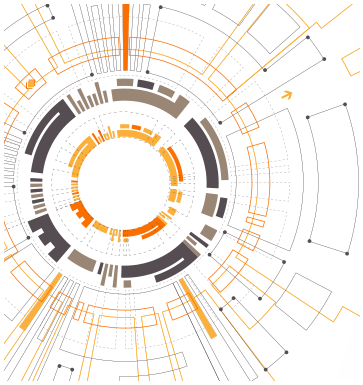
Abstract

This paper aims to synthesize measures of assessment literacy in higher education by forging a connection between two research domains: educational assessment and psychometrics. It begins with a systematic review of assessment literacy measures within the context of higher education published within the last ten years. AL measures, including tests of assessment literacy, self-report measures such as inventories, surveys, and rubrics in assessment literacy studies, were reviewed. Psychometric properties of the measures were evaluated against standards related to validity and reliability. Across a number of 11 measures reviewed, we found that while the reviewed studies demonstrated strong adherence to rigorous validation processes, the psychometric evidence presented for the available measures is neither complete nor up to date, concerning researchers' and educators' needs in terms of assessment. Nearly all measures were grounded in substantial literature reviews and expert evaluations, with five measures providing detailed content validity evidence and several studies reporting good fit indices for internal structure analyses. Reliability evidence was generally robust, with most Cronbach's Alpha coefficients ranging between .79 and .94, and high reliability indices reported in Rasch measurement theory studies. Despite these strengths, the identified gaps highlight the need for establishing psychometrically sound, comprehensive, and up-to-date assessment literacy measures. The paper concludes by discussing the development of enhanced assessment literacy measures that are adaptive to the changing landscape of assessment, and their implications for policy and practice in higher education.



AUTHORS

Beyza Aksu Dunya, Ph.D.
University of Alabama

Mehmet Can Demir, Ph.D.
Bartin University

Stefanie Wind, Ph.D.
University of Alabama

A Systematic Review of the International Assessment Literacy Measures in Higher Education (2013–2023)

Assessment literacy (AL) refers to an individual's understanding and use of assessment concepts and procedures (Coombs & DeLuca, 2022; DeLuca et al., 2019; Popham, 2011). As a construct, AL has a strong theoretical background rooted in well-established standards (American Federation of Teachers [AFT], National Council on Measurement in Education [NCME], & National Education Association [NEA], 1990) and research (e.g., Stiggins, 1991). The definition of AL has evolved over time to extend beyond the knowledge of the Standards for Educational and Psychological Testing (AERA et al., 2014), with a recent emphasis on socio-cultural and historical context that shapes assessment practices and beliefs. Even as the specific definition of AL has evolved over time, researchers have consistently recognized this construct as a central component of effective teaching across contexts, grade levels, and disciplinary areas (e.g., Danielson, 2013; Gotch & French, 2014).

In previous studies related to AL, researchers have mostly focused on the development, measurement, and perceptions related to this construct among in-service and pre-service primary and secondary education teachers (e.g., Alkharusi, 2011; Gotch & French, 2011; Plake et al., 1993; Zhang & Burry-Stock, 2003). Despite this emphasis, it is important to recognize that AL has critical implications in higher education, where assessment tasks are often complex (Friesen, 2022). As is true in primary and secondary educational contexts, the effectiveness of assessment practices within and beyond the classroom in higher education settings depends heavily on faculty involvement and proficiency in assessment efforts (Ray

CORRESPONDENCE
Email
baksu@ua.edu

et al., 2012). Despite the central role of assessment in higher education, faculty members often have little formal training related to assessment (Knapper, 2010), which may result in minimal integration of formative feedback and learner-driven assessments in their instruction (Massey et al., 2020).

Reflecting this lack of training, higher education in general has been repeatedly critiqued for emphasizing exam-based summative measures of student achievement while neglecting other potentially useful assessment techniques (Yorke, 2003). In response, many higher education institutions have launched faculty development units that offer training programs to support faculty's curriculum planning, instruction, and assessment practices (Taylor & Colet, 2010). Those training programs include efforts to enhance faculty members' conceptions and practices in student assessment using approaches that integrate summative measures, formative feedback, and dialogical assessment structures to advance learning (Nicol & Macfarlane-Dick, 2006). Today, higher education institutions incorporate a variety of assessment types, focusing not just on outcomes but also on processes and experiences. They are increasingly embracing technological innovations such as artificial intelligence (AI) to create a more detailed and accurate picture of how institutions are meeting their objectives (Watermark, 2023). Organizations like the Association for the Assessment of Learning in Higher Education (AALHE) offer events and webinars to support ongoing faculty development in AL. Despite these advancements, there is still no widely recognized, commonly used measure of AL specifically designed for faculty members. Establishing what AL means for a faculty member in the contemporary higher education sector and how we measure it is necessary to continue improving educational outcomes. With this in mind, the primary objective is to review and revise the psychometric properties of existing AL tools, drawing on a comprehensive analysis of recent literature in the field. This review not only evaluates these tools against a set of predefined criteria but also identifies gaps and areas for improvement. The findings will inform the development of enhanced AL measures that are adaptive to the changing landscape of assessment. In this context, we propose to explore the concept of *Artificial Intelligence Assessment Literacy*, which integrates AI technologies to refine and optimize the assessment process. This exploration is intended to pave the way for future research and practical implementations that respond effectively to the dynamic nature of educational assessment.

Despite the central role of assessment in higher education, faculty members often have little formal training related to assessment…

Theoretical Framework

Our framework draws extensively from the fundamental conceptual foundations of AL, prominently incorporating Stiggins' (1991) model. This model defines AL as the foundational comprehension of educational assessment combined with the skills required to apply this knowledge effectively across various measures of student achievement. Specifically, Stiggins identifies five standards of effective assessment that assessment-literate individuals should achieve (Stiggins, 1991). We selected Stiggins' model because it provides a comprehensive and widely recognized framework that emphasizes both conceptual understanding and practical application, aligning well with the goals of our study:

1. Formulating Precise Assessment Objectives: Focusing on defining clear intentions for what each assessment aims to evaluate. This involves recognizing the various functions assessments serve at the instructional level, such as discerning individual student needs, evaluating group dynamics within the classroom, organizing students into appropriate learning groups, and assigning grades.

