Instrument Validation for Measuring ChatGPT Usage, Problem-Solving, and Ethical Awareness Among TVET Students

Jacey Mariadass @ Manickam[a], Lim Yeong Chyng[b]

[a] *jacey@puo.edu.my, yclim@puo.edu.my*
[a]Polytechnic Ungku Omar, Jalan Raja Musa Mahadi, Ipoh, 31400, Malaysia
[b]Polytechnic Ungku Omar, Jalan Raja Musa Mahadi, Ipoh, 31400, Malaysia

Abstract

The use of ChatGPT and other generative AI tools in education has increased the need to assess how these tools affect students' academic performance, particularly in areas such as problem-solving and ethical awareness. However, there is a lack of a comprehensive instrument to assess these areas among Malaysian Technical and Vocational Education and Training (TVET) students. This study addresses the gap by developing and validating a questionnaire to evaluate the responsible use of ChatGPT among TVET students. The questionnaire was constructed in three phases: (i) questionnaire development, (ii) content validation by four experts using Aiken's V (threshold ≥ 0.83) and (iii) pilot testing with 43 purposively sampled TVET students to assess internal consistency. Expert evaluations of item relevance, clarity, and simplicity (RCS) led to minor items' refinement. The pilot results showed high reliability (Cronbach's α > 0.80) across all sections. Due to the limited sample size (n = 43), exploratory or confirmatory factor analysis was not conducted. Instead, item-level performance was evaluated using descriptive statistics, corrected item–total correlations, and Cronbach's alpha, which provided sufficient evidence for validation. The validated instrument serves as a foundation that offers a reliable tool for full-scale studies and lays a strong foundation for future research on AI-assisted problem-solving and ethical awareness in Malaysian TVET programs.

Keywords: Instrument validation; Expert reviewer; Pilot study; Aiken's V; Cronbach's alpha; ChatGPT; TVET

1. Introduction

The way students access material, finish assignments, and participate in self-learning has been completely transformed by the introduction of artificial intelligence (AI) tools into the classroom. Among these technologies, ChatGPT has drawn a lot of interest due to its capacity to facilitate writing, offer explanations, and aids in idea generation. AI tools like ChatGPT offer new opportunities to support learning beyond traditional classroom methods in Technical and Vocational Education and Training (TVET) institutions in Malaysia, where students typically engage in hands-on and practical disciplines. However, concern about ethical use and academic integrity is also growing as these technologies become more widely available. Many students may not fully understand the impacts of employing AI-generated content in academic settings, including concerns about plagiarism and responsible usage. These issues need to be considered in TVET settings since students may be exposed to academic writing standards at different levels.

To investigate these rising patterns of ChatGPT use and ethical awareness, it is essential to first establish a valid and reliable measurement instrument. While existing literature highlights the growing use of generative AI in education, there is a noticeable gap in terms of validated instruments that can systematically assess

students' behavior, perceptions, and ethical considerations related to the TVET context. Without a tested instrument, any attempt to measure such constructions may produce inconsistent or misleading results.

This study was conducted to address that gap by focusing on the validation of an instrument designed to measure how TVET students utilize ChatGPT and how they perceive the ethical awareness surrounding its use. The primary aim is not to generalize findings to a broader population, but rather to ensure that the development of the instrument is both reliable and clear, capturing the intended constructions with internal consistency. Through this validation and reliability process, the study contributes to a much-needed measurement instrument that can be used in future large-scale research on ChatGPT usage and ethical awareness in education. Hence, this study was conducted to achieve three research objectives (RO): (i) to validate the item's relevance, clarity and simplicity using Aiken's V; (ii) to assess the internal consistency reliability of the instrument using Cronbach's alpha; and (iii) to evaluate item-level response characteristics and discrimination indices in a pilot study among Malaysian TVET students. Figure 1 shows the framework of the study conducted.

