Expert Validation of a Qualitative Interview Protocol for ESL Instructors' Digital Assessment Literacy in Malaysian Higher Education: A Delphi Study

Ameera Irdena Binti Suyansah[1], Noraini Binti Said [2], Esther binti Jawing[3], Wardatul Akmam Binti Din[4]

[1,2,4]Faculty of Education and Sport Studies, University Malaysia Sabah, Malaysia

[3]The Centre for the Promotion of Language Learning, University Malaysia Sabah, Malaysia

DOI: https://dx.doi.org/10.47772/IJRISS.2025.907000212

Received: 02 July 2025; Accepted: 08 July 2025; Published: 08 August 2025

ABSTRACT

The integration of technology into educational assessment is a critical component of modern pedagogy, yet many educators lack proficiency in digital assessment methodologies. This study addresses the need for a validated instrument to explore the digital assessment literacy (DAL) of English as a Second Language (ESL) instructors in the Malaysian higher education context. The primary objective was to validate a qualitative interview protocol designed to investigate instructors' assessment practices and their level of technology integration. The protocol consists of two instruments, one guided by the Teacher Assessment Literacy in Practice (TALiP) framework and the other by the Substitution, Augmentation, Modification, and Redefinition (SAMR) model. A Delphi method was employed to achieve consensus on the instrument's content validity. Six experts in the fields of digital assessment and ESL, holding PhDs with over five years of relevant experience, were selected to evaluate the interview items. Inter-rater reliability was calculated using the Fleiss' Kappa statistic. The results indicated an "almost perfect" level of agreement among the expert validators, with a Kappa value of 0.81 for the TALiP instrument and 0.83 for the SAMR instrument. Based on expert feedback, several items were refined for clarity, including the separation of double-barreled questions and the rephrasing of complex terminology, which increased the total number of items in the TALiP protocol from 10 to 15. The study concludes that the validated interview protocol is a robust and reliable instrument for exploring the multifaceted dimensions of DAL among ESL instructors. This tool can be effectively utilized by researchers and educational institutions to inform the design of targeted professional development programs aimed at enhancing digital assessment practices in higher education.

Keywords: Digital Assessment Literacy, Instrument Validation, ESL Studies, Delphi Method, Teacher Assessment Literacy

INTRODUCTION

In the education domain, instructors play an essential role in moulding the knowledge and skills of their students. As the need for skilled educators continues to rise, there is a corresponding emphasis on the importance of equipping teachers with strong assessment competencies (Malone, 2013). The capability to evaluate student performance effectively is central to the success of the process of teaching and learning. Instructor-led assessments play a pivotal role in producing the data that helps the instructors evaluate the effectiveness of teaching strategies and determine the extent to which the assessments are aligned with the intended learning outcomes (Bane, 2019). As a result, assessment data can significantly influence teaching practices, including the adaptation of instructional materials and the revision of learning objectives.

Despite its increasing prominence, many teachers remain unfamiliar with digital assessment methodologies. Prior research has underscored that digital assessment remains one of the more challenging aspects of implementing online learning environments (Motala & Menon, 2020). Studies have noted that technology in education has often been restricted to basic functions such as data storage, information retrieval, and the preparation of presentation materials, rather than being fully integrated into assessment practices (Jang & Chen, 2010; Keengwe & Geogina, 2011).

In Malaysia, universities have autonomy in the creation of their syllabi in the tertiary sector. However, they are obliged to adhere to the specific national standards. The responsibility for ensuring quality in higher education is managed by an agency which is the Malaysia Qualifications Agency (MQA) as was established under the Malaysia Qualifications Agency Act 2007. MQA as the main regulatory agency regulates the accreditation and operation of quality assurance in higher education institutions nationally. One of the strengths of MQA's schema is the importance it gives to incorporating technology (Ghanbari & Nowroozi, 2021). With the preparation of computer-based tests, online dialogues, and special media interaction, the tests are prepared not just as mere tests of language proficiency, but as instruments that introduce students to their digital literacy. This focus is demonstrated in the Programme Standard: Language document (MOE, 2024) which always emphasises the need for integration of technology in language courses. Such a future-oriented approach recasts language assessment as malleable instruments that change with technological transitions, ensuring that the students get ready for the needs of a digitalised global environment.

Although the MQA has supported the integration of technology in assessment, it cannot be denied that instructors face difficulties in digital assessment. The difficulties associated with integrating digital assessment are not confined to the school system but are also evident in higher education. A study conducted by Arouri et al., (2025) discovered that both students and faculty members have a positive attitude towards distance learning, however, the students do not want distance learning to be continued after the pandemic as they believe that in-person learning is superior compared to digital learning. This is because the experience of digital learning is not as interactive as physical learning. Several studies have explored barriers to the adoption of digital assessment at the tertiary level. It has been identified that issues such as communication breakdowns, technological limitations, and insufficient student preparedness as major obstacles (MOE, 2020).

