., Lanouette, K., & Shareff, R. L. (2022). Exploring variability during data preparation: A way to connect data, chance, and context when working with complex public datasets. *Mathematical Thinking and Learning*, 24(4), 312–330. <https://doi.org/10.1080/10986065.2021.1922838>
- Witte, V., Schwering, A., & Frischmeier, D. (2025). Strengthening data literacy in K–12 education: A scoping review. *Education Sciences*, 15(1), Article 25. <https://doi.org/10.3390/educsci15010025>

Acknowledgements

This research was funded by the Directorate of Research and Community Service, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, through the 2025 Doctoral Dissertation Research Scheme, Contract No. 002/UN33.8/DPPM/PL/2025.

Author Declaration

The authors declare that this manuscript is an original work, has not been published previously, and is not currently under consideration for publication elsewhere. All authors have made substantial contributions to the study, have reviewed and approved the final version of the manuscript, and agree to be accountable for the accuracy and integrity of the work. The authors declare that there are no conflicts of interest related to the research, authorship, or publication of this article. The funding source had no role in the study design, data collection, data analysis, interpretation of the findings, or preparation of the manuscript.