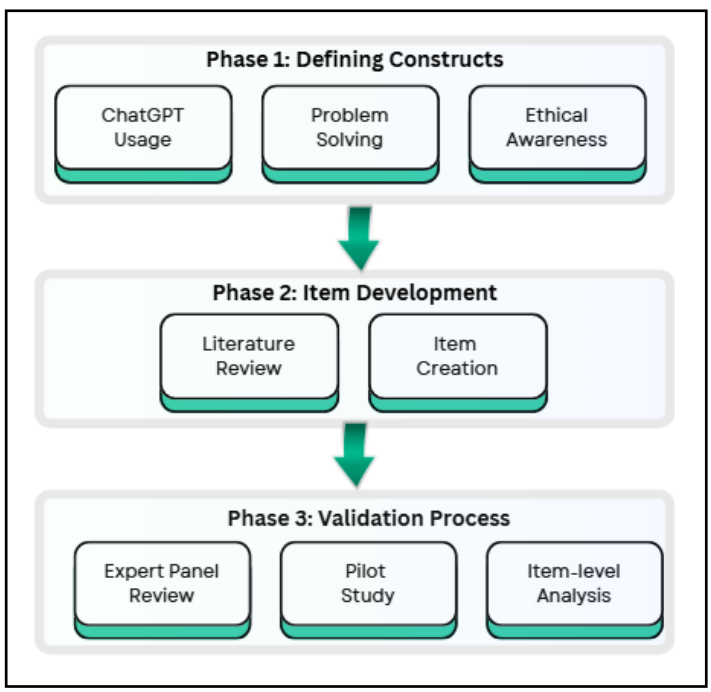

Figure 1: Study Framework

The first phase defines the construct of the study, (i) ChatGPT usage; (ii) problem-solving and (iii) ethical awareness. The second phase involves item development, where items were created based on a comprehensive review of relevant literature on AI-assisted learning, ethical awareness, and problem-solving in Malaysian TVET education contexts. This method ensures that each item reflects the conceptual foundations of the constructs being measured. Third phase is about the validation process. The item develops underwent expert panel review followed by a pilot study involving TVET students. Item-level analysis was conducted to evaluate the performance of each item, determining whether it should be accepted, review or remove for better clarity of instrument to be used in full-scale studies.

2. Literature Review

The integration of AI tools such as ChatGPT in Malaysia's TVET education system offers both opportunities and challenges. Students are increasingly exposed to generative technologies without sufficient ethical guidance, which may lead to misuse or dependency (Amdan, et al., 2024). Developing a validated instrument tailored to the Malaysian TVET context is essential. This would enable educators and researchers to assess not only the extent of AI usage but also students' ethical awareness, thereby contributing to improved teaching strategies and institutional policies that promote responsible AI interaction.

2.1 ChatGPT usage in education

Several studies have attempted to measure students' engagement with ChatGPT, particularly in higher education contexts. An instrument to assess the use of ChatGPT among postgraduates' learners was conducted to measure dimensions like academic writing aid, reliance and efficiency, demonstrating strong reliability and factor structures (Chan and Hu, 2023). However, these tools are not TVET specific and do not address problem-solving and ethical awareness aspects, limiting their generalizability to technical and vocational contexts.

2.2 ChatGPT and problem solving in TVET

The growing popularity of artificial intelligence (AI) in education has generated both excitement and concern, especially with the emergence of generative tools such as ChatGPT. Problem-solving is a key skill in TVET education, where students are trained to handle practical tasks and real-world challenges. With the rise of AI tools like ChatGPT which can assist them in breaking down complex tasks, explore different solution and improve their understanding of difficult topics (Smith et al., 2023). Recent studies show that ChatGPT can support students in generating ideas, analyzing problems, and thinking critically where all of this are essential for effective problem-solving (Chan and Hu, 2023). In Malaysian TVET settings, students are using ChatGPT to assist with assignments and learning activities. However, there is limited research that focuses specifically on how ChatGPT gives impact on problem-solving. This gap shows the need to study this are more closely. To address this, the current study includes problem-solving as one of the main constructs in the instrument. By developing and validating items related to ChatGPT's usage in problem-solving, the study aims to provide useful insights for future teaching and research in TVET.

2.3 ChatGPT and ethical awareness

The necessity for ethical awareness is growing in the context of AI adoption in education. As AI tools like ChatGPT become more common in education, students face new ethical challenges such as plagiarism, misuse of AI-generated content, and lack of proper citation. Studies show that many students use ChatGPT without fully understanding its ethical boundaries, which can lead to academic dishonesty or over-reliance on AI (Titko et al., 2023). Chheang et al. (2023) emphasize the need to embed ethical considerations into AI-assisted learning environments to guide responsible use. This study includes ethical awareness as a key construct in the instrument to assess students' understanding of responsible AI use. By validating items related to ethics, the research supports better educational practices and policy development in TVET.

2.4 Instrument validation

To examine these challenges effectively, researchers must first employ tools that are both valid and reliable. It is essential to create and evaluate the instrument that exactly measure student behavior and awareness. The

validation of research instruments, especially those used in exploratory areas like AI in education, is critical to ensure that data collected accurately reflects the intended constructs. Such an approach has led to a growing recognition of the importance of pilot studies in educational research. This study adopted a multi-phase validation approach involving expert review, pilot testing, and item-level analysis, which are widely recognized as best practices in instrument development (Bond and Fox, 2015; Donate-Beby et al., 2025).

2.4.1 Expert panel review

The main purpose of expert validation is to evaluate content validity, which refers to how well the items in a questionnaire represent the construct being measured. Expert evaluators assist in the assessment of each item's relevance, clarity, and simplicity, thereby guaranteeing that it is in accordance with the intended learning outcomes and theoretical frameworks (Grant and Davis, 1997). The number of experts involved may vary depending on the study context; however, Aiken (1985) suggest that a panel of 3 to 10 experts can be sufficient for preliminary validation, with larger panels improving the reliability of the judgments. Experts are typically chosen based on their subject matter knowledge, experience in related research areas, and familiarity with the target population. Their feedback allows researchers to find items that are unclear, unnecessary, or redundant. Researchers frequently employ quantitative methods like Aiken's to measure expert agreement and identify which items need to be revised or removed (Aiken, 1985).

