BAREKENG: Journal of Mathematics and Its Applications
December 2026 Volume 20 Issue 4 Page 3423-3436
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss4pp3423-3436

COMBINATION OF CONTENT VALIDITY INDEX AND FUZZY DELPHI FOR EXPERT VALIDATION OF A THREE-TIER MATHEMATICS LEARNING DIFFICULTIES DIAGNOSTIC ASSESSMENT

Khaerun Nisa [1*], **Iva Sarifah** [2], **Riyadi** [3]

[1,2,3] *Educational Research and Evaluation Program, Graduate School, Universitas Negeri Jakarta*
[1] *Research Center for Education, National Research and Innovation Agency of Indonesia* (BRIN)
[1,2,3] *Rawamangun Muka Raya Street, East Timur, 13220, Indonesia*
[1] *Jenderal Gatot Subroto Street, South Jakarta 12710, Indonesia*

Corresponding author's e-mail: ***khaerun.nisa@mhs.unj.ac.id***

Article Info

Article History:

Received: 27th October 2025
Revised: 27th April 2026
Accepted: 04th June 2026
Available online: 01st July 2025

Keywords:

Content validity Index;
Diagnostic assessment;
Expert validation;
Fuzzy Delphi method;
Mathematics difficulties;
Three tier multiple choice.

ABSTRACT

The persistent challenges in identifying mathematics learning difficulties require diagnostic instruments that are not only valid but also capable of accurately capturing students' conceptual understanding, reasoning, and confidence. This study aims to compare the Content Validity Index (CVI) and the Fuzzy Delphi Method (FDM) in expert validation of a three-tier diagnostic assessment for mathematics learning difficulties. Theoretically, the study integrates classical content validation theory with fuzzy logic-based expert consensus analysis to strengthen the methodological rigor of instrument development. The research involved five doctoral-level experts who evaluated 21 diagnostic items. The CVI results yielded an excellent S-CVI/Ave value of 0.97, indicating high content validity, while the FDM results confirmed 20 of 21 items met the validity thresholds ($d \leq 0.2$, consensus $\geq 75\%$, $A \geq 0.5$). A key finding shows that the CVI provides clear numerical validation, whereas the FDM captures linguistic uncertainty and confidence variations among experts, producing a more comprehensive validation outcome. The integration of both methods offers a robust framework for expert validation that enhances interpretive depth and accuracy. This study contributes to the field of mathematics education by presenting a validated, reliable, and pedagogically meaningful diagnostic instrument for identifying students' misconceptions and reasoning patterns.

How to cite this article:

K. Nisa, I. Sarifah and Riyadi., "COMBINATION OF CONTENT VALIDITY INDEX AND FUZZY DELPHI FOR EXPERT VALIDATION OF A THREE-TIER MATHEMATICS LEARNING DIFFICULTIES DIAGNOSTIC ASSESSMENT", *BAREKENG: J. Math. & App.*, vol. 20, no. 4, pp. 3423-3436, Dec, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Assessment plays a crucial role in the learning process as it provides feedback for improving instructional quality and evaluating the achievement of learning objectives [1], [2]. In mathematics education, assessment not only measures students' performance but also identifies the underlying causes of learning difficulties that may hinder conceptual understanding and problem-solving ability [3], [4]. The effectiveness of mathematics learning is therefore strongly influenced by the quality of the assessment instruments used to evaluate students' knowledge, reasoning, and confidence [5]. Valid and reliable assessment tools are essential to support diagnostic decisions and to design appropriate learning interventions [6].

In the Indonesian context, mathematics learning difficulties remain a significant concern at the secondary school level. Results from national and international assessments such as Programme for International Student Assessment (PISA) and AKM (*Asesmen Kompetensi Minimum*) have consistently shown that many Indonesian students struggle with mathematical reasoning, problem-solving, and conceptual understanding [7]. These challenges have been further exacerbated by learning loss during and after the COVID-19 pandemic, which has widened the gap in students' cognitive mastery and conceptual readiness [8]. Teachers often rely on traditional multiple-choice tests or summative assessments that emphasize procedural computation rather than diagnostic evaluation [9]. Furthermore, the current national curriculum (Kurikulum Nasional) encourages the implementation of deep learning (*pembelajaran mendalam*) that emphasizes understanding, reasoning, and the ability to connect concepts rather than rote memorization [10]. This shift requires more accurate diagnostic assessment tools capable of identifying students' conceptual understanding and reasoning pathways [11]. However, several studies have reported that diagnostic instruments used in Indonesian schools still lack systematic validation, particularly in terms of content validity and cognitive alignment [12]. This condition highlights the urgent need to develop and validate high-quality diagnostic tools that can help teachers accurately identify students' misconceptions and levels of understanding.

Multiple-choice tests are the most widely used form of assessment because of their efficiency and objectivity [13]. However, conventional multiple-choice tests often fail to reveal students' reasoning processes or the confidence behind their chosen answers [14]. As a result, teachers cannot accurately distinguish between students who genuinely understand the concept and those who answer correctly by guessing [15]. To address this limitation, the three-tier multiple-choice diagnostic assessment was developed. To address this limitation, the three-tier multiple-choice diagnostic assessment was developed. This instrument consists of three components: content knowledge, reasoning ability, and confidence level [16]. The integration of these three components enables educators to diagnose not only what students misunderstand but also why misconceptions occur and how confident students are in their responses. Previous research has shown that the three-tier diagnostic format effectively identifies misconceptions and learning difficulties in mathematics [17], [18], [19].

The development of a diagnostic assessment must be supported by a rigorous content-validation process to ensure that each item accurately represents the intended construct [20]. One widely used method for evaluating expert judgment in content validation is the Content Validity Index (CVI), which quantifies expert agreement on the relevance, clarity, and representativeness of instrument items [21], [22]. However, the CVI method does not fully address the uncertainty inherent in linguistic expert judgments [23]. To overcome this limitation, the Fuzzy Delphi Method (FDM) integrates the Delphi consensus technique with fuzzy logic to manage the vagueness and uncertainty of linguistic assessments [24], [25]. This method transforms experts' linguistic evaluations into triangular fuzzy numbers (TFNs) that capture both the level of agreement and the degree of confidence [26]. Recent studies have demonstrated that the FDM enhances the reliability of validation processes and provides a more accurate representation of expert consensus compared to the traditional Delphi technique [26], [27], [28], [29]. Although many studies have independently applied the Content Validity Index (CVI) and the Fuzzy Delphi Method (FDM) in various educational contexts, few have examined both approaches simultaneously within the domain of diagnostic assessment, particularly in mathematics learning difficulties [20], [30]. Therefore, this study aims to evaluate and compare the effectiveness of the CVI and the FDM in the expert validation of a three-tier diagnostic assessment designed to identify mathematics learning difficulties among Grade X students. By integrating these two validation approaches, this study strengthens methodological rigor and contributes to the development of empirically grounded diagnostic tools aligned with the vision of Kurikulum Nasional to promote deep learning and data-driven educational improvement [31], [32], [33],[34].

