Ensuring Test Validity and Accountability in Omani Higher Education Institutions: Findings, Implications, and a Modified Framework

Ibtisam Alrushaidi, University of Technology and Applied Sciences, Oman

The Barcelona Conference on Education 2023
Official Conference Proceedings

Abstract
This qualitative study examines the roles, obstacles, and strategies employed by educators and policymakers in pursuit of test validity and accountability in Omani higher education. The study provides an insider's perspective on the process of test validation based on in-depth, semi-structured interviews with a small, diverse sample of exam committee members and policymakers. The findings highlight the central role of instructors in assuring the validity of tests, as well as the supporting roles of policymakers and exam committee members. The absence of comprehensive assessment literacy is identified as a significant challenge, prompting the proposal of a customized Accountability Interpretive Use Argument (IUA) framework. This revised IUA supports a collaborative, iterative validation process, highlighting the need for ongoing professional development in the assessment field and the recognition of expertise. While the study focuses on the Omani context, the proposed solutions may be applicable to other educational contexts with similar characteristics. The implications of the research contribute to the ongoing discussion about effective assessment practices in higher education, providing a foundation for future research on test validity and accountability.

Keywords: Test Validity, Interpretive Use Argument (IUA) Framework, Accountability

1. Introduction

Classroom assessment (CA) is a crucial, systematic procedure that allows for the accumulation, analysis, and application of information regarding students learning. It seeks to disclose student capabilities and areas for refinement, monitor progress, designate grades, and facilitate communication with parents (McMillan, 2013, p. 4). The interdependence of language instruction, learning, and evaluation highlights the significance of evaluating students, a fundamental educator responsibility requiring specialized expertise and knowledge. According to multiple studies (Coombe, Vafadar, & Mohebbi, 2020; Latif & Wasim, 2022), the burgeoning field of assessment literacy encompasses these necessary skills and knowledge for designing and evaluating assessments.

Validity remains a crucial metric in assessment, with significant implications for decision-making processes such as course completion, student placement, and accurate diagnoses (Brown & Abeywickrama, 2010). In fact, since the 1960s, scholarly discourse on the validity of language testing and the development of models to operationalize the concept has gained momentum and continues to evolve (Lado, 1961; Cronbach, 1988; Messick,1989; AERA, APA, & NCME, 1999; Chapelle, 1999; Kane, 2001; Embretson, 2007; Hoeve, 2020).

Traditionally, validity is defined as the capacity of a test to measure its intended construct accurately (Heaton, 1975; Henning, 1987; Hughes, 1989). For example, when assessing reading comprehension, the valid test should evaluate the students' reading skills and not their background knowledge. Nonetheless, this definition of validity is more nuanced than it initially appears. Messick (1989) and Glass & Metternich (2020) argue that it is difficult to measure and assess competencies directly and precisely. Moreover, Chapelle (2021) argues that validity is not inherent but rather depends on the expertise of professionals. This demonstrates the need for professionals to establish a shared comprehension of validity, continuously revise these definitions, and assume responsibility for attaining validity and accountability in assessment.

In the realm of higher education in Oman, accountability, symbolizing quality assurance and accreditation, is essential to education policy. Assessment has come to be associated with accountability, which has become associated with the quality of education because of the significant role test results play in determining the quality of outcomes (Smith, 2016). The quality of education is mandated by the Oman Authority for Academic Accreditation and Quality Assurance of Education (OAAAQA), which was established by Royal Decree No. 9/202. This authority employs quality indicators that align with Higher Education Institution (HEI) and program standards. This study specifically meets criterion 2.8 regarding "Assessment Methods, Standards, and Moderation," as outlined in the OAAAQA's Institutional Standards Assessment Manual. In order to fulfill this criterion, the Higher Education Institution (HEI) possesses efficient mechanisms to guarantee the rigorous implementation of assessment procedures, thereby ensuring the validity, reliability, and effectiveness of assessments in upholding academic standards.

Despite the exhaustive description of this criterion, the responsibility for ensuring rigorous assessment procedures varies across Oman's higher education institutions. Notably, the quality of each standard in HEIs is made accessible to the public prior to being archived on the website (https://oaaaqa.gov.om/). The outcomes range from 3 to 0, with a value of 0 indicating that the institution does not meet the specified criterion. The aforementioned website provides a total of twenty-six reports on various institutions of higher education. The

results indicate that one institution did not meet the criteria, six institutions partially met the criteria, 18 institutions met the criteria, and one institution received the highest score of 3. The selection of institutions for this study is based on an analysis of these reports, specifically those with a score of 2 (2 out of 3). This investigation centers on a comprehensive examination of assessment accountability in Oman's higher education institutions.

This study seeks to investigate and comprehend the processes and strategies employed by exam committee members and policymakers in Oman's higher education institutions in order to ensure test validity and accountability. Specifically, this research intends to:

1) Investigate how exam committee members in Oman's higher education institutions ensure test validity;
2) Determine how policymakers assure accountability and evaluate the validity of these institutions;
3) Apply and refine the Interpretive Use Argument (IUA) framework for Accountability within the Omani context.

