Journal of Applied Science, Engineering, Technology, and Education *ISSN: 2685-0591 (online)*
Vol. 8 No. 2 (2026)
https://doi.org/10.35877/454RI.asci

Content Validity of an Instrument Measuring Understanding of Generative Artificial Intelligence (AI) Ethics for Malay Language Assessment

Siti Haryanti Osaman[a] & Mohd Effendi, Ewan Mohd Matore[b,*]

[a]*Faculty of Education, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
[b]*Research Centre of Education Leadership and Policy, Faculty of Education, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*

Abstract

The rapid advancement of Generative Artificial Intelligence (GenAI) has transformed educational assessment, including Malay Language assessment. However, its integration requires strong ethical understanding among teachers to ensure fairness, integrity, and authenticity. This study aims to evaluate the content validity of an instrument measuring teachers' understanding of GenAI ethics in the context of Malay Language assessment. A quantitative design was employed involving five expert panellists who evaluated the instrument in terms of content relevance, construction, and clarity. The instrument comprises six ethical constructs: intellectual property, information authenticity, robustness, recognition of misuse, sociocultural responsibility, and human-centered design. Content validity was assessed using the Content Validity Ratio (CVR) based on Lawshe (1975). The findings show that 50 out of 62 items achieved a CVR value of 1.00 and met the critical threshold of 0.99, indicating strong agreement among experts. The remaining items were refined for clarity and alignment. The instrument demonstrates strong content validity and offers practical value as a tool to support teachers' ethical judgement in assessment. Further validation is recommended.

Keywords: content validity, GenAI, ethics, Malay Language assessment.

Received: 6 March 2026 Revised: 29 July 2026 Published: 31 August 2026

1. Introduction

Content validity is an important part of validity when making a questionnaire instrument because it makes sure that each item accurately and completely represents the construct that is to be measured. Content validity is an important component in instrument development because it ensures that each item accurately represents the construct being measured and supports the psychometric quality of the instrument (Mohd Matore et al., 2021). This is due to the fact that content validity is essential for ensuring that the instrument items are pertinent, exhaustive, and comprehensible in reflecting the measured outcomes (Mokkink et al., 2025). Content validity is crucial in scale development because it ensures that the items in an instrument accurately represent the construct being measured (Estremera & Mendoza - Sarmiento, 2024). Moreover, content validity pertains to the degree to which the items in an instrument sufficiently and pertinently reflect the entire construct domain under assessment (Cohen et al., 2022). This method makes sure that the tools that are made are not only content-valid but also theoretically sound and appropriate for the intended use (Aguirre et al., 2024) .

To ensure the effectiveness of the items, the assessment is usually conducted through an expert panel that evaluates the effectiveness of the items in describing the construct being measured. Content validity was determined through the judgement of a panel of professional and field experts who assessed the importance level of each instrument item based on a combination of identified challenge types (Masuwai et al., 2024). Masdiah and Nurul Fadly (2024) stated that the selection of experts is based on their knowledge, expertise, level of expertise, and relevant professional experience in order to assess content validity. Therefore, expert panel feedback was used to improve the instrument before the pilot study to ensure sufficient psychometric quality (Roebianto et al., 2023).

* Corresponding author.
E-mail address: effendi@ukm.edu.my

Content validity is an important part of making instruments, but past research shows that there isn't much detailed reporting on how to determine content validity, especially in educational research (Raub et al., 2023). It is very important for studies to be more reliable that they report their content validity assessment methods in a clear and systematic way. This is especially true for the development of educational tools (ElKhalil et al., 2025). This study's goal is to systematically describe the content validity assessment process using an expert panel as the first step in developing an instrument. This will help make educational research more evidence-based (Wang & Sahid, 2024). In this sense, the goals of this study are to:

a. To assess the content validity of the Generative Artificial Intelligence (GenAI) ethics understanding instrument through expert panel evaluation.

b. To investigate the suitability of the instrument items based on expert feedback.

2. Literature Review

2.1. Ethics of Generative Artificial Intelligence (GenAI)

The fast growth of Generative Artificial Intelligence (GenAI) technology has big effects on how schools work. GenAI uses algorithms to change and combine data to make different kinds of content that look new and real as if they were made by people (Abdullah et al., 2023). To make sure that people use it responsibly, there needs to be strong regulation and guidelines (Hasibuan & Isnanto, 2025). (Samsudin et al., 2024) say that AI can't form characters very well and needs to be balanced with the role of humans. So, teachers, policymakers, and AI developers need to keep working together to make sure that GenAI is used in a way that is fair, ethical and focused on learning (Kadaruddin, 2023).

This development also demands a deep understanding of ethics to prevent technology from being misused or from compromising authenticity and integrity in the assessment process. According to Williams (2023), the use of GenAI tools in education needs to be accompanied by ethical responsibility to ensure transparency, fairness, and the authenticity of student assessments are maintained. Meanwhile, Cardona et al. (2023) emphasize that AI-based educational transformation must be grounded in the principles of humanity, social responsibility, and academic integrity to avoid imbalances in the use of technology in the classroom. Therefore, this study focuses on the six new principles of AI ethics in the Laine, Minkkinen, and Mäntymäki (2025) model (Laine et al., 2025). The six new principles of AI ethics are respecting intellectual property, accuracy or authenticity of information, system robustness, identification of misuse, socio-cultural responsibility, and human centred design.