2. RESEARCH METHODS

2.1 Research Design

This study employed a comparative quantitative descriptive design to evaluate and compare the outcomes of two expert validation approaches: the Content Validity Index (CVI) and the Fuzzy Delphi Method (FDM). The research aimed to examine the degree of consensus, consistency, and stability among experts in validating a three-tier diagnostic assessment of mathematics learning difficulties. The CVI technique was used to quantify expert agreement on item relevance, while the FDM incorporated fuzzy-logic computation to address linguistic uncertainty and refine consensus measurement. The combination of both methods provided a comprehensive framework for assessing content validity, ensuring methodological rigor prior to field implementation. The methodological workflow of this study is summarized in Fig. 1. The validation procedure consisted of sequential stages including instrument development, expert panel selection, expert evaluation across seven assessment aspects, Content Validity Index (CVI) analysis, Fuzzy Delphi Method (FDM) analysis using triangular fuzzy numbers, comparison of validation outcomes from both approaches, and interpretation of results to determine the final validity status of each diagnostic item. This workflow was designed to ensure transparency and reproducibility of the validation process.

Figure 1. Research workflow of the validation procedure

2.2 Participants

The expert panel consisted of five professionals selected through purposive sampling based on specific inclusion criteria: doctoral qualifications in mathematics education, a minimum of five years of teaching or research experience, and familiarity with diagnostic assessment development. These experts represented diverse academic institutions and research backgrounds to ensure a balanced evaluation of the instrument. Each expert independently reviewed the instrument and provided quantitative and qualitative feedback to ensure impartiality and reliability of the validation process. Table 1 presents the demographic profile of the experts.

Table 1. Demographic Profile of the Experts

Expert Code	Highest Qualification	Field of Expertise	Academic Position	Years of Experience
E1	Ph.D. in Educational Research and Evaluation	Diagnostic Assessment, Learning Difficulties	Senior Researcher	15 years
E2	Ph.D. in Mathematics Education	Curriculum & Instruction in Math	Senior Researcher	12 years
E3	Ph.D. in Educational Research and Evaluation	Cognitive Psychology in Mathematics	Senior Lecturer	10 years
E4	Ph.D. in Educational Research and Evaluation	Mathematics Education, Psychometric, Measurement	Lecturer	8 years
E5	Ph.D. in Educational Research and Evaluation	Mathematics Learning, Conceptual Reasoning	Associate Professor	14 years

2.3 Instrument

The instrument validated in this study was a three-tier diagnostic assessment designed to identify students' misconceptions and learning difficulties in mathematics. Each item consisted of a problem statement, a reasoned response, and a confidence rating tier. The initial instrument included 21 items aligned with the mathematics curriculum and conceptual domains related to algebraic reasoning and problem-solving. The instrument's development was guided by literature on diagnostic assessment and validated learning theories to ensure content representativeness and construct alignment prior to expert review. To improve instrument transparency, a representative example of a three-tier diagnostic item is presented directly in this section together with the mapping between diagnostic items and assessed indicators shown in Table 2. Each diagnostic item consists of three components: the first tier measures content knowledge through multiple-choice responses, the second tier evaluates the reasoning underlying the selected answer, and the third tier captures students' confidence levels. This structure enables identification of misconceptions, reasoning patterns, and certainty levels simultaneously, thereby supporting a more comprehensive diagnosis of students' understanding of mathematical growth models.

2.3.1 Example of a Three-Tier Diagnostic Item

Have you ever consumed yogurt? The sour taste of yogurt is produced by the activity of beneficial bacteria called *Lactobacillus*, which play an important role in human digestion and are widely used in fermentation processes. A researcher investigated the growth of *Lactobacillus* bacteria in a glass of milk at room temperature. The observation results showed that the number of bacteria increased rapidly every hour, as presented in the Table 2 below:

Table 2. Growth of Lactobacillus Bacteria at Room Temperature

Time (hours)	Number of Bacteria
0	500
1	1000
2	2000
3	4000
⋮	⋮
t	?

Based on the observation data, the researcher proposed two possible models of bacterial growth:

Model X:
$f(t) = 500 \times 2^t$

Model Y:
$f(t) = 500 + 1500t$

Tier 1 – Content Knowledge (Answer Selection)

Based on Model X and Model Y, it can be concluded that:

A. Model X is correct, whereas Model Y is incorrect
B. Model X is incorrect, whereas Model Y is correct
C. Both Model X and Model Y are correct
D. Both Model X and Model Y are incorrect

Tier 2 – Reasoning (Justification Selection)

I selected the answer above because:

A. The number of bacteria increases by a constant amount every hour
B. The number of bacteria always doubles from the previous hour
C. Both models produce the same result at the initial time point, so both can be used
D. At the third hour, the number of bacteria increases by only 2000 from the previous hour

Tier 3 – Confidence Level

I am _______ about my answer.

A. Confident
B. Uncertain
C. Not confident

2.3.2 Mapping Between Diagnostic Items and Assessed Indicators

Table 3. Mapping Between Diagnostic Items and Assessed Indicators

Item	Context	Mathematical Indicator Assessed	Tier 1	Tier 2	Tier 3
1	Bacterial growth	Distinguishing exponential vs linear growth models	Concept selection	Reasoning selection	Confidence level
2	Ticket price increase	Comparing arithmetic vs percentage increase models	Concept selection	Reasoning selection	Confidence level
3	Livestock population growth	Identifying linear vs exponential growth patterns	Concept selection	Reasoning selection	Confidence level

Item	Context	Mathematical Indicator Assessed	Tier 1	Tier 2	Tier 3
4	Asset value growth	Evaluating suitability of linear vs exponential models	Concept selection	Reasoning selection	Confidence level
5	Population growth prediction	Evaluating appropriateness of exponential modelling	Concept selection	Reasoning selection	Confidence level
6	Savings vs inflation	Comparing linear savings vs exponential inflation effects	Concept selection	Reasoning selection	Confidence level
7	Temperature vs radioactive decay	Identifying exponential change vs linear change processes	Concept selection	Reasoning selection	Confidence level
8	Fish population growth	Selecting logistic growth model with carrying capacity	Concept selection	Reasoning selection	Confidence level
9	Population growth revision	Constructing exponential growth model from percentage rate	Concept selection	Reasoning selection	Confidence level
10	Social media video views	Modelling exponential growth with percentage increase	Concept selection	Reasoning selection	Confidence level
11	Machine depreciation	Modelling exponential decay processes	Concept selection	Reasoning selection	Confidence level
12	Social media followers	Identifying exponential growth from geometric sequence pattern	Concept selection	Reasoning selection	Confidence level
13	Carbon emission growth	Constructing exponential growth function from percentage rate	Concept selection	Reasoning selection	Confidence level
14	Cloud storage capacity	Interpreting exponential doubling-time growth models	Concept selection	Reasoning selection	Confidence level
15	Influenza case growth	Evaluating consistency between model and graphical data	Concept selection	Reasoning selection	Confidence level
16	Electricity consumption growth	Evaluating suitability of exponential growth models	Concept selection	Reasoning selection	Confidence level
17	Polar ice volume decrease	Identifying exponential decay patterns from graphs	Concept selection	Reasoning selection	Confidence level
18	Tourist growth	Constructing exponential growth prediction models	Concept selection	Reasoning selection	Confidence level
19	Carbon emission reduction	Modelling exponential decay with percentage decrease	Concept selection	Reasoning selection	Confidence level
20	Plastic waste growth	Constructing exponential growth projection models	Concept selection	Reasoning selection	Confidence level
21	Plastic waste growth projection	Constructing an exponential growth model from percentage rate and determining threshold time when a quantity double	Concept selection	Reasoning selection	Confidence level