This research is an exhaustive examination of the validation process and accountability mechanisms in Omani higher education institutions with the intent of proposing a modified IUA framework that improves these processes. Understanding the practices, challenges, and shared understanding of validity among the main stakeholders is emphasized. This research not only contributes to the Omani context, but also to the larger literature on test validity and accountability in higher education.

2. Literature Review

Extensive research has been conducted on the multidimensional concept of validity, eliciting a variety of perspectives and numerous validation methodologies (Brown & Abeywickrama, 2010). In lieu of a dichotomous classification (valid/invalid), validity is viewed as a continuum (Messick, 1989; Reeves & Marbach-Ad, 2016), encouraging researchers to consider the extent to which a test is valid. Importantly, validity is not inherent to a test; rather, it refers to the use of a test for a particular purpose (Sireci, 2007). This perspective suggests that the validity of the same test may vary depending on its intended application. Consequently, understanding validity entails determining whether the correlation between intended and accomplished knowledge justifies the use of measurement for decision-making (Hughes, 2018).

Brown and Abeywickram (2010) emphasize the diversity of the evidence supporting validity, which has led to the identification of various categories of validity. Face validity, for example, refers to the appropriateness of a test for examiners and test-takers, whereas content validity ensures that the test accurately reflects the curriculum on which it is founded. Criterion validity, subdivided into concurrent validity and predictive validity, requires statistical analysis of student scores (Cronbach & Meeh, 1955; Davies, 1968). Construct validity refers to a test with a valid rationale founded on theories. Modern validity theories, however, tend toward a unitary validity concept, rendering it unnecessary to provide evidence for each category of validity (Reeves & Marbach-Ad, 2016).

Carlsen and Rocca (2021) argue that traditional methods for validating abstract constructs may not be accurate. They propose divorcing validity from whether or not the test measures the construct and concentrating more on whether test developers or test users are able to construct a convincing argument for their use. Consequently, a test's validity depends on its

intended use (AERA et al., 1999, cited in Sireci & Faulkner-Bond, 2014) and should not be deemed valid or invalid in and of itself.

Existing literature has a tendency to emphasize psychometric characteristics, reliability, and traditional test validity while disregarding stakeholder input (Im, Shin, & Cheng, 2019). Consequently, traditional quantitative methodologies predominate in validity research (Liskinasih, 2016; Hashemi & Daneshfar, 2018; Furwana, 2019). However, Bonner and Chen (2019) caution that research findings on validity may not be entirely applicable to classroom assessments due to their specific requirements.

Several frameworks for validation have emerged to elucidate the concept of validity and establish systematic validation methods. Notable approaches to testing validity include Argument-Based Validity (ABV) by Kane (2006, 2013, 2017), the Integrated Framework for Construct Validity by Embreston (2007, 2008, 2017), and the recently developed Accountability Interpretive Use Argument (IUA) validity framework by Hoeve (2022). Hoeve's IUA framework, designed with the American context in mind, requires further examination of its applicability in other contexts, such as higher education in Oman, to assure robust, systematic procedures (Hoeve, 2022).

Research employing qualitative methodologies to investigate validity emphasized validity from multiple perspectives. Some studies focused either on the perspectives of test takers (e.g., Cheng & DeLuca, 2011; Sato & Ikeda, 2015; Zhan & Wan, 2016; and Hamid, Hardy, & Reyes, 2019) or test designers (e.g., So, 2014; Buckley-Walker & Lipscombe, 2022; and Al Lawati, 2023), as well as the scrutiny carried out by researchers (e.g. Al Fraidan, 2019; and Bax and Chan, 2019). The perceptions of exam committee members, who play vital roles in ensuring test validity, are rarely addressed in these studies. In addition, studies examining the practices of these stakeholders in Omani higher education institutions are uncommon. Some studies focused on a type of validity (face validity like in the studies of Tsagari, 2014, and Sato and Ikeda, 2015; consequential validity like in the study of Saglam and Tsagari, 2022; construct validity like in the studies of Xie, 2011; Sun, Wan, and Kim, 2022; criterion validity like in the study of Clemente et al., 2022); or different types of validity (e.g., Pellegrino, DiBello & Goldman, 2016; Runalika et al., 2023). Some studies used a particular test validity framework. Al-Buraiki's (2020), for example, study employs Weir's socio-cognitive framework, which was developed in 2005, to analyze the overall validation procedure of the reading questions in the Oman General Education Diploma of English Language Test (GEDELT) for the academic year 2016–2017 using a checklist and document analysis. Weir (2005) delineated five distinct categories of validity, namely: context validity, theory-based validity, scoring validity, consequential validity, and criterion-related validity. Several other studies have incorporated widely recognized and significant contributions towards establishing validity, without adhering to a particular framework. For example, Buckley-Walker and Lipscombe (2022) argue that instructors' assessment processes must be thoroughly examined before analyzing classroom assessment. The educators engaged in discourse that centered on overarching concepts that contribute to the establishment of validity, which include: (1) alignment with curriculum and instruction, (2) catering for student abilities, (3) the scoring rubric; and (4) using CA data to meet students' needs.