2. Focusing on Achievement Targets: Individuals who are literate in the core principles of effective assessment recognize that students must achieve various interrelated objectives. These include mastering content knowledge, acquiring performance skills, and creating high-quality products.

3. Selecting Proper Assessment Methods: Assessment-literate educators understand the appropriate times and methods to employ various assessment techniques within these categories: selected response, essay, performance assessment, and personal communication (i.e. discussions, interviews).

4. Sampling student achievement: Any assessment is essentially a sample from a broader set of potential questions that could be asked if the assessment had no length constraints. It should include a sufficient number of questions to ensure that the assessment accurately represents this wider range of possibilities.

5. Avoiding bias and distortion: An assessment-literate educator must consistntly be alert to various specific technical and practical issues that could lead to inaccuracies and bias in measuring student achievement.

In summary, as introduced by Stiggins (1991), the concept of AL outlines that sound assessments rest upon five essential standards: they originate from and reinforce clearly articulated purposes, emerge from and conform to well-defined achievement targets, rely on the appropriate employment of assessment methods, efficiently sample student performance, and proactively prevent and eliminate bias and distortion. Over time, the idea and meaning of AL has evolved from merely being a collection of skills and knowledge to being viewed as a social practice. Drawing on Bernstein's sociocultural theory, Willis et al. (2013) described AL as a social practice characterized by interactions between teachers and students, encompassing collaborative engagement and shared understanding. More recently, Pastore and Andrade (2019) provided a theoretical definition of AL as a context-dependent concept with multiple components that integrate social, cultural, policy, and professional factors.

The significance of AL within higher education is underscored by its capacity to empower educators with the competence needed to create, administer, and interpret assessments effectively. Our proposed framework's content and structure are shaped by a synthesis of prior studies and our experiences as psychometricians and educational measurement professionals. The overarching research questions for this review are: "*How is AL measured in the higher education context*?" and "*What is the psychometric evidence presented for the AL measures used in higher education settings*?" This review is unique given the specific population that is elaborated with an updated definition of AL. The following specific research questions guided our study:

1. What evidence of *reliability*, if any, was provided for the available AL measures in the higher education context?

2. What evidence of *validity*, if any, was provided for the available AL measures in the higher education context?

Method

We used the principles outlined in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to design and carry out our systematic review. As illustrated in Figure 1, these principles involve four major phases: (1) identification, (2) screening, (3) eligibility, and (4) inclusion (Moher et al., 2009). We discuss our methods specific to each of these phases below.

Database Search

Our search was conducted in databases focused on educational research, namely ERIC (Educational Resources Information Center), PsycINFO, Web of Science (WoS) and Education Full Text. These databases have a wealth of peer-reviewed journals and publications related to higher education assessment. We also searched Google Scholar for additional published work citing Stiggins (1991).

Over time, the idea and meaning of AL has evolved from merely being a collection of skills and knowledge to being viewed as a social practice.

Search Terms

To locate relevant studies, we used and modified keywords according to the specific formats and requirements for each database. Specifically, text databases were systematically queried using a comprehensive set of search terms, which included '*assessment literacy*' and the combined usage of '*assessment literacy*' with '*higher education*' to ensure a thorough exploration of the literature in this domain.

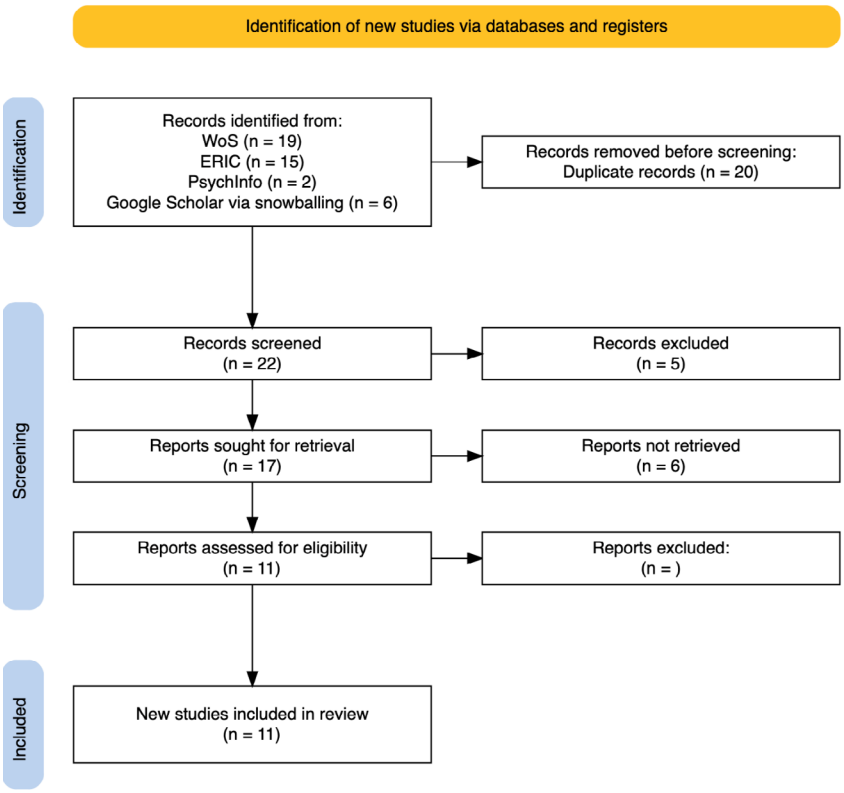


Figure 1
PRISMA flow chart study inclusion process

Inclusion and Exclusion Criteria

We identified a set of inclusion and exclusion criteria that reflected our guiding research questions and objectives. We selected articles that met the following criteria: 1) published in a peer reviewed journal, 2) full text in English was available to the researchers, and 3) the use of the construct directly related to higher education context. Exclusion criteria included: 1) articles that were not published in a peer-reviewed publication (e.g., non-peer reviewed article, dissertation, technical report), and 2) articles not directly related to higher education institutions but to other stakeholders (e.g., in-service teachers, school administrators).