The principle of respecting intellectual property emphasizes that GenAI users must respect the copyrights, original sources, and works of others. According to Bale et al. (2024), the use of GenAI raises privacy and ethical issues, particularly concerning personal data, misinformation, and content misuse. Next, the second principle is accuracy or authenticity of information. This KBG technology is capable of generating convincing answers, but sometimes they are not accurate or supported by real facts. This is proven in the writing of Kenthapadi et al. (2023), Large Language Models (LLMs) often exhibit weaknesses in providing uncertainty-level budgets and are susceptible to data contamination, which can affect user trust in the information they generate. The third principle is robustness, which requires GenAI technology to act consistently across various scenarios and handle unexpected inputs. Mökander et al. (2023) stated that the lack of robustness in the model leads to at least three major risks that is critical system failure, data leakage and command injection attacks.

The fourth ethical principle is identifying misuse, which means finding risks of unethical use of GenAI. GenAI could be used for bad things like making propaganda, spreading false information, and messing up democratic processes like elections, fraud, and making cyberattacks and surveillance activities more effective(Guo et al., 2023). On the other hand, the principle of socio-cultural responsibility stresses how AI should be aware of and respect different cultures, values, and ways of life. It is clear that it also makes the digital divide bigger because not all cultures and organisations are ready to use AI technology equally (Dwivedi et al., 2023). Last but not least, there is the idea of human centred design, which says that GenAI systems should be made with the health, needs, and freedom of human users in mind. When users put too much faith in language models, don't understand what they can do, or use them in unsafe ways by treating them like real people, this interaction can have bad effects Mökander et al. (2023).

2.2. Assessment

The assessment system in Malaysia serves as an indicator of the country's education quality, involving the process of gathering information to evaluate students' proficiency levels and improve learning (Kementerian Pendidikan Malaysia, 2017). Digital assessment in education existed before GenAI and continues to evolve with technological

advancements to enhance learning and assessment experiences (Kizilcec et al., 2024). However, the use of GenAI for valid and reliable performance-based assessment instruments to evaluate actual abilities has not yet been fully realised in education (Jin et al., 2025). According to Amatan et al. (2021), the Content Validity Ratio (CVR) is an effective quantitative method for assessing the content validity of an instrument based on expert agreement on the importance of the items. The CVR method is practical and transparent in assessing content validity based on expert agreement, but it has limitations because it does not consider the rigour and bias of the raters (Mohamat et al., 2022).

Wang & Sahid (2024) recent study on educational instrument development shows that using an expert panel provides strong evidence for content validity, whether in the context of behavioural measurement models, digital knowledge or research culture. Content validity is important to ensure that the instruments in the assessment truly represent the established outcomes, are clearly understood, and can avoid inaccurate conclusions (Mokkink et al., 2025). Currently, educators are more supportive of AI-based assessment and critical thinking, while students show mixed reactions due to concerns about creativity (Kizilcec et al., 2024). Assessment is increasingly shifting towards digital and online approaches, in line with the changing teaching and learning ecosystem (Tahir et al., 2025) .

2.3. 21st Century Technology Competencies and GenAI Ethics in Malay Language Assessment

In 21st-century education, technological competence is an essential element for ensuring the valid, fair and ethical implementation of Malay language assessments in a digital environment. The use of digital technology and Generative Artificial Intelligence (GenAI) in language assessment, particularly for writing and comprehension skills, has the potential to support the evaluation process more efficiently and systematically (Tahir et al., 2025). However, in the study by (Kizilcec et al., 2024), it is emphasised that the use of GenAI in assessment also raises ethical challenges related to the authenticity of information, academic integrity, and excessive reliance on technology. Therefore, Malay language teachers need to have strong technology literacy and GenAI literacy to enable them to make ethical professional judgements when assessing students' learning outcomes (Jin et al., 2025). Integrating GenAI ethical principles such as information accuracy, professional responsibility, and human centred design is crucial to ensure that Malay language assessment practices remain integrated, contextual, and aligned with language education goals (Laine et al., 2025).

3. Methods

According to Creswell & Creswell (2018), quantitative research needs tools that are made in a systematic way and are both valid and reliable. So, it is very important to have content experts involved to make sure that the instrument items really do represent the construct and are based on the right theory. When content experts are involved, it is possible to check that the items are in line with the curriculum, the learning goals, and the target cognitive levels (Kania et al., 2024). Five expert panellists was used to check the content validity of the items in this study. They looked at the items' content, construction and language to see if they were appropriate (Kania et al., 2024). Memon et al. (2025) say that this method picks people based on their skills and knowledge without trying to make generalisations. Professional experts are people who are currently working in the field of study and are chosen based on certain criteria to determine whether instruments are appropriate for psychometric testing(Aguirre et al., 2024). The choice of experts is based on their knowledge of the field, the number of scientific papers they have written, and their relevant work experience (Rubio et al., 2003).

In this study, the professional experts consist of lecturers specializing in the field of education and teacher training, as well as officers in the Ministry of Education who hold a doctoral qualification. The selection process is in line with the concept of content validity, which emphasizes that individuals working in the relevant field are qualified to act as experts or judges to evaluate the content of the instrument and determine the extent to which the developed items can represent the study construct (M. E. Mohd Matore et al., 2017). A total of four professional experts were involved in the content validation process of this instrument. Two of the four experts are from the Ministry of Education Malaysia. (KPM) who serve in the Sports, Co-Curriculum, and Arts Division as well as the Education Policy Planning and Research Division. The two experts hold doctoral degrees in the field of measurement and evaluation. The other two experts are senior lecturers serving at Universiti Teknologi Malaysia in the field of Artificial Intelligence and Universiti Sains Malaysia in the field of Educational Science.

