Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 14, No. 2; 2024, October

ISSN 2476-5880
https://www.ijlt.ir

Exploring the Validity of Applied Linguistics' PhD Program Admission Interviews in Iranian Universities: A Validity Argument Approach

Saman Ebadi[1]*, Rana Rahimi[2], Maryam Salari[3]

ARTICLE INFO

Article History:
Received: September 2023
Accepted: October 2023

KEYWORDS
Fairness
Kane's model of interpretative argument
PhD admission
Interviews
Standardized evaluation criteria
Validity argument approach

ABSTRACT

Using Kane's interpretive argument model and Messick's validity argument approach, this study rigorously examined faculty and PhD candidate's perspectives on PhD admission interviews in Iranian universities. We interviewed 10 professors and PhD interviewees which provided comprehensive insight into nuanced perspectives. We conducted rigorous content analysis to identify prevalent themes, forming a strong foundation for our analysis. This study emphasizes the vital requirement for standardized evaluation criteria, robust support systems, and an enhanced interview process to ensure fair and inclusive admission systems. Additionally, our development of guidelines based on Toulmin's reasoning model underscores the originality of our contribution and its potential to benefit stakeholders and the Ministry of Science, Research, and Technology (MSRT) in Iran. The findings highlighted the importance of standardized criteria, support, and a stronger interview process for fairness and inclusivity in selecting PhD candidates. Faculty stressed clear guidelines to remove subjectivity, while candidates voiced concerns about unclear expectations and proposed added support like mentoring and preparation programs. Based on Toulmin's reasoning model, the study crafted validity argument guidelines for this context. As a result, these proposed changes will impact stakeholders and the MSRT by enhancing the PhD candidate evaluation process and ensuring a fairness and inclusivity. This study provides valuable insights to improve PhD admission procedures at Iranian universities by integrating standardized criteria, enhancing support mechanisms, and fostering fairness in decision-making.

1. Introduction

Assessment results aren't simply categorized as valid or invalid, and validation is an ongoing process that includes collecting, summarizing, and evaluating evidence to determine how well the scores generated by an assessment instrument are associated with their intended meaning and the inferences made about the characteristic being measured (Cizek, 2020**)**. One continuous theme in this view is the test scores interpretation, as highlighted by leading scholars such as Kane (1992), Messick (1989), and Cronbach and Meehl (1955), emphasizing the need to go beyond the mere numerical values

[1]Associate Professor, Department of English Language and Literature, Razi University, Kermanshah, Iran. Email: samanebadi@gmail.com
[2]PhD candidate, Department of English Language and Literature, Razi University, Kermanshah, Iran. Email: r.rahimy@razi.ac.ir
[3]PhD candidate, Department of English Language and Literature, Razi University, Kermanshah, Iran. Email: maryamsalarie981@gmail.com

Cite this paper as: Ebadi, S., Rahimi, R., & Salari, M. (2024). Exploring the validity of applied linguistics' PhD program admission interviews in Iranian universities: A Validity Argument Approach. *International Journal of Language Testing, 14*(2), 1–16. https://doi.org/ 10.22034/IJLT.2023.414511.1286

of test scores and delve into what these scores truly mean in terms of relevant skills and abilities. In light of the priority of valid score interpretations, as Messick (1989) highlighted, score analysis should extend beyond surface-level descriptions of test performance. Instead, it should offer a meaningful comprehension of what a score represents in terms of a profound understanding of the assessed abilities. Cronbach and Meehl (1955) offered that validation is a multifaceted process requiring an inclusive argument rather than a single judgment. The basis of this argument depends on construct validity, as outlined by the American Psychological Association (APA, 1966). This indicates that the interpretation of scores should develop systematically through four key stages: The initial step entails moving from observed behavior, which comprises the raw interview data, to derive an observed score. This score quantitatively reflects the candidate's performance during the interview. In the second step, the rater shifts from the observed score to the "universe score." Then, the candidate's performance is compared to a broader population of candidates, providing essential context. In the third stage, the rater delves into specific assessment criteria and standards, evaluating the candidate's performance against predefined benchmarks. This process ensures objectivity and standardization in the assessment. Finally, the fourth stage culminates in a decision based on the interpretations made in the previous steps. It's important to note that these interpretations are not made in isolation but are derived from a systematic and well-supported argument. These steps must be followed sequentially to ensure the coherence of the interpretative argument, a principle to any performance evaluation, including PhD entrance interviews (APA, 1966). Regarding PhD entrance examinations in Iran, there has been a shift from a decentralized to a semi-centralized admission system (Ahmadi et al., 2015). In this updated system, the Ministry of Science, Research, and Technology (MSRT) conducts a standardized written PhD entrance exam, adding clarity and structure to the admission process. Following successful performance in the written exam, qualified candidates engage in interview sessions at different universities (Rezvani & Sayyadi, 2016). The present study builds upon the research efforts of Rezvani and Sayyadi (2016) and Ahmadi et al. (2015) in the realm of PhD entrance examinations. Rezvani and Sayyadi (2016) took an approach to assess the reliability and credibility of PhD Program Entrance Exams (PPEE) utilizing the validity argument framework. Simultaneously, Ahmadi et al. (2015) focused on the validity of the Iranian PhD entrance exam for Teaching English as a Foreign Language (IPEET), emphasizing test interpretation and consequences.

Furthermore, this study takes a novel approach by combining two established argument-based structures: Kane's (1992) argument model and Bennett's (2010) theory of action. This synthesis allows us to investigate the extent to which our proposed assumptions are supported by empirical evidence. Additionally, the researchers delve into any unintended consequences that might emerge from our validity investigation. This comprehensive exploration adds a valuable dimension to understanding PhD entrance examinations and their impact, making a strong case for their significance in the field. Through a validity argument lens, practitioners can gain valuable insights for improving the validity of PhD interviews. In this academic realm, the current study represents a distinctive effort, focusing on a critical aspect of test score interpretation in the context of PhD entrance interviews. While the significance of valid score interpretations enjoy extensive recognition (Messick, 1989), the specific challenges and complexities associated with their application to PhD interview evaluations in Iran have received limited attention. This study tried to bridge this knowledge gap by examining the interpretation procedures used in PhD interviews in Iran through the lens of Kane's (2013) validity argument framework. In doing so, the researchers aim to justify the alignment between obtained scores and their interpretations and present a robust interpretative argument that can guide the validation process of PhD interviews. The findings from this investigation will inform the development of practical guidelines to address these concerns and provide suggestions for advancing practices in this field. For instance, during PhD interviews in Iran, the behavior elicited from interviewees may not always reflect the intended skills to be assessed, leading to inaccurate scores. Moreover, fairness and consistency in scoring may be compromised if evaluators treat interviewees unfairly or inconsistently.