We have diligently reviewed studies conducted over the past decade, recognizing the significance of updates in the AL construct.

The Reviewing Process

In our investigation, we have diligently reviewed studies conducted over the past decade, recognizing the significance of updates in the AL construct. The screening process included the following stages: 1) eliminating duplicate articles, 2) filtering out articles that did not align with the inclusion criteria based on their titles and abstracts, 3) reviewing full texts and excluding articles that did not meet the criteria, 4) employing a snowballing approach to identify additional relevant articles in Google Scholar, and 5) extracting results from the selected articles. This process returned 11 peer-reviewed articles. Any work employing an AL instrument in the United States higher education setting or an international equivalent was retained for review.

Evaluation of Psychometric Properties

We evaluated the psychometric properties of the identified measures by identifying and interpreting evidence related to validity and reliability. We organized our synthesis and interpretation of validity and reliability-related results using the criteria summarized in Table 1.

Table 1
Psychometric Property Evaluation Criteria

Type	Criteria	Citation
Concurrent Validity Evidence	Correlation > .25	Litwin (1995, p. 45)
Content Validity Evidence	Scale level: S-CVI > .90 or Item level: I-CVI = 1.00 for 3-5 experts; CVI > .78 for 6-10 experts	Polit & Beck (2006, p. 496)
Construct Validity Evidence	CFI > .95 SRMR < .08 RMSE < .05	Hu & Bentler (1999)
Reliability Estimate	Cronbach alpha > .70 for attitude scales or KR-20 or KR-21 > .80 for competency tests	Field (2009) Nunnally (1978) Salvucci et al. (1997)

Note. This table displays evaluation criteria for the validity and reliability aspects of the measures. The details about definition and calculation are presented in the following section. *KR-20/KR-21*: Kuder-Richardson formulas which provide an estimate of proportion of the total variance in the test scores that is attributable to the construct being measured; *CVI*: Content validity index showing the degree of agreement among content experts on each item's relevancy to the intended construct; *SRMR*: Standardized Root Mean-Square Residual which evaluated the discrepancy between observed covariance matrix and model-implied covariance matrix; *RMSE*: Root mean square error which is the average difference between predicted and observed values is used to assess precision of estimates.

Evidence Related to Validity

The current *Standards for Educational and Psychological Testing* (AERA et al., 2014) define validity as the accumulation of evidence to support the interpretation and use of test scores for a particular purpose. This consensus view of validity includes five primary sources of evidence that may be relevant for developing a validity argument to support a particular interpretation and use of test scores, including evidence related to test content, response processes, internal structure, relations to other variables, and consequences of testing.

Researchers operationalize construct validity differently depending on their selected measurement framework…

As we will discuss in our results section, the articles included in our review reported validity-related evidence that reflects qualitative approaches to gathering content-related validity evidence or quantitative techniques to examine the internal structure of the instrument for construct-related validity evidence.

Construct validity is defined as "the degree to which inferences based on test scores are justified in relation to the conceptualization and theoretical framework of the construct that the test is intended to measure" (Messick, 1980, p. 8). Because researchers operationalize construct validity differently depending on their selected measurement framework, we started our analysis related to construct validity evidence by first considering whether researchers used a test-score approach (e.g., Classical Test Theory (CTT) or factor analysis) or scaling model approach (e.g., Item Response Theory (IRT) or Rasch measurement theory); please see Engelhard (2013) for a detailed discussion of these two measurement frameworks.

Among the articles included in our review, those that reflected the test score tradition tended to use a factor-analytic approach to collecting construct-related validity evidence; these analyses also provide validity evidence related to the internal structure of an assessment instrument. We evaluated this evidence using well-established critical values for model fit indices as presented by Hu and Bentler (1999): CFI > .95; SRMR < .08; RMSEA < .05. For the scaling model framework, researchers may use several techniques to evaluate the internal structure of an instrument using various fit statistics. For example, researchers who use Rasch models often use residual-based analyses to examine adherence to model requirements such as unidimensionality or invariance.

Most of the identified measures relied on self-reports where the participants were asked to evaluate their own knowledge, practices, confidence or needs in terms of assessment…

Evidence Related to Reliability

According to the Standards (AERA et al., 2014), reliability refers to the consistency of results over replications of a measurement procedure. For example, replications may include repeated observations of test-takers over multiple items within a form (e.g., internal consistency reliability) or between administrations of a test (e.g., repeated measures reliability). In our research synthesis, we established criteria for interpreting reliability results based on the type of instrument and selected method for evaluating reliability. For example, reliability evidence included in the articles in our review used reliability analyses based on test-score methods, where the authors reported internal consistency statistics such as Cronbach's alpha, KR-20, or KR-21. We interpreted these results using critical values from the literature. Specifically, we used a value of 0.70 for Cronbach's alpha (Nunnally, 1978) and a value of 0.80 for KR-20 or KR-21 (Salvucci et al., 1997). Researchers also reported reliability evidence based on scaling models from Rasch measurement theory. In these analyses, *reliability of separation statistics* describe the precision of location estimates for items, persons, or other facets. As a general guideline, a value exceeding 0.8 typically suggests reproducibility of measures within a Rasch context (Linacre, 2002).

Results

After applying our inclusion and exclusion criteria, we identified a total of eleven instruments. The instruments took three forms as presented in Table 1:

1) Scales, tests or inventories of assessment knowledge or AL (*n*=5),

2) Surveys on attitudes, perceptions or needs on AL (*n*=5), and

3) Rubrics evaluating other faculty members' work on assessment (*n*=1).