2.4 Data Collection and Expert Evaluation Procedure

Data were collected in two stages corresponding to the CVI and FDM analyses. Each diagnostic item was evaluated across seven assessment aspects by five experts. For the CVI analysis, aspect-level ratings were first dichotomized into relevant (ratings 3–4) and not relevant (ratings 1–2) categories following standard CVI procedures. The Item-Level Content Validity Index (I-CVI) was then computed by aggregating expert agreement across all evaluated aspects for each item. Similarly, for the FDM analysis, linguistic ratings assigned to each aspect were converted into triangular fuzzy numbers and averaged across aspects to obtain a composite fuzzy representation for each item prior to threshold calculation and defuzzification. This aggregation procedure ensured that the final CVI and FDM values reflected the overall expert evaluation across all seven aspects rather than a single-aspect judgment.

These seven evaluation aspects were selected to ensure comprehensive coverage of content validity indicators in diagnostic assessment development. Alignment with competencies or learning objectives was used to determine whether each item appropriately represented the intended curriculum targets and conceptual domains. Linguistic clarity assessed whether the wording of each item was understandable, precise, and free from ambiguity for secondary school students. Objectivity examined whether each item avoided multiple interpretations and maintained a clear conceptual focus. Distractor effectiveness evaluated whether incorrect answer options functioned appropriately in identifying students' misconceptions rather than encouraging guessing. Conformity with higher-order thinking skills (HOTS) ensured that the items measured reasoning ability and conceptual understanding rather than procedural recall. Absence of bias assessed whether the items were fair and free from cultural, linguistic, or contextual disadvantages for particular student groups. Finally, relevance of stimulus material evaluated whether the contextual information provided in each item supported meaningful interpretation and facilitated students' reasoning processes. In the second stage, these linguistic evaluations were converted into fuzzy numerical values to facilitate analysis using the Fuzzy Delphi Method (FDM). Expert responses were collected through online forms and compiled for subsequent data analysis. Table 4 presents the complete expert evaluation data for all 21 diagnostic items across the seven validation aspects used in the Content Validity Index (CVI) and Fuzzy Delphi Method (FDM) analyses.

Table 4. Expert Evaluation Data for CVI and FDM Analyses

Item Code	Evaluation Aspect	Expert Evaluation Data					Linguistic Category	TFN (l,m,u)
		1	2	3	4	5		
Item 1	Alignment with learning objectives	4	3	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 2	Linguistic clarity	3	3	4	3	4	Relevant	(0.25, 0.50, 0.75)
Item 3	Objectivity	4	4	4	3	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 4	Distractor effectiveness	4	4	3	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 5	HOTS conformity	3	4	4	3	4	Relevant	(0.25, 0.50, 0.75)
Item 6	Absence of bias	4	4	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 7	Stimulus relevance	4	3	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 8	Alignment with learning objectives	3	3	3	3	4	Relevant	(0.25, 0.50, 0.75)
Item 9	Linguistic clarity	3	4	4	3	4	Relevant	(0.25, 0.50, 0.75)
Item 10	Objectivity	4	4	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 11	Distractor effectiveness	3	3	4	3	4	Relevant	(0.25, 0.50, 0.75)
Item 12	HOTS conformity	4	4	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 13	Absence of bias	4	4	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 14	Stimulus relevance	4	4	4	3	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 15	Alignment with learning objectives	3	4	4	3	4	Relevant	(0.25, 0.50, 0.75)
Item 16	Linguistic clarity	4	4	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 17	Objectivity	4	4	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 18	Distractor effectiveness	4	4	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 19	HOTS conformity	3	4	4	4	4	Relevant	(0.25, 0.50, 0.75)
Item 20	Absence of bias	4	4	4	4	4	Highly Relevant	(0.50, 0.75, 1.00)
Item 21	Stimulus relevance	3	3	3	4	3	Relevant	(0.25, 0.50, 0.75)

Note. TFN = Triangular Fuzzy Number. Expert ratings were given on a four-point Likert scale (1 = not relevant to 4 = highly relevant). Linguistic categories and TFN representations follow the fuzzy scale parameters shown in Table 4.

2.4.1 Content Validity Index (CVI) Analysis

The Content Validity Index (CVI) was employed to evaluate the degree of agreement among experts regarding the relevance and clarity of each item in the instrument. This analysis determines whether each item adequately represents the construct being measured. At the item level, the Item-Level Content Validity Index (I-CVI) was calculated using Eq. (1):

$$I\text{-}CVI = \frac{n_e}{N}, \tag{1}$$

where n_e denotes the number of experts who rated the item as either 3 (Relevant) or 4 (Highly Relevant) based on the four-point scale (see Table 1), and N denotes the total number of experts. The Scale-Level Content Validity Index (S-CVI/Ave) was then computed to assess the overall content validity of the instrument using Eq. (2):

$$\frac{S\text{-}CVI}{Ave} = \frac{\sum I\text{-}CVI}{k}, \tag{2}$$

where k= total number of items evaluated. According to Polit and Beck [35], an item is considered to have satisfactory content validity if $I\text{-}CVI \geq 0.78$, and the instrument was considered valid overall if $S\text{-}CVI/Ave \geq 0.90$.

2.4.2 Fuzzy Delphi Method (FDM) Analysis

The Fuzzy Delphi Method (FDM) was adopted to minimize ambiguity and uncertainty in expert judgments. This method combines the strengths of the classical Delphi technique and fuzzy set theory, enabling a more systematic and quantitative interpretation of experts' opinions. In this study, the experts' evaluations were expressed using four linguistic terms representing the degree of relevance of each item. These linguistic responses were then converted into Triangular Fuzzy Numbers (TFN) to facilitate quantitative analysis. Specifically, each linguistic response was represented as a TFN $M = (l, m, u)$, where l denotes the lowest possible value, m denotes the most representative value, and u denotes the highest possible value. Triangular fuzzy numbers (TFNs) were selected in this study because they provide a simple yet effective representation of linguistic expert judgments and are widely used in educational validation research due to their computational efficiency and interpretability. Compared with other fuzzy representations such as trapezoidal fuzzy numbers, TFNs allow straightforward aggregation of expert opinions while preserving uncertainty information in linguistic assessments. The defuzzification procedure using the average of the three TFN parameters was adopted because it provides a stable and commonly accepted approach for transforming fuzzy consensus values into crisp scores suitable for decision-making in validation studies. The linguistic scales and their corresponding TFN values are presented in Table 5.