While challenges remain, it is undeniable that digital assessments offer substantial advantages, particularly for ESL (English as a Second Language) instructors. Research has shown that with the integration and intervention of technology in assessment, students have had a positive improvement in their assessment (Santhi et al., 2019, Nordin & Swanto, 2023). One significant benefit lies in the expanded capacity to assess multiple competencies simultaneously (Awang, 2022). Digital technologies now enable the construction of assessments capable of evaluating complex skill sets, such as technical computing abilities, critical decision-making, and strategic planning—tasks that were traditionally difficult to measure comprehensively. Simulation-based assessments serve as a notable example of this expanded capability. Not only that, but it also increases the engagement by students and also the collaboration among students, which includes peer assessment (Podsiad & Harvard, 2020; Loureiro & Gomes, 2023).

In light of the rapid advancement of digital assessment tools, it is imperative that Malaysian English instructors are able to develop proficiency in their digital assessment practices. They must ensure that assessments are not only valid and reliable but also closely aligned with the intended learning outcomes. A lack of competence in digital assessment may lead to inaccurate evaluations, negatively impacting students' educational experiences and outcomes. It is crucial for ESL educators to strengthen their digital assessment skills to make informed, accurate instructional decisions. Therefore, it is crucial to have an instrument that is able to explore the digital assessment literacy (DAL) of instructors in the ESL context. With the existence of the instrument, it is able to improve the digital assessment literacy and practices in the ESL domain, therefore creating a better environment. This study seeks to investigate the practices employed by ESL instructors in their assessment in the digital context within the Malaysian tertiary education institutions.

This research paper aims to validate the interview protocol by using the expert judgment in the field of study. This interview protocol aims to explore the instructors' level of digital assessment integration and the elements of assessment literacy in a digital environment. By the end of this research, the quality of the proposed instrument will be evaluated, and improvements will be made based on the suggestions given by the experts.

SAMR Model

The SAMR model is a model that was designed to help educators meaningfully integrate technology into teaching and learning. Developed by Dr. Ruben Puentedura (2013), the SAMR model supports instructors in planning and conducting digital assessments and rethinking how they use technology in the classroom. The model is structured into four progressive stages—Substitution, Augmentation, Modification, and Redefinition. As instructors move up these levels, the use of technology shifts from simply enhancing existing tasks (Substitution and Augmentation) to transforming learning experiences altogether (Modification and Redefinition). However, instructors are not only restricted to following a fixed sequence. The integration of technology can begin at any level that suits their context.

The core purpose of the SAMR model is to inspire ESL instructors to push beyond basic uses of technology and explore more transformative, innovative approaches in their lessons. The integration of the SAMR model in ESL teaching encourages the design of tech-integrated activities that foster student collaboration, motivation, and deeper engagement with learning (Sarker et al., 2019). Rather than prescribing specific tools or apps, the model emphasises how technology is used—encouraging educators to reflect on their practices and consider how tech can truly elevate the learning experience.

TALiP Framework

The TALiP framework which was developed by Xu and Brown (2016) is a comprehensive framework that emphasises the elements of assessment literacy that need to be mastered by instructors in their assessment. The TALiP framework emphasises the dynamic and interactive relationship between the components. It is not a linear model, but instead the components are in constant interaction with each other. The TALiP framework consists of six components interacting with each other, with the knowledge base at the bottom pyramid, followed by teacher conception of assessment, micro- and macro contexts, teacher assessment literacy in practice, teacher learning and identity (re)construction at the top of the pyramid. The knowledge base serves as a dimension where instructors must master different knowledge to apply in their assessment. However, it is worth noting that a knowledge base alone is not sufficient as it is a criterion, not a solution when a problem arises. The second component is the teacher's conception of assessment. The teacher conceptions of assessment are both at the collective level, whereby they are influenced by socio-cultural and institutional contexts; and at the individual level whereby the instructor’s conception of assessment is influenced by the affective domain (the emotional aspect that denotes their emotional inclination towards assessment) and the cognitive domain (what is true and false).

The third component delves into macro- and macro- institutional context. Instructors are pressured by external forces in their assessments, as they are bound by the policies, rules, and regulations. When the external forces exert a heavier pressure on the instructor, the instructor will have less room to exercise their autonomy in their assessment. The fourth component is teacher assessment of literacy in practice. This component focuses on the instructor's actual practice in their assessment while also balancing the stakeholders and their autonomy as an instructor, as they know their students best. The fifth component is the teacher learning. As learning is a lifelong process, instructors are constantly improving and upgrading their methods in assessment to develop their professional knowledge in assessment. This can be achieved in professional development programs, collaboration with peers and reflective practices. Situated at the apex of the pyramid, the last component is the instructor's (re)conceptualization. This relates to how instructors view themselves in the assessment. The instructors see assessment as an integral part of teaching and learning and can make principled and effective assessment decisions that are tailored for their students and their context. This identity is continually shaped and reshaped through the dynamic interplay of all the other components of the framework.