Chapelle (2012) argued that the notion of validity as an argument places significant emphasis on the involvement of the socio-academic community. "*if validity entails demonstrating the meaning of test scores and justifying their use, the issues are how one goes about doing this and who is responsible for getting it done. In other words, what are the rules of the validity*

game?" (p. 21). Chapelle and Lee (2021) present an extensive overview of argument-based validation in the context of language testing. They examine the fundamental elements of a validity argument and explore various factors that may pose a risk to validity, along with strategies to mitigate them. Bai (2020) argues that validity studies in the domain of language testing should take into account the complex and evaluative relationships between factors such as test takers' motivation to learn, their attitudes toward test use, and other test-related elements in relation to their test performance in order to assist test users and other interested parties in making equitable decisions based on test scores, promoting positive outcomes, and ensuring test accountability. That is to say, one of the fundamental elements of test validity, that is directly related to this study, is the multi-part argument about the interpretation and use of the test scores.

2.1. Accountability Interpretation and Use Argument (IUA) Validity Framework (Hoeve, 2022)

Hoeve's framework provides valuable contributions and practical implications for test validation, rectifying a deficiency in Embreston's framework, which disregards the significance of aggregate scores and their implications for the accountability system. As outlined by Hoeve, the IUA framework provides a standard procedure for authenticating tests, beginning with the identification of the intended interpretation and applications of the test and test scores. This emphasizes the significance of contemplating the test's intended purposes and applications during the design phase. The validity of inferences and actions based on test scores should be adequately supported by identifying the evidence required to support these interpretations.

Significantly, the IUA framework acknowledges the accountability system's function in test validation. It suggests collecting evidence for both student-centered and group-centered factors to support the validity of conclusions drawn from accountability indicator data. This indicates that the framework recognizes the importance of both individual and aggregate scores and emphasizes the need for test developers and policymakers to collaborate.

The IUA framework promotes stakeholder collaboration by integrating the accountability system into the validation procedure. This ensures that testing needs and objectives are aligned, resulting in a more consistent and locally pertinent interpretation of test validity, especially in the context of higher education in the Sultanate of Oman.

The integration of Hoeve's Accountability Interpretive Use Argument (IUA) validity framework provides additional insights into resolving the shortcomings of existing models and highlights the need to consider both individual and aggregate scores within the accountability system. This collaborative approach between test developers and policymakers can lead to a more robust and contextually relevant interpretation of test validity.

As Moss (2013) suggests, it is essential to evaluate the framework's applicability in real-world settings. Different stakeholders may have diverse data requirements and interpretations of test validity, which should be addressed in a transparent manner during the validation process. Effective implementation of the IUA framework and attainment of a shared understanding of test validity require policymakers, test developers, and instructors to have a shared understanding of test validity.

In this research, the perspectives of policymakers and instructors will be explored to determine the best methods to implement Hoeve's framework in practice. The purpose of this study is to contribute to the practical application of the IUA framework and its congruence with the context of Omani higher education by examining their perspectives and experiences. This research will cast light on the framework's strengths and weaknesses and provide recommendations for its successful implementation in the field of language assessment in Oman.

3. Methodology

In order to better understand how exam committee members and policymakers in Oman's higher education institutions actually ensure test validity and accountability, the current study employs a qualitative case study approach using semi-structured interviews. The researcher made the decision to use this method of data collection as the qualitative method is better used for comprehending social phenomena like people's views, beliefs, experiences, attitudes, behavior, and interactions, as well as for viewing the data more extensively (Banister et al., 1994; Pathak, Jena, & Kalra, 2013). Thematic analysis was used to identify patterns and themes from the elicited data (Braun & Clark, 2006) in a pragmatic and reflexive manner whilst placing the needs of the local context at the heart of the research (Braun, Clark & Hayfield, 2022).

3.1. Study Design and Participants

This study selected participants from four universities in Oman, including both public and private institutions, to ensure a diverse sample. From the cohort, fifteen individuals were chosen, including ten members of the examination committee and five policymakers. The selected universities all received a score of 2 on criterion 2.8 (Assessment Methods, Standards, and Moderation), ensuring a consistent foundation for handling sensitive data. This diversified yet interrelated participation will facilitate a comprehensive understanding of assessment procedures and cast light on the practices of test validity and accountability within the Omani higher education framework.

Purposeful sampling was used to identify individuals directly involved in shaping the design and validation of teachers' tests and making important decisions based on test scores. In light of the limited number of exam committee members and policymakers within these institutions, fifteen was regarded as a sufficient sample size for attaining data saturation in this context. This is consistent with the opinion of researchers like Bertaux (1981), who contend that fifteen participants are sufficient for qualitative research studies. It is essential to observe that the members of the examination committee also teach at their respective institutions.

The responsibilities of participants within their respective institutions have a significant impact on the assessment procedure. Exam committee members, who are also instructors, perform essential academic and assessment duties that are only stated within the institution and are not generalized to all institutions. Their primary responsibility is to evaluate and approve mid-semester and final examinations administered by instructors.

