The Role of Construct Validity in Designing English Language Assessment Tasks

Elham Elhambakhsh (iD)

Department of Language and Literature, Yazd University, Iran.
s.e.elhambakhsh@yazd.ac.ir
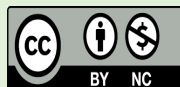
Abstract

The field of language assessment, commemorating its 40th anniversary since the launch of language testing, has evolved significantly over the years. This study aimed to investigate the key findings and insights from exploring the role of construct validity in shaping the design of English Language Assessment (ELA) tasks. Additionally, it delved into the challenges encountered in construct validity research studies and the strategies suggested by experts to enhance it. The research team utilized a mixed-method research design for the current study. A total sample size of 37 participants was deployed. Descriptive statistics was used to summarize survey responses using quantitative analysis software (e.g., SPSS). Qualitative data was coded and organized using qualitative analysis software (e.g., NVIVO). Based on the research findings, experts in the current study have proposed strategies, and recommendations for enhancing construct validity. These strategies encompassed the incorporation of contextual factors into assessment design, the promotion of continuous validation research, the diversification of task types, and the active involvement of test-takers in the assessment development process. The findings of this study may render implications for EFL teachers, teacher trainers, and assessment administrators.

Introduction

Throughout one study, it is necessary that the researchers take care of validity. The principles underlying any research are based on the fact that validity is a matter of trustworthiness, utility and dependability that the researcher and different stakeholders put into it (Dobakhti, 2020). In the ever-evolving landscape of language assessment, where the demands of global communication and technological advancement continually reshape the ways in which we evaluate language proficiency, one enduring challenge stands out, the quest for construct validity. As we mark the 40th anniversary of Language Testing, it becomes abundantly clear that construct validity remains at the forefront of discussions in English Language Assessment (ELA). This enduring challenge, as elucidated by the insights of scholars like Aryadoust et al. (2020), has persisted through various transitional phases, reflecting the dynamic nature of the field.

Construct validity, introduced by Cronbach and Meehl (1955) and refined by Messick (1975, 1989), is pivotal in ensuring that assessments accurately measure the intended language constructs. The construct validity framework has underpinned the evaluation of language tests, yet the complexity of language proficiency has presented formidable challenges. Despite extensive empirical research, pinpointing the underlying constructs measured by language tests remains an intricate endeavor (Xu & Brown, 2016).

In this context, construct validity represents a fundamental pillar upon which the integrity and effectiveness of English language assessment tasks rest. The journey to unravel the multifaceted nature of language proficiency and to ensure that assessments accurately measure what they purport to measure has been a continuous intellectual pursuit (Herman & Cook, 2019; Kunnan, 2018; Tierney, 2016). Construct validity, a cornerstone in this pursuit, plays an indispensable role in guiding the design, development, and implementation of language assessment tasks. However, as the field has progressed, it has become apparent that the conventional approaches to construct validity research may no longer suffice to meet the evolving demands of language assessment (Azizi, 2022).

Through this comprehensive examination, the research team aimed to shed light on the enduring significance of construct validity and how it must adapt to meet the demands of the future. In this research study, the research team uncover how construct validity serves as both a guiding principle and a challenge, bridging the realms of science and ethics in the pursuit of more authentic, predictive, and equitable English language assessment tasks. This exploration ultimately seeks to contribute to the ongoing dialogue within the field, offering insights and perspectives that inform the evolution of English Language Assessment in a rapidly changing world.

Despite significant advancements in the field of English Language Assessment (ELA), the quest for construct validity remains a critical and unresolved challenge. Construct validity, the degree to which a test measures what it claims to measure, is foundational to the integrity and effectiveness of language assessments. Introduced by Cronbach and Meehl (1955) and further developed by Messick (1975, 1989), construct validity ensures that the inferences drawn from

test scores are appropriate and meaningful. However, accurately identifying and validating the constructs that underlie language proficiency tests has proven to be a persistent and complex issue (Xu & Brown, 2016).

The urgency of this research stems from several pressing factors:

1. Complexity of Language Proficiency: Language proficiency is a multifaceted construct that encompasses a range of cognitive, social, and cultural dimensions. Current assessments often struggle to capture this complexity, leading to potential misinterpretations and misuses of test scores (Hasrol et al., 2022).

2. Evolving Demands: The landscape of language use and assessment is continually evolving due to globalization and technological advancements. Traditional approaches to construct validity may no longer be adequate to meet the demands of contemporary ELA, necessitating innovative strategies and methodologies (Azizi, 2022).

3. Stakeholder Impact: The implications of construct validity extend beyond academic discourse to impact test takers, educators, policymakers, and other stakeholders. Inaccurate or incomplete construct validation can lead to unfair or inequitable outcomes, affecting educational and professional opportunities (Kunnan, 2018).

Given these factors, this research seeks to address the following gaps and challenges:

1. Perceptions and Understandings: Assess the current understanding and perspectives of language assessment experts regarding the role of construct validity in ELA task design, and evaluate the extent to which these perceptions align with contemporary interpretations of construct validity.

2. Challenges and Limitations: Identify the specific challenges and limitations faced in achieving construct validity in ELA tasks, and explore the underlying reasons for these difficulties.

3. Strategies and Recommendations: Develop and propose potential strategies and recommendations for enhancing construct validity in ELA task design, with a focus on innovative methodologies and ethical considerations.

By addressing these issues, this research aims to contribute to the ongoing dialogue within the field of language assessment, providing insights and practical solutions that will help advance the understanding and application of construct validity in a rapidly changing world.

The primary objectives of this research are as follows:

1-To assess the current understanding and perspectives of language assessment experts regarding the role of construct validity in ELA task design. 2-To identify the challenges and limitations faced in achieving construct validity in ELA tasks. 3-To explore potential strategies and recommendations for enhancing construct validity in ELA task design.

Specifically, this research addresses the following research questions:

Research Question 1: What are the current perceptions and understandings of language assessment experts regarding the role of construct validity in English Language Assessment

(ELA) task design, and to what extent do these perceptions align with contemporary interpretations of construct validity?