METHODOLOGY

Instrumentation

Before conducting the study, the adapted instruments that will be used in this study undergo the process of validation to ensure that the items from the instruments are applicable in the context of the ESL instructors. A total of six instructors were selected to validate the instruments. The questionnaire was distributed to the experts after receiving their consent. A symmetric Likert scale was chosen as it provides the participants with a balanced choice construction balanced, whereby the direction they choose as an asymmetric scale may lead to forced choices when there is no neutral value (Tsang, 2012; Malhotra et al., 2006). The experts were then asked to fill in the SAMR Model and TALiP framework interview questionnaire which consists of a 5-point Likert scale, ranging from strongly disagree (5) to strongly agree (1). The validation from the validators is reported in this report along with the comments given by the validators.

Assessment Instrument: TALiP Framework

The interview questions are divided into two sections. This section covers the topic of assessment literacy with the guidance of the TALiP framework. Table I displays the interview questions for the TALiP framework which consists of 10 questions that, consist of five criteria which are Identity Construction, Teacher Learning, Institutional & Socio-Cultural, Assessment Literacy in Practice, Teacher Conception of Assessment, and Knowledge Base (Xu and Brown,2016). This section explores the elements of assessment literacy in a digital environment.

TABLE I THE TALIP FRAMEWORK ASSESSMENT INSTRUMENT

Criteria	Indicator
Identity Construction	4. In which way do you think the recent technological advancement have changed or added to the roles and responsibilities of a teacher as an assessor?
	9. What are the teacher’s roles in technology-enhanced- assessment?
	10. How do you define an effective and successful teacher as an assessor in digital environment?
Teacher Learning	7. What was your experience of becoming familiar with the technological tools and learning to use them in the classroom assessment?
	8. How do you attempt to improve your knowledge and skills of using technology in language assessment? how do you keep your knowledge and skills updated?
Institutional & Socio-Cultural	5. Which challenges do you face when you use technology in assessment (e.g. cultural, persona, contextual, administrative, ideological and the like)? Please explain what knowledge and skills you need to deal with them
Assessment Literacy in Practice	6. Are there any moral concerns that you think you should be cultivated regarding the use of technology for proper language and classroom assessment? if so, please provide an example and explain what you do need for establishing them.
	2. Do you think having knowledge and skills in digital assessment is essential for teachers in the 21st century? How? Why?
Teacher Conception of Assessment	3. How comfortable and confident do you consider yourself in using technology for classroom assessment?
Knowledge Base (KB)	1. Please explain what knowledge and skills teachers as classroom assessor need to success to successfully conduct technology enhanced/digital assessment?

Assessment Instrument: SAMR Model

The second section delves into the topic of digital assessment with the SAMR model as the guide. The SAMR model is a model that intends to categorise the teachers' use of integration of technology (Puentedura, 2013). The interview questions for the SAMR framework consist of 12 questions, as shown in Table II. As this section will dissect the instructors' level of digital integration in their assessment, there are no specific criteria. The researcher will divide the instructors' level of digital integration in their assessment in the analysis phase of the research. The experts must fill in the questionnaire based on the 5- 5-level Likert scale, ranging from strongly disagree (5) to strongly agree (1). Experts are also encouraged to give their opinion in the comment section.

TABLE II THE SAMR MODEL ASSESSMENT INSTRUMENT

Criteria	Item
SAMR	1. Talk about your experience using technology in your current role.
	2. When thinking of the SAMR model, describe the level you feel most comfortable integrating technology
	3. In an assessment and learning environment, describe the levels of the SAMR model you most commonly use when integrating technology.
	4. How do you make decisions concerning the level of technology integration?
	5. When integrating technology, explain how the educational environment changes when technology is used in an environment where technology was not previously integrated.
	6. What is your overall opinion of the SAMR model?
	7. When thinking of the bottom half of the model (enhancement) and the top half of the model (transformation), what are your thoughts on the assessment process as you move from the bottom to the top?
	8. Describe how time spent on lesson design changes from the bottom half of the model to the top half of the model?
	9. What are your thoughts of the amount of time spent when creating an assessment at the substitution level versus the redefinition level?
	10. Do you think when integrating technology, teachers should always strive for the redefinition level or otherwise? Explain.
	11. What is your opinion on the time commitment creating lessons at the redefinition level?
	12. Do you believe student learning outcomes would justify the time required to design lessons at the redefinition level?

Expert Validators

The suggested number of expert validators would be between five to eight experts, and should not exceed 10 experts (Yusoff, 2019). Polit and Beck (2006) highlighted that when there are more than 10 experts in the validation process, this would diminish the likelihood of reaching a consensus in the validation process. Therefore, six experts were selected for the validation of the instruments. The expert was chosen based on several criteria which are, they must have a PhD in their respective field for more than five years and above; they must have in-depth knowledge in their respective field at university level; they must have experience of teaching and researching for more than five years and above.

The expert panel were approached by the researcher to receive their permission to participate in the research. After the expert panel has agreed to participate in the research, the researcher officially appointed each expert panel member with a letter of appointment. In the appointment letter, information about the study that details the purpose and instructions of the study was also attached to the letter. The experts were requested to give their professional judgements on the items which need to be rectified. The experts were also asked to give a comment on how to improve the items. This was to ensure that the items are distributed, the items can assist in answering the research questions of this study. The details of the expert panel are explained in the table below.