It seems wise to choose two experts from the Ministry of Education Malaysia who are experts in measurement and evaluation to make sure that the instrument fits with Malaysian educational policies, assessment standards, and teachers' professional concerns. A senior lecturer in Artificial Intelligence is also hired to make sure that the ideas and information about AI are correct. A senior lecturer in Education is also hired to make sure that the tool follows

educational theories, pedagogy and teachers' teaching methods. This helps to make sure that the instrument's content is correct. This study for lay experts had a teacher with a master's degree. We select professional specialists based on their academic qualifications and affiliations with institutions to ensure that the tool complies with educational policies and standards. On the other hand, lay experts were chosen to judge how useful the tool would be in real world AI technology applications, how relevant the content and how clear the technical term.

The deliberate differentiation between professional and public authorities is established to guarantee an equitable validation process. This integration ensures that the instrument is both theoretically sound and useful in the context of Malaysian education. This study was executed utilizing digital methodologies via email or online forms. The expert panel was emailed and sent letters explaining the study's goals and methods in order to get permission to take part. The Faculty of Education at Universiti Kebangsaan Malaysia sends out the official appointment letter and gives the expert panel two to four weeks to finish the exam. Detailed information about the expert panel is displayed in Table 1.

Table 1. Expert Panel Information

No.	Expert Panel	Position	Institution
1	A1	Officer	Sports, Co-curricular and Arts Division, Ministry of Education Malaysia
2	A2	Officer	Educational Planning and Policy Research Division, Ministry of Education Malaysia
3	A3	Lecturer	Universiti Teknologi Malaysia
4	A4	Lecturer	Universiti Sains Malaysia
5	A5	Lecturer	Sunway University Malaysia

3.1. Instrument

This study's survey instrument comprises two primary portions tailored to fulfil the research goals. Section A encompasses items pertaining to the demographic data of educators, including gender, age, teaching experience, educational attainment and familiarity with Generative Artificial Intelligence (GenAI) technology. This information is crucial for identifying the respondents' backgrounds, which may affect their comprehension of ethical problems pertaining to GenAI technology. Section B comprises items that assess educators' comprehension of the six primary components, which represent the novel AI ethical principles outlined in the Model by Laine, Minkkinen, and Mäntymäki (2025). This study identifies the following constructs: respect for intellectual property, accuracy and authenticity of information, robustness and security of AI systems, identification of misuse, sociocultural responsibility and human-centered design (Laine et al., 2025). The development of measurement instruments requires systematic empirical testing to ensure good psychometric properties and measurement accuracy (Sovey et al., 2022).

3.2. Content Validity Ratio Measurement Metric

There are many ways to find out if an instrument's content is valid, and one of the most common is the Content Validity Ratio (CVR). The CVR method, developed by Lawshe C. H. (1975) is designed to gauge the extent of expert consensus regarding the suitability of each item in the instrument, while also assisting in the identification and removal of extraneous items to enhance the overall quality of the instrument (Creswell & Guetterman, 2019). This study utilized CVR analysis to evaluate the content validity of the questionnaire, with the critical CVR value established according to the number of experts, as proposed by Lawshe C. H. (1975). Each expert panel must use a four-point Likert scale to show how much they agree that each item is appropriate.

In essence, the data analysis for this study is based on the CVR formula developed using Microsoft Excel. The CVR calculation formula is as follows:

$$CVR = (ne - N/2) / (N/2)$$

Notes:
ne = number of experts who rated the item as essential
N = total number of experts

To assess content validity, expert panels were asked to provide individual evaluations for each item using a four-point Likert scale: 1 = very inappropriate, 2 = inappropriate, 3 = appropriate, and 4 = very acceptable. In the equation given by Lawshe C. H. (1975) , CVR stands for the content validity ratio for an item, *ne* stands for the number of experts who rate the item on a scale of 3 and 4 based on validity criteria, and N stands for the total number of expert panellist's who are taking part in the content validity evaluation. A CVR rating of +1 means that everyone agrees that the item being tested is appropriate or very appropriate for content validity. When the CVR number is less than

0, it means that less than half of the experts on the panel thought the item was appropriate or very appropriate. When the CVR number is 0, it means that the experts had different opinions about the item, with some saying it was not appropriate and others saying it was. If the CVR value is greater than 0, it means that more than half of the experts gave the item a rating of 3 or 4. A CVR value of 1 means that all the experts on the panel agree that the item is appropriate or very appropriate. Five expert panellists said that the important CVR value for this study is 0.99, as Lawshe C. H. (1975) suggested. Table 3 shows the important CVR value that Lawshe C. H. (1975) suggested.

Table 3. Reference for Critical CVR Values (Lawshe, 1975)

No. of Expert Panelists	Minimum CVR Value
5	0.99
6	0.99
7	0.99
8	0.75
9	0.78
10	0.62
11	0.59
12	0.56
13	0.54
14	0.51
15	0.49
20	0.42
25	0.37
30	0.33
35	0.31
40	0.29

4. Results and Discussion

Table 4 presents the number of constructs and items developed to assess the understanding of GenAI ethics.