On the other hand, the policymakers, as members of the institution's council, are responsible for approving students' evaluations after department councils have given their approval. The diverse yet interdependent roles of these participants guarantee a comprehensive

comprehension of the assessment procedures, casting light on the practices of test validity and accountability in these Omani higher education institutions.

3.2. Data Collection

This qualitative investigation was based on a methodology of semi-structured interviews. The interview questions were derived from Hoeve's (2022) Accountability IUA validity framework, ensuring a solid and trustworthy foundation for the interviews. These queries were then divided into two categories: one for evaluation committee members and one for policymakers. The interviews were conducted online via Zoom and lasted between 40 and 50 minutes, providing a comprehensive look at the experiences and perspectives of the participants.

The interviews were digitally recorded and then transcribed using Otter.ai, an online transcription service, which assisted in converting the spoken words into text and thereby facilitated data analysis. This process was conducted over the course of three months, yielding a large corpus of data for subsequent analysis.

3.3. Data Analysis

NVivo, software for qualitative data analysis, was utilized for the data analysis. Utilizing Braun & Clarke's (2006) six-step procedure, a systematic and exhaustive analysis of the data was conducted. This process began with acquainting oneself with the data, was followed by initial classification, the search for themes, the review of themes, the definition and naming of themes, and ultimately the production of the report.

The initial phase of coding consisted of perusing through the transcripts and labeling significant sections with pertinent codes. Based on their similarities, the codes were then categorized into potential themes. These potential themes were evaluated, refined, and renamed to reflect their underlying concept.

NVivo was used for categorizing and identifying themes, and the entire process was routinely double-checked for consistency and accuracy. This process of double-checking ensured the accuracy of the analysis and enhanced the credibility of the research findings.

4. *Results*

The analysis of the interviews with members of the examination committees and policymakers at Oman's universities has yielded a number of significant findings regarding the approach to testing validity and accountability in these institutions. Five major themes emerged from the data, each revealing significant aspects of the current assessment landscape.

4.1. Assessment Literacy

The first important conclusion concerns the significance of assessment literacy. All participants highlighted the significance of instructors' ability to devise legitimate assessments and accurately interpret test results. According to one participant, *"... a teacher may have a PhD in linguistics, but that doesn't mean that that teacher knows much about education… We presuppose that everybody who teaches in universities is a teacher, that's not*

correct. Very few teachers in education are actually teachers..." (Participant S). Another example is "*.. somebody may have a PhD in education, but if that person doesn't have a PhD in assessment or a PhD in curriculum, it doesn't mean that that person understands how the curriculum works. It doesn't mean that I don't want to say that the person is a bad teacher or something as the person may be the best in the college"* (Participant Q). The comments highlighted the distinction between academic knowledge and pedagogical and assessment expertise. This understanding was shared by all participants, indicating a shared belief in the need for assessment-specific training or professional development. Regarding writing on a test blueprint, which makes it easier to match different skills with the course material and the right type of evaluation, which increases its validity (Patil, et. al., 2015; Raymond & Grande, 2019), all exam committee members mentioned that when they receive exams to be reviewed, they are not attached with blueprints or any certain details like objectives and question types. M. mentioned that " *teachers only bear these details in mind when designing their exams"*. That is due to the lack of guidelines from exam committee members themselves and policymakers to attach blueprints along with exams for reviewing, as stated by one of the exam committee members, *"We can review only what they give us."*

4.2. Professional Development in Assessment

A second major theme was the significance of professional development in assessment. Participants suggested that instructors would benefit from seminars that facilitate discussions on curriculum development, assessment, and other crucial issues. For instance, one participant stated, *"...I believe we should hold seminars that include curriculum-related dialogues about how to plan and evaluate courses. We should not assume that everyone knows this"* (Participant S).

4.3. Teacher Autonomy

Regarding teacher autonomy, a third main motif emerged. Teachers at these universities resisted external evaluation or review of their evaluations, citing the uniqueness of their courses and their specialized knowledge. For instance, one participant stated*, "We are problematic people, we do not like to follow instructions, and teachers do not follow instructions. And we will always respond affirmatively, but my course is unique. This is my coursework"* (Participant M). Another quotation by one participant is "*We are difficult people, teachers, so it's I don't feel comfortable sometimes telling each teacher as a teacher. I find that your exam is too one-sided. Because we don't have rules. So for example, in, we I sent, I sent a couple of times a model, using Bloom's Taxonomy and allotting the marks in line with Bloom's Taxonomy, only 10% from memory, maybe 20% for application, blah, blah, blah, blah. So, if we did this, only very few students would have an A, which would be the normal situation, only a student that has met the whole has gone up the ladder of Bloom's Taxonomy should get A"*. This finding suggests a potential barrier to the institution-wide implementation of standardized practices to ensure test validity.