Table III EXPERT PANEL DESCRIPTIONS

Validators	Position	Descriptions
Validator 1	Associate Professor	An expert in the field of digital assessment as the validator has entered competitions related to digital learning and has been a keynote speaker in conferences related to digital learning. The validator has published articles on digital learning
Validator 2	Associate Professor	An expert in the field of ESL assessment as the validator has been a keynote speaker in conferences related to ESL learning. The validator has published articles on ESL.
Validator 3	Associate Professor	An expert in the field of digital assessment as the validator has published articles on digital learning
Validator 4	Senior Lecturer	An expert in the field of assessment in ESL as the validator has entered competitions related to digital learning. The validator has published articles on assessment in ESL
Validator 5	Senior Lecturer	An expert in the field of digital assessment in ESL as the validator has published articles on ESL teaching and learning.
Validator 6	Lecturer	An expert in the field of digital assessment in ESL as the validator has entered competitions related to digital learning. The validator has published articles on digital learning

Delphi Method

The research method used in the collection of the expert judgement was by using the Delphi method. The Delphi method was originally developed by the RAND Corporation to forecast the priority in the military (Dalkey & Helmer, 1963). Dalkey (1972) defines the experts for the context of the Delphi method as people who are knowledgeable and proficient in the field of study. In the context of this research, the Delphi method is used to gather the expert in the field of digital assessment and ESL for their knowledge and experience to come to a consensus. For this method, the experts are consulted for their input on a predefined set of variables that have been extracted based on the literature findings.

The Delphi method often involves multiple iterative rounds to gradually build consensus among experts. However, for this study, a single round of the Delphi method was deemed sufficient. The decision to conclude after one round was justified by the high degree of initial consensus achieved, as evidenced by the "almost perfect" Fleiss' Kappa values of 0.81 and 0.83 for the respective instruments, as stated in the findings. This high degree of initial agreement among the experts served as the justification for concluding the validation process after one round, as further iterations were unlikely to significantly shift the results, aligning with Chuenjitwongsa's (2017) guidance that the process can conclude once consensus is reached.

Fleiss Kappa

The instrument proposed for expert validation consists of 10 and 12 questions each, which aim to determine the extent of agreement among the expert panel. The Kappa Fleiss statistic was chosen as the statistical measure to evaluate the inter-rater reliability as this statistical measure evaluates the agreement for more than two raters (Fleiss, 1971). The Fleiss Kappa Statistics was calculated with 22 items (N = 22) which were divided into two sections, involving six experts (n = 6), and five rating values (k = 5). The calculation of the probability of agreeing Pr(a) and the chance of agreement Pr(e) was made. The level of agreement between the expert is indicated by the Kappa value (κ) and was calculated using the formula below:

$$\kappa = \frac{Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

The formula above is used to calculate the difference in value between the level of agreement and the agreement by chance. This level of agreement for the Kappa Value (κ) are demonstrated in Table IV.

TABLE IV The Level of Agreement of the Kappa Value

Kappa Value (κ)	Level of agreement
< 0.0	Poor
0.01 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.00	Almost Perfect

ANALYSIS AND DISCUSSION

Analysis of Findings

The data was calculated using the Fleiss Kappa Statistics as it involved more than two raters to determine the extent of agreement among the expert panel. Previously, it has been demonstrated that the Fleiss Kappa Statistic involves 22 items, 10 items for the TALiP framework (N=10) and 12 for the SAMR model (N=12). Both the SAMR model item and the TALiP Framework item involve six experts (n=6), and five rating values (k=5).

The calculation of the probability of agreeing Pr(a) and the chance of agreement Pr(e) for the TALiP Framework's interview question is at the value of 0.8467 and 0.1930 respectively. Both values are crucial to calculating the Kappa value, which measures the level of agreement between the raters. Based on the calculation made below, the kappa value observed is at 0.8100, therefore indicating that there is a strong agreement between the raters. The calculation is as follows.

$$\kappa = \frac{0.8467 - 0.1930}{1 - 0.1930} = 0.8100$$

Another Fleiss Kappa analysis was also done with the SAMR model interview protocol. This interview question involves 12 items (N=12). The pr(e) and pr(a) values are at 0.8667 and 0.2159, respectively. Based on this the kappa value can be calculated.

$$\kappa = \frac{0.8667 - 0.2159}{1 - 0.2159} = 0.8300$$

Based on the calculation above, the level of agreement between experts is at 0.8300. According to the Landis & Koch scale (1977), when the observed kappa value is at 0.81 - 1.00, the agreement range is at an "almost perfect" level. This value is at a satisfactory level and this fulfils the benchmark that has been set by Brennan and Prediger which is at 0.70 (Brennan and Prediger, 1981). However, for the item to be used in applied decision-making, the Kappa value should be in the range of 0.80-0.90 or higher (Ghazali et al., 2011). As the judgment of the expert is at a satisfactory level, the Delphi method underwent one process as the consensus was achieved between the experts. However, adjustments were made to the TALiP Framework and the SAMR model interview protocol as suggested by the experts.