Table 4. Mean Scores of Constructs for Understanding Generative Artificial Intelligence (GenAI) Ethics

Construct	Number of Items	Mean
Respect for Intellectual Property	10	0.92
Accuracy or Information Authenticity (Truthfulness)	10	0.88
Robustness	10	0.84
Recognition of Malicious Uses	10	0.88
Sociocultural Responsibility	10	0.96
Human-Centric Design	12	0.97

The first objective of this study was to assess the content validity of the instrument developed to measure teachers' understanding of Generative Artificial Intelligence (GenAI) ethics in the context of Malay Language assessment. The validation process was conducted through expert panel evaluation using the Content Validity Ratio (CVR) method proposed by Lawshe (1975). A total of 62 items representing six constructs of GenAI ethics were evaluated by five expert panellists consisting of professional experts and a lay expert. The CVR analysis indicated that the majority of the items demonstrated strong content validity. These findings provide strong evidence that the developed instrument adequately represents the intended constructs of GenAI ethics in the context of Malay Language assessment. The high level of agreement among expert panellists suggests that the items are not only relevant and clear but also conceptually aligned with the theoretical framework. This supports previous studies which emphasise that expert judgement plays a critical role in ensuring the validity and quality of measurement instruments (Roebianto et al., 2023; Wang & Sahid, 2024). Beyond its statistical validity, the instrument also demonstrates practical significance in supporting teachers' professional judgement. As a diagnostic and reflective tool, it enables teachers to evaluate their level of ethical understanding when integrating GenAI into assessment practices. This is particularly important in ensuring the authenticity of student work, maintaining academic integrity, and making fair evaluation decisions in increasingly AI- mediated learning environments.

Out of the 62 items evaluated, 50 items achieved the maximum CVR value of 1.00, indicating unanimous agreement among the expert panel regarding the relevance and appropriateness of these items. Meanwhile, 10 items recorded a

CVR value of 0.60 and two items obtained a value of 0.20, reflecting lower levels of agreement among experts. Based on the critical CVR value for five expert panellists (0.99), only items that achieved the maximum CVR value were retained without modification, while the remaining items required revision to improve their clarity and alignment with the intended constructs. Further analysis was conducted by examining the mean score for each construct. The results show that all six constructs recorded high mean values ranging from 0.84 to 0.97, indicating a strong level of expert agreement regarding the suitability of the items in representing the GenAI ethics constructs. The construct of Human- Centric Design recorded the highest mean value (0.97), followed by Sociocultural Responsibility (0.96), Respect for Intellectual Property (0.92), Accuracy or Information Authenticity (0.88), Recognition of Malicious Uses (0.88), and Robustness (0.84).These findings suggest that the developed instrument provides strong evidence of content validity and adequately represents the theoretical framework of GenAI ethics proposed by Laine et al. (2025). The high level of agreement among experts indicates that the instrument items are relevant, comprehensive, and appropriate for measuring teachers' ethical understanding of GenAI in educational assessment. This is consistent with previous studies which emphasise that expert panel evaluation is an effective approach for ensuring that measurement instruments accurately represent the construct domain and minimise the risk of measurement error (Roebianto et al., 2023; Wang & Sahid, 2024).

The second objective of this study was to examine the suitability of the instrument items based on expert feedback obtained during the content validation process. Although the majority of the items achieved high CVR values, several items required revision due to issues related to wording clarity, conceptual precision, and contextual alignment. In addition to identifying items requiring revision, further analysis of expert feedback revealed specific issues related to wording clarity and conceptual precision. For example, certain items initially used broad terms such as "appropriate use of AI" without specifying the assessment context, which led to varied interpretations among experts. These items were refined to include clearer contextual indicators, such as "use of AI-generated content in evaluating students' written responses." Similarly, some items related to information authenticity were considered too general and were revised to explicitly emphasize the need to verify AI-generated outputs against reliable sources. This refinement process is consistent with established scale development principles, which emphasize clarity, specificity, and contextual alignment to ensure measurement accuracy (DeVellis & Thorpe, 2020).

Items with CVR values lower than the critical threshold were revised. Only items that achieved a CVR value of 1.00 were retained, as they met the minimum critical value of 0.99 for five expert panellists. For the construct of Respect for Intellectual Property, eight items achieved the maximum CVR value and were accepted without modification. However, two items required revision as their CVR values did not meet the required threshold. Expert feedback indicated that these items required clearer wording and stronger alignment with ethical practices related to intellectual property protection in the use of GenAI technology. Similarly, in the construct of Accuracy or Information Authenticity, eight items were accepted while two items required revision due to lower CVR values. Expert comments highlighted the need to refine the wording of these items to ensure that they more accurately represent the concept of information authenticity in AI-generated outputs. This finding supports previous research suggesting that ambiguous or overly general item wording may lead to inconsistent interpretation among respondents (Polit & Beck, 2006).

For the construct of Robustness, three items required revision due to insufficient expert agreement. Experts recommended improving the contextual clarity of the items so that they better reflect the reliability and stability of AI systems used in educational assessment settings. In the construct of Recognition of Malicious Uses, several items were also identified for revision to better represent potential risks and misuse scenarios of GenAI technology. The construct of Sociocultural Responsibility demonstrated a high level of expert agreement, with most items accepted without modification. Minor revisions were recommended to ensure that the items adequately reflect cultural sensitivity and the diversity of educational contexts. Meanwhile, the construct of Human-Centric Design showed the strongest level of expert agreement, with only one item requiring revision to further emphasis the role of human control, responsibility, and ethical considerations in the use of AI systems.

Overall, the expert feedback played an important role in improving the clarity, relevance, and conceptual alignment of the instrument items. The revision process ensures that the instrument is theoretically grounded and contextually appropriate for measuring teachers' understanding of GenAI ethics in Malay Language assessment. Such iterative refinement is widely recommended in instrument development studies to enhance the quality of measurement instruments before conducting pilot testing and further psychometric validation (DeVellis & Thorpe, 2020; Roebianto et al., 2023).