2.4.2 Pilot study

Pilot studies offer a practical means of refining the instrument, testing internal consistency, and evaluating item clarity before larger-scale data collection is conducted. They help reveal potential weaknesses in design, identify unclear wording, and provide an opportunity for initial statistical validation, such as calculating reliability coefficients (Alrababah, et al., (2024); Muasya and Mulwa, (2023)). Recent studies have shown the effectiveness of pilot testing across various educational contexts, including complex problem-solving, curriculum development, and digital literacy. The statistical analysis effectively validated the data literacy assessment for educators, achieving a high Cronbach's alpha of 0.977 (Donate-Beby et al., (2025)). Similarly, Anuar, (2024) utilized a pilot methodology to improve research training modules, demonstrating how iterative testing can enhance content clarity and instrument alignment. Mariño, et al., (2024) and Papa et al., (2023) implemented curriculum and assessment instruments in healthcare education to better understand learner answers and refine instructional design. Their work is another significant example. The process they employ illustrates the significant role that well-designed pilot studies play in the credibility of instruments and their potential for future scalability.

Studies by Azmi, et al., (2024); Rodriguez, et al., (2023) and Othman, et al., (2024) have shown that validated tools in TVET and special education settings can produce more precise, focused insights that eventually guide practice and policy. Speidel, et al., (2023) emphasizes how pilot study feedback aids in tool refinement and education innovation. On a broader methodological level, Kunselman, (2024) reviewed best practices for pilot studies, emphasizing that while pilot data should not be overinterpreted, it is vital for testing instrument clarity and internal consistency. Li, et al., (2024); Oppenlaender, et al., (2023) and Smith, et al., (2023) further highlight the necessity of accurate, well-structured instruments by looking at technologically advanced training environments. Similar findings were confirmed by Lee, et al., (2024) who looked at online learning and virtual involvement in vocational and higher education settings.

2.4.3 Item-level analysis

The performance of each item in an instrument needs to be analysed to item-level analysis provides detailed insights into the performance of each item. This process involves evaluating individual items using statistical

techniques such as Cronbach's alpha and item-total correlation, which help determine internal consistency and identify items that may weaken the instrument's reliability. The item-total correlation which assesses the degree to which each item correlates with the overall score. Values of 0.30 or higher indicate acceptable alignment (Nunnally and Bernstein, 1994). Another important measure is "Cronbach's Alpha if Item Deleted," which shows whether removing an item improves the scale's internal consistency (Tavakol and Dennick, 2011).

Exploratory Factor Analysis (EFA) is recommended to uncover the instrument's underlying factor structure if the sample size is large typically 100-300 respondents (Costello and Osborne, 2005). However, in small-sample settings like pilot studies, EFA is not suitable due to low statistical power (De Winter et al., 2009). Instead, small-sample studies should focus on descriptive statistics, item-total correlations, and internal consistency measures such as Cronbach's alpha. These approaches ensure that items function as intended and provide evidence to support future full-scale validation.

The importance of item-level analysis is well-supported in educational measurement literature. Bond and Fox (2015) emphasize that evaluating item performance is essential for achieving fundamental measurement standards, while Pallant (2020) highlights its role in improving scale precision and interpretability. Furthermore, the use of item-total correlation aligns with best practices in survey validation, ensuring that each item contributes meaningfully to the construct it represents.

2.5 Theoretical Framework for Instrument Development

To ensure theoretical depth in the development of the instrument, established framework were adopted to guide item construction. The three core constructs: ChatGPT usage, problem-solving, and ethical awareness items were constructed with reference to AI literacy frameworks, the Affective, Behavioral, Cognitive and Ethical (ABCE) framework, (Ng et al., 2021). The affective includes students' emotions, motivation, and confidence in interacting with AI tools. The behavioral refers to actual engagement with AI technologies, including the use of tools, participation, and collaboration in AI-supported activities. The cognitive focuses on students' knowledge, understanding, and ability to apply critical thinking and problem-solving strategies when interacting with AI. Finally, the ethical addresses students' awareness, evaluation, and reflection on the responsible and fair use of AI, including issues such as bias, misuse, and academic integrity. This framework was empirically validated through confirmatory factor analysis and offers a robust theoretical foundation for developing instruments that assess AI-related competencies. This framework stresses the ability to critically evaluate AI-generated responses and apply them effectively within academic and vocational contexts.

3.0 Methodology

3.1 Research design

This study used a quantitative, survey-based methodology to analyze the reliability, clarity, and usefulness of a structured instrument developed to assess the use of ChatGPT among students in TVET settings. The primary aim of the pilot was to validate the instrument in terms of its internal consistency and to examine its capacity to effectively measure three core constructs: (i) usage patterns of ChatGPT, (ii) the impact of ChatGPT on academic problem-solving, and (iii) students' ethical awareness related to ChatGPT use in academic contexts.

3.2 Participants and sampling

Participants were selected using a convenience sampling method, focusing on accessibility and willingness

to participate. The sample consisted of 43 students from two classes at Polytechnic Ungku Omar (PUO). According to Hertzog, (2008), a sample size ranging from 10 to 30 participants is typically sufficient for initial assessments of survey instruments, particularly when the goal is to refine item wording and structure. For studies assessing internal consistency using metrics such as Cronbach's alpha, Teijlingen and Hundley, (2001) recommend including at least 30 respondents to ensure meaningful reliability results. These students represented a relevant demographic, as they are likely to encounter both the practical applications and ethical implications of ChatGPT tools in their academic and future professional settings.