Table 5. Fuzzy Scale Parameters

Scale	Interpretations	TFN (l, m, u)
1	Not Relevant / Not Clear	(0.00, 0.00, 0.25)
2	Somewhat Relevant	(0.00, 0.25, 0.50)
3	Relevant	(0.25, 0.50, 0.75)
4	Highly Relevant	(0.50, 0.75, 1.00)

Source: Adapted from [25] and [36]

As shown in Table 5, experts were asked to rate each item based on its level of relevance. The responses ranged from *"Not Relevant/Not Clear"* to *"Highly Relevant."* Each linguistic term corresponds to a Triangular Fuzzy Number (TFN) that represents the expert's judgment in a numerical form between 0 and 1. These fuzzy numbers were later used to compute the consensus among experts. The degree of agreement between experts was quantified using a threshold value *(d)* for each item, calculated with Eq. (3):

$$d = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sqrt{\frac{1}{3}\left[\left(l_i - l_j\right)^2 + \left(m_i - m_j\right)^2 + \left(u_i - u_j\right)^2\right]}. \tag{3}$$

An item was considered accepted when the computed threshold value (d) ≤ 0.2 and at least 75% of experts reached consensus [36]. After verifying consensus, a defuzzification process was applied to obtain a crisp (definite) value for each item. The defuzzified value was computed using the average of the three fuzzy parameters as shown in Eq. (4):

$$A = \frac{l + m + u}{3}. \tag{4}$$

This defuzzified value represents the final consensus score for each item. Items with higher defuzzified values indicate greater perceived relevance by the experts, and were therefore prioritized in the subsequent stages of analysis.

2.5 Data Analysis Procedure

The results from both CVI and FDM analyses were compared to determine the level of convergence and consistency between the two validation methods. Items fulfilling both CVI and FDM criteria were categorized as *strongly valid*, while discrepancies were examined qualitatively to interpret variations in expert consensus and the influence of fuzzy logic in refining judgments. The final output of this research was a validated three-tier diagnostic assessment instrument that accurately identifies students' mathematics learning difficulties and provides a reliable tool for further empirical testing in Indonesian high school settings.

3. RESULTS AND DISCUSSION

3.1 Content Validity Index Validity (CVI) Analysis

The results of the Content Validity Index (CVI) analysis, as presented in Table 6, indicate that all twenty-one items achieved I-CVI values ranging from 0.82 to 1.00, surpassing the minimum criterion of 0.78 recommended by Polit and Beck [37]. This finding demonstrates that all items were judged valid in terms of relevance, clarity, and representativeness. The overall S-CVI/Ave value of 0.97 further confirms that the diagnostic assessment instrument possesses excellent content validity according to the classification proposed by Polit and Beck in [37]. These results suggest a high degree of agreement among the five doctoral-level experts regarding the quality and appropriateness of the items. Items with slightly lower I-CVI values (e.g., Items 9, 11, and 15) were reviewed to ensure that minor linguistic ambiguities were refined during subsequent revisions. Nevertheless, all items met or exceeded the minimum acceptance threshold, confirming that the instrument comprehensively represents the construct of mathematics learning difficulties and aligns with the theoretical framework underlying its development. The consistently high I-CVI and S-CVI/Ave values also reflect the clarity and conceptual adequacy of the instrument's content. According to Polit and Beck in [37], CVI values above 0.90 demonstrate excellent inter-rater agreement, implying that experts share a strong understanding of item intent and construct alignment. Similar results have been reported in recent validation studies by Afrizon *et al* [39], where high CVI scores confirmed that instrument items accurately captured the intended domains of learning. In this study, the limited variation in I-CVI values (0.82–1.00) suggests minimal differences in expert interpretation, further emphasizing the instrument's coherence and content relevance.

Table 6. Results of Content Validity Index (CVI) Analysis

Item Code	I-CVI	Criteria (≥ 0.78)	Decision
Item 1	1.00	≥ 0.78	Valid
Item 2	0.94	≥ 0.78	Valid
Item 3	0.94	≥ 0.78	Valid
Item 4	0.94	≥ 0.78	Valid
Item 5	1.00	≥ 0.78	Valid
Item 6	1.00	≥ 0.78	Valid
Item 7	1.00	≥ 0.78	Valid
Item 8	1.00	≥ 0.78	Valid
Item 9	0.89	≥ 0.78	Valid
Item 10	1.00	≥ 0.78	Valid
Item 11	0.82	≥ 0.78	Valid
Item 12	1.00	≥ 0.78	Valid
Item 13	1.00	≥ 0.78	Valid
Item 14	1.00	≥ 0.78	Valid
Item 15	0.94	≥ 0.78	Valid
Item 16	1.00	≥ 0.78	Valid
Item 17	1.00	≥ 0.78	Valid
Item 18	1.00	≥ 0.78	Valid
Item 19	0.94	≥ 0.78	Valid
Item 20	1.00	≥ 0.78	Valid
Item 21	1.00	≥ 0.78	Valid

Furthermore, the high S-CVI/Ave value (0.97) provides empirical support for the overall reliability of expert evaluations. According to Polit and Beck in [22], consistently high S-CVI/Ave values signify robust consensus among evaluators and strengthen the validity argument of the instrument. Similarly, Yusoff [40] emphasized that an S-CVI/Ave value above 0.90 indicates excellent inter-rater agreement and reflects the clarity and coherence of the construct being assessed. Hence, the results of this study affirm that the developed three-tier diagnostic assessment has achieved excellent content validity, serving as a reliable foundation for subsequent validation using the Fuzzy Delphi Method (FDM). The transition from CVI to FDM allows for a more nuanced validation process by addressing the linguistic uncertainty often present in expert judgments and quantifying the degree of consensus with greater precision.