Table 5. Construct 1: CVR Values and Item Status

No.	Construct	Item	CVR Value	Item Status
1	Respect for Intellectual Property	1	1.00	Accepted
		2	1.00	Accepted
		3	1.00	Accepted
		4	1.00	Accepted
		5	1.00	Accepted
		6	0.60	Revised
		7	0.60	Revised
		8	1.00	Accepted
		9	1.00	Accepted
		10	1.00	Accepted

Table 5 indicates a number of experts view the elements designed for construct 1, Respect for Intellectual Property, are good. Of ten items tested, eight received the highest CVR score of 1.00 and were accepted as they were, indicating all the experts in the panel agreed with each of these items fitting the measured construct. On the other hand, items such as Item 6 and Item 7 had CVR values at only 0.60, below the critical value set by the five expert panellists at 0.99. This finding indicates that not all items achieved full consensus among the expert panel. This means the claims and materials should be reviewed again to verify whether they are clear and represent the intellectual property concept. Overall, the results show that most of the items in this construct have excellent content validity. A phased improvement process is nevertheless necessary to make sure each item accurately reflects the constructs consistently and in line with the systematic instrument development methodology proposed by Lawshe C. H. (1975) and applied in the study by Abdul Hadi & Mohd Matore (2025).

Table 6. Construct 2: CVR Values and Item Status

No.	Construct	Item	CVR Value	Item Status
1	Accuracy or Information Authenticity	1	0.60	Revised
		2	1.00	Accepted
		3	1.00	Accepted
		4	1.00	Accepted
		5	1.00	Accepted
		6	0.20	Revised
		7	1.00	Accepted
		8	1.00	Accepted
		9	1.00	Accepted
		10	1.00	Accepted

Based on Table 6, the results of the CVR analysis performed on the construct 2, which is Accuracy or Information Authenticity, present an overall high expert agreement level for this construct. Of these ten items used as part of the construct, eight were considered acceptable, with their maximum possible CVR value of 1.00, signifying a unanimous expert agreement based on their appropriateness as part of the construct discussing accuracy or authenticity of information. Meanwhile, both Item 1 and Item 6 of the construct were accepted with lower CVR values, 0.60 and 0.20, respectively, which are both lower than the minimum acceptable expert agreement value of 0.99, particularly with five expert panellists assessing these items. The low CVR value therefore, indicates an insufficient expert agreement level, thus highlighting the need to reevaluate these items, particularly with regard to their clarity of statement, the accuracy of terms, and their degree of correspondence with the conceptual definition of the construct, generally, or the characteristics under investigation, as supported by the aspects of systematic instrument development as stated by (Polit & Beck, 2006 ; Mohammed Afandi Zainal et al., 2020).

Table 7. Construct 3: CVR Values and Item Status

No.	Construct	Item	CVR Value	Item Status
1	Robustness	1	1.0	Accepted
		2	1.0	Accepted
		3	1.0	Accepted

No.	Construct	Item	CVR Value	Item Status
		4	0.2	Revised
		5	0.6	Revised
		6	1.0	Accepted
		7	0.6	Revised
		8	1.0	Accepted
		9	1.0	Accepted
		10	1.0	Accepted

Table 7 shows the CVR analysis for construct 3, which is called Robustness. It seems that experts generally agree on this. Seven of the ten items that were looked at received the highest CVR score of 1.00 and were accepted as they were. This shows that the expert panel agreed that these items were the best way to measure robustness. Three items (Item 4, Item 5, and Item 7) have CVR values of 0.20 and 0.60, which are below the 0.99 level set by five expert panellists. The low CVR score means that the experts didn't agree enough. This means that these items require further evaluation again to see if they are clear, easy to understand, or fit with the idea of strength. In short, most of the parts of the resilience construct are demonstrates strong validity at measuring what they are supposed to. But items that don't meet the minimum CVR value need to be fixed so that measurements are accurate and consistent (DeVellis & Thorpe, 2020).

Table 8. Construct 4: CVR Values and Item Status

No.	Construct	Item	CVR Value	Item Status
1	Recognition of Misuse	1	0.6	Revised
		2	1.0	Accepted
		3	1.0	Accepted
		4	1.0	Accepted
		5	1.0	Accepted
		6	1.0	Revised
		7	1.0	Accepted
		8	1.0	Accepted
		9	0.6	Revised
		10	0.6	Revised

Based on Table 8, the CVR analysis for construct 4, Recognition of Misuse shows that experts mostly agree on the issue. The expert panel unanimously concurred that six of the ten tested items exhibited a maximum CVR value of 1.00 and were endorsed without modification. This means that the questions were good for showing how to identify abuse. Four items (Item 1, Item 9, and Item 10) had a CVR value of 0.60, and Item 6 was also flagged for review. This means that these items did not meet the CVR threshold of 0.99 for the five expert panellists. This finding indicates the necessity for a re-evaluation of the construct's conceptual definition, clarity of statements, and contextual precision, particularly in articulating more specific and contextual manifestations of generative artificial intelligence abuse. In conclusion, these findings indicate that the majority of items within the abuse identification construct exhibit strong content validity; however, incremental improvements are required for certain items to ensure precise, clear, and consistent measurement in alignment with the goals of instrument development (Connell et al., 2018).

Table 9. Construct 5: CVR Values and Item Status

No.	Construct	Item	CVR Value	Item Status
1	Sociocultural Responsibility	1	1.0	Accepted
		2	1.0	Accepted
		3	1.0	Accepted
		4	1.0	Accepted
		5	1.0	Accepted
		6	1.0	Revised

No.	Construct	Item	CVR Value	Item Status
		7	1.0	Accepted
		8	0.6	Revised
		9	1.0	Accepted
		10	1.0	Accepted

Specifically looking at Table 9, the examination of the Content Validity Ratio (CVR) for construct 5, which is Sociocultural Responsibility, showed that the experts were in very high agreement. Eight of the ten items in the analysis were accepted at the highest level of 1.00. This means that all of the experts agreed that these questions were good at measuring sociocultural responsibility. Two issues specifically Item 6 and Item 8, were selected for change, with Item 8 exhibiting a lower agreement level of 0.60, below the minimal threshold of 0.99 necessary for a five expert panellists. The findings indicate that the components of the sociocultural responsibility construct possess adequate content validity; however, revisions are necessary in specific domains to attain precise measurement, sensitivity, and alignment with the developmental goals of the GenAI ethical instrument (Bujang et al., 2025).