4.4. Teacher Collaboration

However, the fourth finding revealed that a substantial quantity of teacher collaboration is occurring. One participant explained that instructors of the same course collaborate in the test creation process: *"We have a coordinator for the course, so I was the coordinator last term, and we sit together and put together the exam. And we ensure we are aligned with the learning outcomes at the same level as the students."* (Participant M).

4.5. Institutional Framework for Assessment

The need for an institutional framework was the fifth recurring motif. Participants emphasized the significance of a precise, well-communicated set of assessment guidelines or frameworks. They believed that the current practice was less formal and lacked specific directives: "*We are simply managing the situation." Therefore, I continue to assert that we require an institutional framework. That is evident. It must originate from the bottom up, from us through consultation, and also from experts. This is then the framework, which we adhere to"* (Participant F).

The interviews revealed that teachers play a crucial role in policymaking, which is an intriguing finding given that this is traditionally the responsibility of administrators and policymakers. One policymaker participant stated, *"We refer to the teacher's work/test if we suspect that the students' grades are uniform."* The data revealed that instructors have considerable control over their assessments, suggesting that they play a larger role than previously believed in ensuring test validity and accountability.

These results disclose a complex picture of test validity and accountability in universities in Oman. While there are challenges associated with assessment literacy, professional development requirements, and institutional guidelines, there is evidence of effective collaboration among teachers, and teachers play a significant role in policy-making. Future efforts to enhance practices in these areas should take these aspects into consideration.

5. Discussion

This discussion sheds light on the answers to the research concerns, illuminating how exam committee members and policymakers ensure test validity and accountability.

5.1. Ensuring Test Validity: Exam Committee Members' Perspective

The findings indicate that exam committee members encounter substantial obstacles in ensuring test validity due to a variety of factors. First and foremost, the data indicate that not all instructors possess the pedagogical expertise required to design valid and reliable assessments, corroborating the results of previous research (Stiggins, 2004; Xu & Brown, 2017). Members of the examination committee appeared to perceive this difficulty and express the need for additional assessment literacy training and seminars.

In addition, examination committee members appeared to grapple with teachers' resistance to having their assessments reviewed, which is consistent with findings from a larger body of research on professional autonomy and resistance in education (Ingersoll, 2006). This opposition appears to hinder the examination committee's ability to assure the validity of institution-wide assessments.

Nonetheless, it was also discovered that exam committee members and instructors engage in some collaborative processes when constructing exams. This is encouraging and in line with research (Voogt, Pieters, & Handelzalts, 2016) highlighting the benefits of teacher collaboration in devising assessments. That might be related to the different terms held by different people, "reviewing from the exam committee’ and ‘discussing from other teachers".

5.2. Ensuring Test Validity and Accountability: Policymakers' Perspective

The role of policymakers in ensuring test validity appears less clear-cut. The data indicate that policymakers rely significantly on teachers and examination committee members, suggesting a lack of active participation in the validation process. They appeared to be more involved in problem-solving and data collection for quality assurance.

Their reliance on instructors and examination committees may be indicative of systemic deficiencies. Policymakers have expressed the need for additional training to better support instructors and stakeholders in the assessment process, indicating that they may be unprepared to carry out their responsibility of ensuring test validity.

In accordance with research on teacher leadership (York-Barr and Duke, 2004; Danielson, 2007), it has been observed that teachers play a central role in shaping test validity policies. Even though this finding is encouraging, there is cause for concern if teachers lack the assessment literacy required to make informed decisions about test design and validity.

5.3. Proposed Interpretive Use Argument (IUA) Framework for Accountability

This study proposes an Accountability Interpretive Use Argument (IUA) framework specific to the Omani context in order to resolve these challenges and improve the assessment procedure. This framework recognizes the central role of instructors in ensuring test validity, the supporting roles of policymakers and exam committee members, and the critical need for ongoing professional development in the assessment field. The IUA framework encourages an iterative validation process initiated by teachers and supported by policymakers and exam committee members, with a focus on effective communication and collaboration to ensure valid assessments and accountability.

Implementing the IUA framework could potentially improve the administration of accountability and test validity in Oman's higher education system. It could cultivate a culture of accountability that respects disciplinary norms and local customs, thereby encouraging continuous progress. Future research should seek to validate and alter this framework for use in a variety of educational settings.

This study examines the extant obstacles and possible solutions for ensuring test validity and accountability in Oman's higher education institutions. By addressing these challenges and implementing the proposed framework, institutions will be able to improve the validity of assessments, cultivate stakeholder collaboration, and promote effective assessment practices that are aligned with global standards. This study contributes to the field of educational assessment in Oman and possibly beyond by providing valuable insights and proposing a context-specific assessment framework.

6. Pedagogical Implications

The findings of this study and the existing literature suggest a number of universally applicable, yet Omani-specific pedagogical implications:

1. The critical role of teachers in test design and ensuring validity necessitates an in-depth understanding of assessment procedures (Sultana, 2019; Stiggins, 2004). Therefore, emphasis should be placed on introducing professional development programs geared toward enhancing the assessment literacy of teachers.

