International Journal of Assessment Tools in Education
2024, Vol. 11, No. 3, 424–444
https://doi.org/10.21449/ijate.1440210
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Harmonizing perspectives to understand attitudes: A mixed methods approach to crafting an assessment literacy attitude scale

Beyza Aksu Dünya[1,2], Stefanie A. Wind[2], Mehmet Can Demir[3*]

[1]Bartın University, Faculty of Education, Department of Educational Sciences, Bartın, Türkiye
[2]University of Alabama, College of Education, Tuscaloosa, AL, USA
[3]Bartın University, Faculty of Education, Department of Educational Sciences, Bartın, Türkiye

ARTICLE HISTORY

Received: Feb. 20, 2024
Accepted: Apr. 29, 2024

Keywords:
Assessment literacy,
Higher education,
Scale development,
Mokken scale analysis,
Mixed-methods.

Abstract: Assessment literacy's vital role in faculty effectiveness within higher education lacks sufficient tools for measuring faculty attitudes on this matter. Employing a sequential mixed-methods approach, this study utilized the theory of planned behavior to develop the Assessment Literacy Attitude Scale (ALAS) and evaluate its psychometric properties within the U.S. higher education context. The qualitative phase involved a literature review of relevant studies and existing self-report measures, interviews with stakeholders, and panel reviews to shape initial item development. Following the establishment of a conceptual foundation and a comprehensive overview of the scale's construction, our study advanced to the quantitative stage that involves factor analytical and item response theory approaches using data from 260 faculty across three public universities in the U.S. Exploratory factor analysis (EFA) was employed initially to obtain preliminary insights into the scale's factorial structure and dimensionality. Confirmatory factor analysis (CFA) was subsequently applied with separate data and the findings largely supported the conclusions from the EFA. Exploratory and confirmatory factor analyses resulted in 15 items loading across two factors in a good model fit range. Finally, we used nonparametric item response theory (IRT) techniques based on Mokken Scale Analysis (MSA) to evaluate individual items for evidence of effective psychometric properties to support the interpretation of ALAS scores, including monotonicity, scalability, and invariant item ordering. The newly-developed scale shows promise in assessing faculty attitudes toward enhancing their assessment literacy.

1. INTRODUCTION

In the realm of higher education, faculty members consistently face challenges related to student assessment (Jankowski & Marshall, 2017; Medland, 2019; Sadler, 2017). The complexities associated with assessment practices are further intensified by the current demand for accountability in higher education (Caspersen & Smeby, 2018; Liu et al., 2012; Scholl & Olsen, 2014). The shift toward outcome-based education (Adam, 2004; Coates, 2016; Singh & Ramya, 2011) alongside changes in quality assurance and accreditation standards (Williams, 2016) has spurred a renewed emphasis on assessment literacy (Biggs & Tang, 2011; Dann, 2014;

*CONTACT: Mehmet Can Demir ✉ mehmetdemir@bartin.edu.tr Bartın University, Faculty of Education, Department of Educational Sciences, Bartın, Türkiye.

e-ISSN: 2148-7456

Eubanks, 2019; Wolf et al., 2015). The rapid advancements in artificial intelligence (AI) and its influence on assessment also necessitate faculty members to adapt and enhance their assessment methods to embrace these changes (McMurtrie, 2023). Despite this emphasis, assessment literacy remains an area in need of improvement in higher education (Pastore, 2022).

As a practical response to the growing and evolving importance of assessment literacy among faculty, higher education institutions have made recent strides in providing various faculty development opportunities to enhance assessment literacy among faculty. One of the objectives of such programs is to encourage changes in faculty members' routine assessment practices (Hines, 2009; Holmboe et al., 2011). To foster faculty members' willingness to adopt these practices, it is also important to explore their attitudes towards enhancing their assessment literacy. Understanding faculty attitudes towards their own assessment literacy and motivation to improve in this area can inform the development and improvement of in professional development initiatives. For example, institutions can tailor workshops, seminars, or training programs to address specific areas for improvement, leading to continuous growth in assessment practices. Tailored professional development opportunities play a vital role in ensuring that faculty members involved in assessment are well-prepared to carry out their responsibilities proficiently (Horst & Prendergast, 2020).

Accrediting bodies often require higher education institutions to provide evidence of effective assessment practices (Dill, 2007). Having a well-established scale to measure faculty attitudes and demonstrate changes in their attitudes towards assessment literacy can provide evidence that an institution is committed to fostering a culture of continuous improvement in assessment. By using a psychometrically sound scale to measure faculty members' attitudes towards enhancing their assessment literacies, higher education institutions can effectively make data-driven decisions in designing targeted professional development initiatives and assess their impact on faculty attitudes. Recognizing the link between attitudes and behavior (Ajzen, 1991), such a scale can also offer evidence of institutions' dedication to fostering a culture of continuous improvement in assessment practices. Understanding faculty attitudes towards enhancing assessment literacy allows institutions to tailor professional development programs to cater to individual strengths and areas for improvement. Customized initiatives aligned with faculty attitudes can support improvement in engagement and motivation, resulting in improved retention of knowledge and skills, making the training more productive and impactful.

1.1. Assessment Literacy and Higher Education

Assessment literacy in higher education is a multifaceted construct with a comprehensive scope, encompassing various dimensions and components. In general terms, it is envisioned as an individual's thorough understanding of assessment requirements (Zhu & Evans, 2022). This multidimensional construct includes a spectrum of skills, knowledge, and dispositions (Pastore & Andrade, 2019). Initially, assessment literacy was narrowly defined to establish a common language regarding assessment terminology. However, the conceptual boundaries have expanded significantly over time. For example, Price et al. (2012) broadened the definition by incorporating principles, methods, standards, and feedback. Subsequent contributions by Xu and Brown (2016) introduced the identification of assessors as a crucial component, emphasizing knowledge, conceptions, and practice. Evans (2016) further enriched the concept by incorporating an affective domain, highlighting the inclusion of staff and student entitlement in assessment literacy. Furthermore, the cognitive dimension of assessment literacy, as detailed by Balloo et al. (2018), underscores the significance of making assessment criteria explicit and transparent, thereby clarifying the requirements of assessment tasks. In essence, assessment literacy in higher education encompasses a diverse range of dimensions, mirroring its rich and evolving nature. Pastore and Andrade's (2019) three-dimensional model further elucidates assessment literacy, highlighting conceptual knowledge, the practical application of this knowledge to support learning, and a socio-emotional dimension.

Recent studies, such as those by Kremmel and Harding (2020), have extended the scope of assessment literacy to encompass socio-cultural values, personal beliefs, and attitudes. This evolution demonstrates a transition from a foundational understanding rooted in terminology and knowledge domains to a more intricate and holistic concept that integrates socio-cultural and personal dimensions. Within higher education, assessment literacy entails managing assessment practices and upholding standards and fairness (Zhu & Evans, 2022).

Significantly, the acknowledgment of personal attitudes and beliefs emerges as a crucial element within the assessment literacy framework. Assessment literacy is being influenced by personal beliefs, attitudes, and conceptions on assessment (O'Neill et al., 2023). Cultivating assessment literacy and promoting faculty ownership of this evolving definition necessitates the recognition and understanding of the nuanced individual attitudes and beliefs held by educators. By doing so, initiatives aimed at enhancing assessment literacy can be precisely tailored to align with the diverse perspectives and values that faculty bring to the educational setting. This approach fosters a more inclusive and effective stance toward assessment practices. In this context, the Theory of Planned Behavior (TPB) (Ajzen, 1991), with its well-grounded principles, holds substantial potential for guiding these efforts.