Most of the identified measures relied on self-reports where the participants were asked to evaluate their own knowledge, practices, confidence or needs in terms of assessment. Details about psychometric properties of the measures are presented in Table 2 and discussed below.

Evaluation of Psychometric Properties

In this section, we summarize the reported psychometric properties of the identified measures as they relate to validity and reliability.

Validity Evidence. Among the identified measures for our review, researchers reported validity evidence related to content validity and construct validity using internal structure analyses. Content validity evidence was provided in detail for a total of five measures. No detail regarding development of item content was provided for two measures (DiLoreto et al., 2017 and McGrath et al., 2020). Almost all the measures that were reviewed were grounded on a substantial literature review process followed by expert reviews. The process for identification of content and development of items involved both inductive (i.e., focus groups) and deductive evidence in one study (Alonzo et al., 2019). In two studies (Mokshein et al., 2015, 2019), content was grounded on AFT et al. (2009), but no information was provided related to expert review in these studies. Overall, no quantitative index was presented as evidence of content validity.

Five studies reported statistical evidence related to internal structure in the form of either EFA or PCA results. Among them, Alonzo et al. (2019) conducted a CFA as well as

Table 2
Psychometric Property Evaluation Criteria

Measure Name	Construct of Interest	Instrument & Item Characteristics	Respondents	Test Content or Content Validity	Internal Consistency /Reliability	Construct Validity
Adapted version of the Classroom Assessment Literacy Test (CAL) (Mertler, 2003)	Classroom assessment literacy skills	35 dichotomously scored multiple choice items	Regular teaching faculty in public and private higher education institutions in Pakistan	Expert opinion was collected qualitatively. No quantitative evidence was provided.	Cronbach alpha = .89	Not reported
Confidence in assessment survey for faculty members (Massey et al., 2020)	Conceptions and confidence in assessment	23 scale items in confidence in assessment and 2 open-ended items in conceptions of assessment	27 instructors at a two-year college in TX, USA	The instrument adapted from DeLuca et al., (2013) but no quantitative content validity evidence was provided	Cronbach alpha for the confidence in assessment scale items was found .84 and .68 for two subscales, namely assessment approaches and assessment praxis	Principal Component Analysis (PCA) resulted in a minimum subscale loading equals to .40.
Modified Conceptions of Assessment III (CoA-III) survey (DiLoreto et al., 2017)	Attitude about assessment & knowledge about assessment types	2 open-ended items were added to the modified CoA-III scale (Q1. What does the term assessment mean to you? Q2. What types of activities come to mind when you think of the term assessment?)	156 faculty members from 10 higher education institutions from the Southeast U.S.	Not reported	Not reported	The original CoA-III scale was validated through second-order CFA model based on a different population (teachers) but no evidence was provided for the modified survey
Assessment practice inventory for teacher educators – Section B (MAPITE) (Mokshein et al., 2015)	Practice with respect to assessment literacy standards	The inventory had five sections. Section B directly related to AL practices so was involved in this review. It includes 10 items on a 5-point scale.	254 faculty members from a teacher education university in Malaysia	Content was grounded on AFT, NCME & NEA (1990). No information related to expert review was provided.	Section B included two factors. Cronbach alpha for factor 1= .863 and for factor 2= .786	EFA results yielded all factor loadings > .60 Fit values were not reported.
Rasch-validated version of MAPITE Section B (Mokshein et al., 2019)	Practice with respect to assessment literacy standards	One item was dropped as the results of the Rasch validation	763 faculty members from various teacher education institutions in Malaysia	Content was grounded on AFT, NCME & NEA (1990).	Person reliability index = .84 Item reliability index = .91	Variance explained by Rasch measures = %53.1 (more than 20% variance explained is needed for accurate estimation, Reckase, 1979)
Assessment literacy survey for medical education faculty (McGrath et al., 2020)	Assessment literacy and practices	A survey with 4 open-ended items on assessment literacy (e.g., What do formative and summative assessment mean to them), 7 open-ended items particularly on peer assessment	35 faculty members from the departments of Medicine and Biomedical	No detail was provided regarding how the survey items were developed and survey content were determined	Not Applicable	Not Applicable
Questionnaire on formative and summative assessment (type 1) (Davies & Taras, 2018)	Assessment literacy on summative and formative assessment	A 44-item questionnaire with Yes-No response format. Some items required written comment (e.g., give a definition of formative assessment)	100 faculty members in The U.K.	Items were pilot tested with 5 faculty members when they were initially developed (Taras, 2008)	Not Applicable	Not Applicable

Table 2
Psychometric Property Evaluation Criteria, continued

Measure Name	Construct of Interest	Instrument & Item Characteristics	Respondents	Test Content or Content Validity	Internal Consistency /Reliability	Construct Validity
SALRubric-summative assessment literacy (Edwards, 2017)	Summative assessment literacy and knowledge	A rubric with descriptors on ten dimensions (i.e., knowledge on purpose of summative assessment, interpretation of results, fairness) developed for five levels of expertise	Academic and teaching staff in New Zealand	Inductive evidence (questionnaire, interviews, summative assessment tasks used by teachers) was used to ensure content validity	The evaluation criteria was reviewed by assessment experts to ensure the evaluation criteria measures what it purports to measure	Two senior academics in assessment independently scored data using SALRubric. No statistics were provided regarding the degree of agreement.
Academic SBA Practices Tool – ASBAPT (Alonzo et al. 2019)	Standard based assessment literacy of academics	21 items with 6 theoretical dimensions	410 academics in public universities in Philippines	8 experts reviewed the tool. Items and content were determined using inductive (focus groups) and deductive (literature review) evidence	Cronbach alpha values for the six subscales as follows: .92, .88, .89, .90, .92, .88, .94	EFA results for the 6-factor model: RMSEA= 0.02, SRMR= 0.03, CFI= 0.95, TLI= 0.96 CFA Results: RMSEA= 0.03, CFI= 0.98, TLI= 0.97
Language assessment literacy survey (Kremmel, 2020)	Perceived need on language assessment literacy	71 self-report items on a 5-point scale from No knowledgeable to extremely knowledgeable. Item format example: "How knowledgeable do people in your chosen group/profession need to be about…"	138 language assessment researchers, 198 language assessment developers and 645 language instructors	Content was grounded on a well-known theory (Taylor, 2013). 6 expert reviews on the initial instrument and 2 additional expert reviews on the piloted instrument were taken.	Cronbach alpha values for the identified nine subscales were ranged from .85 to .96	EFA results only included eigenvalues on a large sample. No fit values were reported.
Questionnaire on Language Assessment Literacy and Assessment Training Needs (Sayyadi, 2022)	Perceived need for language assessment literacy and received training on it	22 items on a 3-point scale from not at all to advanced on received training (if you received training on…) and perceived training needs (if you need training on…) on LAL	68 university instructors on English in Iran	The questionnaire was adapted from a well-known instrument (Vogt & Tsagari, 2014) developed for teachers and piloted with four instructors	Cronbach alpha value of .92 was reported	Not reported