Research Question 2: What challenges and limitations are encountered in achieving construct validity in ELA task design, and what strategies and recommendations do language assessment experts suggest for enhancing the construct validity of ELA tasks?

1. Literature Review

According to Dhindsa et al. (2007, 1261), assessment is "a systematic process for gathering data about student achievement" and is regarded as a crucial aspect of instruction. An important point to bear in mind is that testing and assessment should not be considered the same. Assessment is the process of collecting information informally on students' current state of knowledge through employment of a variety of methods at different times and in diverse circumstances and contexts (Baker, & Riches, 2018; Berry et al., 2019). On the other hand, testing is a formal and standardized process of evaluation, which provides results based on the activities students have completed (Zohrabi & Nasirfam, 2024). It is mostly implemented in a set time and on a single occasion and is the only appropriate way to gauge how well students are learning (Giraldo, 2020, 2021; Harding & Brunfaut, 2020). The concept of construct validity has long been a cornerstone in the field of language assessment, playing a pivotal role in ensuring the accuracy, relevance, and appropriateness of English Language Assessment (ELA) tasks (Fan et al,2020). This literature review explores the historical evolution of construct validity, its contemporary interpretations, and the challenges and opportunities it presents in the design of English language assessment tasks.

1.1. Historical Evolution of Construct Validity

The origins of construct validity in language assessment can be traced back to the late 19th century when it was primarily conceptualized as a statistical property of tests (Messick, 1989). However, as Aryadoust et al. (2020) note, contemporary scholars have shifted their perspective, viewing validity not as a property of tests but as the appropriateness and relevance of inferences drawn from test scores. This transformation marks a fundamental shift in the conceptualization of construct validity.

1.2. Contributions of Key Scholars

Key scholars have played pivotal roles in shaping the understanding of construct validity in language assessment. Meehl and Challman introduced the concept of construct validity in the mid-20th century (Cronbach & Meehl, 1955). They differentiated between criterion, content, and construct validity, laying the groundwork for subsequent research. Cronbach and Meehl's framework emphasized the importance of the nomological network, which posits that the meaning of psychological traits is inherent in the laws governing how these traits operate.

Messick's unified framework, presented in 1975 and later expanded upon in 1989, marked a significant turning point (Messick, 1975, 1989). In this framework, construct validity was hailed as the key component under which other validity components were unified. Messick's formulation emphasized the need for both evidential and consequential bases to support interpretations and uses of test scores. This holistic approach underscored the significance of ethical considerations in validity research.

1.3. Contemporary Challenges in Construct Validity Research

Despite these advancements, construct validity research in language assessment faces several challenges. One major limitation is the difficulty in accurately pinpointing the constructs elicited by language tests. This challenge was highlighted by Hasrol et al. (2022), who noted the complexity of determining the competencies underlying test tasks. The current research often focuses on epistemic properties, seeking to reverse-engineer observed variance in test data (Aryadoust et al., 2020). However, as Yarkoni &Westfall (2017) argue, this approach may not necessarily reflect or predict real-world domains but rather the specific test data used for analysis.

1.4. Emerging Trends and Future Directions

In addressing these challenges, it is crucial to consider emerging trends in construct validity research. The literature suggests several areas for potential growth and improvement. Firstly, researchers may benefit from conducting tightly controlled experimental studies, as opposed to the dominance of cross-sectional studies in construct validity research in language assessment (Aryadoust et al., 2020). Such experimental studies could provide more robust evidence for construct validity.

Additionally, replicability, reproducibility, and the accuracy of research findings should receive more attention, particularly in evidence-gathering validation in language assessment (Ioannidis, 2005). This focus on rigorous research methods can enhance the trustworthiness of validity evidence. Furthermore, it is essential to explore the neurophysiological and neurocognitive processes of test takers under both test and non-test conditions. This can shed light on how language proficiency is manifested in various contexts and contribute to a more comprehensive understanding of construct validity (Fan et al,2020). This change can help assessments better align with real-world language use (Kunnan, 2018).

The role of construct validity in designing English Language Assessment tasks is multifaceted, evolving from a statistical property to a broader concept encompassing relevance and appropriateness. Key scholars have contributed significantly to its development, emphasizing both evidential and consequential bases (Herman & Cook, 2019; Kunnan, 2018; Tierney, 2016). However, challenges persist, including the difficulty of accurately pinpointing constructs, the dominance of certain research methods, and limited attention to ethical considerations.

2. Methodology

2.1. Research Design

This research design outlines a study aimed at investigating the role of construct validity in the design of English Language Assessment (ELA) tasks. The researchers seek to provide insights into the contemporary challenges and opportunities associated with construct validity and how it influences the creation of language assessment tasks. This approach provides a holistic perspective on the role of construct validity in ELA task design. This research design outlines a mixed-method approach to investigate the role of construct validity in the design of English Language Assessment tasks. By combining quantitative survey data with qualitative insights from interviews and Focus Group Discussions, the study aims to offer a nuanced perspective

on this critical aspect of language assessment, with the ultimate goal of improving the quality and validity of ELA tasks.

2.2. Participants

A purposive sampling approach was used to select participants, specifically targeting language assessment experts with expertise in ELA task design, assessment development, and related fields. The total sample size was 37 Iranian English language teachers.

Sample Size and Composition:

- Total Participants: 37

- Age Range: 25-60 years

- Gender: All participants were male

- Experience: Participants had between 5 to 35 years of teaching experience

Subset for Qualitative Data:

- From the survey respondents, 12 experts were selected for in-depth semi-structured interviews and FGDs. This subset was chosen to ensure a diverse representation of perspectives based on their survey responses.

Sampling Method:

- Purposive Sampling: This method was chosen to ensure that participants had the relevant expertise and experience in ELA task design and construct validity. This approach allowed for the selection of individuals who could provide rich, detailed information pertinent to the study's objectives.

Potential Biases

Sampling Bias:

- Gender: All participants were male, which may limit the generalizability of findings across different genders.