3.3 *Instrumentation*

The initial questionnaire was develop based on ABCE framework to ensure construct validity and contextual relevance. All items were carefully adapted to align with the Malaysian Technical and Vocational Education and Training (TVET) context and organized into three constructs: ChatGPT Usage (12 items), Problem-Solving (8 items), and Ethical Awareness (14 items). Responses were measured using a 4-point Likert scale (1 = Strongly Disagree to 4 = Strongly Agree). This scale was selected to encourage more decisive responses and avoid neutrality. The decision to use a four-point Likert scale was based on guidance from previous studies, such as those by Allen and Seaman, (2007) who emphasized that even-numbered scales can reduce central tendency bias and force respondents to take a clearer stance. To ensure content validity, the instrument was reviewed by three subject matter experts for RCS prior to pilot testing.

3.3.1 Content validity

To establish content validity, this study adopted Aiken's V coefficient as the primary metric for evaluating expert agreement on item quality. A panel of experts reviewed the initial draft questionnaire before distributing it for pilot testing. The panel consisted of 4 academic and field experts in educational measurement, TVET, and AI in education. The purpose of this review was to evaluate how well the content aligned with the research objectives. Each expert was asked to rate all items based on three criteria: RCS using a 4-point Likert scale as in Table 1.

Table 1. Three Criteria of Content Validity Assessment

Score	Relevance (R)	Clarity (C)	Simplicity (S)
1	Not relevant	Very unclear or confusing	Very complex and hard to understand
2	Slightly relevant	Somewhat unclear	Somewhat complex
3	Highly relevant	Clear and understandable	Simple and easy to understand
4	Extremely relevant	Very clear and precise	Very simple and straightforward

The Aiken's V coefficient was calculated using the following formula:

$$V = \frac{\sum(r - l)}{n(c - 1)}$$

Where:

- r = rating given by each expert
- l = lowest possible rating on the scale
- n = number of experts
- c = number of rating categories

Microsoft Excel is used to compute the Aiken's V coefficient where the expert ratings were entered. The value for Aiken's V coefficient was generated for each item across the three dimensions (RCS).

Even though the statistically derived critical value for this formation at a 5% significance level is 0.75 (Aiken, 1985), a more demanding, conventional threshold was implemented to guarantee a high level of expert agreement. Items with an Aiken's V coefficient of 0.83 or higher, as this value is widely acknowledged as a standard for establishing strong content validity in academic research, particularly within the Malaysian context (Nurjanah et al., 2023).

Furthermore, following the recommendations of Penfield and Giacobbi (2004), a 95% score confidence interval was calculated for each item's V-score. This approach provided a more robust and statistically sound basis for item acceptance, ensuring that the true content validity of each item was highly likely to meet the established threshold.

The expert panel's feedback was used to revise and refine questionnaire items to ensure accuracy and appropriateness for the target population. The revised questionnaire was only finalized for distribution via Google Forms for the pilot study after the feedback.

3.3.2 Pilot study for reliability

A pilot study was conducted to assess the reliability and clarity of the questionnaire. As described in the Participants and Sampling section, a group of 43 students was involved in this pilot testing. An online questionnaire administered via Google form was distributed to students. This mode was suitable for pilot testing and instrument validation within a limited timeframe and resource setting. Participants completed the instrument and provided feedback on item clarity and comprehension. Data collected were analyzed using descriptive statistics. To ensure the reliability of the instrument used in this study, internal consistency was measured using Cronbach's alpha coefficient for each section of the instrument Bond and Fox, (2015). Table 2 presents the interpretation of Cronbach's alpha values.

Table 2. Interpretation of Cronbach's Alpha

Cronbach's Alpha (α) Value	Reliability Interpretation
0.8 until 1.0	Very good and effective with a high degree of consistency
0.7 until 0.8	Good and acceptable
0.6 until 0.7	Acceptable
< 0.6	Items need to be repaired
< 0.5	Items need to be dropped

Source: Bond and Fox (2015)

The responses obtained from the online questionnaire were analyzed quantitatively according to the items in sections B, C, and D. The data were analyzed using Statistical Package for the Social Sciences (SPSS) version 26.

3.3.3 Item-level response

To determine whether each item aligned well with its intended construct, corrected item–total correlations were calculated. According to Nunnally and Bernstein (1994), items with a corrected item–total correlation below 0.30 may not contribute meaningfully to the internal consistency of the scale and are generally considered for revision or removal. For this study, a data-driven decision rule was applied to guide the refinement process. A corrected item-total correlation was computed for each item by correlating its scores with the total score of its respective scale, with the item's own score excluded from the total. This analysis

was performed using statistical software.

4.0 Results and Findings

This section reports the results of both expert validation and pilot study reliability analysis of the instrument developed to measure students' perceptions of ChatGPT in three constructs: Usage of ChatGPT, Problem-Solving, and Ethical Awareness. The expert review using Aiken's V ensured content validity, while the pilot study was conducted to assess the internal consistency reliability of the instrument concerning the three constructs.