an EFA and reported that the results from both analyses showed good fit according to our criteria. In other studies that employed factor analytic methods to collect construct validity evidence, minimum factor loadings were reported as .40 for Massey et al. (2020) and .60 for Mokshein et al. (2015). The dimensional structure of the measures was analyzed and reported in several studies. For example, Mokshein et. al. (2019) investigated dimensionality for the MAPITE using a Rasch measurement approach and found that Rasch measures based on data collected through explained 53.1% of the variance, supporting a unidimensional structure. Kremmel (2020) investigated dimensionality using EFA, where the number of dimensions was determined based on eigenvalues. However, fit statistics were not reported in either study. No statistical evidence was provided for a total of four studies, including the adapted version of Mertler's (2003) assessment literacy inventory. In one study, no further statistical analysis was conducted with the new sample, but fit statistics for the original scale were reported (DiLoreto et. al., 2017).

Reliability Evidence. Among the identified measures, researchers reported reliability using either internal consistency statistics based on CTT or reliability of separation statistics based on Rasch measurement theory models. Cronbach Alpha coefficients were reported in a total of six studies. Among the Cronbach alpha coefficient values that were reported, only the assessment praxis dimension in Massey et al. (2020) had a relatively low reliability index value against our criteria, while the other values were reported between .79 and .94. The validation study of MAPITE (Mokshein et. al., 2019) based on Rasch measurement theory

suggested that the person reliability index and item reliability index values were high, .84 and .91, respectively. Lastly, a total of three measures were not accompanied with any reliability evidence.

Synthesis of the Reviewed Studies

Our systematic review of the literature on AL measures in higher education has revealed several critical insights, highlighting both strengths and gaps in the existing instruments. We identified eleven instruments classified into three categories: scales, tests, or inventories of assessment knowledge or literacy (n=5); surveys on attitudes, perceptions, or needs related to AL (n=5); and rubrics evaluating faculty members' assessment work (n=1). The predominant reliance on self-report measures indicates a strong focus on subjective evaluations of knowledge, practices, confidence, and needs in assessment.

Validity Evidence. The psychometric evaluation of these instruments revealed mixed results. While five measures provided detailed content validity evidence, two lacked specific information regarding item content development. Most measures were grounded in substantial literature reviews and expert evaluations, indicating a rigorous approach to establishing content validity. However, the absence of a quantitative index for content validity in many studies highlights a gap that needs to be addressed. Construct validity evidence, primarily through factor analytic methods, was reported in five studies. While some studies, like Alonzo et al. (2019), reported good fit indices, others did not provide sufficient statistical evidence.

Reliability Evidence. The reliability analysis revealed that six studies reported Cronbach Alpha coefficients, with most values falling between .79 and .94, indicating acceptable internal consistency. The MAPITE study by Mokshein et al. (2019) reported high person and item reliability indices using Rasch measurement theory. However, three measures lacked any reported reliability evidence.

General Analysis. The current instruments, while grounded in solid methodological foundations, often fall short of providing up-to-date content. This gap is particularly significant given the evolving educational landscape and the increasing integration of innovative technologies such as artificial intelligence (AI). The advancement of AI presents new opportunities to refine and optimize assessment processes, making it imperative to explore the concept of *Artificial Intelligence Assessment Literacy*. Educators need to be literate in integrating, using, and tracking the effects of AI in assessment to fully leverage these technologies. In conclusion, our review highlights the critical need for ongoing research and development in the field of AL. Establishing robust and adaptive AL measures will contribute significantly to improving educational outcomes and aligning assessment practices with the rapidly changing landscape of higher education.

Redefining Assessment Literacy and Measuring It

In light of recent advancements in technology, particularly with the rise of generative AI (GAI), there is a pressing need to redefine AL. Scholars have suggested that we need to reconsider the kinds of assessment tasks instructors assign to make them 'AI-resistant' by reducing the likelihood that GAI can complete the entire assignment task (Moorhouse et al., 2023). While there is an increasing amount of advice available to instructors on how to modify their assignment tasks in the GAI world (e.g., blogs, newsletters), many instructors need to look to their institutions for guidance and direction regarding GAI (Moorhouse et al., 2023).

The advancement of AI presents new opportunities to refine and optimize assessment processes, making it imperative to explore the concept of Artificial Intelligence Assessment Literacy.

For this reason, a comprehensive framework on Artificial Intelligence Assessment Literacy (AI AL) must be developed, and resources must be provided to faculty. This framework should include guidelines on creating AI-resistant assignments, training on the use of AI tools in assessment, and strategies for leveraging AI to enhance learning outcomes. By equipping faculty with these resources, institutions can help ensure that assessment practices remain effective and relevant in the face of rapidly evolving technological capabilities.