Table 10. Construct 6: CVR Values and Item Status

No.	Construct	Item	CVR Value	Item Status
1	Human Centred Design	1	1.0	Accepted
		2	1.0	Accepted
		3	1.0	Accepted
		4	1.0	Accepted
		5	1.0	Accepted
		6	1.0	Accepted
		7	1.0	Accepted
		8	1.0	Accepted
		9	0.6	Revised
		10	1.0	Accepted
		11	1.0	Accepted
		12	1.0	Accepted

Table 10 shows that the experts in the Content Validity Ratio (CVR) study for Construct 6, or Human Centred Design, all agree very strongly. The maximum CVR value for eleven of the twelve items that were reviewed was 1.00, and they were all approved as they were. This means that the expert panel all agreed that the things were good examples of human centred design. Only Item 9 had a CVR score of 0.60, which meant it needed to be reviewed because it didn't meet the minimum score of 0.99 set by five expert panellists. This lower CVR value suggests that the item needs to be reviewed for clarity of the statement, emphasis on the role of human users, or alignment with the conceptual definition of human centred design in the context of GenAI ethics, as it indicates a lack of expert agreement on the item. The study results show that most items have strong content validity. However, based on expert feedback, some items need to be improved to make sure that the instrument meets user needs, is relevant to the context, and follows the rules of human centred design in the ethical use of AI (Zare et al., 2025).

The instrument developed for the six GenAI ethics constructs (62 items) shows strong evidence of content validity when the majority of items achieve the maximum CVR (1.00). Overall, 50 items were accepted while 12 items needed to be reviewed because they did not meet the established critical threshold; the CVR distribution also showed that only 2 items were at 0.20 and 10 items at 0.60, indicating that certain items were still unclear/not aligned with the construct and thus required phased improvements. These findings support the assertion in recent literature that content validity needs to assess relevance, comprehensiveness, and comprehensibility so that the instrument truly represents the construct domain and reduces the risk of incorrect measurement conclusions (Mokkink et al., 2025). In contemporary educational instrument studies, expert panel evaluations and systematic item reviews typically result in the majority of items being “retained” with high content validity evidence, while a small number of items require refinement in terms of terminology, context, and construct focus in line with instrument development practices that emphasise phased reviews before subsequent psychometric testing (Abdul Hadi & Mohd Matore, 2025).

All six constructs show a consistent and strong content validity trend, with the majority of items achieving CVR=1.00. However, a small number of items in each construct require focused review. Constructs 1 and 2 require improvements in terms of term clarity and alignment with actual practices, particularly regarding intellectual property and

information authenticity. Constructs 3 and 4 require specificity in wording and situational scope so that the indicators are clearer and not too general, especially in aspects of robustness and identification of misuse. Construct 5 emphasises the need for sensitivity to the sociocultural context to avoid bias, while Construct 6 only requires minimal refinement to strengthen user autonomy and wellbeing. Therefore, the proposed review is focused and incremental, without compromising the fundamental strength of the instrument.

5. Conclusion

This study has demonstrated that the developed instrument possesses strong content validity in measuring teachers' understanding of Generative Artificial Intelligence (GenAI) ethics within the context of Malay Language assessment. The high level of agreement among expert panellists indicates that the instrument is conceptually sound, relevant, and aligned with the theoretical framework of GenAI ethics. Beyond its role as a measurement tool, the instrument offers meaningful practical value by supporting teachers' ethical awareness and professional judgement in integrating GenAI into assessment practices. By linking ethical constructs with real classroom decision-making, the instrument contributes to bridging the gap between theoretical understanding and practical implementation. In the context of the Malaysian education system, the instrument has the potential to inform teacher training programmes and policy development related to digital and AI-integrated assessment. Specifically, it may serve as a diagnostic and evaluative tool to assess teachers' readiness and guide targeted professional development initiatives. The instrument also provides a structured framework that can assist educators in identifying ethical risks, such as issues related to information authenticity, intellectual property, and potential misuse of GenAI, thereby promoting more responsible assessment practices. However, further empirical validation is necessary to ensure its effectiveness and generalisability across different educational settings. Future research should focus on pilot testing involving Malay Language teachers, followed by advanced psychometric analyses such as Exploratory Factor Analysis or Rasch Measurement Model to establish construct validity and reliability. Additionally, large-scale implementation studies are recommended to examine how the instrument can be systematically integrated into teacher training and continuous professional development programmes. Overall, this study provides a foundational step towards developing a contextually relevant and practically meaningful instrument that supports the ethical integration of GenAI in Malay Language assessment, thereby contributing to more responsible, fair, and culturally sensitive educational practices in the era of digital transformation.

Acknowledgements: The authors acknowledge Universiti Kebangsaan Malaysia for funding this research.

References

Abdul Hadi, A. A., & Mohd Matore, M. E. E. (2025). Content Validity Assessment of Malaysian Teachers' Professional Judgement Instrument. *International Journal of Organizational Leadership*, *14*(First Special Issue 2025), 689–704. https://doi.org/10.33844/ijol.2025.60512

Abdullah, H. S. V., Alias, A. K., Tasir, Z., & Sharef, N. M. (2023). *MQA Advisory Note No.22023 - AI Generatif* (Issue 2).