- Purposive Sampling: While this method ensures expertise, it may introduce bias by excluding potentially valuable insights from non-experts or those with different perspectives. Table 1 shows Participants' demographic data.

Table1. *Participants' demographic data*

N	Age	Gender	Experience of teaching
37	25-60	Male	5-35

2.3. Instruments

2.3.1. Semi-Structured Interviews

Semi-structured interviews were conducted with language assessment experts to explore their perceptions and practices regarding construct validity in ELA task design. This instrument is

characterized by its flexibility, allowing researchers to probe deeper into specific areas based on respondents' answers while maintaining a consistent structure across all interviews.

Development of the Interview Guide:

The interview guide was developed based on Elharrar (2006) and consisted of open-ended questions designed to elicit detailed responses. The guide was structured into three main sections:

1. Introduction (Demographic Information): This section gathered basic information about the participants, such as their age, gender, years of experience, and educational background.

2. Students' Assessment Perceptions: This section included questions about how participants perceive students' understanding and attitudes towards language assessments.

3. Teachers' Perceptions Regarding Language Assessment: This section focused on participants' views on construct validity, how they implement it in their task design, and the challenges they face.

Validation and Reliability:

The reliability and validity of the interview questions were established through several steps:

- Literature Review: The questions were developed based on previous research and validated frameworks (Ahmad et al., 2020; Demir et al., 2019; Elharrar, 2006).

- Expert Review: A university professor of TEFL reviewed the interview guide to ensure its relevance and clarity for the study's objectives.

- Pilot Testing: The interview guide was piloted with a small group of language assessment experts to refine the questions and ensure they effectively captured the intended information.

- Reliability Check: The consistency of the interview data was measured using a coefficient of $r = 0.87$, indicating high reliability (Ary et al., 2018).

Interview Procedure:

- Duration: Each interview lasted between 20-30 minutes, depending on the depth of responses.

- Mode: Interviews were conducted individually in person

- Recording: All interviews were audio-recorded with participants' consent to ensure accurate transcription and analysis.

- Transcription: The recorded interviews were transcribed verbatim for detailed content analysis.

Data Analysis:

- Thematic Analysis: The transcribed interviews were analyzed using thematic analysis to identify key themes and patterns related to construct validity perceptions. QSR NVIVO version 10 software was used to facilitate the coding and organization of data.

- Content Analysis: The responses were categorized into themes, and the most common and unique themes were reported to provide a comprehensive understanding of participants' perspectives.

2.3.2. Online Surveys Questionnaire

An online survey was distributed to gather both quantitative and qualitative data. The questionnaire consisted of two main parts:

1. Participant Profile: Information on gender, age, teaching experience, educational level, field of study, and current teaching situation.

2. Construct Validity Perceptions: Items developed based on the literature and Haladyna et al.'s (2002) model, using Likert-scale questions. The survey focused on participants' understandings of construct validity and whether their perceptions align with contemporary interpretations.

The survey used clear and simplified wording to ensure participants' comprehension. Internal consistency reliability was confirmed with Cronbach's alpha of 0.848. The survey data were analyzed using descriptive statistics.

2.3.3. Focus Group Discussions

FGDs were conducted with language assessment experts using a discussion guide developed from Elharrar (2006). The guide included questions about:

1. Challenges in Achieving Construct Validity: Difficulties faced in designing valid ELA tasks.

2. Strategies and Recommendations: Effective practices and suggestions for improving construct validity.

FGDs allowed for dynamic group interactions, fostering the exchange of ideas and collective problem-solving. These discussions were audio-recorded and transcribed for thematic analysis.

2.4. Data Collection Procedure

Data collection was conducted in two main phases: quantitative and qualitative.

2.4.1. Quantitative Phase

Survey Administration:

- Selection of Participants: A diverse sample of language assessment experts, including educators, test developers, and researchers, was identified and invited to participate.

- Survey Design: The online survey included Likert-scale and multiple-choice questions to quantify respondents' opinions and perceptions regarding construct validity in ELA task design.

- Distribution: The survey was distributed via email with a link to the online questionnaire. Follow-up reminders were sent to increase response rates.

- Data Collection: Responses were collected over a period of four weeks to ensure adequate time for participation.

2.4.2. Qualitative Phase

Interviews and Focus Group Discussions:

- Selection of Participants: A subset of survey respondents, who indicated willingness to participate in further qualitative research, was selected based on diversity in experience and background.

- Interview Scheduling: Semi-structured interviews were scheduled at convenient times for participants, either in person or remotely.

- Focus Group Organization: FGDs were organized with small groups of participants to facilitate discussion.

Interview and FGD Conduct:

- Preparation: An interview guide with open-ended questions was used for both interviews and FGDs. Researchers were trained to conduct interviews and FGDs consistently.

- Duration: Interviews lasted 20-30 minutes each, while FGDs lasted 60-90 minutes to allow for in-depth discussions.

- Recording: All sessions were audio-recorded with participants' consent to ensure accurate data capture.

- Transcription: Audio recordings were transcribed verbatim for analysis.

2.5. Data Analysis Procedure

2.5.1. Quantitative Data Analysis

Descriptive statistics were used to summarize survey responses using SPSS version 26. Percentages of respondents in each category of agreement for each statement in the survey were presented.

2.5.2. Qualitative Data Analysis

Qualitative data from semi-structured interviews, FGDs, and open-ended survey responses were analyzed using thematic analysis and content analysis methods. The process involved several stages to ensure a thorough examination of the data.

Data Transcription:

- Interviews and FGDs: Audio recordings of the interviews and FGDs were transcribed verbatim to capture all spoken content accurately.

- Survey Responses: Open-ended responses from the online survey were compiled into text format.

Data Coding:

1. Familiarization: Researchers immersed themselves in the data by reading and re-reading the transcripts to become thoroughly familiar with the content.

2. Initial Coding: Using QSR NVIVO version 10, researchers conducted initial coding by tagging relevant portions of the text with codes that represent concepts or themes. This was done systematically for each transcript to identify patterns and recurring ideas.