2. Given policymakers' critical role in ensuring test validity and accountability, their participation in initiatives to improve assessment literacy is essential. Their participation can contribute to the validation of test design and scoring procedures, as well as cultivate a nuanced comprehension of the complexities of test validity and accountability.

3. Validation and accountability procedures require effective communication and collaboration between instructors, exam committee members, and policymakers. This research demonstrates that a lack of precise guidelines and communication hinders the validation process. These issues can be mitigated by adopting a systematic validation approach in which responsibilities are clearly defined and understood.

4. Encouraging Parallel Validation Processes: Although instructors frequently use their own validation methods, a parallel, collective validation process should be encouraged to maintain consistency in assessment criteria without compromising individual autonomy.

5. Iterative IUA Validity Framework Implementation: According to Hoeve (2022), the Accountability Interpretive Use Argument (IUA) validity framework should be iterative. Teachers should initiate the validation procedure, and the process should include a continuous feedback cycle. Policymakers should validate the process before delineating the consequences, whereas evaluators can validate in the opposite direction, with the two groups meeting in the middle to determine the consequences.

These implications, which resolve the identified challenges in the Omani context, can have far-reaching benefits in the field of education. When instituting adjustments to pedagogical practices, it is essential to take into account the specific context and requirements of educational institutions.

7. Conclusion

This study illuminates the crucial role instructors play in ensuring test validity and accountability in Oman's higher education sector. In addition, it emphasizes the need for knowledgeable and well-prepared policymakers who can guide the process in collaboration with instructors and exam committee members. The proposed modified Accountability Interpretive Use Argument (IUA) framework provides an innovative strategy for promoting knowledge exchange, nurturing consensus, and augmenting assessment validity in Oman's higher education sector.

The fluid nature of validity, which is a dynamic process requiring the active participation of various stakeholders, is a key finding of this study. The active involvement of instructors as evaluators and policymakers as validators is essential for optimal test validity. In addition, the validation process should be conceptualized as a collaborative, two-way endeavor in which decisions and repercussions are deliberated upon after extensive consultation.

Based on these findings, this study advocates for instructors, exam committee members, and policymakers to participate in assessment practices-centered seminars. Such seminars could promote enhanced comprehension, stimulate validation practices, and promote assessment uniformity. As Chapelle (2021) suggests, it is possible to cultivate a culture of test validity

that adheres to disciplinary standards, regional traditions, and the philosophy of continuous improvement.

The study also suggests the introduction of assessment qualification certificates as a means of recognizing the proficiency of those involved in the process of test validation. In accordance with Oman's Vision 2040, such recognition could further professionalize the education sector by highlighting essential pedagogical competencies.

Although the context of this study was uniquely Omani and the sample size was relatively small, it provides the groundwork for future research in other contexts. Additional research could substantiate the applicability and adaptability of the Accountability Interpretive Use Argument (IUA) framework across diverse educational environments and geographies.

Funding

This paper is funded by the University of Technology and Applied Sciences, Rustaq College of Education.

Abbreviations:
ABV: Argument-Based Validity approach
CA: Classroom assessment
HEIs: Higher Educational Institutions
HoD: Head of Department
IUA: Accountability Interpretive Use Argument
OAAAQA: Oman Authority for Academic Accreditation and Quality Assurance of Education

References

AERA, APA, & NCME (1999). Standards for educational and psychological testing. Washington, D.C.

Al-Buraiki, S. (2020). Establishing the Validity of the Reading Questions in a Centralized Test Using Weir Socio-Cognitive Framework. Journal of Educational and Psychological Studies (JEPS), 14 (4), pp. 642-655. https://doi.org/10.53543/jeps.vol14iss4pp642-655

Bai, Y. (2020). The relationship of test takers' learning motivation, attitudes towards the actual test use and test performance of the College English Test in China. Lang Test Asia 10, 10. https://doi.org/10.1186/s40468-020-00108-z

Banister, P., Burman, E., Parker, I., Taylor, M. & Tindall, C. (1994). Qualitative methods in psychology: a research guide. Open University Press。

Bertaux, D. (1981). From the life-history approach to the transformation of sociological practice. Biography and society: The life history approach in the social sciences, 29-45.

Bonner, S., & Chen, P. (2019). Systematic classroom assessment: An approach for regulated learning and self-regulation. Routledge. https://doi.org/10.4324/9781315123127

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., Clarke, V., & Hayfield, N. (2022). 'A starting point for your journey, not a map': Nikki Hayfield in conversation with Virginia Braun and Victoria Clarke about thematic analysis. Qualitative Research in Psychology, 19(2), 424-445. https://doi.org/10.1080/14780887.2019.1670765

Brown, H. D., & Abeywickrama, P. (2010). Language assessment: Principles and classroom practices (2nd ed.). Pearson Longman.