In Iran's PhD interviews, evaluators can assign varying scores to the same candidate, as they may focus on different aspects of performance. For example, while one evaluator may pay attention to a candidate's M.A. thesis, another may prioritize other qualities. Validity theory focuses on the accuracy of score-based interpretations and decisions for all individuals in the population of interest. On the other hand, fairness analyses center around group differences and variations in the accuracy and

appropriateness of interpretations and decisions across different groups, which are defined in terms of race/ethnicity, gender, age, and other relevant factors (Kane,2010). According to Gipps and Stobart (2009), fairness should be viewed within a sociocultural context, similar to the shift in how validity is understood. Rather than an add-on concept, fairness should be integrated within validity arguments as they both recognize the social aspects of assessment and address concerns about bias. So, the significance of this study lies in its recognition of the importance of addressing validity issues and promoting fairness in assessment practices, particularly in PhD interviews, to improve the accuracy and meaning of inferences drawn from the results. Specifically, it seeks to answer the research question about the validity of these interviews as an admission tool based on Kane's framework.

1- To what extent do Iran's PhD entrance interviews align with Kane's model of interpretative argument, as outlined in the Method section?
2- From the participants' perspectives, what practical guidelines can enhance the validity of PhD entrance interviews in Iran?

2. Review of Literature

2.1. Validity

Historically, in the late 19th century, validity was primarily perceived as a statistical characteristic of tests. However, in contemporary times, most scholars view validity not as an inherent quality of the tests themselves but rather as a measure of the suitability and significance of the conclusions drawn from test scores (Chapelle & Voss, 2021; Shahmirzadi,2023). Validity concept has a rich history with diverse perspectives, frameworks, and terminology (Haertel & Herman, 2005). It's often portrayed as playing a gatekeeping role in scientific inquiry (Johnson et al., 2008). In the context of modern validity theory and the Standards for Educational and Psychological Testing (AERA, 2014; Kane, 2013), the primary focus is on the validation process rather than solely on the instrument being validated. According to Messick's unified framework (1989), construct validity assumes a central role by integrating various validity components. Messick emphasized the importance of considering evidential consequential factors to accurately interpret and effectively utilize test scores when establishing construct validity. The conceptual understanding of construct validity requires empirical evidence to substantiate and strengthen the understandings and implications derived from test score utilization (Messick, 1989).

2.1.1. Types of Validity

Traditionally, validity has been categorized into three types: content, criterion-related, and construct validity (Brown, 2004). Content validity involves assessing the representativeness or sampling adequacy of the content within a measuring instrument (Kerlinger, 1973). Content validity involves making a subjective assessment of the significance of a measurement (Brown, 2004). Recognizing the limitations of content validity, Messick (1989) introduced construct validity, which focuses on specific domains of language ability intended for evaluation (Bachman & Palmer, 1996). Construct validity demonstrates experimentally that a test measures the construct it claims to assess (Brown, 2003). Additionally, criterion validity assesses how well one measure predicts an outcome for another (Taherdoost, 2016), involving complex analyses of the relationship between test scores and criteria (Gleser & Cronbach, 1965). Messick introduced the term "construct validity" because content validity evidence often lacks validation due to a lack of test scores or the performances upon which such scores are based (American Psychological Association, 1954).

2.1.2. Argument-Based Validation

Kane expanded Messick's (1989) framework by introducing an enhanced argument-based validity approach. In collaboration with Chapelle and Voss (2021), Kane advocated for using Toulmin's (2003) informal logic principles in data collection and analysis for validation. This approach encompasses backing, evidence, warrants, counterevidence, and qualifiers. Cronbach (1989) emphasized the importance of a validity argument, focusing on evidence collection supporting or challenging the interpretation of test scores. One method developed for deriving accurate conclusions from test scores is Evidence Centered Design (Mislevy et al., 2003), which emphasizes meticulous test design considering logical reasoning throughout the assessment process. This methodology emphasizes meticulously designing tests while considering the logical progression of reasoning throughout the assessment process. Determining the validity of a test relies on discussions surrounding the intended

interpretation and practical applications of the test scores (Kane, 2016). Kane (2006) outlines principles for developing a valid argument. Firstly, the argument centers around test scores' intended meaning and use, forming an interpretive argument. Secondly, the validity argument considers both technical and social aspects to define the meaning of the scores. Lastly, the ultimate aim of the interpretive argument is to determine the usefulness of the test score for a particular purpose. So, validation is an ongoing process beyond collecting evidence to assess test scores and their applications. It also requires judgments about those interpretations' credibility (Kane, 2006). Therefore, validity can be seen as a construct built by examining theoretical and empirical evidence (Chappelle et al., 2010).

Toulmin's model of argumentation has four main components: claim, evidence, warrant, and backing. The claim is the main point or conclusion, supported by relevant evidence. The warrant justifies the claim, and the backing represents the underlying assumptions supporting the warrant (Toulmin,1958). Additionally, qualifiers can define the scope and limitations of the claim's validity. Due to its simplicity, Toulmin's model has been widely adopted in research examining students' argumentation use (Cavagnetto, 2010; Erduran et al., 2004; Zohar & Nemet, 2002). Toulmin's model suggests an argument consists of a claim, evidence, a warrant, backing, and qualifiers. These principles will shape the data collection process by prioritizing gathering relevant evidence that either supports or challenges the claims made during interviews. The analysis will examine the justification behind these claims. The validity of interviews will be evaluated by assessing the coherence and strength of the arguments made by interviewees. Adopting this framework ensures the reliability and validity of the collected data, contributing to a thorough evaluation of interview validity.