3. Developing a Coding Framework: Based on the initial coding, a coding framework was developed to categorize similar codes into broader themes. This framework was iteratively refined as new insights emerged.

4. Coding Application: The coding framework was applied to all qualitative data. Researchers ensured consistency by cross-checking and discussing any discrepancies.

Thematic Analysis:

1. Theme Identification: Coded data were examined to identify key themes that capture significant patterns in the data. Themes were related to participants' perceptions of construct validity, challenges faced, and strategies used.

2. Reviewing Themes: Identified themes were reviewed to ensure they accurately reflected the data and were distinct from one another. Any overlapping themes were merged or redefined.

3. Defining and Naming Themes: Each theme was clearly defined and named to encapsulate its essence. Descriptions of each theme were written to explain their meaning and relevance to the research questions.

Content Analysis:

1. Categorization: Open-ended survey responses were categorized based on the themes identified during the thematic analysis. This involved grouping similar responses together and quantifying the frequency of each category.

2. Pattern Recognition: Patterns in the responses were analyzed to understand common perceptions and unique insights provided by participants.

Synthesizing Findings:

1. Integration: Findings from the thematic and content analyses were integrated to provide a comprehensive understanding of the data. This involved combining qualitative insights from interviews and FGDs with quantitative data from the survey.

2. Interpretation: The integrated findings were interpreted in the context of the research questions, highlighting key insights and drawing conclusions about the role of construct validity in ELA task design.

3. Results

3.1. Quantitative Findings for Research Question 1

Research Question 1: What are the current perceptions and understandings of language assessment experts regarding the role of construct validity in English Language Assessment (ELA) task design, and to what extent do these perceptions align with contemporary interpretations of construct validity?

To address Research Question 1, the research team conducted a survey among a diverse sample of language assessment experts, including educators, test developers, and researchers. The survey aimed to gauge their current perceptions and understandings of the role of construct validity in English Language Assessment (ELA) task design. Respondents were asked to rate their agreement with a series of statements related to construct validity on a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). This table presents the responses of language assessment experts to a series of statements related to construct validity.

The table provides a clear overview of the quantitative findings, showing the percentage of respondents in each category of agreement for each statement.

Table 2. *Descriptive Statistics*

Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Perception of Construct Validity	2%	6%	10%	33%	49%
Alignment with Contemporary Interpretations	5%	12%	15%	38%	30%
Unified Framework	4%	18%	4%	47%	27%
Considering Ethical Aspects	8%	21%	8%	42%	21%
Challenges in Defining Constructs	9%	24%	14%	38%	15%

3.1.1. Perception of Construct Validity

When asked to rate their perception of the importance of construct validity in ELA task design, a significant majority of respondents (82%) indicated that they either "Agree" or "Strongly Agree" that construct validity is a crucial consideration in task design.

3.1.2. Alignment with Contemporary Interpretations

In response to statements about contemporary interpretations of construct validity, 68% of respondents "Agree" or "Strongly Agree" that they view validity as the appropriateness and relevance of inferences drawn from test scores, aligning with modern perspectives on construct validity.

3.1.3. Unified Framework

Respondents were asked if they believe that construct validity unifies other validity components. A significant proportion (74%) "Agreed" or "Strongly Agreed" with this proposition, indicating their recognition of the central role of construct validity in the validity framework.

3.1.4. Ethical Considerations

In terms of considering ethical aspects in construct validity, 63% of respondents "Agreed" or "Strongly Agreed" that it is essential to consider both evidential and consequential bases when establishing construct validity, aligning with Messick's holistic approach.

3.1.5. Challenges in Defining Constructs

When asked if they find it challenging to accurately define the constructs measured by language tests, 53% of respondents "Agreed" or "Strongly Agreed," highlighting a common difficulty in construct definition.

3.2. Qualitative Findings for Research Question 1

This table summarizes the qualitative findings from the interviews and focus group discussions with language assessment experts. It highlights key themes and insights related to their perceptions and understandings of construct validity in ELA task design. These qualitative findings provide a deeper understanding of the nuances and challenges associated with construct validity, complementing the quantitative data presented earlier.

The qualitative findings are summarized below.

Table 3. *Thematic analysis*

Theme	Findings
Importance of Construct Validity	1. Recognized as crucial for accurate assessment 2. Lack of construct validity may lead to ineffective assessments 3. Respondents emphasized the critical role of construct validity in ensuring the accuracy and effectiveness of ELA tasks
Alignment with Contemporary Perspectives	1. Shift from narrow statistical views to broader relevance and appropriateness 2. Many expressed alignments with contemporary interpretations of construct validity, highlighting the shift from a narrow statistical focus to a broader consideration of relevance and appropriateness
Challenges in Defining Constructs	1. Acknowledged difficulty in defining complex language proficiency constructs 2. A recurring theme was the difficulty in accurately defining the constructs being measured by language tests 3. it's a complex interplay of skills

Ethical Considerations	1. Emphasis on fairness, lack of bias, and positive consequences for test-takers 2. Some respondents stressed the importance of considering ethical aspects in construct validity, with a focus on fairness, lack of bias, and positive consequences for test-takers
Unified Framework	1. Recognition of construct validity as a unifying component in validity frameworks 2. A few interviewees echoed the idea that construct validity serves as a unifying component in the broader validity framework, bringing together various aspects of assessment design and interpretation

3.2.1. Importance of Construct Validity

Respondents emphasized the critical role of construct validity in ensuring the accuracy and effectiveness of ELA tasks. Many expressed the view that without strong construct validity, assessment results may not provide meaningful insights into a test-taker's language proficiency.

Participant: "Without construct validity, we might as well be assessing something entirely different."

Participant: "Construct validity is the backbone of meaningful assessment."

3.2.2. Alignment with Contemporary Perspectives

Several interviewees indicated that they align with contemporary interpretations of construct validity, highlighting the shift from a narrow statistical focus to a broader consideration of the relevance and appropriateness of inferences drawn from test scores.