4.1 Content Validity Evaluation using Expert Ratings

Table 3 shows the summarization of Aiken's V result for section B. This section examined the use of ChatGPT for academic tasks. Based on the Aiken's V analysis, 9 out of 10 items exceeded the threshold of 0.83 and were accepted. Items B1: I use ChatGPT to help generate ideas for my assignments and item B6: I rely on ChatGPT more than traditional study resources (books, lectures) received strong expert agreement, with values ranging from 0.92 to 1.00. Only one item that is B10: I have learned how to craft effective prompts to receive accurate responses scored below the cutoff on the relevance dimension. After further discussion with the experts, this item was omitted from the final instrument, as its content was conceptually covered by other items in the construct. Thus, 9 validated items were retained to measure ChatGPT's role in supporting academic tasks

Table 3. Summarization of Aiken's V Result for Section B

Item Statement	Aiken's V (Relevance)	Aiken's V (Clarity)	Aiken's V (Simplicity)	Decision (Cutoff = 0.83)
B1: I use ChatGPT to help generate ideas for my assignments.	1.00	1.00	1.00	Accept
B2: I use ChatGPT to solve specific problems related to my courses.	0.92	0.92	0.92	Accept
B3: I use ChatGPT to understand difficult concepts related to my courses.	1.00	0.92	0.92	Accept
B4: I use ChatGPT to find information	0.92	0.83	0.83	Accept
B5: ChatGPT supports my self-learning by providing personalized responses.	0.92	0.92	0.92	Accept
B6: I rely on ChatGPT more than traditional study resources (books, lectures)	1.00	0.92	1.00	Accept
B7: I use ChatGPT to help edit my writing.	0.92	0.92	0.92	Accept
B8: I use ChatGPT as a study tool during revisions or exam preparations.	0.92	0.92	0.92	Accept
B9: I use ChatGPT to clarify instructions or questions in assignments.	0.83	0.83	0.83	Accept
B10: I have learned how to craft effective prompts (memberi arahan / soalan) to ChatGPT to receive accurate responses.	0.75	0.92	0.83	Revise

Table 4 shows the summarization of Aiken's V result for Section C. This section focused on the use of ChatGPT's among students for problem-solving. The Aiken's V results showed that 6 out of 8 items met the threshold of 0.83, confirming their validity. High-scoring items are CI: ChatGPT helps me solve complex problems by breaking them into smaller parts and item C6: ChatGPT encourages me to approach problems from different perspectives, with values ranging from 0.92 to 1.00. Two items, C4 and C7 fell below the cutoff in certain dimensions. After expert discussion, these items were omitted because their ideas evaluating solutions and using ChatGPT as support rather than replacement were already represented in other items. Accordingly, 6 validated items were retained to represent the construct of problem-solving.

Table 4. Summarization of Aiken's V Result for Section C

Item Statement	Aiken's V (Relevance)	Aiken's V (Clarity)	Aiken's V (Simplicity)	Decision (Cutoff = 0.83)
C1: ChatGPT helps me solve complex problems by breaking them into smaller, more manageable parts.	1.00	0.92	0.92	Accept
C2: Using ChatGPT allows me to consider a wider rangen of potential solutions to problems.	0.92	0.92	0.92	Accept
C3: I use ChatGPT to generate different approaches to solving a problem.	0.83	0.83	0.83	Accept
C4: ChatGPT helps me to evaluate the pros and cons of different solutions.	0.75	0.92	0.92	Revise
C5: Using ChatGPT has improved my critical thinking skills in problem-solving.	0.83	0.83	0.83	Accept
C6: ChatGPT encourages me to approach problems from different perspectives.	1.00	0.92	0.92	Accept
C7: I use ChatGPT as a tool to support - not to replace my problem-solving efforts.	0.83	0.67	0.83	Revise
C8: Using ChatGPT gives me relevant tech skills for the current job market.	0.83	0.83	0.83	Accept

Table 5 shows the summarization of Aiken's V result for Section D. This section measured ethical awareness in using ChatGPT. The Aiken's V analysis indicated that 12 out of 14 items exceeded the cutoff value, with values ranging from 0.83 to 1.00. Item D6: I know the difference between using ChatGPT for learning support and using it to cheat in academic tasks and item D8: I believe using ChatGPT without proper citation is a form of plagiarism were rated highly, indicating strong expert consensus. Two items, D7 and D14 fell below the cutoff on clarity. Following consultation with the experts, these items were omitted as their constructs were already reflected in other accepted items. As a result, 12 validated items were retained to measure ethical awareness when using ChatGPT.