1.2. Purpose

The purpose of this study is to develop the Assessment Literacy Attitude Scale (ALAS) and evaluate its psychometric properties as a tool for measuring faculty attitudes towards assessment literacy enhancement. Specifically, we focused on the following research questions:

1. What is the internal structure of the initial set of ALAS items?
2. What modifications or refinements can be made to improve the psychometric properties of the ALAS based on results from the Exploratory Factor Analysis?
3. What is the degree of reproducibility of the ALAS items' structure?
4. What is the degree to which ALAS items conform to invariant item ordering principles?
   4.1. What is the degree of monotonicity exhibited by individual items within the scale?
   4.2. How scalable are the items within the scale, and what does this reveal about their ability to discriminate between different levels of the latent trait?
   4.3. Does the scale exhibit invariant item ordering, indicating consistent item difficulty across different levels of the latent trait?
5. Does ALAS yield sufficient reliability evidence?

1.3. Theoretical Framework

We grounded development of our scale items to TPB (Ajzen, 1991). The TPB stands as a highly influential framework for predicting and explaining human behavior, as outlined by Ajzen (1991, 2001). It has demonstrated successful applications in diverse domains, including professional development and adult and lifelong learning, where it proves valuable in comprehending the link between attitude and behavior (Archie et al., 2022; Dunn et al., 2018; Kao et al., 2018; Madigan & Kim, 2021). The central emphasis of the theory lies in an individual's intention to carry out a specific behavior, which, in our context, pertains to participating in activities and adopting programs and strategies to enhance assessment literacy. Intentions are regarded as the driving force behind behavior. Generally, a more robust intention to undertake a behavior correlates with a higher likelihood of successfully completing the action (Ajzen, 1991). The theory posits that the intention to adopt a behavior involves several psychological stages. These include developing a positive attitude towards the behavior, forming beliefs about the behavior's value, influenced by others' approval or disapproval, and engaging in the behavior based on perceived competency or the absence of constraints.

Attitude toward the behavior pertains to the extent of positive or negative evaluation of the behavior under consideration. This implies that a more nuanced and specific attitude serves as a more accurate predictor for the targeted outcome behavior in question (Ajzen & Timko, 1986). In the context of assessment literacy, discerning the extent to which faculty members appraise their active involvement in enhancing assessment literacy becomes crucial for forecasting their intent to participate in faculty development activities/initiatives related to assessment. Subjective norms encompass the perceived social standards that impact whether individuals sense external pressure to engage in a particular behavior. Multiple research studies affirm a positive correlation between perceived norms and behaviors among adults (Hora & Anderson, 2012; Knauder & Koschmieder, 2019; Rimal & Real, 2003). Finally, perceived behavioral control pertains to an individual's assessment of the ease or difficulty associated with performing a particular behavior, considering any constraints that may exist. The more challenging individuals perceive it to be to initiate or complete the behavior, the less likely they are to develop strong intentions to do so. Key issues related to perceived behavioral control in the context of enhancing assessment literacy would involve exploring whether faculty members believe that developing assessment literacy is within their sphere of influence/competency.

2. METHOD

2.1. Instrument Development

In accordance with the principles outlined by Creswell and Plano Clark (2011) for the development of an exploratory instrument, the initial step in our instrument development procedure involved establishing a clear definition of the construct. This process aimed to identify the main themes, constructs, and available instruments related to assessment literacy in the context of higher education. Drawing insights from existing literature, particularly in the context of assessment literacy within higher education, we started by understanding how researchers defined this construct in previous studies. Additionally, we engaged in semi-structured face-to-face interviews with two individuals affiliated with the faculty development office at the U.S. university where the research was initiated. Through these interviews, we sought to gain perspectives on the attitudes of faculty members towards assessment literacy, further enriching our exploration. The scale's initial development primarily centered on the dimensions of the theory of planned behavior framework, as it succinctly explains the motivational influences on behavior (Conner & Armitage, 1998). To establish an initial set of items aligned with the adopted theory, we drew upon the themes identified through an analysis of both the relevant literature and the insights gained from interviews. During this item generation process, we focused on ensuring sufficient content coverage. Each item underwent scrutiny for language and content appropriateness, considering clarity, length, and relevance to the target population. Furthermore, we assessed each item for potential biases, leading or suggestive phrasing, loading (encouraging automatic answers), and double-bareled content. These efforts resulted in a final set of 29 items. Our hypothesis posited that each of the 29 items would fall within one of the three domains of the TPB.

We presented the 29-item scale to two experts in assessment and measurement, who are currently working as faculty in different public universities in the U.S. They examined each item for relevance, accuracy, and representativeness, using a three-category rating scale (1= should be deleted, 2= requires revision, and 3= can be used) to affirm content and face validity. We also sought their suggestions on the number of response categories. We incorporated input from these two experts to complete the refinement of the instrument before its administration to the developmental sample for a think-aloud session (discussed in the next section). Following the panel reviews that identified redundancy as the primary concern, we subsequently downsized the initial instrument from 29 items to a more streamlined version containing 20 items.

To investigate examinee response processes for the new items, we conducted concurrent think-aloud sessions using the pilot instrument consisting of 20 items. In these sessions, three participants engaged in real-time verbalization of their thoughts and reactions while responding to the items via Zoom (2023). The think-aloud sessions followed the principles outlined by Padilla and Leighton (2017), where participants were requested to articulate their thoughts without interruption or leading questions from the interviewer. This psychological method, aligned with the Standards (AERA et al., 2014), is designed to capture data on human information processing and responses. The primary objective was to gain insights into participants' cognitive processes, allowing us to refine the instrument based on their feedback. Consequently, adjustments were made to the wording of some items to enhance clarity, informed by the observations and feedback obtained during these sessions.

2.2. Participants

The target population for this study was faculty members who have experience teaching or are currently teaching in either graduate or undergraduate classes in the U.S. setting. To reach a representative sample from the target population, we employed a combination of two non-probabilistic sampling techniques: convenience sampling and snowball sampling (Cochran, 1977). The only demographic information we collected was academic discipline, as this characteristic was important for the focus of our study. By limiting the demographic variables to academic discipline, we aimed to prioritize the relevance and specificity of the findings to the academic and teaching contexts under investigation.

The study involved a total of 260 faculty members from three U.S. public universities. In the initial round of data collection during the summer of 2023, participants were recruited from faculty at two large public universities in the southern region. For our initial analysis using Exploratory Factor Analysis (EFA), a total of 136 individuals responded to the instrument out of the 1687 faculty invited (8.06% completion rate). Despite the substantial data collection efforts and response rates, 33 individuals did not complete the instrument due to various reasons. Participants included 29 faculty from Natural and Applied Sciences (28.2%), 33 from Social Sciences (32%), 27 from Humanities (26.2%), and 14 from Business (13.6%). A subsequent round of data collection, utilizing the same sampling approach, was conducted for the Confirmatory Factor Analysis (CFA) and phases. The second round of data collection took place in the fall of 2023, involving faculty from a different public university in the Midwest. Over 2000 faculty were invited. The second sample included 197 respondents, with 40 failing to complete the instrument. Faculty representation was as follows: 51 from Natural and Applied Sciences (32.5%), 37 from Social Sciences (23.5%), 43 from Humanities (27.4%), 21 from Business (13.4%), and 5 from other (or multiple) academic disciplines (3.2%).

2.3. Data Collection Procedures

The study was conducted in the United States, where higher education institutions actively prioritize the continuous development of faculty member programs to enhance assessment literacy. Upon obtaining Institutional Review Board (IRB) approval, face-to-face interviews were conducted with faculty development professionals on campus, delving into their perspectives. Subsequently, the think-aloud process described earlier was implemented via Zoom with three participants, who offered valuable qualitative insights. Following these qualitative phases, the scale administration employed a web-based recruitment strategy through Qualtrics (Dillman et al., 2014). Faculty members were invited to participate via a weblink, which was disseminated across various channels, including college faculty listservs, ResearchGate, and LinkedIn. Moreover, participants were encouraged to share the survey link within their professional networks through social media, creating a snowball sampling effect that fostered broader participation.