Suggestions of Change for Qualitative Instrument

TALiP interview Question

Based on the results for the TALiP interview question, there have been suggestions of improvement made by the experts to further refine the quality of the interview protocol. No items were removed. However, there has been an addition of five questions as suggested by the experts. The alteration of the interview protocol is further explained below. Therefore, the number of questions in the interview protocol has been increased from 10 items to 15 items.

One of the suggestions for change for the interview questions can be seen for item 1 “Please explain what knowledge and skills teachers as classroom assessors need to success to successfully conduct technology-enhanced/digital assessment?”. The suggestion that was made by the expert was for the question to be divided into two questions to avoid confusion for the interviewee. Therefore, the question has been divided into two parts “Please explain what knowledge teachers must master as a classroom assessor to successfully conduct technology-enhanced/digital assessment?” and “Please explain what skills teachers must master as a classroom assessor to successfully conduct technology-enhanced/digital assessment?”

For item 2 “Do you think having knowledge and skills in digital assessment is essential for teachers in the 21st century? How? Why?” whereby the experts suggested that the questions be broken down so that the interviewees will not only answer one question. The alteration to it was made to “Do you think having knowledge and skills in digital assessment is essential for teachers in the 21st century? Why do you think so? And how can the teacher showcase it?”. When the question is constructed in such a way, the interviewer can understand the reason for the interviewee’s view then proceed to look into how digital assessment was implemented.

For item 5 “which challenges do you face when you use technology in assessment (e.g cultural, persona, contextual, administrative, ideological and the like) Please explain what knowledge and skills you need to deal with them”, expert 2 suggested the breakdown of the question that looks into the challenges and what are knowledge and skills needed to deal with the challenges. Meanwhile, expert 4 suggested that the items be broken down into questions which focus on the challenges for institutional matters and socio-cultural aspects as clumping the elements into one question will cause the questions to be too vast. By doing so it will address the elements of teacher conception of assessment and institutional and socio-cultural elements in the TALiP Framework. Therefore the question was broken down into “What challenges do you face when you use technology in assessment?” followed by “Do you face challenges in the aspect of socio-cultural? (e.g cultural, persona, contextual, admin, ideological and the like)?”)

For item 8, “How do you attempt to improve your knowledge and skills of using technology in language assessment? How do you keep your knowledge and skills updated”, expert 2 suggested that the question be separated into two. The new questions are “How do you attempt to improve your knowledge and skills in using technology in language assessment?” and “How do you keep your knowledge and skills updated”. A visual representation of the revision made is explained in the Table V below.

TABLE V TALiP Framework Item original and revised items as suggested by experts.

Original Item (Before)	Expert Suggestion Summary	Revised Item(s) (After)	
1. Please explain what knowledge and skills teachers as classroom assessors need to success to successfully conduct technology-enhanced/digital assessment?	Divide the question into two separate items to distinguish between "knowledge" and "skills" and avoid confusion.	Please explain what knowledge teachers must master as a classroom assessor to successfully conduct technology-enhanced/digital assessment?	Please explain what skills teachers must master as a classroom assessor to successfully conduct technology-enhanced/digital assessment?
2. Do you think having knowledge and skills in digital assessment is essential for	Break down the question to first elicit the interviewee's view	Do you think having knowledge and skills in digital assessment is	Why do you think so? And how can the teacher showcase it?

teachers in the 21st century? How? Why?	("Why?") and then ask for practical examples ("How?").	essential for teachers in the 21st century?	
5. which challenges do you face when you use technology in assessment (e.g cultural, persona, contextual, administrative, ideological and the like) Please explain what knowledge and skills you need to deal with them	Break the question down to address challenges generally, then focus specifically on institutional and socio-cultural aspects to align with the TALiP framework.	What challenges do you face when you use technology in assessment?	Do you face challenges in the aspect of socio-cultural? (e.g cultural, persona, contextual, admin, ideological and the like)?
8. How do you attempt to improve your knowledge and skills of using technology in language assessment? How do you keep your knowledge and skills updated	Separate the question into two distinct parts to address (1) improvement efforts and (2) staying updated.	How do you attempt to improve your knowledge and skills in using technology in language assessment?	How do you keep your knowledge and skills updated?

SAMR interview Question

Based on the results of the SAMR interview question, there has been a suggestion of improvement made by the experts to further refine the quality of the interview protocol. No items were removed, and there were no additional questions added in the interview protocol. However, there are some changes made in the phrasing and the order of the question. The alteration of the interview protocol is further explained below. Therefore, the number of questions in the interview protocol has been maintained at 12 questions.

For item 2, “When thinking of the SAMR model, describe the level you feel most comfortable integrating technology”, expert 2 suggested that the word “thinking” be replaced with the word “considering”. Meanwhile, expert 5 suggested that the word “comfortable” be replaced with “confident”. This is because expert 5 mentioned that the word comfortable is not the most suitable word to be used in the interview protocol.