2.2.1. PhD Entrance Interviews in Iran

PhD entrance interviews constitute a critical component of the admission process for PhD programs in Iran (Ahmadi et al., 2015). These interviews aim to evaluate candidates' academic and research potential and their suitability for the program (Ahmadi et al., 2015). While the interview format may vary across institutions, it typically involves face-to-face conversations between candidates and one or more faculty members or stakeholders (Derakhshan et al., 2021). Interviews can be conducted in person or online, often via video conferencing. They serve as a valuable assessment tool for various purposes, including evaluating job applicants, admitting PhD program candidates, or conducting research studies (McDaniel et al., 1994). However, interviewer bias can affect their validity (Dorussen et al., 2005), yet with training in standardized scoring, interviews reveal valuable insights into candidates' potential. Fair tests adhere to recognized ethical and administrative standards, ensuring impartiality and lack of bias (Kheirzadeh et al., 2015). It is crucial to administer standardized tests consistently and according to instructions to enable accurate and comparable score interpretations (Zieky, 2006). This standardization guarantees equal opportunities for all test takers to showcase their abilities. Additionally, maintaining test security is vital to preventing unfair advantages.

In conclusion, fair tests uphold integrity, objectivity, and validity (Kheirzadeh et al., 2015). Moreover, interviews can provide insight into a candidate's communication skills and personality traits, which may not be visible in other assessment methods, such as tests or written application materials. This study evaluates how well Iran's PhD entrance interviews align with Kane's model of interpretative argument and gathers insights for improving their validity. Specifically, it seeks to answer the research question about the validity of these interviews as an admission tool based on Kane's framework. The current study delves into an area influenced by trailblazers such as Kane, Messick, Cronbach, and Meehl, whose insights continue to resonate. Also, this study is motivated by a dedication to improving the comprehension of test results in this context, ultimately aiming to promote equitable and uniform assessment procedures. The main challenges revolve around inconsistencies resulting from the absence of standardized criteria and the use of various evaluation methods. To address these issues, the researchers suggest using Kane's validity argument framework as a practical way to evaluate PhD interviewers. This approach is intended to improve the dependability and validity of PhD interviews within the field of Applied Linguistics in Iran.

3. Method

3.1. Participants and Setting

This account attempted to qualitatively analyze the views of faculty members and PhD candidates of Applied Linguistics regarding the validity of PhD entrance interview processes employing

Kane's model of interpretative argument. Data were gathered through semi-structured interviews, and content analysis was employed wherein the most frequently recurring themes were coded. The present study involved 10 participants since a saturation point was reached where including more participants would not add new information. The participants included four university professors and six PhD candidates recently participating in PhD entrance interviews, which helped to hear their presumably opposing views. The candidates, two males and four females, had received their English language-related MA degrees (i.e., English Language and Literature, Teaching English as a Foreign Language, English Translation) from Iranian universities. They aged between thirty and thirty-seven and had varying experience with the English language, but they all held English language-related positions (e.g., university lecturing, translating, writing, tutoring, etc.). The participants provided informed consent before the interviews and were assigned pseudonyms throughout the study to ensure anonymity.

3.2. Instrumentation

Individual semi-structured interviews were used, which, according to Kvale and Brinkmann (2009), are valuable since they can render "knowledge claims that are so powerful and convincing in their own right that they carry the validation with them, like a strong piece of art" (p.252). The 12 interview questions were developed based on Kane's interpretative argument model to determine PhD entrance interviews' validity and explore ways to alleviate limitations. The questions targeted scoring, generalizability, extrapolation, and implications in the interviews. The questions were developed in consultation with an expert in TEFL who, having years of experience and lots of knowledge on the subject, was significantly helpful in minimizing bias in the questions, ensuring their relevance and increasing clarity. Additionally, the questions were piloted through consultation with several junior PhD candidates (so their expertise would match that of the target population) before the actual interview sessions.

The open-ended questions (see Appendix A for a sample) allowed the participants to express themselves through a reflective mindset. Interviews allow interaction between the interviewer and the interviewee and consist of several stages: an initial warm-up phase to establish a comfortable setting, placement-type questions to assess the interviewees, further probing to explore their knowledge, eliciting responses to validate subsequent interpretations, and finally, providing feedback to address any factors that may impact decisions (Castillo-Montoya, 2016). In this study, demographic information was obtained, followed by a question-and-answer procedure (further questions were posed for deeper exploration of ideas), which lasted for about twenty-five minutes on average. The main ideas were reviewed at the end of each session to ensure that the participants had fully expressed themselves and agreed with the researchers' interpretations. All interview sessions were recorded, transcribed verbatim, fed into, and thematically analyzed using NVIVO 20.

3.3. Procedures

The interviews, conducted in Persian, adhered to Kane's model of interpretative argument. Guided by the research questions, data were analyzed using an inductive approach, thoroughly reading the transcripts and segmenting them into meaningful units, as Merriam and Tisdell (2015) outlined. Data were transcribed verbatim and fed into NVIVO. The researchers first coded the data individually, the extracted themes were discussed, and intercoder reliability of 90 percent was obtained through negotiation. Answers to each interview question were grouped to be coded more consistently, and meaningful units (i.e., sentences or groups of sentences aligned with the research questions) were extracted. The codes referred to by more than half of the participants were considered as the criterion. Such codes were deemed frequent, and their relevance to the research questions was explored; therefore, off-topic themes were dismissed. Significantly, the researchers realized no new themes were emerging by comparing new data to the existing themes.

3.4. The Framework

The framework employed in this context is derived from the reasoning model proposed by Toulmin et al. (1979), further developed and utilized by Kane (2006). This model has been verified and streamlined by Gotch and Perie (2012) to enable its practical implementation while preserving the fundamental elements of the argument framework. This framework is suitable for evaluating PhD

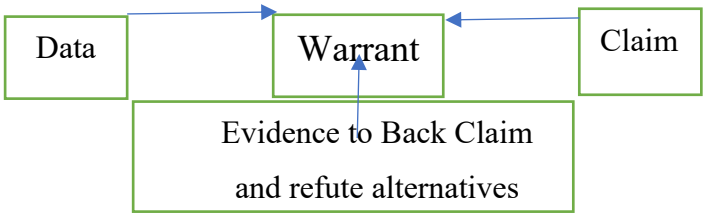
admission interviews' validity for its comprehensive coverage. It includes multiple aspects of validity, including content, construct, criterion-related, and consequential validity, which ensures a comprehensive analysis of validity in the context of PhD entrance interviews in Iran. The scope of the present account necessitates an exploration of each aspect and its corresponding steps by drawing on the works of Chuisano et al. (2022), Aryadoust (2023), and Nguyen (2022).