The scale, structured into three sections, began with a comprehensive informed consent presentation in the initial section. Subsequently, participants engaged with the main set of scale items. The final segment prompted participants to indicate their discipline grouping (e.g., humanities, applied sciences, etc.). Participants responded to scale items using an ordinal four-category Likert-type scale ranging from strongly disagree to strongly agree (*1= strongly disagree, 2= disagree, 3= agree, 4= strongly agree*).

Throughout the data collection process, participants were assured of the confidentiality and exclusive research-oriented use of their responses and scale data. To uphold participant privacy, the assessment procedure did not request any personal identifiers, such as names or other identifiable information.

2.4. Data Analysis

For the instrument development stage of the study, the data analysis approach included content analysis of existing literature. This method ensured the provision of content evidence of validity for the instrument. Following expert reviews, a standardized statistical approach was applied to gather evidence related to content validity based on expert review. Specifically, we calculated inter-rater agreement statistics to evaluate the reliability of the expert review process.

Both factor analytic and non-parametric Item Response Theory (IRT) based approaches were employed to gather evidence related to the internal structure of the ALAS items. We provide details about our analysis related to each approach in the following sections.

2.4.1. *Phase 1: Exploratory factor analysis*

Prior to executing EFA, the assumptions of sampling adequacy (Kaiser-Meyer-Olkin [KMO] test) for evaluating sample size sufficiency and Bartlett's Test of Sphericity to ensure adequate item correlation were examined. EFA was performed using responses from 103 participants to the 20-item version of the ALAS (Table 1). We performed the analysis using the "psych" package for R (R Core Team, 2023; Revelle, 2023).

Table 1. *20-item version of ALAS.*

I1. I continually strive to enhance my assessment literacy.

I2. I must stay current with the latest assessment methods to fulfill my teaching responsibilities.

I3. I believe that improving my assessment literacy is crucial to enhance student learning outcomes.

I4. I feel motivated to learn more about assessment strategies to better teach my students.

I5. I am open to exploring new assessment techniques to improve my teaching practices.

I6. I believe that having strong assessment literacy is important for being an effective faculty member.

I7. I believe that increasing my assessment literacy will help me to better meet the needs of a diverse student population.

I8. I view increasing my assessment literacy as a continuous process, rather than a one-time task.

I9. Frequent conversation with colleagues improves my assessment practices.

I10. I value learning new concepts about assessment.

I11. Faculty professional development in assessment is necessary for quality instruction.

I12. I would like to complete more training in assessment in the future.

I13. I would only take an assessment training if it was required by my department.

I14. I plan to continue learning new techniques about assessment.

I15. I participate in professional development activities regarding assessment.

I16. I seek out opportunities to increase my assessment literacy.

I17. Learning innovative assessment approaches is valuable.

I18. I strive to use different applications and technology in assessment.

I19. Learning new tools and information in assessment is part of my professional development.

I20. I think faculty in higher education should have substantial knowledge in assessment.

During the EFA, we utilized specific criteria for retaining items. These criteria included: (a) a measure of sampling adequacy (KMO) of 0.5 or higher for each item, according to Field (2000), (b) a statistically significant Bartlett's Test of Sphericity value (Pett et al., 2003) and (c) adhering to Howard's (2016) recommendation that each item should have a minimum loading of 0.40 onto its primary factor, a maximum loading of 0.30 onto other factors, and a minimum difference of 0.20 loading between the primary factor and other factors.

The overall KMO value was equal to 0.914 and Bartlett Test of Sphericity value was statistically significant—indicating that the item responses could be explored using EFA ($\chi^2 = 1314.921$, df=190, p<0.001). For most items, the standardized univariate skewness and kurtosis values fell outside the range of ±1.96. The multivariate normality was checked by Mardia's test for multivariate normality and the multivariate skewness and kurtosis values were statistically significant (*p<0.001*), which indicates univariate and multivariate non-normal response distributions. As such, we applied EFA using the minimum residuals method suggested by Kline (1994).

2.4.2. *Phase 2: Confirmatory factor analysis*

To avoid overfitting in scale development studies, it is recommended to conduct CFA on a separate sample to confirm the structure of the proposed scale that resulted from an EFA (Fokkema & Greiff, 2017). Therefore, CFA was performed using responses from a new sample of 157 participants. The analysis was performed using the "lavaan" package for R (Rosseel, 2012).

Upon examination of the univariate and multivariate normality values, it was found that most of the standardized univariate skewness and kurtosis values fall out range of ±1.96 range. The multivariate normality was checked by Mardia's test for multivariate normality and the multivariate skewness and kurtosis values were statistically significant (p<0.001). Consequently, we applied the diagonally weighted least squares (DWLS) (Muthén, 1993) estimation method for the CFA since it is a suitable estimator with small samples and non-normal distributions.

Numerous fit indices are utilized in the CFA domain, with the Comparative Fix Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) being the most prevalent. The recommended criteria for best and good fit according to Hu and Bentler (1999) are as follows: CFI and TLI should be equal to or greater than 0.95 for best fit and equal to or greater than 0.90 for good fit; RMSEA should be less than or equal to 0.05 for best fit and less than or equal to 0.08 for good fit; and SRMR should be equal to or greater than 0.05 for best fit and equal to or greater than 0.10 for good fit. The internal consistency of the scale was evaluated by computing Cronbach's alpha (α; Cronbach, 1951) and McDonald's omega (ω; McDonald, 1999).

2.4.3. *Phase 3: Nonparametric item response theory: Mokken scale analysis*

Our first step in exploring the ALAS under nonparametric IRT framework was to examine the items for evidence of psychometric quality using basic item analysis statistics. First, we examined the frequency of responses in each rating scale category across items to ensure that we could apply our selected item analysis techniques to the data. Then, we examined item responses for evidence of internal consistency using inter-item and corrected item-total correlations. We conducted these analyses within the factors identified in the earlier analysis phases.

Next, we evaluated the scaling properties of the ALAS items within the identified factors using MSA (Mokken, 1971), which is a theory-driven nonparametric approach to item response theory (IRT). We used MSA to evaluate the ALAS items for several reasons. First, MSA includes several graphical and statistical techniques that provide an exploratory perspective into the degree to which items exhibit fundamental scaling properties while maintaining an ordinal level of measurement. This is particularly useful in scale development studies for which the construct or a set of items is not well-understood, and the use of an interval-level scale may be inappropriate or unnecessary (Meijer et al., 2015). Although it is nonparametric, MSA is characterized by clear ordering requirements for the relationship between item and person characteristics (i.e., the item response function; IRF) that reflect invariant measurement. In contrast to a-theoretical nonparametric IRT techniques (e.g., kernel smoothing) (Mazza et al., 2014; Ramsay & Silverman, 2005), these requirements provide a framework in which to evaluate item quality. Finally, because MSA is nonparametric, it can be used with relatively small examinee sample sizes, such as the current sample.