Participant: "We've moved beyond just crunching numbers; it's about making assessments that matter."

3.2.3. Challenges in Defining Constructs

A recurring theme was the difficulty in accurately defining the constructs being measured by language tests. Participants noted that language proficiency is complex and multifaceted, making it challenging to pinpoint the precise components that need to be assessed.

Participant: "Language proficiency isn't just about vocabulary and grammar; it's a complex interplay of skills."

Participant: "Defining language constructs can feel like chasing smoke; they're elusive."

3.2.4. Ethical Considerations

Some respondents emphasized the importance of considering ethical aspects in construct validity. They discussed the need to ensure that assessments are fair, unbiased, and have positive consequences for test-takers.

Participant: "We have a responsibility to ensure that assessments don't disadvantage anyone unfairly."

Participant: "Ethics are not just a checkbox; they're integral to validity."

3.2.5. Unified Framework

A few interviewees echoed the idea that construct validity serves as a unifying component in the broader validity framework. They mentioned that construct validity helps bring together various aspects of assessment design and interpretation.

Participant: "It's like the glue that holds everything together, making sure it all makes sense."

3.3. Quantitative Findings for Research Question 2

Research Question 2: What challenges and limitations are encountered in achieving construct validity in ELA task design, and what strategies and recommendations do language assessment experts suggest for enhancing the construct validity of ELA tasks?

Research Question 2 aimed to investigate the challenges and limitations encountered in achieving construct validity in English Language Assessment (ELA) task design and to explore the strategies and recommendations language assessment experts suggest for enhancing construct validity in ELA tasks. The quantitative findings related to this research question are summarized below:

Table 4. *Descriptive Statistics*

Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Challenges in Achieving Construct Validity					
Complexity of Language Proficiency	3%	7%	23%	46%	21%
Cultural and Contextual Factors	5%	11%	25%	40%	19%
Balancing Rigor and Accessibility	2%	6%	20%	50%	22%
Recommendations for Enhancing Construct Validity					
Incorporating Contextual Factors	2%	5%	15%	45%	33%
Continuous Validation Research	1%	4%	12%	43%	40%
Diverse Task Types	3%	6%	21%	47%	23%
Involving Test-Takers	2%	5%	19%	45%	29%

4.3.1. Challenges and Limitations in Achieving Construct Validity

3.3.1.1. Complexity of Language Proficiency

A majority of respondents (67%) agreed or strongly agreed that one of the primary challenges in achieving construct validity in ELA task design is the inherent complexity of language

proficiency. This complexity stems from the multifaceted nature of language skills, including speaking, listening, reading, and writing.

3.3.1.2. Cultural and Contextual Factors

Approximately 59% of respondents acknowledged the influence of cultural and contextual factors as a challenge in construct validity. They indicated that the cultural and sociocultural contexts in which language is used can introduce variability in language assessment.

3.3.1.3. Balancing Rigor and Accessibility

A substantial proportion (72%) of respondents agreed or strongly agreed that finding a balance between rigor and accessibility in ELA task design poses a significant challenge to achieving construct validity. They expressed concerns about the potential exclusion of certain test-taker populations in pursuit of validity.

3.3.2. Strategies and Recommendations for Enhancing Construct Validity

3.3.2.1. Incorporating Contextual Factors

Around 78% of respondents supported the idea of incorporating contextual factors, such as cultural and sociocultural elements, into ELA task design as a strategy for enhancing construct validity.

3.3.2.2. Continuous Validation Research

A majority (83%) of respondents suggested that conducting continuous validation research, including piloting and field-testing tasks, is a critical strategy for improving construct validity. They emphasized the need for ongoing refinement and validation of assessment instruments.

3.3.3.3. Diverse Task Types

Approximately 70% of respondents recommended diversifying the types of tasks used in ELA assessments. They believed that incorporating a broader range of task formats, including real-world tasks, would contribute to construct validity.

3.3.3.4. Involving Test-Takers

A significant proportion (74%) of respondents advocated for involving test-takers in the assessment development process. They argued that test-takers' perspectives and feedback are valuable for enhancing construct validity and the user-friendliness of assessments.

3.4. Qualitative Findings for Research Question 2

Research Question 2 aimed to investigate the challenges and limitations encountered in achieving construct validity in English Language Assessment (ELA) task design and to explore the strategies and recommendations language assessment experts suggest for enhancing construct validity in ELA tasks. The qualitative findings from semi-structured interviews with experts are summarized below:

Table 5. *Thematic analysis*

Theme	Findings
Challenges and Limitations in Achieving Construct Validity	
Complexity of Language Proficiency	Respondents highlighted the complexity of language proficiency as a major challenge in construct validity
Cultural and Contextual Factors	The impact of cultural and contextual factors on construct validity was a recurring theme, emphasizing the need to consider diverse contexts.
Balancing Rigor and Accessibility	Participants acknowledged the delicate balance between rigor and accessibility, a challenge in ensuring that assessments are both valid and inclusive
Strategies and Recommendations for Enhancing Construct Validity	
Incorporating Contextual Factors	A prevailing recommendation was to incorporate contextual factors into assessment design, ensuring assessments align with cultural and situational contexts
Continuous Validation Research	Experts stressed the importance of ongoing validation research, including piloting and field-testing tasks, to keep assessments relevant and aligned with evolving language proficiency
Diverse Task Types	Diversifying the types of tasks used in assessments, including real-world scenarios, was recommended to enhance construct validity
Involving Test-Takers	Several experts advocated for involving test-takers in assessment development, viewing their feedback as invaluable for user-centered improvements

3.4.1. Challenges and Limitations in Achieving Construct Validity

3.4.1.1. Complexity of Language Proficiency

Many interviewees highlighted the inherent complexity of language proficiency as a major challenge. They emphasized that language skills involve intricate interactions between speaking, listening, reading, and writing, making it challenging to develop assessments that comprehensively capture these nuances. As two participants noted, "Language proficiency is like a puzzle with multiple pieces that need to fit perfectly." "Language proficiency is a dynamic interplay of skills; capturing it fully is like chasing shadows."