3.2 Result of Fuzzy Delphi Method (FDM) Analysis

The results of the Fuzzy Delphi Method (FDM) analysis are presented in Table 7. Based on the evaluations of five experts, the threshold (d) values for all items ranged from 0.00 to 0.25, while the percentage of consensus varied between 75% and 100%. These findings indicate that expert judgments were highly consistent across all items. Out of the 21 items analyzed, 20 fulfilled the FDM validity criteria ($d \leq 0.2$, consensus ≥ 75%, and $A \geq 0.5$), suggesting that most items were clearly understood and strongly agreed upon by the experts. Only Item 8 displayed a slightly higher threshold value ($d = 0.25$) and was thus identified for potential revision to enhance its clarity and contextual accuracy. The defuzzified mean (A) values ranged from 0.50 to 0.75, indicating that experts generally perceived the items as relevant and representative of the intended constructs. The rejection of Item 8 under the FDM criteria, despite meeting the CVI threshold, indicates that expert agreement on this item was close to the decision boundary and showed greater dispersion in linguistic evaluations compared with other items. Although the numerical agreement in CVI remained acceptable, the fuzzy consensus analysis revealed a slightly higher threshold value (d = 0.25), suggesting variability in expert perceptions regarding the clarity or contextual suitability of the item. This result demonstrates the sensitivity of the FDM approach in detecting subtle differences in expert opinion that may not be fully captured by conventional agreement indices such as CVI. These results are consistent with findings reported by Abdul Rahman *et al.* [27] and Hasim *et al.* [28], who observed that defuzzified mean values within this range demonstrate strong consensus and confirm the adequacy of item content. The narrow range of d-values and the high consensus percentages observed in this study further reinforce the stability of expert agreement, in line with the fuzzy-based consensus principles described by Roldán López de Hierro *et al.*[24]. The use of the FDM in this study not only quantified consensus but also addressed linguistic uncertainty that often accompanies qualitative expert judgments. By transforming linguistic evaluations into triangular fuzzy numbers (TFNs), the FDM enabled a more precise representation of both the level of agreement and the degree of confidence among experts [41]. This capability distinguishes the FDM from traditional Delphi techniques, which rely solely on descriptive or statistical aggregation and may not fully capture expert ambiguity. Similar to the results of Pazilah *et al.* [29] and Yusoff [40], the present findings demonstrate that the FDM yields a more refined and reliable measure of expert consensus in educational validation studies.

Table 7. Results of Fuzzy Delphi Method (FDM) Analysis

Item Code	Threshold (d)	Consensus (%)	Defuzzified Mean (A)	Decision
Item 1	0.00	100.0	0.75	Accepted
Item 2	0.15	100.0	0.60	Accepted
Item 3	0.15	100.0	0.65	Accepted
Item 4	0.10	100.0	0.55	Accepted
Item 5	0.15	100.0	0.60	Accepted
Item 6	0.10	100.0	0.70	Accepted
Item 7	0.10	100.0	0.55	Accepted
Item 8	0.25	75.0	0.60	Rejected
Item 9	0.20	100.0	0.50	Accepted
Item 10	0.15	100.0	0.60	Accepted
Item 11	0.10	100.0	0.70	Accepted
Item 12	0.10	100.0	0.75	Accepted
Item 13	0.15	100.0	0.70	Accepted
Item 14	0.15	100.0	0.60	Accepted
Item 15	0.10	100.0	0.70	Accepted
Item 16	0.15	100.0	0.65	Accepted
Item 17	0.15	100.0	0.60	Accepted
Item 18	0.10	100.0	0.75	Accepted
Item 19	0.15	100.0	0.60	Accepted
Item 20	0.15	100.0	0.65	Accepted
Item 21	0.10	100.0	0.70	Accepted

Overall, the high level of consistency observed across threshold, consensus, and defuzzified mean values confirms the robustness of the expert review process. The majority of items were validated with strong agreement, underscoring the clarity, coherence, and construct alignment of the three-tier diagnostic assessment instrument. Consequently, the use of the FDM effectively complements the CVI findings and provides an advanced layer of validation that integrates expert certainty and agreement. The convergence of these results affirms the methodological rigor and reliability of the developed instrument for diagnosing mathematics learning difficulties in secondary education.

3.3 Comparison Between CVI and FDM Analysis

The comparison between the Content Validity Index (CVI) and the Fuzzy Delphi Method (FDM) analyses, illustrated in Fig. 2 and summarized in Table 8, reveals a strong convergence between the two validation approaches. All twenty-one items were rated *Valid* according to the CVI criteria, while twenty items were *Accepted* under the FDM thresholds. This high degree of correspondence indicates that both methods produced consistent expert judgments regarding the relevance, clarity, and representativeness of the items. The overall S-CVI/Ave value of 0.97, classified as excellent content validity, closely corresponds with the average defuzzified mean ($A = 0.65$), suggesting a strong alignment between the traditional and fuzzy-based validation techniques.

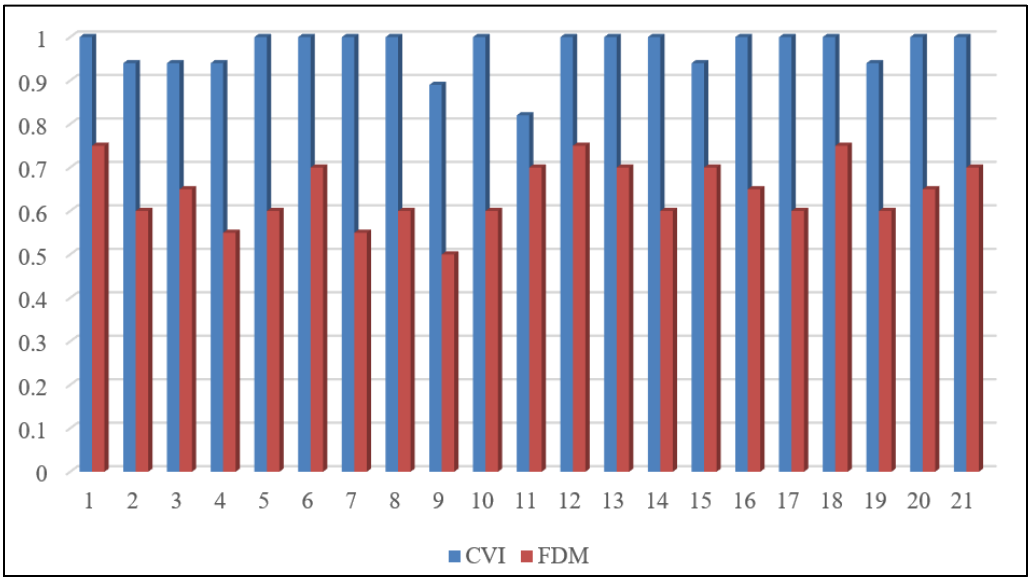

Figure 2. Comparison Results Between CVI and FDM Analysis

Table 8. Comparison of CVI and FDM Validation Results

Item Code	CVI Decision	FDM Decision	Final Decision
1–7, 9–21	Valid	Accepted	Strongly Valid
8	Valid	Rejected	Revision Required