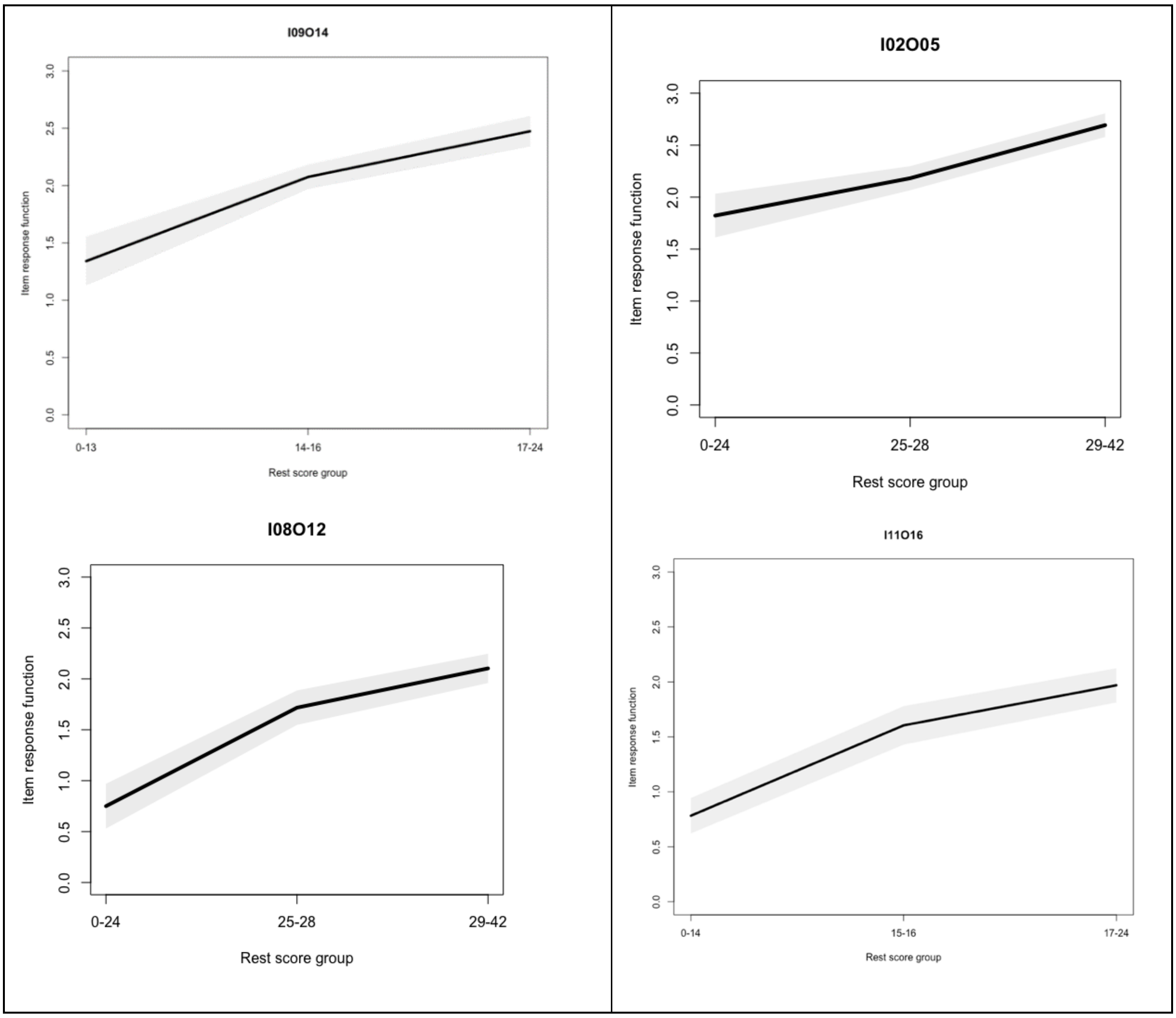
Typical MSA procedures are based on two nonparametric scaling models: (1) the Monotone Homogeneity Model (MHM), and (2) the Double Monotonicity Model (DMM). These models are characterized by ordering requirements that facilitate item analysis. The MHM is based on three requirements: (1) unidimensionality: a single latent variable is sufficient to explain most of the variation in item responses, (2) local item independence: After controlling for the primary latent variable, there are no meaningful associations between item responses (i.e., responses are statistically independent), and (3) monotonicity: participants' average responses for individual items are non-decreasing as their locations on the latent variable increase. The DMM shares the same requirements as the MHM and adds a fourth requirement: invariant item ordering (IIO): items have the same relative difficulty order for all participants. We used techniques based on these models to examine three major indicators of measurement quality for the ALAS items. From the MHM, we examined evidence of item monotonicity and item scalability. From the DMM, we examined evidence of invariant item ordering (IIO). We conducted the MSA analyses using the "mokken" package for R (van der Ark, 2007, 2012). Details on the specifics and procedures for testing these requirements are outlined below.

2.4.3.1. Item Monotonicity. For individual items, monotonicity occurs when participants' average ratings on an item are non-decreasing as their locations on the latent variable increase. Unlike parametric IRT models for which participant locations on the latent variable are estimated using an interval scale, MSA uses an ordinal nonparametric indicator of person locations based on total scores. Specifically, item monotonicity is evaluated using item-specific restscores, which are total scores minus participant scores on the item of interest. Typical procedures for evaluating item monotonicity include combining participants with equal or adjacent restscores into restscore groups with approximately balanced sample sizes in each group to improve statistical power for evaluating item properties.

Figure 1 illustrates item monotonicity at the overall item level using nonparametric IRFs for two example items from the ALAS. In each plot, the x-axis shows examinee rest-score groups, and the y-axis shows the rating scale, which was re-scaled to start at zero. The nonparametric IRFs show the average ratings for each item within restscore groups, and light shading around the line shows a 95% confidence interval. Both items exhibited adequate monotonicity because the average ratings are non-decreasing as rest-scores increase. In addition to graphical displays, researchers can also evaluate monotonicity using one-sided statistical hypothesis tests that evaluate whether monotonicity holds between pairs of adjacent restscore groups.

We examined monotonicity for each item within each of the the ALAS factors using graphical displays of nonparametric IRFs similar to Figure 1 as well as statistical hypothesis tests.

Figure 1. *Examples of plots for evaluating item monotonicity.*



2.4.3.2. Item Scalability. In the context of MSA, scalability refers to the degree to which response patterns associated with individual or groups of items support a consistent interpretation of item ordering across persons. Specifically, scalability coefficients describe the degree to which item responses are free from Guttman errors, or unexpected response patterns given item and person ordering on the latent variable. MSA procedures include scalability coefficients for individual items (H_i), pairs of items (H_{ij}), and sets of three or more items *(H*). Researchers typically interpret scalability coefficients as an indicator of overall item quality and fit to the MHM (Sijtsma & Molenaar, 2002). In scale development studies, researchers often focus on scalability coefficients for individual items, which can be calculated as:

$$H_i = 1 - \frac{\sum_{j\neq i} F_{ij}}{\sum_{j\neq i} E_{ij}} \qquad (1)$$

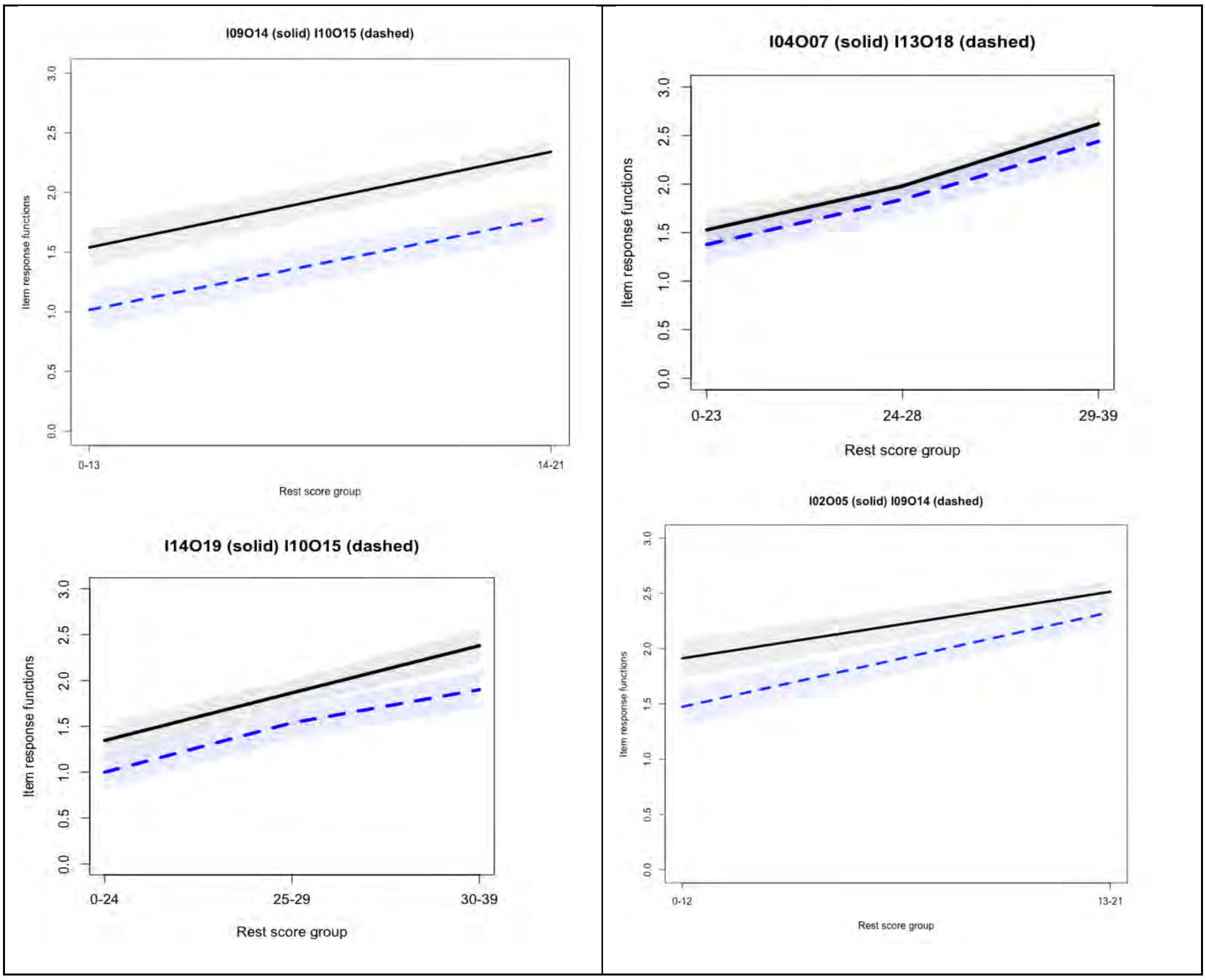
where F_{ij} is the observed frequency of Guttman errors associated with item i in combination with all other items in the scale, and E_{ij} is the expected frequency of Guttman errors for item i based on marginal independence. Researchers typically interpret item scalability coefficients with values greater than or equal to $H_i = 0.30$ as evidence of meaningful contribution to a scale (Meijer & Baneke, 2004; Sijtsma & van der Ark, 2017).

2.4.3.3. Invariant Item Ordering. The last major category of MSA item analysis is Invariant Item Ordering (IIO). This property is related to the DMM, and it describes the degree to which items exhibit a consistent relative difficulty ordering for participants with different locations on the latent variable. IIO has important theoretical and practical implications that are relevant for scale development. When items adhere to IIO, there is a single and consistent hierarchy of items that does not depend on examinees' location on the latent variable. In practice, researchers typically evaluate IIO for rating scale items such as those included in the ALAS using Manifest IIO analyses (Ligtvoet et al., 2010), which involve evaluating pairs of nonparametric IRFs for evidence of non-intersection across examinee restscores specific to the item pair.

For example, Figure 2 shows plots for evaluating IIO using items from the ALAS. For these IIO analyses, restscores are calculated specific to the item pair of interest, using examinee total scores across all items minus their scores on the two items being evaluated. In each plot, separate IRFs are plotted for each item that show examinees' average response to each item within rest-score groups calculated using their total score on all of the ALAS items except IO9014 and I10O15. These two items adhere to IIO because the IRF for IO9014 (solid line) is always higher—indicating higher average ratings—compared to the IRF for I10O15 (dashed line). Adherence to IIO for these two items suggests that participants always endorse IO9014 more readily than I10O15. IIO also holds for the item pair made up of I02O05 and I09O14.

Figure 1 illustrates item monotonicity at the overall item level using nonparametric IRFs for two example items from the ALAS. In each plot, the x-axis shows examinee rest-score groups, and the y-axis shows the rating scale, which was re-scaled to start at zero. The nonparametric IRFs show the average ratings for each item within restscore groups, and light shading around the line shows a 95% confidence interval. Both items exhibited adequate monotonicity because the average ratings are non-decreasing as rest-scores increase. In addition to graphical displays, researchers can also evaluate monotonicity using one-sided statistical hypothesis tests that evaluate whether monotonicity holds between pairs of adjacent restscore groups.

Figure 2. *Examples of plots for evaluating invariant item ordering.*



3. FINDINGS

3.1. Initial Scale Construction Results

In our examination of existing measures and studies on assessment literacy in higher education, the literature review yielded a pool of 40 candidate items. Through extensive discussions and careful consideration, some items were excluded from the pool for various reasons, such as redundancy. Following this refinement process, we initiated expert reviews, commencing with a set of 29 candidate items. The main themes identified from the content review, encompassing perceived necessity and the rationale for enhancing assessment literacy, were identified in conjunction with insights from face-to-face interviews with two experts. Additionally, identified several instruments related to skill improvement, such as the Effective Lifelong Learning Inventory (ELLI) (Crick et al., 2004), but noted that these were not specifically designed to measure assessment literacy.

After collecting expert item-level ratings, we computed Cohen's Kappa (Cohen, 1960) as a chance-corrected measure of inter-rater agreement for each criterion. According to Landis and Koch's (1977) guidelines, we achieved an almost perfect agreement level, surpassing 0.80 across the criteria of relevance, accuracy, and representativeness. This result indicates a high degree of consensus. Notably, during this process, nine items were identified and subsequently removed from consideration. This decision was rooted in the rating of 1 given by two raters on at least one of the criteria, ensuring a stringent and consistent approach to item selection.

3.2. EFA Results

The overall KMO value was equal to 0.914 and Bartlett Test of Sphericity value was statistically significant—indicating that the item responses could be explored using EFA (χ^2 = 1314.921, *df*= 190, *p*<0.001). For most items, the standardized univariate skewness and kurtosis values fell outside the range of ±1.96 (see Table 2). and the multivariate skewness and kurtosis values were statistically significant (*p*<0.001), which indicates univariate and multivariate non-normal response distributions. As such, we applied EFA using the minimum residuals method suggested by Kline (1994) with oblique rotation which allows for correlation between the latent factors.

Table 2. *Descriptive statistics for EFA.*

Item	Mean	Mdn	SD	Skewness	Kurtosis	Item	Mean	Mdn	SD	Skewness	Kurtosis
I1	2.954	3	0.802	-0.581	0.128	*I11*	3.075	3	0.832	-0.642	-0.109
I2	2.926	3	0.782	-0.467	0.007	*I12*	2.861	3	0.901	-0.423	-0.557
I3	3.159	3	0.791	-0.759	0.274	*I13*	2.269	2	0.953	0.358	-0.746
I4	2.925	3	0.843	-0.529	-0.175	*I14*	3.074	3	0.68	-0.82	1.753
I5	3.333	3	0.684	-1.073	1.954	*I15*	2.636	3	0.84	-0.098	-0.546
I6	3.056	3	0.818	-0.729	0.263	*I16*	2.704	3	0.788	-0.351	-0.152
I7	3.206	3	0.774	-0.873	0.653	*I17*	3.167	3	0.634	-1.047	3.366
I8	3.299	3	0.69	-0.999	1.716	*I18*	2.981	3	0.785	-0.439	-0.167
I9	2.843	3	0.888	-0.337	-0.622	*I19*	2.907	3	0.746	-0.54	0.391
I10	3.167	3	0.69	-0.752	1.211	*I20*	3.259	3	0.661	-0.733	1.147

The EFA results supported a two-factor structure, unlike the grounded TPB with 3 factors. Also, the Velicer's minimum average partial (MAP) (Velicer et al., 2000) test and parallel analysis supported two-factor structure (Figure 3). The two factors jointly captured 52.5% of the variance in the set of items and were positively correlated with each other (*r*= 0.66). Based on Howard's (2016) rule, five items (#1, #2, #4, #9, and #13) were removed (see Table 3). After removal of the items, the EFA was re-run and the factor structure and loadings were similar. Cronbach α values were equal to 0.91 and 0.88 for factor #1 and factor #2, respectively—

suggesting strong internal consistency. For discriminant validity, average variance extracted (AVE) values for factors were acceptable (*0.49 and 0.60*) and composite reliability (CR) values or factors were good (*0.89 and 0.90*) (Fornell & David, 1981). Also, the heterotrait-monotrait ratio of correlations (HTMT) was lower than 0.85 threshold (*0.82*) and this indicates the structure has sufficient discriminant validity (Henseler et al., 2015). Following the interpretation of these results and necessary revisions, a distinct sample was employed for CFA.