3.4.1.2. Cultural and Contextual Factors

A recurring theme was the impact of cultural and contextual factors on construct validity. Several experts noted that language is deeply influenced by the cultural and sociocultural contexts in which it is used. This complexity requires assessment developers to consider these factors, ensuring that assessments are relevant and fair for diverse test-taker populations. One interviewee stated, "You can't divorce language from culture; they're intertwined." And other interviewee noted, "Languages are living entities; they evolve in context, and assessments should reflect that."

3.4.1.3. Balancing Rigor and Accessibility

Interviewees recognized the delicate balance between rigor and accessibility in assessment design. They expressed concerns about assessments being too challenging, potentially excluding certain groups, or being too lenient, compromising validity. Striking the right balance was described as a constant challenge. An expert commented, "We need rigorous assessments, but not at the cost of alienating test-takers." And other interviewee stated, "Validity is essential, but not at the expense of fairness and accessibility."

3.4.2. Strategies and Recommendations for Enhancing Construct Validity

3.4.2.1. Incorporating Contextual Factors

A prevailing recommendation from interviewees was the incorporation of contextual factors into assessment design. They suggested that assessments should be sensitive to cultural and situational contexts, ensuring that tasks are meaningful and relevant to test-takers' experiences. As one participant emphasized, "Context matters; it's not just about the language itself, but how it's used." And other interviewee noted, "To be valid, assessments must resonate with the lived **experiences** of test-takers."

3.4.2.2. Continuous Validation Research

Experts stressed the importance of continuous validation research, including piloting and field-testing tasks. They emphasized the need for ongoing refinement and validation of assessment instruments to ensure that they align with the evolving landscape of language proficiency. An interviewee stated, "Validation isn't a one-time event; it's an ongoing process to keep assessments relevant." And other interviewee noted, "Validation is a continuous journey; we must adapt to the changing landscape of language."

3.4.2.3. Diverse Task Types

Interviewees recommended diversifying the types of tasks used in ELA assessments. They argued that incorporating a broader range of task formats, including real-world tasks and authentic communication scenarios, would contribute to construct validity. An expert noted, "Assessments should mirror real-life language use; it's about practicality." And other interviewee noted, "Real-life language use is diverse; our assessments should reflect that diversity."

3.4.2.4. Involving Test-Takers

Several experts advocated for involving test-takers in the assessment development process. They viewed test-takers as valuable sources of feedback, highlighting the need to consider their perspectives to enhance construct validity. An interviewee expressed, "Test-takers are the end-users; their insights are invaluable in making assessments better." And other interviewee stated, "Test-takers are the ones navigating our assessments; they should have a say in how they're designed."

In conclusion, the qualitative findings from Research Question 2 provide a deeper understanding of the challenges faced in achieving construct validity in ELA task design and the strategies and recommendations proposed by language assessment experts. These insights reflect the complexities of language assessment and underscore the importance of considering cultural, contextual, and user-centered factors in assessment development to enhance construct validity.

4. Discussion

The quantitative findings of the current investigation suggest that the surveyed language assessment experts generally recognize the importance of construct validity in ELA task design. They align with contemporary interpretations of construct validity, acknowledging its role as the appropriateness and relevance of inferences drawn from test scores. Furthermore, respondents largely agree that construct validity unifies other validity components, underscoring its central position in validity frameworks. Despite this recognition, a substantial number of respondents acknowledged challenges in accurately defining the constructs measured by language tests, indicating a need for further research and guidance in this area. These findings provide a valuable quantitative snapshot of language assessment experts' perceptions and understanding of construct validity, which can inform discussions and recommendations for enhancing the validity of ELA tasks.

The quantitative findings shed light on the challenges and strategies associated with achieving construct validity in ELA task design. These findings reflect the perspectives of language assessment experts and provide valuable insights into the complexities of this aspect of assessment development. Finding delved into the challenges faced by language assessment experts in achieving construct validity in English Language Assessment (ELA) task design and explored the strategies and recommendations they suggest for enhancing construct validity. The qualitative findings provide valuable insights into the complexities of this aspect of language assessment. In this discussion, we examine these challenges and strategies, drawing upon relevant literature where applicable.

The findings highlight that language assessment experts overwhelmingly recognize the critical importance of construct validity in ELA task design. A substantial proportion of respondents strongly agreed that construct validity is a fundamental consideration when creating assessments. This recognition aligns with the foundational role that construct validity plays in ensuring that assessment results accurately reflect the language skills and competencies being measured. As one participant aptly stated, "Without construct validity, we might as well be assessing something entirely different." This emphasis on the significance of construct validity

underscores its role as a guiding principle in the development of meaningful language assessments. It resonates with Messick's (1989) assertion that validity is not just an abstract concept but a cornerstone of ethical and effective assessment practice. The findings reinforce the idea that construct validity is central to ensuring that assessments are more than just exercises in measurement; they are meaningful tools for evaluating language proficiency.

The thematic analysis also reveals that a majority of language assessment experts align with contemporary interpretations of construct validity. They emphasize the importance of moving beyond narrow statistical views and embracing a broader perspective that considers the relevance and appropriateness of inferences drawn from test scores. This shift in perspective reflects the evolution of construct validity from a mere statistical property of tests to a more holistic consideration of the implications of test scores. The alignment with contemporary perspectives reflects the field's adaptability to the changing demands of language assessment. It echoes the sentiment that assessments should not be confined to quantitative metrics but should encompass the meaningfulness of inferences drawn from test results (Aryadoust et al., 2020). As one participant aptly stated, "We've moved beyond just crunching numbers; it's about making assessments that matter."

A recurring theme in the qualitative findings is the challenge of defining constructs in language assessment. Respondents acknowledged the complexity of language proficiency, which involves a multifaceted interplay of skills and competencies. This difficulty in pinpointing precise components that need to be assessed echoes the sentiment expressed by Hasrol et al. (2022), who noted that determining the competencies underlying test tasks can be elusive. This challenge in defining constructs suggests that construct validity research remains a dynamic and evolving field. It highlights the need for continued exploration and refinement of construct definitions to align assessments more closely with the intricate nature of language proficiency. As assessments become more nuanced and context-specific, construct validity will play a pivotal role in ensuring that they accurately capture the diverse dimensions of language ability (Nematzadeh,2018).