Only one item (Item 8) was identified for Revision Required under the FDM analysis due to a slightly higher threshold value ($d = 0.25$), suggesting minor differences in expert interpretation. Despite this variation, both methods yielded comparable outcomes for most items, confirming the robustness of the validation process. The integration of CVI and FDM thus offers a complementary perspective: while the CVI provides a quantitative measure of expert agreement, the FDM extends this analysis by incorporating fuzzy logic to address the uncertainty and subjectivity inherent in linguistic expert judgments. This complementary use of CVI and FDM is consistent with findings from recent validation studies. For example, Afrizon *et al.* [39] effectively combined both methods to validate a physics diagnostic instrument, demonstrating that FDM enhances the interpretive depth of CVI-based evaluations. Similarly, Abdul Rahman *et al.* [27] and Hasim *et al.* [28] found that FDM contributes to improved consensus stability and a more accurate representation of expert agreement. Meanwhile, Mastam and Zaharudin [33] focused on applying FDM as a stand-alone technique but reported comparable advantages in capturing expert consensus reliability. In the context of this study, the close alignment between the CVI and FDM results reinforces the validity and reliability of the developed three-tier diagnostic assessment instrument. The integration of both methods strengthens methodological rigor by ensuring that content validity is supported not only by quantitative expert agreement but also by measures of consensus consistency and stability. Consequently, the developed instrument can be considered both conceptually coherent and empirically robust for diagnosing mathematics learning difficulties in secondary education. The convergence of these findings further supports the methodological soundness of employing a dual validation framework in educational measurement research. The findings of this study indicate that the FDM approach provides additional sensitivity particularly in situations where expert judgments are close to decision thresholds or where linguistic interpretations vary despite acceptable numerical agreement levels. While the CVI effectively quantifies agreement on item relevance, the FDM captures uncertainty embedded in linguistic evaluations and identifies subtle variations in expert consensus stability. Therefore, the complementary use of CVI and FDM is especially useful when validation decisions involve borderline agreement levels or multidimensional diagnostic instruments requiring nuanced interpretation of expert feedback.

4. CONCLUSION

This study successfully achieved its objective of comparing the effectiveness of the Content Validity Index (CVI) and the Fuzzy Delphi Method (FDM) in validating a three-tier diagnostic assessment instrument designed to identify mathematics learning difficulties. The results of the CVI analysis showed that all 21 items satisfied the validity criteria, with Item-Level Content Validity Index (I-CVI) values ranging from 0.82 to 1.00 and an overall Scale-Level Content Validity Index (S-CVI/Ave) of 0.97**,** indicating excellent content validity and strong agreement among experts. Meanwhile, the Fuzzy Delphi Method confirmed that 20 out of 21 items (95.24%) met the acceptance criteria with threshold values $d \leq 0.2$**,** expert consensus levels ranging from 75% to 100%**,** and defuzzified mean values $A \geq 0.5$. Only one item required revision due to a slightly higher threshold value identified through fuzzy consensus analysis. These findings demonstrate that both CVI and FDM provide complementary validation evidence, where CVI offers a clear quantitative representation of expert agreement and FDM refines the interpretation by addressing linguistic uncertainty in expert judgments. The integration of both approaches therefore strengthens the methodological rigor and supports the validity of the developed diagnostic assessment instrument. Despite these encouraging findings, several limitations should be acknowledged. First, the expert validation process involved only five experts, which may limit the generalizability of the results across broader academic and institutional contexts. Second, the validation relied primarily on structured expert judgment, which may still introduce subjectivity despite the application of quantitative agreement indices. Third, the study focused mainly on content validity evidence and did not include a detailed empirical analysis of students' learning difficulties through large-scale field implementation. Fourth, the developed diagnostic instrument was not compared with other existing diagnostic assessment instruments, which could provide additional insight into its relative effectiveness and measurement performance. Future research is therefore recommended to involve a larger and more diverse panel of experts to enhance representativeness and strengthen validation evidence. Further empirical testing using approaches such as Item Response Theory (IRT) or Rasch modeling is also necessary to evaluate the psychometric characteristics of the instrument across different student populations and educational contexts. In addition, comprehensive classroom-based implementation is required to investigate students' learning difficulties more deeply and to examine the diagnostic effectiveness of the instrument in authentic learning environments. Comparative studies involving other established diagnostic assessment instruments are also suggested to evaluate the relative strengths of the developed tool. Finally, the integration of multiple analytical approaches is expected to enrich interpretation of validation results and support broader application of the instrument in mathematics education research.

Author Contributions

Khaerun Nisa: Conceptualization, Formal Analysis, Methodology, Writing -Original Draft, Writing -Review and Editing, Visualization, Data Curation. Iva Sarifah: Supervision, Funding Acquisition, Writing-Review and editing, Validation. Riyadi: Supervision, Writing- Editing, Validation. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research has been funded by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia through the 2025 Doctoral Dissertation Research Grant Scheme (Grant No. 2/UN39.14/C3/DT.05.00/PPS-PDD/PL/2025).

Acknowledgment

This work was supported by the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia through a doctoral research grant. The authors also acknowledge the Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) and the Sekolah Pascasarjana, Universitas Negeri Jakarta, for their valuable support throughout this research.

Declarations

The authors declare no conflicts of interest to report study.

Declaration of Generative AI and AI-assisted technologies

The authors declare that no generative AI or AI-assisted technologies were used in the preparation of this manuscript, including for writing, editing, data analysis, or the creation of tables and figures.

REFERENCES

[1] M. S. Ibarra-Sáiz, G. Rodríguez-Gómez, and D. Boud, "THE QUALITY OF ASSESSMENT TASKS AS A DETERMINANT OF LEARNING," *Assess Eval High Educ*, vol. 46, no. 6, 2021. doi: https://doi.org/10.1080/02602938.2020.1828268.

[2] M. A. Adarkwah, "THE POWER OF ASSESSMENT FEEDBACK IN TEACHING AND LEARNING: A NARRATIVE REVIEW AND SYNTHESIS OF THE LITERATURE," *SN Social Sciences*, vol. 1, no. 3, 2021. doi: https://doi.org/10.1007/s43545-021-00086-w.

[3] A. I. Säfström *et al.*, "DEVELOPING A DIAGNOSTIC FRAMEWORK FOR PRIMARY AND SECONDARY STUDENTS' REASONING DIFFICULTIES DURING MATHEMATICAL PROBLEM SOLVING," *Educational Studies in Mathematics*, vol. 115, no. 2, 2024. doi: https://doi.org/10.1007/s10649-023-10278-1.

[4] S. Ndiung and M. Jediut, "PENGEMBANGAN INSTRUMEN TES HASIL BELAJAR MATEMATIKA PESERTA DIDIK SEKOLAH DASAR BERORIENTASI PADA BERPIKIR TINGKAT TINGGI," *Premiere Educandum : Jurnal Pendidikan Dasar dan Pembelajaran*, vol. 10, no. 1, 2020. doi: https://doi.org/10.25273/pe.v10i1.6274.

[5] M. J. Gierl, O. Bulut, Q. Guo, and X. Zhang, "DEVELOPING, ANALYZING, AND USING DISTRACTORS FOR MULTIPLE-CHOICE TESTS IN EDUCATION: A COMPREHENSIVE REVIEW," *Rev Educ Res*, vol. 87, no. 6, 2017. doi: https://doi.org/10.3102/0034654317726529.