Discussion

In this study we examined the reported psychometric evidence of existing AL measures developed within the international higher education context over the past decade (2013-2023).

Based on psychometric evaluation of existing measures against a set of criteria, we concluded that the available psychometric evidence supporting these measures is not strong. Despite the importance of understanding AL levels, perceptions, and practices of faculty members in higher education being widely recognized, our findings raise doubts about the preparedness of these measures to meet such demands.

First, we examined the literature on measures of AL for evidence related to validity. In terms of content validity, the majority of the reviewed measures were grounded on a substantial literature review process followed by expert reviews, although these procedures were not accompanied by statistical indices such as the Content Validity Index (Lawshe, 1975). However, a noteworthy critique of the existing measures pertains to their content. These measures lack coverage of current topics such as digital AL, fairness in the era of Artificial Intelligence (AI), and fairness related to Assessment for Learning. This deficiency can be attributed to their content development and validation processes being based on outdated standards that no longer adequately encompass these areas. One significant practical implication of this study regarding this finding is the need to update and expand the content of AL measures, while ensuring the collection of robust content validity evidence that aligns with current standards. This entails seeking consensus from content experts, which can be quantified using appropriate indices like CVI.

Many of the reviewed measures primarily concentrate on individuals' *perceived* knowledge or skills, while there is a notable scarcity of research examining individuals' actual skills, knowledge, or competency. On the other hand, self-report measures are limited in terms of their susceptibility to response biases and social desirability effects (Fisher, 1993). Among the reviewed measures, five of them included direct statements (i.e., "I can"). Self-reports are used most often since they are easiest to administer and shown to correlate with actual assessment practices (Kelly, 2020) despite their well-researched limitations. In future work, we suggest that researchers consider using objective assessments of individuals' knowledge, literacy, and practices in assessment rather than their subjective perceptions in future studies.

Several studies acknowledged the limitation of sample size and sample characteristics as potential constraints on the validity of their findings (Massey et. al., 2020; McGrath et. al., 2020). It is crucial to consider both the size and characteristics of the sample when evaluating the validity and reliability of scores obtained from measures. However, it is worth noting that the Rasch model, which is a member of IRT-based models, can address limitations regarding small sample size (Linacre, 1994), since these models are known for their robustness to handle small sample sizes and can be effectively utilized for validation purposes in measurement processes with limited samples. Of the reviewed studies, only one measure employed the Rasch framework for validation purposes. Future studies on developing and adapting measures in higher education assessment can utilize a scaling model approach more often to address sampling limitations.

Many of the reviewed measures primarily concentrate on individuals' perceived knowledge or skills, while there is a notable scarcity of research examining individuals' actual skills, knowledge, or competency.

Another concern about the reviewed measures was related to the presented reliability evidence. We observed that none of the reviewed studies took into account the dimensional structure when calculating reliability indices. However, it is crucial to consider the dimensionality of the measure to determine the appropriate type of reliability coefficient to report. For multidimensional scales, reporting composite reliability provides a smaller margin of error compared to reporting separate Cronbach alpha values for individual sub-scales (Cronbach et al., 1965); in other words, reliability may be overestimated. Therefore, future research studies aiming to develop AL scales, including those for assessing attitudes, should carefully consider the dimensionality of the measure when calculating and presenting reliability evidence. By doing so, researchers can enhance the accuracy and precision of the reliability estimates, providing more robust evidence of the measures' internal consistency and stability over time.

The results of the review also suggested that there is a lack of evidence regarding validity and reliability evidence for the modified instruments, which were adapted from existing measures through the addition, omission, or revision of items. However, any adaptation may raise concerns about the trustworthiness of the modified instruments if they lack the necessary validation and reliability evidence to support their use. It is crucial for

researchers to provide comprehensive evidence of the validity and reliability of modified instruments to ensure the robustness of their results (AERA et al., 2014).

None of the reviewed measures presented evidence related to relationships with external variables (e.g., concurrent validity; Lin & Yao, 2014). However, the newly developed measures could report their association with relevant measures such as student outcomes to provide evidence of concurrent validity (Murphy & Davidshofer, 1988).

Lastly, providing institutional support and resources for ongoing research into AL measures is essential for their continuous development. This support can include funding for validation studies, access to advanced technologies, and dedicated time for faculty to engage in research activities. By investing in these resources, institutions can foster innovation in assessment practices and ensure that AL measures are grounded in the latest educational research and methodologies.

By adopting these strategies, institutions can ensure that AL measures are robust, comprehensive, and well-aligned with contemporary educational needs. This proactive approach will help institutions stay ahead of emerging trends and challenges in education, ultimately leading to improved teaching and learning outcomes.

Higher education institutions and faculty development centers can utilize well-established measures to inform the design and implementation of targeted training programs aimed at enhancing faculty's understanding, attitude, and skills in assessment.

Limitations

This review has several limitations that must be acknowledged. First, our review focused on published studies that met specific inclusion criteria, which may have resulted in the exclusion of relevant unpublished or non-peer-reviewed works, such as dissertations, that could offer additional insights into the psychometric properties of AL measures. This selection bias may limit the generalizability of our findings. Second, the rapid advancement of AI technologies is reshaping assessment requirements and expectations. While our study discusses the definition of *AI assessment literacy* and the development of enhanced AL measures, it does not fully encompass the dynamic and rapidly evolving landscape of educational assessment.

Lastly, while we have highlighted both the strengths and gaps in the existing measures, the review itself is limited by the quality and depth of the original studies.

Despite these limitations, this review aims to provide insights into the current state of AL measures and discuss avenues for future research aimed at developing psychometrically sound, comprehensive, and up-to-date tools that are adaptive to the changing landscape of assessment.