1. Content Validity: It refers to the extent to which an assessment measures the intended constructs adequately. The following steps are typically followed to evaluate content validity: a. Define the content domain, b. Expert panel review: Assemble a group of experts in the domain to review the assessment items. c. Item analysis, d. Revision and refinement.
2. Construct Validity: It assesses whether an assessment measures the theoretical construct or trait it intends to measure. The steps for evaluating construct validity include a. Theoretical framework of the Construct, b. Hypothesis generation about the expected relationships between the assessed construct and other variables, c. Data collection from a sample of participants, d. Data analysis, e. Interpretation of the results.
3. Criterion-Related Validity: This assesses how an assessment predicts or correlates with an external standard criterion. The steps for evaluating criterion-related validity are: a. Select criteria, b. Data collection, c. Data analysis, d. Interpretation.
4. Consequential Validity: It focuses on the broader impact of using an assessment, including its intended and unintended effects on individuals and society. The steps for evaluating consequential validity include: a. Identify consequences of using the assessment, such as the impact on test-takers; b. Evaluation framework including ethical considerations, fairness, accessibility, and the potential for bias or adverse impact; c. Data collection, d. Data analysis and interpretation, e. Decision-making about the assessment.

Given Kane's (2006) framework, the interviewers utilized the data to determine the degree to which evidence supports decisions. In the present account, the researchers ensured the relevance, comprehensiveness, and sufficiency of the interview questions and responses as the criteria for evaluating the evidence and decision-making. The researchers also accounted for possible reasons that could contradict the intended conclusions by comparing the emergent themes with new data. They collected new data by conducting follow-up interviews with other candidates and realized no new themes were emerging. This approach requires interviewers to consider potential reasons that could disprove the intended inferences. The validity argument framework allows evaluations even without a robust theory explaining an underlying construct (Chapelle et al., 2010). Instead of solely relying on a well-defined theory, this framework emphasizes building a coherent and evidence-based argument to support the validity of an assessment.

Nevertheless, it has a number of limitations, such as subjective interpretation of the evidence, limited guidance on how to interpret an evaluation's results, potential biases in decision-making, etc. Therefore, additional framework validation might be required for PhD entrance interviews. This framework can be challenging to comprehend, so the researchers used more straightforward language when reporting the results to facilitate understanding. For example, the extracted themes have been clarified by using explanations, categorizations, and the inclusion of interview extracts.

Figure 1.
Argument Framework Simplified by Gotch and Perie (2012)



The components of the argument framework include the following:

1. Data: It refers to the evidence that supports the claim. It can be quantitative or qualitative data, observations, research findings, etc., that adds credibility to the argument (Dorsey,2022). 2. Warrant:

The warrant connects the data to the claim and justifies why the data supports the claim. It explains the logical relationship between the evidence and the claim, demonstrating how the data leads to the conclusion (Johnson, 2022). 3. Claim: The claim is the main conclusion being argued for. It states the viewpoint the argument aims to establish as true or valid. The claim should be supported by the data and warrant to make a convincing argument (Dorsey, 2022).

In short, this framework simplifies constructing an argument by emphasizing the importance of providing relevant and reliable data, establishing a logical connection between the data and the claim through a warrant, and making a clear claim supported by the evidence.

4. Results

This study assessed the alignment between Iran's PhD entrance interviews of Applied Linguistics and Kane's model of interpretative argument. Additionally, practical recommendations are provided to enhance the quality of the PhD interviews. This section presents the most frequently recurrent themes detected through careful content analysis.

4.1. Perceptions of the Interview Processes

4.1.1. Content Validity

4.1.1.1. Content Knowledge

The following points were made when the participants were asked about the type of knowledge targeted during the interviews:

- General and specialized language knowledge are important in effective communication, academic engagement, knowledge integration, and professional growth and guaranteeing a PhD candidate's success. Some interviewees highlight the importance of specialized knowledge during PhD courses.
- The participants argued that the interviews lacked construct validity because they failed to capture the full range of abilities and characteristics necessary for success in the PhD program. Sarina, a PhD candidate, said:

Writing and other general English skills should be evaluated because even a person's publications do not accurately indicate their true abilities. General English proficiency is very determining in PhD programs. It can estimate the degree to which a candidate can handle the workload of reading, comprehending, and producing academic texts.

- Candidates also complained that some of the questions were too detailed. The following extract by Nazanin, a PhD candidate, best represents this theme:

During the interview, they asked me very detailed, unimportant questions. There were lots of more important topics they could have touched upon. I had not prepared myself for such details.

- PhD stakeholders allocate some of the interview scores for organized writing. However, the candidates referred to the challenges posed by the short time allotted for producing writing samples on the interview day. By Paniz's (one of the candidates) verbal communication:

Effective academic writing is crucial for publishing scholarly articles, but I often find myself where academic writing requires significant time and effort. I feel overwhelmed to dedicate the time to producing high-quality academic written work. So, PhD stakeholders should overcome time constraints by offering writing workshops or courses.

- Interviewers of the PhD entrance exam expressed their beliefs regarding deciding whether one candidate is more knowledgeable than others. They left comments:

I look for candidates who understand the subject well. They should show that they've thought deeply about the key ideas. It's essential to see if candidates have used their knowledge in real-life situations, but it can be hard to assess this because we may not have access to their experiences, and it takes time.

4.1.1.2. Variation in Content Coverage and Interests

The participants were also invited to elaborate on the content coverage of the interview questions. What follows reflects their ideas.

While the lack of content coverage of PhD interview questions was found to be one of the essential themes, the participants acknowledged inconsistencies in content coverage and faculty

members' interests and priorities. Regarding this matter, the remarks made by Sarina are particularly persuasive:

While it is generally expected that the content coverage of PhD interview questions aligns with standard criteria, I have observed variations across different academic institutions and departments, which reflect the specific interests and priorities of the faculty members conducting the interviews.

4.1.1.3. Lack of Enough Attention to the Thesis

The thesis holds significance, encompassing profound understanding, academic proficiency, and research expertise. However, participants suggest that the thesis is not adequately emphasized in interviews. Nazanin's statement highlights the crucial role of mastering a thesis in enhancing research expertise.