Brumen, M., & Cagran, B. (2011). Teachers' perspectives and practices in assessing young foreign language learners in three Eastern European countries. Education, 3-13, 39(5), 541-559. https://doi.org/10.1080/03004279.2010.488243

Buckley-Walker, K., Lipscombe, K. (2022). Validity and the design of classroom assessment in teacher teams. Aust. Educ. Res. 49, 425–444. https://doi.org/10.1007/s13384-021-00437-9

Clemente, F. M., Praça, G., Oliveira, R., Aquino, R., Araújo, R., Silva, R., Sarmento, H., & Afonso, J. (2022). A systematic review of the criterion validity and reliability of technical and tactical field-based tests in soccer. International Journal of Sports Science & Coaching, 17(6), 1462–1487. https://doi.org/10.1177/17479541221085236

Chapelle, C. A. & Lee, H. (2021). Conceptions of Validity. In G. Fulcher & F. Davidson (Eds.) The Routledge Handbook of Language Testing (2nd Ed.). Routledge pp. 21-33. https://doi.org/10.4324/9781003220756

Chapelle, C. A. (1999). Validity in language assessment. Annual Review of Applied Linguistics,19, 254-272.

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.) The Routledge Handbook of Language Testing. Routledge pp. 21-33. Accessed on: 15 May 2023. https://www.routledgehandbooks.com/doi/10.4324/9780203181287.ch1

Cheng, L. & DeLuca, C. (2011). Voices From Test-Takers: Further Evidence for Language Assessment Validation and Use. Educational Assessment, 16:2, 104-122, DOI:10.1080/10627197.2011.584042

Coombe, C., Davidson, P., O'Sullivan, B., & Stoynoff, S. (2012). The Cambridge Guide to Second Language Assessment. Cambridge: Cambridge University Press.

Coombe, C., Vafadar, H. & Mohebbi, H. (2020). Language assessment literacy: what do we need to learn, unlearn, and relearn? Lang Test Asia 10, 3. https://doi.org/10.1186/s40468-020-00101-6

Cronbach, L.J. (1988). Five perspectives on validation argument. In H. Wainer and H. Braun(eds.) Test validity (pp. 3-17). Hillsdale, NJ: L.Erlbaum.

Danielson, C. (2007). Enhancing professional practice: A framework for teaching (2nd ed.). Alexandria, VA: ASCD.

Davies, A. (ed.) (1968). Language testing symposium: A Psycholinguistic Approach. Language and Language Learning [Series], No. 21. London: Oxford University Press. Partial It. transl. in: Amato, A. (a cura di), Il testing nella didattica linguistica. Roma: Bulzoni, 1974.

Denman, C., Al-Mahrooqi, R. (2018). Teachers' Attitudes Toward Alternative Assessment in the English Language Foundation Program of an Omani University. In: Al-Mahrooqi, R., Denman, C. (eds) English Education in Oman. English Language Education, vol 15. Springer, Singapore. https://doi.org/10.1007/978-981-13-0265-7_4

Embretson, S. E. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? Educational Researcher, 36(8), 449–455. http://www.jstor.org/stable/4621099

Furwana, D. (2019). Validity and Reliability of Teacher-Made English Summative Test at Second Grade of Vocational High School 2 Palopo. Language Circle: Journal of Language and Literature, 13.

Glass, R. & Metternich, J. (2020). Method to measure competencies - a concept for development, design and validation. Procedia Manufacturing, 45, pp. 37–42 https://doi.org/10.1016/j.promfg.2020.04.056

Hamid, M.O., Hardy, I. & Reyes, V. (2019). Test-takers' perspectives on a global test of English: questions of fairness, justice and validity. Lang Test Asia 9, 16 (2019). https://doi.org/10.1186/s40468-019-0092-9

Hashemi, A., & Daneshfar, S. (2018). A Review of the IELTS Test: Focus on Validity, Reliability, and Washback. IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics).

Hoeve, K. B. (2022). A validity framework for accountability: educational measurement and language testing. Lang Test Asia 12, 3. https://doi.org/10.1186/s40468-021-00153-2

Hughes, D. (2018). Psychometric Validity: Establishing the Accuracy and Appropriateness of psychometric measures. In P. Irwing, T. Booth, D. J. Hughes (Eds.) The Wiley Handbook of Psychometric Testing: A Multidisciplinary Approach to Survey, Scale and Test Development John Wiley & Sons Ltd. https://doi.org/10.1002/9781118489772

Im, GH., Shin, D. & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. Lang Test Asia 9, 14. https://doi.org/10.1186/s40468-019-0089-4

Ingersoll, R. M. (2006). Who controls teachers` work? Harvard University Press.

Kane, M. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38 (4),319-342.

Lado, R. (1961). Language testing. New York: McGraw-Hill.

Latif, M.W., Wasim, A. (2022). Teacher beliefs, personal theories and conceptions of assessment literacy: a tertiary EFL perspective. Lang Test Asia 12, 11. https://doi.org/10.1186/s40468-022-00158-5

Liskinasih, Ayu. (2016). The Validity Evidence of Toefl Test as Placement Test. Jurnal Ilmiah Bahasa dan Sastra Unikama, vol. 3, no. 2, pp. 173-180.

Lundahl, C. (2009). Varför nationella prov? Framväxt, dilemman, möjligheter. Lund: Studentlitteratur AB.