Future Work

Overall, the findings of this review study highlight the need for further research on refining existing AL measures in diverse higher education contexts. This can also be considered as a new research avenue for developing psychometrically sound measures of AL in higher education. In practical terms, the study emphasizes the significance of integrating psychometrically-sound AL measures when planning training and professional development initiatives in higher education. Higher education institutions and faculty development centers can utilize well-established measures to inform the design and implementation of targeted training programs aimed at enhancing faculty's understanding, attitude, and skills in assessment. This may entail offering customized resources, workshops, or ongoing support to assist faculty in effectively designing and implementing assessments that align with learning objectives, promote student engagement, and provide meaningful feedback. By incorporating these measures and implementing comprehensive training programs, institutions can foster a culture of AL among faculty, thereby enhancing the overall quality of assessment practices in higher education.

References

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

Alkharusi, H. (2011). An Analysis of the internal and external structure of the teacher assessment literacy questionnaire. *International Journal of Learning*, *18*(1), 515-528. https://doi.org/10.18848/1447-9494/CGP/v18i01/47461

Alonzo, D., Mirriahi, N., & Davison, C. (2019). The standards for academics' standards-based assessment practices. *Assessment & Evaluation in Higher Education*, *44*(4), 636-652. https://doi.org/10.1080/02602938.2018.1521373

American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), & National Education Association (NEA) (1990). *Standards for Teacher Competence in the Educational Assessment of Students*. Retrieved from https://eric.ed.gov/?id=ED323186

American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), & National Education Association (NEA) (2009). *Assessment Literacy Standard*.

Coombs, A., & DeLuca, C. (2022). Mapping the constellation of assessment discourses: a scoping review study on assessment competence, literacy, capability, and identity. *Educational Assessment, Evaluation and Accountability*, *34*(3), 279-301. https://doi.org/10.1007/s11092-022-09389-9

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. J. , Schoneman, P., & McKie, D. (1965). Alpha coefficient for stratified parallel tests. *Educational & Psychological Measurement*, *25*(2), 291-312. https://doi.org/10.1177/001316446502500201

Danielson, C. (2013). *The framework for teaching evaluation instrument*. The Danielson Group.

Davies, M.S. & Taras, M. (2018). Coherence and disparity in assessment literacies among higher education staff. *London Review of Education*, *16*(3), 474–490. https://doi.org/10.18546/LRE.16.3.09

DeLuca, C., Chavez, T. & Cao, C. (2013). Establishing a foundation for valid teacher judgement on student learning: The role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 107–126. https://doi.org/10.1080/0969594X.2012.668870

DeLuca, C., Coombs, A., MacGregor, S., & Rasooli, A. (2019). Toward a differential and situated view of assessment literacy: Studying teachers' responses to classroom assessment scenarios. *Frontiers in Education*, *4*(94), 1-10. https://doi.org/10.3389/feduc.2019.00094

DiLoreto, M. A., Pellow, C., & Stout, D. L. (2017). Exploration of conceptions of assessment within high-stakes US culture. *International Journal of Learning, Teaching and Educational Research*, *16*(7), 1-9.

Edwards, F. (2017). A rubric to track the development of secondary pre-service and novice teachers' summative assessment literacy. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 205-227. https://doi.org/10.1080/0969594X.2016.1245651

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Field, A. (2009). Discovering statistics using SPSS (3rd edition). *Sage*.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*(2), 303-315. https://doi.org/10.1086/209351

F Flaugher, R. (2000). Item pools. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.), *Computerized adaptive testing* (pp. 37–59). Routledge. https://doi.org/10.4324/9781410605931

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a computer adaptive test for depression based on the Dutch-Flemish version of the PROMIS item bank. *Evaluation & the Health Professions*, *40*(1), 79-105. https://doi.org/10.1177/0163278716684168

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of life Research*, *14*, 2277-2291. https://doi.org/10.1007/s11136-005-6651-9

Fonseca-Pedrero, E., Menéndez, L. F., Paino, M., Lemos-Giráldez, S., & Muñiz, J. (2013). Development of a computerized adaptive test for schizotypy assessment. *PLoS One*, *8*(9), e73201. https://doi.org/10.1371/journal.pone.0073201

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: a review and illustration with the MMPI-2 Computerized Adaptive Version. *Psychological Assessment*, *19*(1), 14-24. https://doi.org/10.1037/1040-3590.19.1.14

Friesen, D. W. (2022). *Towards a situated view of assessment literacy for higher education* [Unpublished master's thesis]. [University of Saskatchewan]. *ProQuest Dissertations & Theses Global*.

Gorgun, G., & Bulut, O. (2023). Incorporating test-taking engagement into the item selection algorithm in low-stakes computerized adaptive tests. *Large-scale Assessments in Education*, *11*(1), 1-21. https://doi.org/10.1186/s40536-023-00177-5

Gotch, C. M., & French, B. F. (2011). *Development and validity evidence for the teacher educational measurement literacy scale* [Conference presentation]. National Council on Measurement in Education, New Orleans, LA, United States.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, *33*(2), 14–18. https://doi.org/10.1111/emip.12030

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Kelly, M. P., Feistman, R., Dodge, E., St. Rose, A., & Littenberg-Tobias, J. (2020). Exploring the dimensionality of self-perceived performance assessment literacy (PAL). *Educational Assessment, Evaluation and Accountability*, *32*, 499-517. https://doi.org/10.1007/s11092-020-09343-7

Knapper, C. (2010). Plus ça change… educational development past and future. *New directions for teaching and learning*, *122*, 1-5. https://doi.org/10.1002/tl.392

Kozierkiewicz-Hetmańska, A., & Nguyen, N. T. (2010, September). A computer adaptive testing method for intelligent tutoring systems. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 281-289). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kraska, J., Bell, K., & Costello, S. (2023). Graded response model analysis and computer adaptive test simulation of the depression anxiety stress scale 21: Evaluation and validation study. *Journal of Medical Internet Research*, *25*, e45334. https://doi.org/10.2196/45334

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly*, *17*(1), 100-120. https://doi.org/10.1080/15434303.2019.1674855

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel psychology*, **28**(4), 563-575. https://psycnet.apa.org/doi/10.1111/j.1744-6570.1975.tb01393.x

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Lin, WL. & Yao, G. (2014). Concurrent Validity. In *Encyclopedia of Quality of Life and Well-Being Research*. https://doi.org/10.1007/978-94-007-0753-5_516

Litwin, M. S., & Fink, A. (1995). *How to measure survey reliability and validity*. Sage.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382-386.