I believe that cultivating research expertise goes beyond acquiring knowledge; it encompasses developing unique skills and competencies that enable researchers to navigate the complexities of their field. I think research expertise empowers individuals to contribute to their field and allows them to conduct rigorous investigations that advance knowledge.

4.1.2. Construct Validity

Construct Irrelevance Variance

Some valuable insights shed light on the importance of tackling construct irrelevant variance and its influence on student scores, manifested in the words of Melika:

Stress and an unfamiliar environment significantly impacted my ability to showcase my true potential. It made it hard for me to think, remember important details, and express myself effectively since I spent some time getting familiar with the new environment and new faces.

Mina highlighted the effect of professors' inquiries about priority on the selection process:

A notable issue is that some professors inquire about the priority assigned to their university, neglecting to consider overall performance. This practice raises concerns regarding bias in the selection process.

4.1.3. Criterion Related Validity

Scoring Rubric

Our discussion on whether a specific rubric for calculating scores was followed revealed the need for a structured framework for evaluation based on a predefined set of criteria. According to one of the faculty members:

We should design a more meticulous scoring rubric. We should know exactly how to rate a candidate's performance. Our rubric is too general right now, and a more well-designed rubric is necessary to determine what proportion of a score should be allocated to a certain performance aspect. Like the IELTS band descriptors, we should know exactly how to assign scores to those with certain features.

4.1.4. Consequential Validity

4.1.4.1 Lack of Equitable Treatment

When asked about the fairness of PhD interviews, most participants replied that equitable treatment results in perceived impartiality. As elaborated by one of the candidates, Amir:

I noticed how easy their questions were when I talked to the other candidates. I could answer almost all of them. But when I entered the room, I encountered many difficult questions I had never heard of. It seemed they had made their decisions. I believe I have potential and my resume speaks for itself, but I am sure no candidate would have been able to answer that sort of detailed questions.

Melika, another candidate, also mentioned:

I deserved to be asked some questions. I went in last, and I did not receive any attention. I guess they were exhausted. I talked about myself for the most part. It was humiliating. This was not the case with the other candidates, which is unfair. I even touched upon my thesis to trigger some questions, but it was useless.

4.1.4.2. Collective Interviews

The participants favored collective interviews because of their efficiency, collaboration, diversity, and fairness. As supported by PhD entrance authorities, group interviews provide an opportunity to give everyone the benefit of the doubt and decide on evaluating his performance when other interviewers are posing questions. Amir's statement is noteworthy:

Group interviews can be beneficial because they allow diverse members to contribute their unique insights. This helps create a more thorough and well-rounded assessment of the candidates. Having different viewpoints represented reduces bias and promotes inclusivity, leading to better decision-making.

4.1.4.3. Lack of Access to Academic Profiles

The lack of access to the candidates' academic profiles was a point that both parties (interviewers and interviewees) believed had strengthened the validity of the interview processes. During the interviews, all PhD interview holders expressed that, beforehand, they did not have access to the academic records of the PhD candidates. As one of the interviewers claimed:

I think it could potentially lead to bias and unfair assessment. I focus solely on each candidate's performance, qualifications, and potential without being influenced by bias based on academic achievements, which might impact my judgment without being an accurate indicator of ability.

Another piece of evidence supporting the notion that not having access to academic records beforehand promotes fair assessment was highlighted by Sarah. Sarah emphasized evaluating candidates based on their demonstrated skills and potential rather than relying on previous academic achievements.

I agree. It increases a more equitable evaluation process for all PhD candidates. Also, it allows them to delve into each candidate's unique strengths, skills, and contributions.

4.2. Ideas to Promote Challenges

Regarding the second research question, the following practical guidelines were found to improve the PhD interview process.

4.2.1. A Focus on Language Skills

A focus on language skills is a common theme among practical guidelines. According to the viewpoints of Mahtab:

PhD interviewers consider language proficiency as an important evaluation criterion during interviews. I think language proficiency illuminates the path to delivering effective presentations.

4.2.2. Accounting for Sum Scores

In certain countries, decisions are made based on the sum of scores. In Iran, however, the mean is considered the basis of evaluation. The impact of sum scores is a guideline introduced by Sina:

In some countries, the assessment process focuses on the sum of scores rather than the mean. This approach carries weight and should be considered when evaluating candidates, as it can significantly impact the overall assessment and selection process.

4.2.3. Equity and Fairness

Ensuring all candidates receive fair treatment, regardless of personal or demographic factors, can guarantee an average amount of equitable treatment.

Aligning the interests and priorities of faculty members with the evaluation criteria was also found to increase equity across departments. As Maryam pointed out:

The stakeholders must establish uniform guidelines for interview questions that apply to all academic departments. This will ensure that the content covered in interviews is consistent across departments. It is important to align the interests and priorities of faculty members with the evaluation criteria to reduce discrepancies in question-asking.

4.2.4. Enhancing Electronic Registration

Challenges with electronic registration are another factor demonstrated in the words of PhD interviewees. As expressed by Negin:

In the final week before the interview, interviewees must allocate ample time for preparation. It is essential to give adequate focus to comprehending and utilizing the electronic registration system proficiently, thus guaranteeing a fruitful interview experience.

4.2.5. Time Management

Time management and longer intervals between interviews were mentioned as important guidelines. Sina stated that:

As an interviewee, I believe it is imperative to emphasize effective time management during the interview process. Allowing longer intervals between interviews is recommended to ensure candidates have time to prepare and perform to the best of their abilities.

4.2.6. Reintroducing Old Examination Systems

Most PhD stakeholders agree with reintroducing written exams and oral interviews from previous periods to review candidates. This consensus stems from the belief that such a comprehensive assessment method effectively evaluates candidates' knowledge, skills, and suitability, fostering the selection process.

The themes and subthemes identified through thematic analysis in the "Results" section are summarized in Table 1. Each theme is supported by participant quotations that offer evidence and context. The themes include a variety of perceptions of the interviewing procedure and improvement recommendations.

Table 1.