The thematic analysis underscores the importance of ethical considerations in construct validity. Participants stressed the ethical responsibility of assessment developers to ensure fairness, lack of bias, and positive consequences for test-takers (Bachman& Palmer,2012). This perspective aligns with Messick's (1989) assertion that construct validity encompasses both evidential and consequential bases, with ethics and fairness being integral components. Ethical considerations regarding fairness and inclusivity were prominently featured. Experts emphasized that assessments should be designed to be fair and accessible to all test-takers, regardless of their background. This resonates with the ethical dimension of construct validity, where the consequences of test use must be just and fair (Messick, 1989).

The data indicate that language assessment experts recognize several challenges in achieving construct validity. The complexity of language proficiency emerges as a predominant challenge, emphasizing the need for assessments to comprehensively capture the diverse dimensions of language skills. Additionally, the influence of cultural and contextual factors and the balance between rigor and accessibility are perceived as significant challenges, reflecting the need for assessments to be sensitive to diverse test-taker populations and

contexts. Respondents offer a range of strategies and recommendations to enhance construct validity in ELA tasks. The widespread support for incorporating contextual factors underscores the importance of considering the broader sociocultural and situational contexts in which language is used. Continuous validation research, including piloting and field-testing, is seen as a crucial practice to refine and validate assessment instruments continually. Diversifying task types and involving test-takers in the assessment development process are recommended to improve the authenticity and user-friendliness of ELA assessments.

One of the most prominent challenges identified by experts is the inherent complexity of language proficiency. This complexity arises from the multifaceted nature of language skills, including speaking, listening, reading, and writing. As noted by participants, language proficiency is not a monolithic entity but a dynamic interplay of these skills. This complexity aligns with the literature's recognition of language as a complex, multifaceted construct (Bachman & Palmer, 2010). Addressing this challenge requires a nuanced approach to assessment design. Language assessments must strive to comprehensively capture these multifaceted skills, ensuring that the test-taker's abilities are accurately represented. This aligns with the principle of construct underrepresentation, which suggests that construct validity is threatened when important aspects of the construct are not adequately assessed (Messick, 1989).

The influence of cultural and contextual factors emerged as another critical theme. Language is deeply intertwined with cultural and sociocultural contexts (Chapelle & Voss, 2021). Assessments that do not consider these factors risk compromising construct validity. The recognition of this challenge underscores the need for assessments to be context-sensitive, relevant, and culturally fair (Chalhoub & O'Sullivan, 2020). Experts recommended incorporating contextual factors into assessment design to enhance construct validity. This aligns with the principle of consequential validity, which emphasizes the need to consider the impact of test use on individuals from different cultural backgrounds (Messick, 1989). Achieving this balance between cultural sensitivity and linguistic proficiency remains a complex task.

The challenge of balancing rigor with accessibility was a recurring concern. This challenge reflects the tension between ensuring the validity of assessments and making them inclusive for diverse test-taker populations. Striking the right balance is essential to prevent assessments from being overly challenging or too lenient. This challenge aligns with the literature on test fairness and differential item functioning (Camilli & Shepard, 1994). Ensuring fairness and equity in assessments is a crucial aspect of construct validity. Construct-irrelevant variance, such as demographic factors, should be minimized to maintain construct validity (Messick, 1989).

Assessments should evolve to reflect changes in language proficiency and educational contexts. Continuous validation research, including piloting and field-testing, ensures that assessments remain aligned with construct validity principles. Contextual sensitivity and authenticity were highlighted as critical aspects of assessment design. Experts advocated for assessments that mirror real-world language use and situations. This aligns with the principles of authenticity and contextualization in language assessment (Borsboom & Wijsen, 2016).

Construct validity has long been recognized as the linchpin of test development and validation in language assessment. It serves as the cornerstone for ensuring that assessments accurately measure the intended language constructs. As the complexity of language proficiency became increasingly apparent, so did the challenges in unpacking these multifaceted constructs. The construct validity framework, initially introduced by Cronbach & Meehl (1955) and further elaborated by Messick (1975, 1989), has provided a critical foundation for assessing the quality of language tests. However, despite significant advancements in empirical construct validity research, including the development of various validation methodologies, accurately pinpointing the constructs elicited by language tests remains a formidable challenge. The multifaceted and context-dependent nature of language use defies simple reductionist approaches. As Hasrol et al. (2022) aptly noted, determining the competencies underlying performance on language tests remains an intricate and indirect process.

Conclusions

The research underscores the ongoing importance of construct validity in language assessment. The insights gained from experts' perceptions and experiences can guide assessment developers in designing tasks and assessments that better align with the underlying language constructs. This can lead to more accurate and reliable assessments of language proficiency. The emphasis on involving test-takers in the assessment development process has implications for user-centered design. Engaging test-takers can lead to assessments that are more relevant, authentic, and fair. Recognizing the influence of cultural and contextual factors on language use highlights the importance of cultural sensitivity in assessment design. Language assessments should aim to be inclusive and culturally fair, ensuring that they do not disadvantage specific groups. The recommendation for continuous validation research and iterative assessment design highlights the dynamic nature of language proficiency. Assessments should evolve alongside changes in language use and educational contexts. This emphasizes the need for ongoing research and development in the field of language assessment. The acknowledgment of opportunities presented by technology and artificial intelligence (AI) suggests that these innovations can play a pivotal role in enhancing construct validity. AI-driven assessment tools and neurocognitive research can offer new avenues for understanding and assessing language proficiency.