For item 6. “When integrating technology, explain how the educational environment changes when technology is used in an environment where technology was not previously integrated.”, expert 2 suggested for the question to be paraphrased. Concurrently, expert 4 suggested that the question be rephrased and that the questions be divided into two. This is because the question becomes confusing for the participant as there are many noun clauses. By breaking the question into two questions, it will be easier for the participants to digest the question. Consequently, the question was broken into, “How does the educational environment change when technology is now integrated in an environment that was not previously integrated?” and “What do you do differently in digital assessment as compared to traditional assessment”.

For item 7, “What is your overall opinion of the SAMR Model in digital assessment?”, there were no significant changes that were made. However, for item 7, expert 2 stated that this question should be asked earlier in the questionnaire. Therefore, this question was moved from item 7 and became item 2.

For item 8, “ When thinking of the bottom half of the model (enhancement) and the top half of the model (transformation), what are your thoughts on the learning process as you move from the bottom to the top?”, the question underwent paraphrasing. expert 2 suggested that the question be altered to “What are your thoughts on the assessment process as you move from the bottom (enhancement) to the top (transformation)?”. Expert 4 added that the participants should be given the model prior to the interview and be given a briefing on the model. A visual representation of the revision made is explained in the Table VI below.

TABLE VI SAMR Model Item original and revised items as suggested by experts.

<table>
<tr><th>Original Item (Before)</th><th>Expert Suggestion Summary</th><th colspan="2">Revised Item(s) (After)</th></tr>
<tr><td>2. When thinking of the SAMR model, describe the level you feel most comfortable integrating technology</td><td>Replace "thinking" with "considering" and "comfortable" with the more suitable term "confident".</td><td colspan="2">When considering the SAMR model, describe the level you feel most confident integrating technology.</td></tr>
<tr><td>6. When integrating technology, explain how the educational environment changes when technology is used in an environment where technology was not previously integrated.</td><td>Paraphrase and split the question to reduce confusion caused by multiple noun clauses.</td><td>How does the educational environment change when technology is now integrated in an environment that was not previously integrated?</td><td>What do you do differently in digital assessment as compared to traditional assessment?</td></tr>
<tr><td>7. What is your overall opinion of the SAMR Model in digital assessment?</td><td>Move this question earlier in the sequence to function as a general introductory item.</td><td colspan="2">What is your overall opinion of the SAMR Model in digital assessment? (moved to item 2)</td></tr>
<tr><td>8. When thinking of the bottom half of the model (enhancement) and the top half of the model (transformation), what are your thoughts on the learning process as you move from the bottom to the top?</td><td>Rephrase to focus specifically on the "assessment process" rather than the broader "learning process" for better alignment with the study's goals.</td><td colspan="2">What are your thoughts on the assessment process as you move from the bottom (enhancement) to the top (transformation)?</td></tr>
</table>

DISCUSSION

This study set out to validate two qualitative interview instruments. They are based on the TALiP framework and the SAMR model. It aims to explore digital assessment literacy (DAL) among ESL instructors in Malaysian higher education. By involving six subject matter experts and applying the Delphi method, the research achieved consensus on the clarity, relevance, and alignment of the instrument items with the constructs under investigation. The resulting Fleiss' Kappa value is 0.81 for the TALiP-based instrument and 0.83 for the SAMR-based instrument, respectively. This indicates an “almost perfect” level of agreement among raters Landis & Koch (1977), validating both the structure and conceptual coherence of the interview protocols. Several amendments were made to the interview protocol, including splitting the double-barreled questions, clarifying the terminology, rearrangement of questions, and adding context-specific prompts. The revisions made are crucial as they ensure that the protocol will elicit examples of practice rather than just the instructor's beliefs about digital assessment.

The high level of agreement achieved through expert validation underscores the robustness of the TALiP framework in capturing multifaceted dimensions of assessment literacy, including institutional, socio-cultural, and teacher learning domains. This supports the assertions by Xu and Brown (2016) that teacher assessment literacy is not merely technical but contextually embedded and dynamically constructed through practice. In particular, the strong agreement on indicators related to identity construction and assessment literacy in practice suggests that ESL instructors are increasingly aware of their evolving roles in digitally mediated learning environments.

Meanwhile, the SAMR-based interview instrument effectively captured expert consensus on the spectrum of technology integration in assessment design, ranging from substitution to redefinition. The item-level agreement suggests that the SAMR model remains a relevant and intuitive scaffold for evaluating the depth of digital technology integration among educators which is in the range of 0.80-0.90 or higher (Ghazali et al., 2011). Notably, expert feedback led to improvements in item phrasing and sequence, indicating that even widely accepted models benefit from localized adaptation and linguistic clarity when applied in research contexts.

Several refinements were made to the instruments based on expert suggestions. For example, multiple double-barreled questions were split into two more precise items. This not only improves interview clarity but also aligns with best practices in qualitative instrument construction (Menold & Raykov, 2021) Additionally, breaking down complex items into narrower socio-cultural or institutional themes reflects a more nuanced application of the TALiP framework and prevents oversimplification of layered constructs.

While the Kappa values were statistically strong, it is worth noting that the iterative refinement process also revealed tensions within expert interpretations of certain constructs. This is in particular with how socio-cultural challenges were framed. This variation aligns with findings by Motala and Menon (2020), who observed that institutional readiness for digital transformation varies significantly across higher education contexts, even within the same national system. Such findings emphasize the need for future studies to consider institutional diversity when deploying assessment instruments across universities.