Table 5. Summarization of Aiken's V Result for Section D

Item Statement	Aiken's V (Relevance)	Aiken's V (Clarity)	Aiken's V (Simplicity)	Decision (Cutoff = 0.83)
D1: I avoid using ChatGPT to copy answers directly into my assignments.	0.92	0.92	0.92	Accept
D2: I check the accuracy of ChatGPT's responses before using them.	0.92	0.92	0.92	Accept
D3: I am concerned that using ChatGPT too much might reduce my own learning and critical thinking.	1.00	0.83	0.83	Accept
D4: I always review and verify the information provided by ChatGPT based on my learning and experience before using it in my academic work.	1.00	0.92	0.92	Accept
D5: I understand ChatGPT is a learning support, not a replacement for my own work.	1.00	0.83	0.83	Accept
D6: I know the difference between using ChatGPT for learning support and using it to cheat in academic tasks.	1.00	1.00	1.00	Accept
D7: Using ChatGPT for assignments is ethically acceptable.	0.83	0.67	0.75	Revise
D8: I believe using ChatGPT without proper citation is a form of plagiarism.	1.00	1.00	1.00	Accept
D9: I am aware ChatGPT may provide inaccurate information.	0.92	0.92	0.92	Accept
D10: I am aware of my institution's policies regarding the use of AI tools like ChatGPT.	0.92	0.83	0.92	Accept
D11: I am aware that ChatGPT could interpret my question wrongly, which will end up giving the wrong answer/solution.	0.92	0.92	0.92	Accept
D12: I have received training on using ChatGPT effectively and ethically in education.	0.83	0.83	0.83	Accept
D13: I think students should receive training on responsible use of AI tools.	0.92	0.83	0.83	Accept
D14: I have discussed the ethical use of AI tools like ChatGPT with my lecturers or peers.	0.83	0.75	0.83	Revise

The expert validation confirmed that the items were content valid and appropriate for assessing the targeted constructs. Only minor revisions were required to strengthening the clarity and focus of the questionnaire for use in the pilot study.

4.2 Pilot Study for Reliability

Prior to the content validation phase, a pilot study was executed to assess the internal consistency reliability of the instrument concerning the three constructs: ChatGPT utilization for academic assignments, problem-solving, and ethical awareness. Forty-three students were engaged in the pilot study.

Table 6 shows the result for the pilot study conducted. The results of Cronbach's alpha indicated high internal consistency across all constructs validating that the items consistently assessed their intended structures.

Table 6. Result for Pilot Study

Section	N of Items	Cronbach's Alpha
B: ChatGPT Usage	9	0.925
C: Impact on Problem-Solving	6	0.933
D: Ethical Awareness	12	0.915
Overall Instrument	**27**	**0.945**

4.3 Item level response

Item-level analysis was conducted to assess the contribution of each item within the three constructs: (i) ChatGPT usage, (ii) problem-solving and (iii) ethical awareness. Corrected item–total correlations and "Cronbach's Alpha if Item Deleted" were examined to evaluate internal consistency and identify whether any items should be considered for removal.

Table 7 shows the item-level analysis for ChatGPT usage. The result shows the corrected item–total correlations ranged from 0.534 to 0.893, which are all above the recommended cutoff value of 0.30 (Nunnally and Bernstein, 1994). This indicates that all items demonstrated acceptable levels of correlation with the overall scale, contributing meaningfully to the measurement of the construct. Cronbach's Alpha if Item Deleted values ranged between 0.906 and 0.928, suggesting that removal of any item would not significantly improve the internal consistency of the construct. The overall Cronbach's alpha for ChatGPT usage construct was 0.925, which reflects a high level of internal consistency reliability. Therefore, all 9 items under ChatGPT construct were retained for the final instrument, confirming their adequacy in capturing the intended dimensions of students' usage of ChatGPT for their academic task.

Table 7. Item Level Response for ChatGPT Usage.

ITEM	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	R TABLE	Corrected > R – VALID Corrected < R – NOT VALID
B1	24.86	17.504	.534	.928	0.301	VALID
B2	24.86	16.790	.838	.913	0.301	VALID
B3	24.79	17.169	.675	.920	0.301	VALID
B4	24.79	17.503	.662	.922	0.301	VALID
B5	24.91	16.134	.716	.918	0.301	VALID
B6	25.23	14.802	.788	.915	0.301	VALID
B7	25.09	15.515	.739	.917	0.301	VALID
B8	24.93	15.971	.823	.911	0.301	VALID
B9	24.91	15.563	.893	.906	0.301	VALID

The findings shows that students responded consistently to the items measuring Usage of ChatGPT, indicating that the construct is well-represented by the selected items. The high internal reliability demonstrates that the items collectively capture different but related aspects of how students engage with ChatGPT in their academic work. This implies that students perceive the usage of ChatGPT as a structured and meaningful activity, and the instrument is suitable for measuring their behaviors and patterns of use.

Table 8 shows the item-level analysis for problem-solving. The result shows the corrected item–total correlations ranged from 0.726 to 0.856, all of which are above the recommended cutoff point of 0.30 (Nunnally and Bernstein, 1994). This indicates that each item is positively correlated with the total construct and contributes meaningfully to the overall measurement of students' problem-solving skills. The values for "Cronbach's Alpha if Item Deleted" ranged between 0.914 and 0.932, suggesting that the deletion of any single item would not result in a significant increase in the reliability of the scale. The overall Cronbach's alpha for the Problem-Solving construct was 0.933, which indicates excellent internal consistency. Consequently, all 6 items under the Problem-Solving construct were retained for the final version of the instrument.