[6] M. M. Wooten, A. M. Cool, E. E. Prather, and K. D. Tanner, "COMPARISON OF PERFORMANCE ON MULTIPLE-CHOICE QUESTIONS AND OPEN-ENDED QUESTIONS IN AN INTRODUCTORY ASTRONOMY LABORATORY," *Physical Review Special Topics - Physics Education Research*, vol. 10, no. 2, 2014. doi: https://doi.org/10.1103/PhysRevSTPER.10.020103.

[7] Kemendikbudristek, "PISA 2022 DAN PEMULIHAN PEMBELAJARAN DI INDONESIA," *Laporan Pisa Kemendikbudristek*, 2023.

[8] OECD, "PISA 2022 RESULTS FACTSHEETS INDONESIA," *OECD (Organisation for Economic Co-operation and Development) Publication*, 2023.

[9] T. Fan, J. Song, and Z. Guan, "INTEGRATING DIAGNOSTIC ASSESSMENT INTO CURRICULUM: A THEORETICAL FRAMEWORK AND TEACHING PRACTICES," *Language Testing in Asia*, vol. 11, no. 1, 2021. doi: https://doi.org/10.1186/s40468-020-00117-y.

[10] K. BSKAP, "TAHAPAN IMPLEMENTASI KURIKULUM MERDEKA DI SATUAN PENDIDIKAN.," *Kemendibudristek*, 2022.

[11] B. R. Bryant *et al.*, "THE EFFECTS OF A TIER 3 INTERVENTION ON THE MATHEMATICS PERFORMANCE OF SECOND GRADE STUDENTS WITH SEVERE MATHEMATICS DIFFICULTIES," *J Learn Disabil*, vol. 49, no. 2, 2016. doi: https://doi.org/10.1177/0022219414538516.

[12] P. Djiwandono and D. Ginting, "EVALUATING RESEARCH REPORTS ON THE QUALITIES OF TESTS OF ENGLISH LANGUAGE SKILLS IN INDONESIAN SCHOOLS: A SYSTEMATIC REVIEW," *Language Education & Assessment*, vol. 8, no. 1, p. 2237, Jun. 2025. doi: https://doi.org/10.29140/lea.v8n1.2237.

[13] A. C. Butler, "MULTIPLE-CHOICE TESTING IN EDUCATION: ARE THE BEST PRACTICES FOR ASSESSMENT ALSO GOOD FOR LEARNING?," 2018. doi: https://doi.org/10.1016/j.jarmac.2018.07.002.

[14] A. Zuhriyah, A. Suprijono, and A. Kasdi, "NEW MULTIPLE CHOICE QUESTIONS IN CRITICAL THINKING ASSESSMENT: AS A WAY TO DISTINGUISH BETWEEN LOGICAL ANSWERS AND GUESS ANSWERS," *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, 2019. doi: https://doi.org/10.29322/ijsrp.9.10.2019.p9422.

[15] R. M. Abu-Ghazalah, D. N. Dubins, and G. M. K. Poon, "DISSECTING KNOWLEDGE, GUESSING, AND BLUNDER IN MULTIPLE CHOICE ASSESSMENTS," *Applied Measurement in Education*, vol. 36, no. 1, 2023. doi: https://doi.org/10.1080/08957347.2023.2172017.

[16] Y. S. Pramesti, H. Mahmudi, and I. Setyowidodo, "USING THREE-TIER TEST TO DIAGNOSE STUDENTS' LEVEL OF UNDERSTANDING," in *Journal of Physics: Conference Series*, 2021. doi: https://doi.org/10.1088/1742-6596/1806/1/012013.

[17] A. S. Pramasdyahsari, D. Viviana, S. Sugiyanti, B. P. Pant, and I. Ratnayake, "CONSTRUCTION OF THREE-TIER DIAGNOSTIC TEST FOR IDENTIFYING STUDENT MISCONCEPTIONS IN EXPONENTIAL PROPERTIES," *Jurnal Riset Pendidikan Matematika*, vol. 11, no. 1, pp. 64–76, 2024. doi: https://doi.org/10.21831/jrpm.v11i1.73561.

[18] S. Prayitno, A. Arjudin, and H. Hapipi, "ANALYZING GEOMETRY MISCONCEPTION OF PROSPECTIVE TEACHERS USING THREE-TIER DIAGNOSTIC TESt," in *Proceedings of the 1st Annual Conference on Education and Social Sciences (ACCESS 2019)*, Altantis Press, 2020. doi: https://doi.org/10.2991/assehr.k.200827.031.

[19] C. M. H. Laeli, Gunarhadi, and Muzzazinah, "THE 3 TIERS MULTIPLE-CHOICE DIAGNOSTIC TEST FOR PRIMARY STUDENTS' SCIENCE MISCONCEPTION," *Pegem Egitim ve Ogretim Dergisi*, vol. 13, no. 2, 2023. doi: https://doi.org/10.47750/pegegog.13.02.13.

[20] D. S. Anggara and C. Abdillah, "CONTENT VALIDITY ANALYSIS OF LITERACY ASSESSMENT INSTRUMENTS," *Cakrawala Pendidikan*, vol. 42, no. 2, 2023. doi: https://doi.org/10.21831/cp.v42i2.55900.

[21] A. Roebianto, S. I. Savitri, I. Aulia, A. Suciyana, and L. Mubarokah, "CONTENT VALIDITY: DEFINITION AND PROCEDURE OF CONTENT VALIDATION IN PSYCHOLOGICAL RESEARCH," *TPM Test Psychom Methodol Appl Psychol*, vol. 30, no. 1, 2023. doi: https://doi.org/10.4473/TPM30.1.1.

[22] N. H. Jamin, S. Surat, and W. M. R. Wan Mohammad, "TEACHERS' KNOWLEDGE IN HOTS (TKHOTS) INSTRUMENT: CONTENT VALIDITY INDEX," *International Journal of Academic Research in Progressive Education and Development*, vol. 11, no. 1, 2022. doi: https://doi.org/10.6007/ijarped/v11-i1/12179.

[23] K. Beck, "ENSURING CONTENT VALIDITY OF PSYCHOLOGICAL AND EDUCATIONAL TESTS – THE ROLE OF EXPERTS," *Frontline Learn Res*, vol. 8, no. 6, 2020. doi: https://doi.org/10.14786/flr.v8i6.517.

[24] A. F. R. L. de Hierro, M. Sánchez, D. Puente-Fernández, R. Montoya-Juárez, and C. Roldán, "A FUZZY DELPHI CONSENSUS METHODOLOGY BASED ON A FUZZY RANKING," *Mathematics*, vol. 9, no. 18, 2021. doi: https://doi.org/10.3390/math9182323.