From a practical standpoint, the validated interview protocols offer scalable and reliable tools for understanding ESL instructors' digital assessment practices in tertiary education. These tools can be used by researchers, program designers, and university administrators to inform targeted professional development initiatives. Furthermore, the alignment of expert feedback with theoretical models such as TALiP and SAMR indicates a promising avenue for integrating reflective assessment literacy practices with emerging technologies in the ESL domain.

Ultimately, the findings affirm the value of structured expert validation—particularly when developing instruments for rapidly evolving domains such as digital assessment. By anchoring instrument design in well-established theoretical frameworks and rigorously verifying content validity, this study contributes to the growing body of literature supporting the contextualization and localization of digital assessment research within ESL education.

CONCLUSION

The research was carried out to confirm a qualitative interview schedule aimed at investigating the digital assessment literacy (DAL) of ESL educators in higher education in Malaysia. The expert validation of the content validity of the two tools that were being studied (which are again based on the TALiP framework and the SAMR model) proved to be rigorously carried out during the research. The results affirm the statements that only one entry of the Delphi method was needed to get agreement from the six expert validators. The existence of a very large inter-rater reliability is signified by a Fleiss Kappa of 0.81 in the TALiP instrument and a Fleiss Kappa of 0.83 in the SAMR instrument yielding an almost perfect consistency. The end result of the recommendations of the experts was the improvement of the final protocols through the division of the double-barreled questions, definition of terms, and the introduction of five more indicators to the TALiP instrument to make it clearer and in-depth.

The main outcome of the given study lies in the creation of a sound and stable instrument that even researchers, university administrators, and curriculum developers should be able to utilize and learn about the digital assessment work of ESL teachers and gauge its effectiveness. The validated tools present a viable way to communicate specific competence-enhancement initiatives targeting increases in the competence of digital assessments. Moreover, the research supports the importance of designing instruments of assessment with high levels of theoretical justification, to guarantee the conceptual integrity of the digital assessment field that is facing a rapid change. While this study successfully validated the interview protocol, several avenues for future research are apparent.

The immediate next step is to deploy the validated protocol with a larger and more diverse sample of ESL instructors across Malaysia to develop a comprehensive understanding of their digital assessment literacy practices. Furthermore, future studies could conduct a comparative analysis of DAL between different types of higher education institutions (e.g., public vs. private, research-intensive vs. teaching-focused universities) to identify contextual factors that influence digital assessment practices.

Another valuable line of inquiry would be to investigate the impact of this protocol on professional development; researchers could explore how insights gained from using this instrument can directly inform the design and evaluate the effectiveness of targeted training programs for ESL instructors. These future studies would not only address the specific challenges and training needs of educators but also enhance the quality of digital learning environments in Malaysia and beyond.

FUNDING INFORMATION

This research was supported by University Malaysia Sabah (UMS) through the UMSGreat Research Grant Scheme (GUG0663-1/2024).

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

INFORMED CONSENT

All the participants involved in this study were briefed and provided their consent prior to the participation of this study. They were fully informed of the purpose of the study, the procedures involved, their right to withdraw from the study if they wish to without being held any consequences, and the measures taken to ensure that their confidentiality and anonymity remained intact.

DATA AVAILABILITY

The data supporting this study’s findings are available upon reasonable request from the corresponding author. Due to confidentiality considerations, some data may not be publicly accessible.

REFERENCE LIST

1. Adegbenro, J. B., Gumbo, P. M. T., & Olakanmi, E. E., (2017). In-Service Secondary School Teachers’ Technology Integration Needs in an ICT- Enhanced Classroom. *TOJET: The Turkish Online Journal of Educational Technology*, 16(3)
2. Arouri, Y. M., Alshaboul, Y. M., Hamaidi, D. A., & Alshaboul, A. Y. (2025). Higher education instructors’ and students’ attitudes toward distance learning. *International Journal of Evaluation and Research in Education (IJERE)*, *14*(3), 1949. https://doi.org/10.11591/ijere.v14i3.29383
3. Awang, M.I., (2021). The Digitalisation of Learning Assessment. *The Proceeding Books of the 4th International Conference of Multidisciplinary Research 2021*, 4 (1),
4. Bane, J. A. (2019). Academic integrity in the online classroom. *eLearn*, *2019*(8). https://doi.org/10.1145/3343412.3343233
5. Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. https://doi.org/10.1177/001316448104100307
6. Butler-Henderson, K. & Crawford, J. (2020). A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. Computers & Education, 159, 104024. https://doi.org/10.1016/j.compedu.2020.104024
7. Chuenjitwongsa, S. (2017). How to: Conduct a Delphi Study. Cardiff University https://www.cardiff.ac.uk/__data/ assets/pdf_file/ 0010/1164961/how_to_conduct_a_delphistudy.pdf