Figure 3. *Scree plot.*

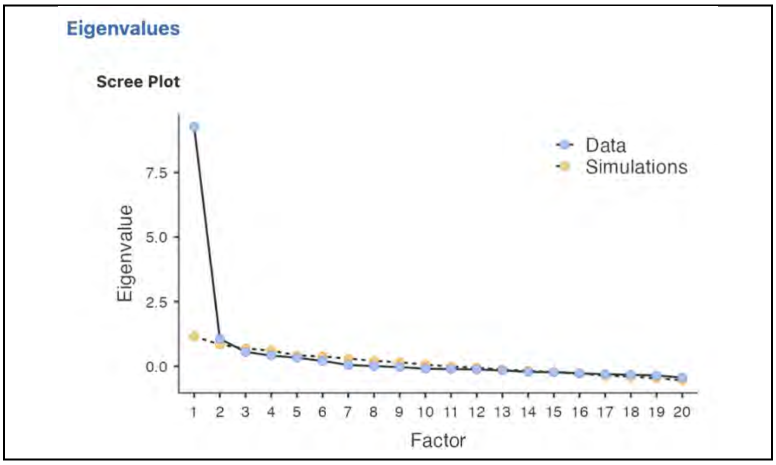


Table 3. *Factor loadings for EFA.*

	Item	Factor #1	Factor #2
I19	Learning new tools and information in assessment is part of my professional development.	0.837	
I14	I plan to continue learning new techniques about assessment.	0.832	
I5	I am open to exploring new assessment techniques to improve my teaching practices.	0.801	
I8	I view increasing my assessment literacy as a continuous process, rather than a one-time task.	0.774	
I18	I strive to use different applications and technology in assessment.	0.794	-0.138
I10	I value learning new concepts about assessment.	0.610	0.185
I16	I seek out opportunities to increase my assessment literacy.	0.569	0.260
I17	Learning innovative assessment approaches is valuable.	0.542	0.234
I15	I participate in professional development activities regarding assessment.	0.410	
I1	*ally strive to enhance my assessment literacy.*	*0.366*	*0.347*
I11	Faculty professional development in assessment is necessary for quality instruction.	-0.170	0.846
I3	I believe that improving my assessment literacy is crucial to enhance student learning outcomes.		0.840
I6	I believe that having strong assessment literacy is important for being an effective faculty member.		0.822
I12	I would like to complete more training in assessment in the future.	0.124	0.638
I20	I think faculty in higher education should have substantial knowledge in assessment.	0.134	0.624
I7	I believe that increasing my assessment literacy will help me to better meet the needs of a diverse student population.	0.206	0.621
I4	*I feel motivated to learn more about assessment strategies to better teach my students.*	*0.425*	*0.486*
I2	*I must stay current with the latest assessment methods to fulfill my teaching responsibilities.*	*0.314*	*0.468*
I9	*Frequent conversation with colleagues improves my assessment practices.*	*0.259*	*0.279*
I13	*I would only take an assessment training if it was required by my department.*		*-0.269*

Note: italicized items were removed

3.3. CFA Results

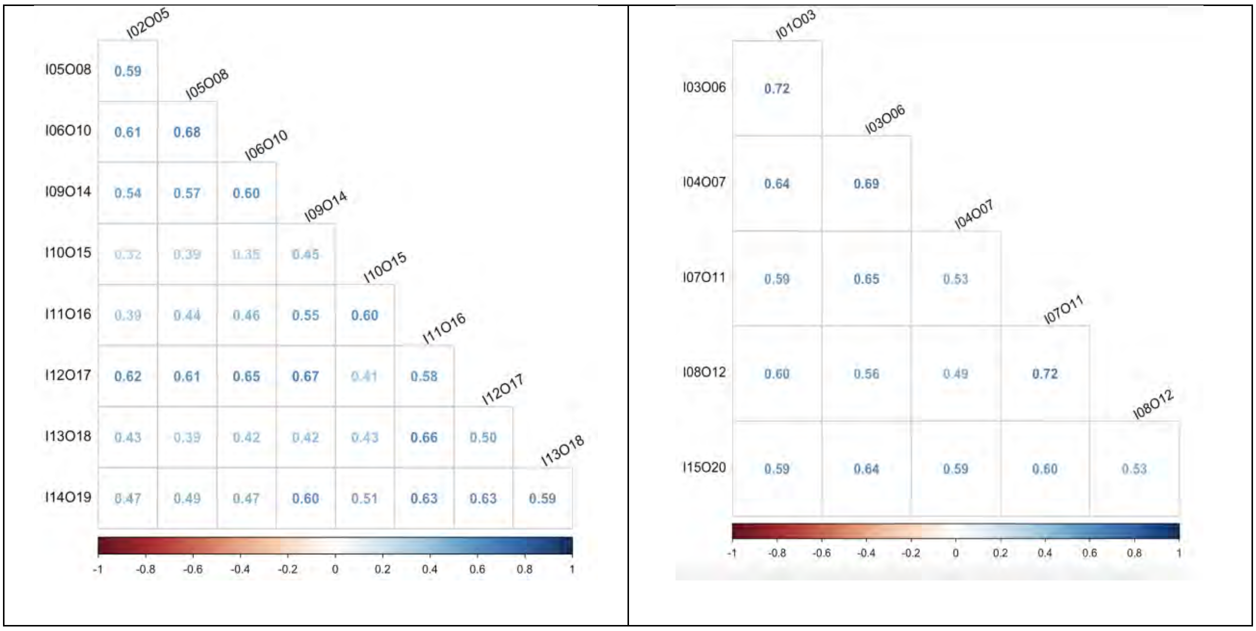
The descriptive statistics and reliability coefficients for the two dimensions of the scale are presented in Table 4. With the CFA sample, the Spearman correlation coefficients (ρ) between all pairs of items were within the range of $0.32 \leq \rho \leq 0.72$ and all of these values were statistically significant ($p<0.001$). Matrix representations of the correlations are provided in Figure 4.

Table 4. *Descriptive statistics and reliability coefficients.*

Factor	Item*	Mean	Mdn	*SD*	Skewness	Kurtosis	α	ω	Mean (F)	*SD* (F)
#1	I02O05	3.287	3	0.651	-0.789	1.387	0.917	0.921	2.944	0.569
	I05O08	3.28	3	0.696	-0.789	0.719				
	I06O10	3.141	3	0.74	-0.715	0.581				
	I09O14	3	3	0.716	-0.743	1.096				
	I10O15	2.478	3	0.764	-0.098	-0.349				
	I11O16	2.51	3	0.773	-0.074	-0.355				
	I12O17	3.089	3	0.664	-0.766	1.709				
	I13O18	2.879	3	0.827	-0.321	-0.466				
	I14O19	2.854	3	0.749	-0.216	-0.284				
#2	I01O03	2.955	3	0.728	-0.537	0.458	0.916	0.917	2.939	0.646
	I03O06	3.019	3	0.791	-0.589	0.099				
	I04O07	3.032	3	0.729	-0.552	0.407				
	I07O11	2.904	3	0.791	-0.616	0.275				
	I08O12	2.647	3	0.833	-0.478	-0.268				
	I15O20	3.064	3	0.74	-0.679	0.65				

*: I02O05 stands for Item #2 (Old Item #5), I05O08 stands for Item #5 (Old Item #8), etc.