Table 8. Item Level Response for Problem-Solving.

ITEM	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	R TABLE	Corrected > R – VALID Corrected < R – NOT VALID
C1	15.35	4.566	.856	.914	0.301	VALID
C2	15.30	5.073	.824	.922	0.301	VALID
C3	15.30	4.359	.808	.923	0.301	VALID
C4	15.35	4.804	.830	.918	0.301	VALID
C5	15.42	4.678	.835	.917	0.301	VALID
C6	15.37	4.715	.726	.932	0.301	VALID

The findings demonstrate that students responded consistently to the items measuring Problem-Solving with ChatGPT, highlighting that the construct is well-represented by the selected items. The high internal reliability suggests that the instrument successfully captures different aspects of how students perceive ChatGPT as a problem-solving tool in their learning activities. This implies that ChatGPT is regarded not only as a support tool for solving academic tasks but also as a structured aid that aligns with students' approaches to tackling problems effectively.

Table 9 shows the item-level analysis for ethical awareness. The result shows the corrected item–total correlations ranged from 0.350 to 0.891, which are all above the minimum threshold of 0.30 (Nunnally and Bernstein, 1994). This indicates that every item demonstrated a sufficient correlation with the total construct and contributed meaningfully to the measurement of students' ethical awareness. The "Cronbach's Alpha if Item Deleted" values ranged between 0.899 and 0.921, confirming that the removal of any single item would not substantially increase the overall reliability. The overall Cronbach's alpha for the Ethical Awareness construct was 0.915, signifying excellent internal consistency reliability. Based on these results, all 12 items under the Ethical Awareness construct were retained for the final version of the instrument.

Table 9. Item Level Response for Ethical Awareness.

ITEM	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted	R TABLE	Corrected > R – VALID Corrected< R – NOT VALID
D1	35.00	22.857	.350	.920	0.301	VALID
D2	34.88	20.581	.727	.904	0.301	VALID
D3	35.02	21.071	.590	.911	0.301	VALID
D4	34.86	20.790	.796	.902	0.301	VALID
D5	34.98	21.690	.419	.921	0.301	VALID
D6	34.84	20.663	.891	.899	0.301	VALID
D7	35.00	20.857	.658	.907	0.301	VALID
D8	34.72	20.730	.795	.902	0.301	VALID
D9	34.88	20.439	.755	.903	0.301	VALID
D10	34.86	20.980	.753	.904	0.301	VALID
D11	35.12	21.581	.533	.913	0.301	VALID
D12	34.86	20.742	.737	.904	0.301	VALID

The findings reveal that students responded consistently to the items measuring Ethical Awareness in relation to ChatGPT usage. The high internal consistency shows that the construct is well-captured by the items, covering three constructs of students' perceptions and attitudes toward responsible use of ChatGPT. Similar findings are found in Ng et al. (2023), who validated an AI literacy questionnaire using affective, behavioural, cognitive, and ethical constructs. This suggests that students recognize ethical considerations such as honesty, originality, and fairness as an integral part of using AI tools in their academic activities.

5.0 Conclusion

Both expert validation and pilot study results confirmed the reliability and validity of the instrument. The expert review (Aiken's V) ensured that the items were conceptually relevant, clear, and non-redundant, while the pilot study demonstrated excellent internal consistency across all constructs, with Cronbach's alpha values exceeding 0.90. These results are consistent with prior validation studies in education and technology (Teijlingen and Hundley, 2001; Muasya and Mulwa, 2023), which stress the importance of conducting expert validation and pilot testing before deploying an instrument in a main study.

The rising adoption of AI tools in education, particularly within the TVET sector, this study seeks to contribute to the field by validating a survey instrument designed to assess ChatGPT usage, impact on problem-solving and ethical awareness among TVET students. Through careful instrument validation, the study aims to ensure that the instrument is reliable, clear, and suitable for future large-scale research. In doing so, the study lays a foundation for future research and contributes to ongoing efforts to foster ethical and effective AI engagement within educational settings.

Declarations

AI tools such as ChatGPT were used to support idea brainstorming and structure refinement throughout the research and writing process. QuillBot was employed for grammar and spelling corrections.

Acknowledgements

The author would like to express sincere gratitude to the TVET students from Polytechnic Ungku Omar who participated in the pilot study, providing valuable feedback. Special thanks are also extended to the expert reviewers for their constructive insights during the instrument design and validation phases, which significantly enhanced the quality of the questionnaire developed.

References

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131–142. https://doi.org/10.1177/0013164485451012

Allen, I. E., & Seaman, C. A. (2007). Likert scales and data analyses. Quality Progress, 40(7), 64–65.