8. Dalkey N.C. The Delphi method: an experimental study of group opinion. In: Dalkey NC, Rourke DL, Lewis R, Snyder D, editors. Studies in the quality of life: Delphi and decision-making. Lexington, MA: Lexington Books; 1972. P. 13–54.
9. Dalkey, N.C., & Helmer, O. (1963). An Experimental Application of the Delphi Method to the Use of Experts. *Management Science*, *9*(3), 458–467. http://www.jstor.org/stable/2627117
10. Eaton, E. (2020). Academic integrity during COVID-19: Re□ections from the University of Calgary. ISEA, 48(1), 80–85
11. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https:// doi.org/10.1037/h0031619
12. Ghazali, A.M., Johare, R. and Masrek, M.N. (2011). The Study Of Records Management Competencies By Applying Kappa Coefficient In Coding Process For Inter-Coder Reliability. *Business & Management Quarterly Review, 2(4),* 33-4
13. Jang, S.J., & Chen, K.C. (2010). From PCK to TPACK: Developing a Transformative Model for Preservice Science Teachers. *Journal of Science Education and Technology, 19,* 553-564.
14. Judi, H. M. (2020). Integrity and Security of Digital Assessment: Experiences in Online Learning. *Global Business and Management Research: An International Journal*, *14*(1), 97–107.
15. Keengwe, J., & Georgina, D. (2011). The digital course training workshop for online learning and teaching. *Education and Information Technologies*, *17*(4), 365–379. https://doi.org/10.1007/s10639-011-9164-x
16. Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310
17. Loureiro, P., & Gomes, M. J. (2023). Online peer assessment for learning: Findings from higher education students. Education Sciences, 13(3), 253. https://doi.org/10.3390/educsci13030253
18. Malaysian Qualification Agency (2020).Guide on Compliance Evaluation for Self-Accreditation Universities. Management and Handbook]. Putrajaya: Ministry of Education Malaysia.
19. Malaysian Qualification Agency (2024). Malaysian Qualification Framework (2nd). Putrajaya: Ministry of Education Malaysia.
20. Malhotra, N. K., Vriens, R., & Grover, R. (2006). Questionnaire Design and Scale Development. In *The Handbook of Marketing Research* (pp. 176–202). essay, Sage Publication .
21. Malone, M. E. (2013). The Essentials of Assessment Literacy: Contrasts between testers and users. *Language Testing*, *30*(3), 329–344.https://doi.org/10.1177/0265532213480129
22. Mat Hussin, A., Idris, I.S., Misnan, N.I., (2020). How Does It Challenge In Higher Education? A Case Study. *International Journal of Technical Vocational and Engineering Technology iJTvET],* 2(2), 38-49.
23. McHugh, M. L. (2012). Interrater Reliability: The kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/ 10.11613/ bm.2012.031
24. Menold, N., & Raykov, T. (2021). On the relationship between item stem formulation and criterion validity of multiple-component measuring instruments. *Educational and Psychological Measurement*, *82*(2), 356–375. https://doi.org/ 10.1177/00131644 20988169
25. Motala, S., & Menon, K. (2020). In Search of the ‘New Normal’: Reflections on Teaching and Learning during COVID-19 in a South African University. Southern African Review of Education with Education with Production, 26, 80-99.
26. Podsiad, M., & Havard, B. (2020). Faculty acceptance of the Peer Assessment Collaboration Evaluation Tool: A Quantitative Study. *Educational Technology Research and Development*, *68*(3), 1381–1407. https://doi.org/10.1007/s11423-020-09742-z
27. 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, 489–497. https://doi.org/10.1002/nur.20147
28. Puendetura, R. (2013, June 29). *SAMR: Moving from enhancement to transformation - hippasus*. SAMR Enhancement To Transformation. http://www.hippasus.com/rrpweblog/archives/2013/05/29/SAMREnhancementToTransformation.pdf

29. Raman, A. & Rathakhrishnan, M. (2018). FROG VLE: Teachers' Technology Using Utaut Model. International Journal of Mechanical Engineering and Technology (IJMET), 9(3) pp. 529–538
30. Santhi, D., Suherdi, D., & Musthafa, B. (2019). ICT and Project-Based Learning in a rural school: An EFL context. Proceedings of the Third International Conference on Sustainable Innovation 2019 – Humanity, Education and Social Sciences (IcoSIHESS2019). https://doi.org/10.2991/icosihess-19.2019.5
31. Sarker, M.N.I., Min, W., Qian, C., Alam. G.M.M, & Dan, L., (2019). Leveraging Digital Technology for Better Learning and Education: A Systematic Literature Review. International Journal of Information and Education Technology, 9(7), 453-461
32. Shraim, K. (2019). Online examination practices in Higher Education Institutions: Learners' Perspectives. *Turkish Online Journal of Distance Education*, *20*(4), 185–196. https://doi.org/10.17718/tojde.640588
33. Tsang, K. K. (2012). The use of midpoint on Likert Scale: The implication for educational research. *Hong Kong Teachers Centre Journal*, *11*, 121–130.
34. Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A Reconceptualization. *Teaching and Teacher Education*, *58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010
35. Yusoff, M. S. (2019). ABC of content validation and content validity index calculation. Education in Medicine Journal, 11(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6