Figure 4. *Inter-item correlation matrices.*



	I02O05	I05O08	I06O10	I09O14	I10O15	I11O16	I12O17	I13O18
I05O08	0.59							
I06O10	0.61	0.68						
I09O14	0.54	0.57	0.60					
I10O15	0.32	0.39	0.35	0.45				
I11O16	0.39	0.44	0.46	0.55	0.60			
I12O17	0.62	0.61	0.65	0.67	0.41	0.58		
I13O18	0.43	0.39	0.42	0.42	0.43	0.66	0.50	
I14O19	0.47	0.49	0.47	0.60	0.51	0.63	0.63	0.59

	I01O03	I03O06	I04O07	I07O11	I08O12
I03O06	0.72				
I04O07	0.64	0.69			
I07O11	0.59	0.65	0.53		
I08O12	0.60	0.56	0.49	0.72	
I15O20	0.59	0.64	0.59	0.60	0.53

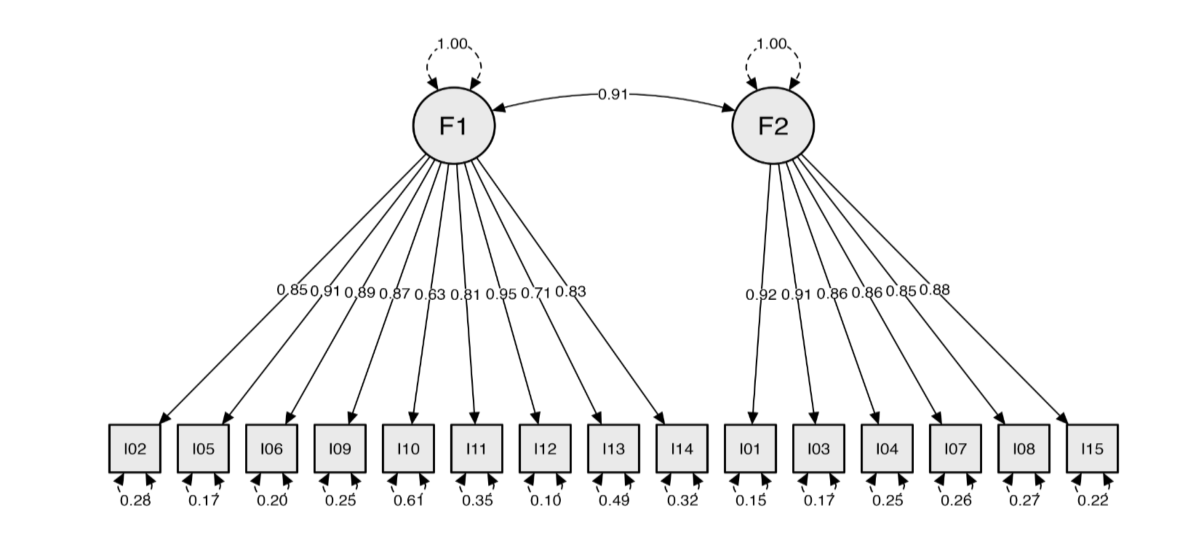
To be sure of the number of dimensions of the model, CFA was conducted with one-factor and two-factor models. The comparison was in favor of two-factor model and the results were given in Table 5.

Table 5. *Comparison of one-factor and two-factor models.*

Model	χ^2	*df*	CFI	RMSEA	SRMR	ΔX²	Δdf	ΔCFI	ΔRMSEA
One-factor	262.077	90	0.992	0.112	0.076	49.46***	1	0.003	0.017
Two- factor	212.617	89	0.995	0.095	0.070				

According to the CFA results, the model was statistically significant (χ^2= 212.617, *df*= 89, *p*<0.001). While some of the fit indices were in the good fit range (CFI= 0.995, TLI= 0.994), others were in acceptable fit range (SRMR= 0.070) or mediocre (slightly below the acceptable fit range; RMSEA= 0.095) (Hu & Bentler, 1999). Factor loadings for individual items ranged from 0.63 to 0.95, the correlation between factors was equal to 0.91, and all of the values were statistically significant (*p*<0.001). Figure 5 illustrates these results using a path diagram.

Figure 5. *Path diagram.*



3.4. MSA Results

Preliminary data screening revealed that the inter-item and corrected item-total correlations were positive for all items within each factor. Moreover, internal consistency statistics indicated acceptable levels of internal consistency reliability (Factor 1: α= 0.92; ωhierarchical,= 0.95; Factor 2: α= 0.92; ωhierarchical, total =0.94). Together, these results support further analysis of the ALAS items using a scaling approach.

Within each factor, all of the ALAS items adhered to monotonicity and IIO with no statistically significant violations. As shown in Table 6, all items had positive scalability coefficients. For Factor 1, individual item scalability coefficients ranged from Hi = 0.55 (*SE* = 0.07) for item I10O15 5 to Hi = 0.72 (*SE = 0.04*) for item I09O14. The overall scalability coefficient for the ALAS items in Factor 1 was equal to H = 0.67 (*SE = 0.04*). For Factor 2, individual item scalability coefficients ranged from Hi = 0.68 (*SE = 0.05*) for item I04O07 to Hi = 0.75 (*SE = 0.04*) for item I08O12. The overall scalability coefficient for Factor 2 was equal to H = 0.73 (*SE = 0.04*).

Table 6. *Item scalability coefficients.*

Factor #1			Factor #2		
Item	Item Scalability (H_i)	Standrard Error	Item	Item Scalability (H_i)	Standard Error
I02O05	0.67	0.05	I01O03	0.74	0.04
I05O08	0.66	0.05	I03O06	0.75	0.04
I06O10	0.67	0.05	I04O07	0.69	0.05
I09O14	0.72	0.04	I07O11	0.73	0.04
I10O15	0.55	0.07	I08O12	0.75	0.04
I11O16	0.70	0.04	I15O20	0.70	0.05
I12O17	0.74	0.04			
I13O18	0.61	0.06			
I14O19	0.71	0.04			

3.5. Summary of the Findings

In reconciling the outcomes of factor analysis and MSA, we identified and confirmed two factors, deviating from the initially hypothesized three within the TPB framework. Nevertheless, it is crucial to note that these two factors align with the core constructs of the theory. The explanation for each factor is provided below.

Factor 1: Attitude in Learning (new approaches, tools, etc.). This factor appears to align closely with the "attitude" construct in the TPB. Participants' attitude in learning new approaches and tools likely encompasses their personal evaluations of the benefits and drawbacks associated with adopting new assessment approaches. This factor could influence faculty's inclination to embrace assessment literacy practices.

Factor 2: Perceived Importance of Assessment Literacy (AL). This factor can be seen as a combination of the subjective norms and perceived behavioral control constructs from the TPB. The perceived importance of assessment literacy may reflect social influences (subjective norms) where educators gauge the significance of assessment literacy based on external factors such as colleagues' opinions or institutional priorities. Additionally, the perceived importance of assessment literacy could encompass a sense of control over the behavior (perceived behavioral control), as faculty may believe that developing assessment literacy is a critical factor within their sphere of influence.