The Themes and Subthemes Identified Through Thematic Analysis

Theme and Frequency	Subthemes and Frequency	Example Quotes from Participants
Perceptions of the Interview Processes	Content Knowledge (8)	"General and specialized language knowledge is crucial for success." - Participant 1
		"Interviews lacked construct validity, failing to capture essential abilities." – Sarina
		"Some interview questions lack importance and detail." – Nazanin
		"PhD Stakeholders should overcome time constraints for writing samples." – Paniz
Variation in Content Coverage (8)	Inconsistent Content Coverage and Interests (8)	"Content coverage varies across academic institutions and departments." – Sarina
Lack of Enough Attention to the Thesis (7)	Importance of the Thesis (6)	"Mastering a thesis enhances research expertise." – Nazanin
Construct Irrelevance Variance (10)	Impact of Stress and Environment on Interview Outcomes (7)	"Stress and unfamiliar environment impacted my interview performance." – Melika
	Influence of Professors' Inquiries on the Selection Process (6)	"Some professors inquire about priority, raising concerns of bias." – Mina
Scoring Rubric (7)	Lack of a Meticulous Scoring Rubric (7)	"A more well-designed rubric is necessary for accurate evaluation." – Stakeholder
	Lack of Equitable Treatment (8)	"When I talked to other candidates, their questions were easy. But when I entered the room myself, I encountered difficult questions." – Amir
		"I deserved to be asked a few questions. I did not receive any attention." – Melika
Collective Interviews (10)	Preference for Group Interviews (10)	"Conducting group interviews fosters inclusivity and better decision-making." – Amir
Lack of Access to Academic Profiles (6)	Fair Assessment without Access to Academic Records (6)	"Not having access to academic records promotes fair assessment." - Navid

Ideas to Address Challenges	Focus on Language Skills (9)	"Consider language proficiency as an important criterion." – Mahtab
	Accounting for Sum Scores (5)	"Take sum scores into account during evaluation." – Sina
	Equity and Fairness (7)	"Ensure unbiased and equal treatment for all student groups." – Participant
		"Establish uniform guidelines for interview questions across departments." – Maryam
	Enhancing Electronic Registration (7)	"Address concerns associated with electronic registration." – Negin
	Time Management (5)	"Emphasize effective time management and longer intervals between interviews." – Sina
	Reintroducing Old Examination Systems (5)	"Reintroduce written exams and oral interviews for a comprehensive assessment." – Stakeholder

5. Discussion

The present account qualitatively explored the validity of PhD entrance interviews of Applied Linguistics in Iran using Kane's validity framework. Rezvani and Sayyadi (2016) explored the washback effect of the PhD entrance exam. However, the present study aimed to build on previous investigations and further explore other potential challenges. Unlike previous studies that mainly explored the ideas of PhD candidates (e.g., Ahmadi et al., 2015; Derakhshan et al.,2021), data were collected through individual semi-structured interviews with PhD candidates and faculty members recently participating in PhD entrance interview sessions. Data were transcribed verbatim and thematically analyzed through NVIVO, and the most frequently repeated themes were reported.

The study aimed to explore the participants' perceptions of the validity of the interview processes. Results revealed that lack of equitable treatment and interviewer bias (which impact impartiality advocated by Kane's interpretative argument model), inconsistent content coverage and interests on the part of the jury, lack of a meticulous scoring rubric, lack of enough attention to a candidate's thesis, the impact of construct irrelevant factors (e.g., stress) could threaten the validity. The study also aimed to explore ideas to minimize the impact of these threats. It was suggested that equal treatment of the candidates, accounting for general English skills, time management, accounting for sum scores, and enhancing electronic registration could effectively remove the previously mentioned threats. In line with Ahmadi et al. (2015), the findings suggest that the PhD interview processes do not align well with Kane's model.

One notable issue highlighted in the interviews is the variations in content coverage during PhD interviews across different academic departments. In alignment with Darabi Bazvand & Ahmadi (2020) who revealed that PhD test tasks in Iran are not representative of the PhD program objectives, it became evident that, during the interviews, faculty members' interests and priorities play a role in determining the questions, so the candidates must be well-prepared to address diverse interview styles. The findings revealed major threats to content and construct validity among the constituents of Kane's model (Chapelle & Voss, 2021). The results show that GE skills and specialized knowledge are not tapped upon well. Moreover, oscillation in paying attention to a candidate's thesis across departments also threatens the interviews' content validity.

Another challenge that threatens construct validity is the problem of time constraints for academic writing. PhD candidates often feel overwhelmed by the significant time and effort required to produce high-quality written work. This highlights the need for support mechanisms such as writing workshops or courses, which, according to Jusslin and Hilli (2023), help candidates manage their time effectively and enhance their writing skills. One of the faculty members humbly mentioned:

I agree that writing needs time. One cannot expect a candidate to write a well-organized proposal in half an hour or less. We used to expect such a thing and realized that writing should be

measured differently. We can, for example, give them a standardized IELTS writing task and allot enough time.

This variance in styles and interests of interview holders originates from the lack of a specific knowledge base to shape the foundations of the interview questions and the lack of a predefined set of PhD interview evaluation criteria in Iran. In concordance with Ebadi and Dovaise (2015), we recommend that interview questions be extracted from "an agreed-upon domain of knowledge" (p.449) to increase the criterion validity of the process.

According to Dorussen et al. (2005), the validity of interviews as an assessment tool can be influenced by various factors, such as interviewer bias. The participants emphasized the importance of each student group receiving unbiased treatment, which aligns with the concept of impartiality and consequential validity advocated by Kane's interpretative argument model. Bias can also serve as a source of stress (a construct irrelevant factor), which, according to Ebadi and Bashiri (2021), is decreased during virtual interviews for various reasons. Moreover, based on the insufficient data provided during the interviews, they also fail to account for a candidate's success or lack thereof, in PhD programs. Therefore, consequential validity, another key concept in Kane's model, is not well-addressed during these interviews.

Moreover, as pointed out by the participants, the lack of familiarity with the jury's research backgrounds and priorities made it twice as hard for the participants to prepare themselves. This can be attributed to insufficient preparation time for the interview sessions (Ahmadi et al., 2015). In Iran, the candidates have only a couple of weeks to prepare themselves, and they usually spend their time completing their registration and preparing for their trips to the destination universities.

The major contribution of the present account is the practical suggestions it makes for enhancing the Applied Linguistics interview processes. Previous accounts have, more or less, provided us with insight into the extent of validity of the interviews (e.g., Ahmadi et al., 2015; Bazvand & Ahmadi, 2020; Kiany et al., 2013; Rezvani & Sayadi, 2016;), but they have failed to render practical solutions to existing challenges. This study has a number of implications for policymakers, stakeholders, MSRT, and interview holders.