Alrababah, S. A., Wu, H., & Molnár, G. (2024). A Pilot Study for Measuring Complex Problem-Solving in Jordan: Feasibility, Construct Validity, and Behavior Pattern Analyses. SAGE Open, 14(2). https://doi.org/10.1177/21582440241249884

Amdan, M., Janius, N., Jasman, M., & Kasdiah, M. (2024). Advancement of AI-tools in learning for technical vocational education and training (TVET) in Malaysia (empowering students and tutor). International Journal of Science and Research Archive, 12(01), 2061–2068. https://doi.org/10.30574/ijsra.2024.12.1.0971

Anuar, N. (2024). Development of a Research Design Module for Industry Professionals: A Pilot Study. Journal of Creative Practices in Language Learning and Teaching, 12(3), 111–128. https://journal.uitm.edu.my/ojs/index.php/CPLT/article/view/2644

Azmi, A., Jamaludin, M. F., Anuar, S., & Ali, M. F. M. (2024). The Role of Educator's Research Perceptions in Shaping Attitudes: Insights from Technical and Vocational Education. International Journal of Research and Innovation in Social Science, 8(10), 170-184.

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.). Routledge.

Chan, Cecilia & Hu, Wenjie. (2023). Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education. 10.48550/arXiv.2305.00290.

Chheang, V., et al. (2023). Towards anatomy education with generative AI-based virtual assistants. arXiv preprint. https://arxiv.org/abs/2306.17278

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research & Evaluation, 10(7)

De Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis and the Kaiser criterion. International Journal of Psychological Research, 4(1), 1-10.

Donate-Beby, B., García-Pe˜nalvo, F.S., Amo-Filva, D. & Aguayo-Mauri, S. (2025). Filling the gap in K-12 data literacy competence

assessment: Design and initial validation of a questionnaire. Computers in Human Behavior Reports. https://doi.org/10.1016/j.chbr.2024.100583

Grant, L., & Davis, H. (1997). Conducting a content validity study in social work research. Social Work Research, 21(4), 211-215.

Hertzog M.A. (2008). Considerations in determining sample size for pilot studies. Res Nurs Health. 31:180-191.

Kunselman, M.A. (2024). A brief overview of pilot studies and their sample size justification. Fertil Steril, 121(6). https://doi.org/10.1016/j.fertnstert.2024.01.040

Lee, Y., et al. (2024). Evaluating the effectiveness of online learning platforms in vocational training: A pilot evaluation. Vocational Education Review, 12(1), 35–47.

Li, J., Feng, S., Yan, Y., Lee, C.C., & Ong, Y.S. (2024). Virtual co-pilot: Multimodal large language model-enabled quick-access procedures for single pilot operations. arXiv preprint. https://arxiv.org/abs/2403.16645

Mariño, R., Capurro, D., & Merolli, M. (2024). Pilot implementation of a telehealth course for health professions students. BMC Medical Education, 24, Article 129. https://doi.org/10.1186/s12909-024-05931-z

Muasya, J., N. & Mulwa, P., K. (2023). Pilot study, a neglected part of qualitative and quantitative research process: Evidence from selected PhD thesis and dissertations. Higher Education Research. https://doi.org/10.11648/j.her.20230804.11

Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2023). Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. British Journal of Educational Technology. Advance online publication. https://doi.org/10.1111/bjet.13411

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

Nurjanah, et al. (2023). The application of Aiken's V method for evaluating the content validity of instruments that measure the implementation of formative assessments. Journal of Educational Research and Evaluation, 12(2), 125-133

Oppenlaender, J., Abas, T. & Gadiraju, U. (2023). The state of pilot study reporting in crowdsourcing: Challenges and directions. arXiv preprint. https://arxiv.org/abs/2312.08090

Othman, N.N.J.N., Kamaruzaman, F.M., Rasul, M.S. (2024). A pilot study on the perception of special education students towards work ability and job profile. International Journal of Modern Education, 6(1), 22–31. https://gaexcellence.com/ijmoe/article/view/1888

Pallant, J. (2020). SPSS Survival Manual (7th ed.). McGraw-Hill Education.

Papa, R., Sixsmith, J., Giammarchi, C. et al. (2023). Health literacy education at the time of COVID-19: Pilot experiences from European schools. BMC Medical Education, 23, Article 608. https://doi.org/10.1186/s12909-023-04608-3

Penfield, R. D., & Giacobbi, P. R. (2004). Applying Aiken's V: A content validity index for item-level content validity. Measurement in Physical Education and Exercise Science, 8(4), 213–225.

Rodriguez, L., et al. (2023). Implementing gamification in technical education: A pilot approach. Technical Education Quarterly, 15(4), 22–34.

Smith, R., et al. (2023). Integrating AI tools in higher education: A pilot study. Journal of Educational Technology, 20(3), 101–115.

Speidel, R., Wong, T.K.Y., Al-Janaideh, R. et al. (2023). Nurturing child social-emotional development: evaluation of a pre-post and 2-month follow-up uncontrolled pilot training for caregivers and educators. Pilot Feasibility Study, 9, 148. https://doi.org/10.1186/s40814-023-01357-4

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Teijlingen, V., E. & Hundley, V. (2001). The importance of pilot studies. Social Research Update, 35, 1–4. https://sru.soc.surrey.ac.uk/SRU35.html

Titko, J., Steinbergs, K., Achieng, M., & Uzule, K. (2023). Artificial intelligence for education and research: Pilot study on perception of academic staff. Virtual Economics, 6(3), 7-19.