[25] Z. Zeng, X. Li, and S. W. Tho, "THE APPLICATION OF FUZZY DELPHI METHOD FOR THE DEVELOPMENT OF STEM TEACHING MODEL," *International Journal of Evaluation and Research in Education*, vol. 14, no. 3, pp. 2061–2069, Jun. 2025. doi: https://doi.org/10.11591/ijere.v14i3.29780.

[26] A. F. R. L. de Hierro, M. Sánchez, D. Puente-Fernández, R. Montoya-Juárez, and C. Roldán, "A FUZZY DELPHI CONSENSUS METHODOLOGY BASED ON A FUZZY RANKING," *Mathematics*, vol. 9, no. 18, Sep. 2021. doi: https://doi.org/10.3390/math9182323.

[27] F. B. Abdul Rahman, Z. Mustafa, and A. F. Kharuddin, "EMPLOYING FUZZY DELPHI TECHNIQUE TO VALIDATE MULTIPLE INTELLIGENCE BASED INSTRUCTIONAL TEACHING MODULE FOR PRESCHOOL CHILDREN," *Southeast Asia Early Childhood Journal*, vol. 10, no. 1, 2021. doi: https://doi.org/10.37134/saecj.vol10.1.6.2021.

[28] M. A. Hasim, J. Jabar, A. Sufian, N. F. Ibrahim, and F. A. Khalid, "EMPLOYING FUZZY DELPHI TECHNIQUES TO VALIDATE THE COMPONENTS AND CONTENTS OF E-LEARNING ANTECEDENTS AND USAGE BEHAVIOR TOWARDS E-LEARNING PERFORMANCE," *European Journal of Educational Research*, vol. 12, no. 1, 2023. doi: https://doi.org/10.12973/eu-jer.12.1.467.

[29] F. N. Pazilah, H. Hashim, and M. M. Yunus, "REFINING THE TPACK FRAMEWORK: A FUZZY DELPHI APPROACH TO 21ST-CENTURY COMPETENCY AND SELF-EFFICACY CONSTRUCTS," *Cogent Education*, vol. 11, no. 1, 2024. doi: https://doi.org/10.1080/2331186X.2024.2428886.

[30] M. F. Baharuddin, M. N. Masrek, and S. M. Shuhidan, "CONTENT VALIDITY OF ASSESSMENT INSTRUMENT FOR INNOVATIVE WORK BEHAVIOUR OF MALAYSIAN SCHOOL TEACHERS," *International Journal of Scientific and Technology Research*, vol. 9, no. 4, 2020.

[31] E. Karim, J. Ahmad, and K. Osman, "FUZZY DELPHI METHOD FOR CONTENT VALIDATION OF INTEGRATED SCIENCE PROCESS SKILLS INSTRUMENT," *International Journal of Academic Research in Business and Social Sciences*, vol. 7, no. 6, 2017. doi: https://doi.org/10.6007/ijarbss/v7-i6/3037.

[32] N. M. Mastam and R. Zaharudin, "THE FUZZY DELPHI METHOD: VALIDATION OF THE PROFESSIONAL COMPETENCIES ELEMENTS IN CAREER READINESS INSTRUMENT FOR STUDENTS WITH DISABILITIES," *International Journal of Academic Research in Progressive Education and Development*, vol. 13, no. 4, Nov. 2024. doi: https://doi.org/10.6007/ijarped/v13-i4/23817.

[33] N. M. Mastam, R. Zaharudin, and M. N. Md Noor, "THE FUZZY DELPHI METHOD: VALIDATION OF THE PERSONAL QUALITIES AND ABILITIES ELEMENTS IN CAREER READINESS INSTRUMENT FOR STUDENTS WITH DISABILITIES," *International Journal of Academic Research in Progressive Education and Development*, vol. 13, no. 3, Aug. 2024. doi: https://doi.org/10.6007/ijarped/v13-i3/22402.

[34] A. H. bin Abd Rahman, N. binti Mat Norwani, and R. binti Yahaya, "CONTENT VALIDITY OF ACCOUNTING LEARNING SELF EFFICACY INSTRUMENT AMONG MATRICULATION ACCOUNTING STUDENTS," *International Journal of Academic Research in Progressive Education and Development*, vol. 12, no. 3, 2023. doi: https://doi.org/10.6007/ijarped/v12-i3/18741.

[35] D. F. Polit, C. T. Beck, and S. V. Owen, "FOCUS ON RESEARCH METHODS: IS THE CVI AN ACCEPTABLE INDICATOR OF CONTENT VALIDITY? APPRAISAL AND RECOMMENDATIONS," *Res Nurs Health*, vol. 30, no. 4, 2007. doi: https://doi.org/10.1002/nur.20199.

[36] M. A. Jailani and C. K. Loy, "THE APPLICATION OF FUZZY DELPHI METHOD IN CONTENT VALIDITY ANALYSIS," in *20th International Conference on Cognition and Exploratory Learning in Digital Age, CELDA 2023*, 2023. doi: https://doi.org/10.33965/celda2023_202306l011.

[37] D. F. Polit and C. T. Beck, "THE CONTENT VALIDITY INDEX: ARE YOU SURE YOU KNOW WHAT'S BEING REPORTED? CRITIQUE AND RECOMMENDATIONS," *Res Nurs Health*, vol. 29, no. 5, 2006. doi: https://doi.org/10.1002/nur.20147.

[38] I. W. Prananto, Y. Rakhmawati, and T. Pamungkas, "CONTENT VALIDITY RATIO (CVR), CONTENT VALIDITY INDEX (CVI) AND CONFIRMATORY FACTOR ANALYSIS (CFA) IN MATHEMATICS LEARNING INDEPENDENCE INSTRUMENTS," *Kontinu: Jurnal Penelitian Didaktik Matematika*, vol. 6, no. 2, 2022. doi: https://doi.org/10.30659/kontinu.6.2.116-132.

[39] R. Afrizon, L. E. Mohtar, and M. S. M. Azmi, "DEVELOPMENT OF THE FUZZY DELPHI INSTRUMENT FOR IDENTIFYING ELEMENTS OF THE PHYARECI-STEM LEARNING MODULE," *EDUCATUM Journal of Science, Mathematics and Technology*, vol. 12, no. 12, pp. 33–44, Apr. 2025. doi: https://doi.org/10.37134/ejsmt.vol12.sp.4.2025.

[40] M. S. B. Yusoff, "ABC OF CONTENT VALIDATION AND CONTENT VALIDITY INDEX CALCULATION," *Education in Medicine Journal*, vol. 11, no. 2, 2019, https://doi.org/10.21315/eimj2019.11.2.6.

[41] R. M. Rodríguez, Á. Labella, and L. Martínez, "AN OVERVIEW ON FUZZY MODELLING OF COMPLEX LINGUISTIC PREFERENCES IN DECISION MAKING," *International Journal of Computational Intelligence Systems*, vol. 9, 2016, https://doi.org/10.1080/18756891.2016.1180821.