Massey, K. D., DeLuca, C., & LaPointe-McEwan, D. (2020). Assessment literacy in college teaching: Empirical evidence on the role and effectiveness of a faculty training course. *To Improve the Academy: A Journal of Educational Development*, *39*(1). http://dx.doi.org/10.3998/tia.17063888.0039.109

McGrath, M. F., Scott, L., & Logue, P. (2020). Peer assessment in Irish medical science education: Exploring staff assessment literacy and assessment practice. *Practitioner Research in Higher Education*, *13*(1), 37-56.

Mertler, C. A. (2003). *Pre-service versus in-service teachers' assessment literacy: Does classroom experience make a difference?* [Conference presentation]. Mid-Western Educational Research Association, Columbus, OH, United States.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., ... & Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement (Chinese edition). *Journal of Chinese Integrative Medicine*, *7*(9), 889-896.

Mokshein, S. E., Lebar, O., Yunus, J. Y., Rahmat, A., Dollah, M. U., Muhammad, A., Mansor, N. A., Mahmood, A., & Noor, N. M. (2015). Development and validation of assessment practice inventory for teacher educators. *Asian Journal of Assessment in Teaching and Learning*, *5*, 25-43. Retrieved from https://ojs.upsi.edu.my/index.php/AJATeL/article/view/2036

Mokshein, S. E., Ahmad, H., Lebar, O., Dollah, M. U., Yunus, J., Rahmat, A., & Ahmed, H. H. (2019). Validation of the Malaysian-Based Assessment Practice Inventory for Teacher Educators (MAPITE) using Rasch model. *Journal of Engineering and Applied Sciences*, *14*(9), 2783-2798.

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 100151. https://doi.org/10.1016/j.caeo.2023.100151

Murphy, K. R., & Davidshofer, C. O. (1988). *Psychological testing: Principles & applications* (4th ed.). Prentice-Hall.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Nunnally, J. C. (1978) *Psychometric Theory* (2nd edition). McGraw-Hill.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*(4), 679. https://psycnet.apa.org/doi/10.1037/0021-9010.78.4.679

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. *Teaching and Teacher Education*, *84*, 128-138. https://doi.org/10.1016/j.tate.2019.05.003

Petersen, M. A., Aaronson, N. K., Arraras, J. I., Chie, W. C., Conroy, T., Costantini, A., ... & European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group. (2018). The EORTC CAT Core—The computer daptive version of the EORTC QLQ-C30 questionnaire. *European Journal of Cancer*, *100*, 8-16. https://doi.org/10.1016/j.ejca.2018.04.016

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, *12*(4), 10-12. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489-497. https://doi.org/10.1002/nur.20147

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265-273. https://doi.org/10.1080/08878730.2011.605048

Ray, C. M., Peterson, C. M., & Montgomery, D. M. (2012). Perceptions of college faculty concerning the purpose of assessment in higher education. *Journal of Human Subjectivity*, *10*(1), 77-102.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230. https://doi.org/10.3102/10769986004003207

Salvucci, S., Walter, E., Conley, V., Fink, S., & Mehrdad, S. (1997). *Measurement error studies at the National Center for Education Statistics* (NCES 97-464). U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Sayyadi, A. (2022). In-service university-level EFL instructors' language assessment literacy and training needs. *Profile Issues in Teachers Professional Development*, *24*(1), 77-95. https://doi.org/10.15446/profile.v24n1.93676

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT–PD project. *Journal of Personality Assessment*, *93*(4), 380-389. https://doi.org/10.1080/00223891.2011.577475

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*(7), 534-539.

Suskie, L. (2004). *Assessing student learning: A common sense guide*. Anker Publishing Company, Inc.

Taras, M. (2008). Summative and formative assessment: Perceptions and realities. *Active Learning in Higher Education*, *9*(2), 172-192. https://doi.org/10.1177/1469787408091655

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing*, *30*(3), 403-412. https://doi.org/10.1177/0265532213480338

Taylor, K. L. & Colet, N. (2010). Making the shift from faculty development to educational development. In A. Saroyan & M. Frenay (Eds.), *Building teaching capacities in higher education* (pp. 139-167). Stylus.

Thompson, N. A. (2011). *Advantages of computerized adaptive testing* (CAT). https://assess.com/docs/Advantages-of-CAT-Testing.pdf

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1), 1.

Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, *11*(4), 374-402. https://doi.org/10.1080/15434303.2014.960046

Watermark (2023). *The importance of assessment in higher education*. https://www.watermarkinsights.com/resources/blog/importance-of-assessment-in-higher-education

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Willis, J., Adie, L., & Klenowski, V. (2013). Conceptualising teachers' assessment literacies in an era of curriculum and assessment reform. *The Australian Educational Researcher*, *40*, 241-256. https://doi.org/10.1007/s13384-013-0089-9

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education*, *45*, 477-501. https://doi.org/10.1023/A:1023967026413

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, *16*(4), 323-342. https://doi.org/10.1207/S15324818AME1604_4

Appendix A.
Reading List for CAT

1. What is CAT? http://www.iacat.org/what-is-cat
2. Elements of Adaptive Testing https://doi.org/10.1007/978-0-387-85461-8
3. First Adaptive Test http://www.iacat.org/first-adaptive-test
4. Some Current Issues in CAT http://www.iacat.org/some-current-issues-cat
5. Computer-Adaptive Testing: A Methodology whose Time Has Come https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d41985bb9c543b94c60fc7dc6ab5e0ca31b8362f
6. The Impacts of Computer Adaptive Testing from a Variety of Perspectives https://doi.org/10.3352/jeehp.2017.14.12