Building on the findings and implications, several avenues for further research in the field of language assessment can be explored. Future research can delve deeper into the involvement of test-takers in assessment design. Investigating the impact of user-centered approaches on assessment outcomes, including construct validity, can provide valuable insights. Given the emphasis on cultural factors, research can explore the cross-cultural validity of language assessments. This includes investigating how assessments perform across different cultural and linguistic backgrounds. Further research can examine the integration of AI and technology in language assessment. This includes the development and validation of AI-driven assessment tools and their impact on construct validity. Longitudinal studies can be conducted to track changes in language proficiency over time and assess the validity of assessments in capturing these changes. This can provide insights into the dynamic nature of language proficiency. Research can explore innovative task types and formats that better align with construct validity

principles. This includes the development and validation of new assessment tasks that mirror real-world language use. Further research can investigate the ethical dimensions of language assessment, including fairness, bias mitigation, and the impact of assessments on various demographic groups.

References

Ahmad, S., Sultana, N., & Jamil, S. (2020). Behaviorism vs constructivism: A paradigm shifts from traditional to alternative assessment techniques. *Journal of Applied Linguistics and Language Research*, 7(2), 19-33.

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.

Aryadoust, V., Zakaria, A., Lim, M. H., & Chen, C. (2020). An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. *Frontiers in Psychology,* 11, 1941. https://doi.org/10.3389/fpsyg.2020.01941

Azizi, Z. (2022). Fairness in assessment practices in online education: Iranian University English teachers' perceptions. *Language Testing in Asia,* 12(14), 1-17. https://doi.org/10.1186/s40468-022-00164-7

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baker, B. A., & Riches, C. (2018). The development of EFL examinations in Haiti: Collaboration and language assessment literacy development. *Language Testing*, 35(4), 557–581.

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment literacy mean to teachers? *ELT Journal*, 73(2), 113–123.

Borsboom, D., & Wijsen, L. D. (2016). Frankenstein's validity monster: The value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice,* 23(2), 281-283. https://doi.org/10.1080/0969594X.2016.1141750

Camilli, G., & Shepard, L. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage.

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical development and integrated arguments*. British Council Monographs.

Chapelle, C., & Voss, E. (Eds.) (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin,* 52(4), 281-302. https://psycnet.apa.org/doi/10.1037/h0040957

Demir, M., Tananis, C. A., & Trahan, K. W. (2019). Evaluation of alternative assessment methods used in elementary schools. *Egitim ve Bilim*, 44(197)

Dhindsa, H. S., Omar, K., & Waldrip, B. (2007). Upper secondary Bruneian science students' perceptions of assessment. *International Journal of Science Education*, 29(10), 1261-1280.

Dobakhti, L. (2020). The process of enhancing validity, reliability, and ethics in research. *Iranian Journal of Applied Language Studies*, 12(2), 59-88. 10.22111/IJALS.2020.5978

Elharrar, Y. (2006). Teacher assessment practices and perceptions: the use of alternative assessments within the Quebec educational reform. *Unpublished thesis. Montreal: Université du Québec à Montréal*.

Fan, X., Liu, X., & Johnson, R. (2020). A mixed method study of ethical issues in classroom assessment in Chinese higher education. *Asia Pacific Education Review,* 21, 183-195.

https://doi.org/10.1007/s12564-019-09623-y

Fraenkel, R. J., Wallen, E. N., & Hyun, H. H. (2012). *How to design and evaluate research in education (8th ed.)*. McGraw-Hill

Giraldo, F. (2020). A post-positivist and interpretive approach to researching teachers' language assessment literacy. *Profile: Issues in Teachers' Professional Development*, 22(1), 189–200.

Giraldo, F. (2021). A reflection on initiatives for teachers' professional development through language assessment literacy. *Profile: Issues in Teachers' Professional Development*, 23(1), 197–213.

Harding, L., & Brunfaut, T. (2020). Trajectories of language assessment literacy in a teacher-researcher partnership: Locating elements of praxis through narrative inquiry. In: Poehner, M. E. and Inbar-Lourie, O. (Eds.), *Towards a re-conceptualization of second language classroom assessment* (pp. 61–81). Springer International Publishing.

Hasrol, B. S., Zakaria, A., & Aryadoust, V. (2022). A systematic review of authenticity in language assessment. *Research Methods in Applied Linguistics,* 1(3), 100023. https://doi.org/10.1016/j.rmal.2022.100023

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333.

Herman, J., & Cook, L. (2019). Fairness in classroom assessment. In S. M. Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 243-264). Routledge.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0040168

Isbell, D. R., Kremmel, B., & Kim, J. (2023). Remote proctoring in language testing: Implications for validity, fairness, and justice. *Language Assessment Quarterly*, 20(5), 469–487.

Jiang, Y. C., Jong, S. Y., Lau, W. F., Chai, C. S., Liu, S. X., & Park, M. Y. (2022). A scoping review on flipped classroom approach in language education: Challenges, implications and an interaction model. *Computer Assisted Language Learning*, 35(6), 1218–1249.

Kunnan, A. J. (2018). *Evaluating language assessments*. Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). American Council on Education and Macmillan.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist,* 30(10), 955–966. https://doi.org/10.1037/0003-066X.30.10.955

Nematzadeh, A. (2018). Construct Irrelevant Factors and Test Validity: Investigating the Relationship among Gender, Age, Mother Tongue, Field of Study and TOEFL IBT ® Results. *Journal of Foreign Language Research*, 8(1), 139-166. doi: 10.22059/jflr.2018.242996.405

Ockey, G. J., Chukharev-Hudilainen, E., & Hirch, R. R. (2023). Spoken dialogue systems and their potential for aiding in the assessment of interactional competence. *Language Assessment Quarterly*, 20(5), 377–398.

Tierney, R. (2016). Fairness in educational assessment. In M. A. Peters (Ed.). *Encyclopedia of educational philosophy and theory* (pp. 1-6). Springer Singapore.

Xu, Y., & Brown, G. (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

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 12(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zohrabi, M. & Nasirfam, F. (2024). The use of assessment for learning rather than assessment of learning in EFL context. *Applied Research on English Language,* 13(2), 1-30. https://doi.org/10.22108/are.2024.140274.2210