Aguirre, V. C. de S. P., Dória, J. P. da S., Melo, R. A. de, Fernandes, F. E. C. V., & Mola, R. (2024). Content Validity of an Instrument for Assessing Skin Lesions in a Mother and Child Hospital. *Rev Enferm UFPI*, *13*(1), 1–11. https://doi.org/10.26694/reufpi.v13i1.4579

Amatan, M., K Han, C. G., & Pang, V. (2021). Kesahan kandungan soal selidik faktor konteks, input dan proses terhadap penerimaan pelaksanaan elemen pendidikan STEM dalam pengajaran dan pembelajaran guru menggunakan nisbah kesahan kandungan (CVR). *International Journal of Advanced Research in Future Ready Learning and Education*, *23*(1), 10–22.

Bujang, M., Alias, B. S., & Mansor, A. N. (2025). Evaluating the Content Validity of Instruments for Measuring Principal's Ethical Leadership Practices Using the Content Validity Ratio (CVR) Method. *Journal of Social Sciences & Humanities*, *22*(1), 99–108. https://doi.org/https://doi.org/10.17576/ebangi.2025.2208.08

Cardona, M. A., Rodríguez, R. J., & Ishmael, K. (2023). *Artificial Intelligence and the Future of Teaching and Learning Insights and Recommendations Artificial Intelligence and the Future of Teaching and Learning* (Issue May). https://tech.ed.gov

Cohen, R. J., W. Joel Schneider, & Tobin, R. M. (2022). *Psychological Testing and Assessment An Introduction to Tests and Measurement* (10th ed.). McGraw Hill.

Connell, J., Carlton, J., Grundy, A., Taylor Buck, E., Keetharuth, A. D., Ricketts, T., Barkham, M., Robotham, D.,

Rose, D., & Brazier, J. (2018). The importance of content and face validity in instrument development: lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *27*(7), 1893–1902. https://doi.org/10.1007/s11136-018-1847-y

Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative and Quantitative Approaches. In *SAGE Publications* (Fifth edit). https://doi.org/10.2307/328794

Creswell, J. W., & Guetterman, T. C. (2019). Educational Research:Planning, Conducting, and Evaluating Quantitative and Qualitative Research. In *Educational Principles and Practice in Veterinary Medicine*.

DeVellis, R., & Thorpe, C. (2020). *Scale development: Theory and applications* (fifth edit). Sage Publications. Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A.,

Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*(March), 1–63. https://doi.org/10.1016/j.ijinfomgt.2023.102642

ElKhalil, R., Masuadi, E., Bayoumi, R., AlMekkawi, M., Ahmed, L. A., Al-Rifai, R. H., & Elbarazi, I. (2025). Content validity of the Mental Health Literacy Scale for perinatal use based on expert and patient input. *Scientific Reports*, *15*(1), 1–11. https://doi.org/10.1038/s41598-025-17964-5

Estremera, M. L., & Mendoza -Sarmiento, M. A. (2024). Content Validity and Reliability of Questionnaires: Trends, Prospects and Innovation in the Digital Research Epoch Introduction. *ASEAN Innovative and Transformative Education Journal*, *1*(1), 1–10. www.aitej.gov.ph

Guo, D., Chen, H., Wu, R., & Wang, Y. (2023). AIGC challenges and opportunities related to public safety: A case study of ChatGPT. *Journal of Safety Science and Resilience*, *4*(4), 329–339. https://doi.org/10.1016/j.jnlssr.2023.08.001

Hasibuan, M. S., & Isnanto, R. (2025). *Generatif AI* (Januari, Issue January). Darmajaya (DJ) Press. https://www.researchgate.net/publication/388221462_Buku_Gen_AI_Final_20_jan_2025

Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test.

Computers and Education: Artificial Intelligence, *9*, 100436. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100436

Kadaruddin, K. (2023). Empowering Education through Generative AI: Innovative Instructional Strategies for Tomorrow's Learners. *International Journal of Business, Law, and Education*, *4*(2), 618–625. https://doi.org/10.56442/ijble.v4i2.215

Kania, N., Kusumah, Y. S., Dahlan, J. A., Nurlaelah, E., Gürbüz, F., & Bonyah, E. (2024). Constructing and providing content validity evidence through the Aiken's V index based on the experts' judgments of the instrument to measure mathematical problem-solving skills. *REID (Research and Evaluation in Education)*, *10*(1), 64–79. https://doi.org/10.21831/reid.v10i1.71032

Kementerian Pendidikan Malaysia. (2017). *Dasar Pendikan Kebangsaan*. *Edisi keempat*.

Kenthapadi, K., Lakkaraju, H., & Rajani, N. (2023). Generative AI meets Responsible AI: Practical Challenges and Opportunities. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 5805–5806. https://doi.org/10.1145/3580305.3599557

Kizilcec, R. F., Huber, E., Papanastasiou, E. C., Cram, A., Makridis, C. A., Smolansky, A., Zeivots, S., & Raduescu,

C. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Computers and Education: Artificial Intelligence*, *7*(November 2023), 1–11. https://doi.org/10.1016/j.caeai.2024.100269

Laine, J., Minkkinen, M., & Mäntymäki, M. (2025). Understanding the Ethics of Generative AI: Established and New Ethical Principles. *Communications of the Association for Information Systems*, *56*(January), 1–24. https://doi.org/10.17705/1CAIS.05601

Lawshe C. H. (1975). A Qualitative Approach to Content Validity. *Personnel Psychology*, *28*, 563–575.