As evidenced by our results, Kane's framework can assist practitioners in incorporating the validity argument in PhD interviews. The practical implications presented in this account can constructively help increase the validity of the interview sessions. They can promote fairer evaluation and, consequently, decision-making based on the evaluation results.

5. Conclusion

This study aimed to explore the validity of PhD entrance interviews of Applied Linguistics in Iran and suggest ways to promote the existing threats to validity. Results revealed that policymakers should employ an argument-based validity framework for evaluating PhD interviews and improve validity. By implementing the implications and fostering a culture of continuous improvement, PhD entrance stakeholders can work towards enhancing fairness, inclusivity, standardized evaluation criteria, and improved support mechanisms. They should prioritize developing well-defined rubrics for objective and fair evaluation, regularly updating and revising them.

The findings underscore the integral role of language skills in PhD candidates' success; therefore, interview holders should incorporate the evaluation of such skills. For example, candidates can be asked to complete standardized writing tasks in due time. Ahmadi et al. (2015) also briefly suggested this but failed to render applicable guidelines for measuring writing. Policymakers can incorporate PhD interview preparation courses into MA programs so candidates will master the skills.

Furthermore, as the participants suggested, a well-defined evaluation rubric establishes an organized framework that illuminates the path for assessing a candidate's performance, which ensures consistency, transparency, objectivity, and fairness by outlining predetermined criteria for scoring. More specifically, the MSRT in Iran should consider designing better scoring rubrics that effectively distinguish subtle differences in performance by breaking down the scoring rubrics into subcategories that assess different aspects of GE proficiency, academic knowledge, and skills. This can result in a more accurate evaluation and selection of those with higher GE proficiency levels. It should be noted that all the faculty members who participated in this study advocated the view that the decentralized (Ahmadi et al., 2015) approach to PhD candidates' selection that Iranian universities previously

followed proved to be more successful. They highly recommend the MSRT take steps to create similar selection processes again.

According to the findings, respondents preferred conducting interviews collectively due to its efficiency, collaboration, and fairness. This format enables comprehensive assessments of candidates' abilities, aligning with the principle of fairness by mitigating individual biases and ensuring transparent and unbiased evaluations. Based on this finding, it is recommended that interviewers avoid individual interviews to increase validity in the process. Ebadi and Bashiri (2021) also tackled fairness in an attempt to evaluate virtual PhD admission interviews, wherein a collective interview approach was followed during the COVID-19 pandemic, and most participants deemed the interviews fair.

The interviewers unanimously expressed that they do not have access to the academic records of PhD candidates before conducting interviews. This promotes fairness by ensuring assessments are based on individual performance rather than solely on academic achievements. It is highly recommended that policymakers keep the PhD candidates' background secret until after the interview sessions and finalizing performance scores.

Although this study's implications extend far beyond its limitations, it is noteworthy that this attempt failed to explore the ideas of other stakeholders involved. Future studies can investigate the ideas of stakeholders such as high-rank admission authorities, policymakers, university lecturers, and PhD graduate students who might be involved in the process as observers or interact with the new cohort of students. A major delimitation of the present account was its geographical scope. It mainly focused on PhD interview processes in Iran. Therefore, the findings may not be generalizable to settings with different PhD admission processes. Future researchers should render comparative accounts of PhD entrance procedures in different settings to inform the practitioners and help them decide on the most valid procedure. The present study also fails to account for the predictive validity of the interviews. Therefore, longitudinal studies in the future can follow a candidate's success, or lack thereof, in the PhD programs. Mixed-methods studies can provide information on the perceived challenges from a broader range of participants.

In conclusion, it is implied that validation is not an endpoint but a process, and expressing that a test has been 'validated' means that the process has been met. The duty of validation is not to hold an interpretation but to discover what might be incorrect. The findings of this study indicate that it is crucial to view these results as an opportunity for improvement rather than failure. By employing the validity framework, researchers can detect the potential shortcomings of the interviews and identify solutions to promote the validity level that might be expected of such life-changing evaluation instruments. This increases the evaluation's impartiality, inclusiveness, and overall fairness. By addressing the identified gaps and implementing the recommended changes, institutions can work towards creating a more robust interview process, ultimately benefiting the evaluation of PhD candidates.

Declaration of Conflicting Interests

The researchers, hereby, declare that there are no conflicts of interest associated with the content of this article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-profit sectors.

References

AERA, A. NCME, American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ahmadi, A., Darabi Bazvand, A., Sahragard, R., & Razmjoo, A. (2015). Investigating the validity of PhD entrance exam of ELT in Iran in light of argument-based validity and theory of action. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, *34*(2), 1-37. https://doi.org/10.22099/jtls.2015.3581

Aryadoust, V. (2023). The vexing problem of validity and the future of second language assessment. *Language Testing*, *40*(1), 8-14. https://doi.org/10.1177/02655322221125204

Association, A. P., Association, A. E. R., & Education, N. C. o. M. i. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. American Psychological Association

American Psychological Association, American Educational Research Association, National Council on Measurement in Education, & American Educational Research Association. (1966). *Standards for educational and psychological tests and manuals*. American Psychological Association.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*(2-3), 70-91. https://doi.org/10.1080/15366367.2010.508686

Brown, H. D., & Abeywickrama, P. (2004). *Language assessment: Principles and classroom practices.* Pearson.

Brown, R. P. (2003). Measuring individual differences in the tendency to forgive: Construct validity and links with depression. *Personality and Social Psychology Bulletin*, *29*(6), 759-771. https://doi.org/10.1177/0146167203029006008

Castillo-Montoya, M. (2016). Preparing for interview research: The interview protocol refinement framework. *Qualitative Report*, *21*(5), 811-831. http://nsuworks.nova.edu/tqr/vol21/iss5/2

Cavagnetto, A. R. (2010). Argument to foster scientific literacy: A review of argument interventions in K–12 science contexts. *Review of Educational Research*, *80*(3), 336-371. ttps://doi.org/10.3102/0034654310376953

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3-13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A., & Voss, E. (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press. https://doi.org/10.1017/9781108669849.002

Chuisano, S. A., Anderson, O. S., Weirauch, K., Roper, R., Phillips, J., McCabe, C., & Sadovnikova, A. (2022). An application of Kane's validity framework to evaluate formative and summative assessment instruments for tele-simulations in clinical lactation. *Simulation in Healthcare*, *17*(5), 313-321. https://doi.org/10.1097/SIH.0000000000000653

Cizek, G. J. (2020). *Validity: An integrated approach to test score meaning and use*. Routledge.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy: Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 147–171). University of Illinois Press.