McMillan, J. H. (2013). Why we need research on classroom assessment. In J. H. McMillan (Ed.), SAGE handbook of research on classroom assessment (pp. 3–16). SAGE Publications Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). Macmillan.

Moss, P. A. (2013). Validity in Action: Lessons from Studies of Data Use. Journal of Educational Measurement. 50(1) pp.91-98. https://doi.org/10.1111/jedm.12003

Patil, S. Y., Gosavi, M., Bannur, H. B., & Ratnakar, A. (2015). Blueprinting in assessment: A tool to increase the validity of undergraduate written examinations in pathology. International journal of applied & basic medical research, 5(Suppl 1), S76–S79. https://doi.org/10.4103/2229-516X.162286

Pathak, V., Jena, B., & Kalra, S. (2013). Qualitative research. Perspectives in Clinical Research, 4(3), 192. https://doi.org/10.4103/2229-3485.11538

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. Educational Psychologist, 51(1), 59–81. http://dx.doi.org/10.1080/00461520.2016.1145550

Plake, B. S. (1993). Teacher assessment literacy: teachers' competencies in the educational assessment of students. Mid-Western Educational Researcher, 6(2), 21.

Raymond, M. R., & Grande, J. P. (2019). A practical guide to test blueprinting. Medical Teacher, 41(8), 854–861. https://doi.org/10.1080/0142159X.2019.1595556

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: a primer for discipline-based education researchers. CBE life sciences education, 15(1), rm1. https://doi.org/10.1187/cbe.15-08-0183

Runalika, R.; Gautham Melur, S.; Mariamma, P. & Gururaj, G. (2023). Face, content, criterion and construct validity assessment of a newly developed tool to assess and classify work–related stress (TAWS– 16). PLoS ONE, Vol. 17 Issue 1, p1-11. 11p. https://doi.org/10.1371/journal.pone.0280189

Saglam, G. & Tsagari, D. (2022). Evaluating Perceptions towards the Consequential Validity of Integrated Language Proficiency Assessment. Languages 7: 65. https://doi.org/10.3390/languages7010065

Sato, T. & Ikeda, N. (2015). Test-taker perception of what test items measure: a potential impact of face validity on student learning. Language Testing in Asia **5**, 10. https://doi.org/10.1186/s40468-015-0019-z

Smith, W. C. (2016). An introduction to the global testing culture. In W. C. Smith (Ed.), The global testing culture: Shaping educational policy, perceptions, and practice (pp. 7–24). Oxford: Symposium Books.

Sireci, S & Faulkner-Bond, M. (2014). Validity evidence based on test content. Psicothema, 26, 1, 100-107. https://doi.org/10.7334/psicothema2013.256.

So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale International English Test. Language Assessment Quarterly, 11:3, 283-303, DOI: 10.1080/15434303.2014.936936

Sultana, N. (2019). Language assessment literacy: an uncharted area for the English language teachers in Bangladesh. Lang Test Asia, 9, 1. https://doi.org/10.1186/s40468-019-0077-8

Sun, H. & Zhang, J. (2022). Assessment literacy of college EFL teachers in China: Status quo and mediating factors. Studies in Educational Evaluation, 74, https://doi.org/10.1016/j.stueduc.2022.101157

Sun, T., Wang, C. & Kim, S.Y. (2022). Psychometric properties of an English Writing Self-Efficacy scale: aspects of construct validity. Read Writ 35, 743–766. https://doi-org.squ.idm.oclc.org/10.1007/s11145-021-10206-w

Stiggins, R. (2004). New assessment beliefs for a new school mission. Phi Delta Kappan, 86(1), 22-28.

Tsagari, D. (2014). 'Investigating the face validity of Cambridge English First in the Cypriot context'. Research Notes 57: 23–31. Available online: http://www.cambridgeenglish.org/images/177881-research-notes-57-document.pdf (accessed on 15 May 2023).

Voogt, J. M, Pieters, J. M. & Handelzalts, A. (2016). Teacher collaboration in curriculum design teams: effects, mechanisms, and conditions. Educational Research and Evaluation, 22:3-4, 121-140, DOI: 10.1080/13803611.2016.1247725

Weir, C. J. (2005). Language test validation: An evidence-based approach. Oxford: Palgrave.

Xie, Q. (2011)。 Is Test Taker Perception of Assessment Related to Construct Validity?, International Journal of Testing, 11:4, 324-348, DOI: 10.1080/15305058.2011.589018

Xu, Y., & Brown, G. T. (2017). University English teacher assessment literacy: A survey-test report from China. Papers in Language Testing and Assessment, 6 (1), 133-158.

York-Barr, J., & Duke, K. (2004). What Do We Know about Teacher Leadership? Findings from Two Decades of Scholarship. Review of Educational Research, 74(3), 255–316. http://www.jstor.org/stable/3516026

Zhan, Y., & Wan, Z. H. (2016). Test Takers' Beliefs and Experiences of a High-stakes Computer-based English Listening and Speaking Test. RELC Journal, 47(3), 363–376. https://doi.org/10.1177/0033688216631174