Masuwai, A., Zulkifli, H., & Hamzah, M. I. (2024). Evaluation of content validity and face validity of secondary school Islamic education teacher self-assessment instrument. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2308410

Memon, M. A., Ramayah Thurasamyd, Ting, H., & Cheah, J.-H. (2025). Convenience Sampling: A Review and

Guidelines for Quantitative Research. *Journal of Applied Structural Equation Modeling*, *9*(June), 1–15. https://doi.org/10.47263/JASEM.9(2)01

Mohamat, R., Sumintono, B., & Hamid, H. S. A. (2022). Analisis Kesahan Kandungan Instrumen Kompetensi Guru untuk Melaksanakan Pentaksiran Bilik Darjah Menggunakan Model Rasch Pelbagai Faset (Content Validity Analysis of an Instrument to Measure Teacher's Competency for Classroom Assessment Using Many Facet R. *Jurnal Pendidikan Malaysia*, *47*(01), 1–15. https://doi.org/10.17576/jpen-2022-47.01-01

Mohammed Afandi Zainal, Mohd Effendi@Ewan Mohd Matore, Wan NorShuhadah W Musa, & Noor Hashimah Hashim. (2020). Kesahan Kandungan Instrumen Pengukuran Tingkah Laku Inovatif Guru Menggunakan Kaedah Nisbah Kesahan Kandungan (CVR). *Akademika*, *90*(Isu Khas 3), 43–54. http://ejournals.ukm.my/akademika/article/view/42052

Mohd Matore, M. E. E., Zainal, M. A., Mohd Noh, M. F., Khairani, A. Z., & Abd Razak, N. (2021). The Development and Psychometric Assessment of Malaysian Youth Adversity Quotient Instrument (MY-AQi) by Combining Rasch Model and Confirmatory Factor Analysis. *IEEE Access*, *9*(Mi), 13314–13329. https://doi.org/10.1109/ACCESS.2021.3050311

Mohd Matore, M. E., Idris, H., Normawati, A. R., & Khairani, A. Z. (2017). Kesahan Kandungan Pakar Instrumen IKBAR Bagi Pengukuran AQ Menggunakan Nisbah Kesahan Kandungan. *Proseeding of International Conference On Global Education V (ICGE V), Padang Indonesia [10-11 April 2017]*, *April*, 979–997.

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing Large Language Models: A Three-Layered Approach. In *SSRN Electronic Journal* (Vol. 4, Issue 4). Springer International Publishing. https://doi.org/10.2139/ssrn.4361607

Mokkink, L. B., Herbelet, S., Tuinman, P. R., & Terwee, C. B. (2025). Content validity: judging the relevance, comprehensiveness, and comprehensibility of an outcome measurement instrument – a COSMIN perspective. *Journal of Clinical Epidemiology*, *185*, 1–5. https://doi.org/10.1016/j.jclinepi.2025.111879

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/https://doi.org/10.1002/nur.20147

Raub, S. A., Mansor, M., Ishak, R., Pengurusan, F., Pendidikan, U., & Idris, S. (2023). Kesahan Kandungan Instrumen Pengukuran Pengurusan Pengetahuan Organisasi Pembelajaran Menggunakan Kaedah Nisbah Kesahan Kandungan (CVR). *Jurnal Pendidikan Bitara UPSI*, *16*(2), 59–70.

Roebianto, A., Savitri, S. I., Aulia, I., Suciyana, A., & Mubarokah, L. (2023). Content Validity: Definition and Procedure of Content Validation in Psychological Research. *TPM - Testing, Psychometrics, Methodology in Applied Psychology*, *30*(1), 5–18. https://doi.org/10.4473/TPM30.1.1

Rubio, D. M. G., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifyng content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94–104. https://doi.org/10.1093/swr/27.2.94

Samsudin, S. binti, Saleh, H. bin M., & Ahmad, A. S. bin. (2024). Persepsi Bakal Guru Terhadap Kesan Aplikasi kecerdasan Buatan (AI) dalam Pengajaran dan Pembelajaran. *International Journal of Educational Research on Andragogy and Pedagogy*, *2(1)*, 112–124. https://ijerap.com/index.php/ijerap/article/view/27

Sovey, S., Osman, K., & Matore, M. E. E. M. (2022). Rasch Analysis for Disposition Levels of Computational Thinking Instrument Among Secondary School Students. *Eurasia Journal of Mathematics, Science and Technology Education*, *18*(3), 2–15. https://doi.org/10.29333/ejmste/11794

Tahir, M. H. M., Saputra, S., Othman, S., Shah, D. S. M., Sulaiman, S. H., Azhar, M. A., & Mohandas, E. S. (2025).

Online Assessment in Higher Education: A Systematic Review. *Multidisciplinary Reviews*, *9*(1), 1–10. https://doi.org/10.24059/olj.v27i1.3398

Wang, F., & Sahid, S. (2024). Content validation and content validity index calculation for entrepreneurial behavior instruments among vocational college students in China. *Multidisciplinary Reviews*, *7*(9). https://doi.org/10.31893/multirev.2024187

Williams, R. T. (2023). The ethical implications of using generative chatbots in higher education. *Frontiers in Education*, *8*(January). https://doi.org/10.3389/feduc.2023.1331607

Zare, Z., Khosravi, M., Marzaleh, M. A., Jalali, F. S., Izadi, R., & Naseh, H. (2025). Psychometric evaluation of an instrument measuring artificial intelligence utilization in decision-making domains of healthcare organizations. *Scientific Reports*, *15*(1), 1–12. https://doi.org/10.1038/s41598-025-20753-9