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

Darabi Bazvand, A., & Ahmadi, A. (2020). Interpreting the validity of a high-stakes test in light of the argument-based framework: Implications for test improvement. *Journal of Research in Applied Linguistics*, *11*(1), 66-88. https://doi.org/10.22055/RALS.2020.15417

Derakhshan, A., Qafouri, M., & Faribi, M. (2021). An investigation into the demotivating and remotivating factors among Iranian MA and PhD exam candidates of TEFL. *Two Quarterly Journal of English Language Teaching and Learning University of Tabriz*, *13*(27), 81-112. https://doi.org/10.22034/elt.2021.45627.2377

Dorsey, D. W., & Michaels, H. R. (2022). Validity arguments meet artificial intelligence in innovative educational assessment: A discussion and look forward. *Journal of Educational Measurement*, *59*(3), 389-394. https://doi.org/10.1111/jedm.12330

Dorussen, H., Lenz, H., & Blavoukos, S. (2005). Assessing the reliability and validity of expert interviews. *European Union Politics*, *6*(3), 315-337. https://doi.org/10.1177/1465116505054835

Ebadi, S., & Bashiri, S. (2021). Psychological benefits and challenges of Ph.D. entrance exam virtual interviews during COVID-19 pandemic: Does gender play a role? *Frontiers in Psychology*, *12*, 1-13. https://doi.org/10.3389/fpsyg.2021.800715

Ebadi, S., & Dovaise, M. S. (2015). Evaluating Ph.D. candidates' interviews: A validity argument approach. *Iranian EFL Journal, 11*(2), 438-450.

Erduran, S., Simon, S., & Osborne, J. (2004). Tapping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, *88*(6), 915-933. https://doi.org/10.1002/sce.20012

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In *Educational assessment in the 21st century: Connecting theory and practice* (pp. 105-118). Springer. https://doi.org/10.1007/978-1-4020-9964-9_6

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, *30*(4), 395-418. https://doi.org/10.1007/bf02289531

Gotch, C. M., & Perie, M. (2012). *Using validity arguments to evaluate the technical quality of local assessment systems* [Paper presentation]. American Educational Research Association, Vancouver, British Columbia, Canada.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. *Yearbook of the National Society for the Study of Education*, *104*(2), 1-34. https://doi.org/10.1111/j.1744-7984.2005.00023.x

Johnson, E. C., Kirkhart, K. E., Madison, A. M., Noley, G. B., & Solano-Flores, G. (2008). The impact of narrow views of scientific rigor on evaluation practices for underrepresented groups. *Fundamental Issues in Evaluation*, *32*(2), 261-293.

Johnson, R. C. (2022). *Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context* [Unpublished doctoral dissertation]. Macquarie University.

Jusslin, S., & Hilli, C. (2023). Supporting Bachelor's and Master's students' thesis writing: A rhizo-analysis of academic writing workshops in hybrid learning spaces. *Studies in Higher Education*, *48*(9), 1-18.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177-182. http://ltj.sagepub.com

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. *Educational Measurement*, *4*(2), 17-64.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198-211. https://doi.org/10.1080/0969594X.2015.1060192

Kerlinger, F. N. (1973). *Foundations of behavioral research: Educational, psychological and sociological inquiry*. Holt Rinehart and Winston.

Kheirzadeh, S., Marandi, S. S., & Tavakoli, M. (2015). Test administration conditions of the general English section of the Iranian national PhD entrance exam: Are the PhD exam candidates satisfied? *International Journal of Language Testing*, *5*(2), 151-167.

Kiany, G. R., Shayestefar, P., Ghafar Samar, R., & Akbari, R. (2013). High-rank stakeholders' perspectives on high-stakes University entrance examinations reform: Priorities and problems. *Higher Education*, *65*, 325-340.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing*. sage.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*(4), 599-616.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (pp. 13-103). Macmillan.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Nguyen, H. T. M. (2022). Reshaping content validation: A case study of a reading achievement test for third-year English majors. *International Journal of Language Testing*, *12*(1), 26-58. https://doi.org/10.1080/0969594X.2015.1060192

Rezvani, R., & Sayyadi, A. (2016). Washback effects of the new Iranian TEFL Ph.D. program entrance exam on EFL instructors' teaching methodology, class assessment, and syllabus design: A qualitative scrutiny. *Journal of Instruction and Evaluation*, *9*(33), 159-180. https://doi.org/10.17507/tpls.0605.27

Shahmirzadi, N. (2023). Validation of a language center placement test: Differential item functioning. *International Journal of Language Testing, 13* (1), 1-17. https://www.doi.org/10.22034/ijlt.2022.336779.1151

Taherdoost, H. (2016). Validity and reliability of the research instrument; How to test the validation of a Questionnaire/Survey in a research. *International Journal of Academic Research in Management, 5*(3), 28-36. https://doi.org/10.2139/ssrn.3205040

Toulmin, S., Rieke, R. D., & Janik, A. (1979). *An introduction to reasoning*. New York and London: Macmillan.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press. .https://doi.org/10.1017/cbo9780511840005

Zieky, M. (2006). Fairness reviews in assessment. In M. Downing & M. Haladyna (Eds.), *Handbook of test development* (pp. 359-376). Lawrence Erlbaum Associates.

Zohar, A., & Nemet, F. (2002). Fostering students' knowledge and argumentation skills through dilemmas in human genetics. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *39*(1), 35-62.

Appendix A

A sample of the interview questions is provided below.

1. How is General English (GE) proficiency evaluated during the interviews?
2. What proportion of the interview score is advocated to GE and what amount is advocated to content knowledge?
3. How would you evaluate the fairness of the interview questions across groups of candidates?
4. How does the time of the interview, appearance, academic profile, etc. impact an interviewer's judgments?
5. Is a certain scoring rubric followed during the interviews?
6. Which type of interview renders a more comprehensive evaluation? individual or collective?
7. How do registration processes impact a candidate during her/his journey?