4. DISCUSSION

This study aims to develop a psychometrically-sound assessment literacy attitude scale for educators, especially in higher education sector. We addressed a series of questions to fulfill this aim. In addressing the first research question, the EFA results revealed a two-factor structure, deviating from the anticipated three factors posited by the TPB. This unexpected outcome underscores the complexity inherent in the domain of assessment literacy enhancement. Despite this departure from the anticipated structure, the findings suggest a reinterpretation of the TPB in our context. The elements within the TPB framework persist in significance, despite their reconfigured arrangement, with attitudes toward adopting new assessment approaches reflecting individual inclinations and the perceived importance of assessment literacy. This encapsulates both subjective norms and perceived behavioral control. The realignment underscores the complex nature of fostering assessment literacy within the TPB framework in the context of higher education.

Subsequent to the EFA, a careful inspection of factor loadings led to the removal of five items (Research question [RQ] #2). Two items, one of which was negatively-worded, were discarded due to their failure to load onto any factor. The negative formulation, as repeatedly cautioned in the literature (Chang, 1995; Cole et al., 2019; Sliter & Zickar, 2014; Wright & Masters, 1982), raised concerns about the item's interpretability by respondents and fit of the data. Notably, cautionary evidence from a study by Sliter and Zickar (2014) employing IRT framework highlighted that the scales comprising only positively-worded items yielded more information at the peak of the information curve and across a wider range of the trait scale. Moreover, trait scores were estimated with smaller standard errors under such conditions. In addition, think-aloud session insights revealed discrepancies in the length of participant discussions for the other removed item, emphasizing the importance of incorporating qualitative methods in the scale development process (Morell & Tan, 2009; Zhou, 2019).

The proposed factor structure, derived from the EFA findings, underwent scrutiny via CFA with a distinct sample using the 15-item version of the ALAS instrument (RQ #3). The CFA results displayed commendable fit indices, characterized by favorable CFI and TLI values. However, a slight deviation was observed in the RMSEA, resting at 0.095. In conducting both the EFA and CFA, we adhered to the commonly recommended guideline of a minimum of 5 respondents per item (Tabachnick & Fidell, 1996). However, despite meeting this criterion, our RMSEA, a measure assessing how well the model reproduces the observed data, fell within the medium range. This outcome underscores the influence of sample size on fit indices, as larger sample sizes tend to yield more precise estimates, potentially leading to lower RMSEA values. This observation emphasizes the significance of considering sample size implications in interpreting CFA outcomes and points towards a prospective avenue for future research to explore the robustness of the factor structure across diverse sample sizes.

Within the domain of scale development studies, the integration of both classical test theory (CTT) and item response theory (IRT) stands as a crucial and frequently employed practice for a comprehensive assessment of psychometric properties (Dilek & Akbaş, 2022; McKown et. al. 2023; Wright & Jenkins-Guarnieri, 2023). In alignment with this methodological paradigm, our study underscores the importance of using multiple measurement approaches to elucidate the underlying constructs of our assessment literacy instrument. The synergistic use of EFA, CFA, and MSA, was designed to establish a robust foundation for comprehending the factor structure and measurement properties. In pursuit of this objective, we extended our inquiry to MSA (RQ #4), further enriching the depth of our psychometric evaluation. All of the ALAS items demonstrated monotonicity and IIO without statistically significant violations, thereby enhancing the interpretability of our assessment literacy instrument.

5. CONCLUSION

Our study contributes a distinctive perspective on the application of Theory of Planned Behavior (TPB) constructs within the realm of advancing assessment literacy in higher education. The findings underscore the perceived significance of enhancing assessment literacy in facilitating the adoption of faculty development programs, innovative assessment methodologies, and tools. This perceived significance is propelled by various factors, including social influence stemming from institutional priorities and the recognition of assessment as a pivotal determinant of faculty influence. Higher education institutions can capitalize on faculty perceptions by strategically elevating the place of assessment among institutional priorities. This emphasis should be tangibly manifested through a multifaceted approach, including targeted assessment workshops in diverse formats, specialized training modules, and hands-on practice sessions. To enhance accessibility and support, institutions may consider incorporating artificial intelligence tools, such as chatbots, to provide prompt assistance in assessment-related queries. Such a technologically-driven support system, available 24/7, may empower faculty members with real-time guidance and resources, fostering a dynamic and responsive culture of assessment literacy.

While this study has yielded valuable insights, it is crucial to acknowledge its limitations. Firstly, the reliance on a relatively small sample size, comprising faculty members from public sector R1 universities, may restrict the generalizability of findings to broader populations within higher education, particularly in diverse contexts like private universities and teaching-based institutions. Future studies should encompass more varied university settings to ensure a comprehensive understanding. The limited scope of participants might not fully capture the diverse perspectives and experiences prevalent in larger academic environments. In the context of assessment practices, a gap in the literature still pertains to the relationship between faculty members' planned assessment enhancement behavior and their attitudes. This aspect requires further exploration and research to enrich the existing body of knowledge. Additionally, the use of self-report measures introduces a potential source of bias, as participants may respond based on perceived beliefs rather than providing objective assessments of their behavior. While the study makes significant contributions to the comprehension of assessment literacy and faculty development, these limitations underscore the necessity for future research with larger and more diverse samples. Incorporating objective measures, such as an assessment literacy level test, will further enhance the robustness of the findings.

Acknowledgments

This research was financially supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under grant number 1059B192201517.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: University of Alabama, 23-04-6561.

Contribution of Authors

Beyza Aksu Dünya: Funding, Literature review, Conception, Design, Investigation, Methodology, Item writing, Data collection, Receiving experts' opinions, and Writing-original draft. **Stefanie A. Wind**: Statistical analysis, Supervision, and Critical review. **Mehmet Can Demir**: Literature review, Conception, Methodology, Data interpretation, Statistical analysis, and Writing-original draft.

Orcid

Beyza Aksu Dünya https://orcid.org/0000-0003-4994-1429
Stefanie A. Wind https://orcid.org/0000-0002-1599-375X
Mehmet Can Demir https://orcid.org/0000-0001-7849-7078

REFERENCES

Adam, S. (2004). *Using learning outcomes: A consideration of the nature, role, application and implications for European education of employing "learning outcomes" at the local, national and international levels.* Paper presented at the Bologna Seminar, Heriot-Watt University, Edinburgh United Kingdom. http://www.aic.lv/ace/ace_disk/Bologna/Bol_semin/Edinburgh/S_Adam_Bacgrerep_presentation.pdf Accessed on 16 November 2023.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

Ajzen, I. (2001). Nature and operation of attitudes. *Annual Review of Psychology*, *52*(1), 27-58. https://doi.org/10.1146/annurev.psych.52.1.27

Ajzen, I., & Timko, C. (1986). Correspondence between health attitudes and behavior. *Basic and Applied Social Psychology*, *7*(4), 259-276. https://doi.org/10.1207/s15324834basp0704_2

Archie, T., Hayward, C.N., Yoshinobu, S., & Laursen, S.L. (2022). Investigating the linkage between professional development and mathematics instructors' use of teaching practices using the theory of planned behavior. *Plos One*, *17*(4), e0267097. https://doi.org/10.1371/journal.pone.0267097

Balloo, K., Norman, M., & Winstone, N.E. (2018, January). Evaluation of a large-scale inclusive assessment intervention: a novel approach to quantifying perceptions about assessment literacy. In *The Changing Shape of Higher Education-Can Excellence and Inclusion Cohabit?: Conference Programmme and Book of Abstracts*. University of Southern Queensland. https://srhe.ac.uk/arc/conference2018/downloads/SRHE_Conf_2018_Programme_Papers.pdf

Biggs, J., & Tang, C. (2011). Train-the-trainers: Implementing outcomes-based teaching and learning in Malaysian higher education. *Malaysian Journal of Learning and Instruction*, 8, 1-19.

Caspersen, J., & Smeby, J.C. (2018). The relationship among learning outcome measures used in higher education. *Quality in Higher Education*, *24*(2), 117-135. https://doi.org/10.1080/13538322.2018.1484411

Chang, L. (1995). Connotatively consistent and reversed connotatively inconsistent items are not fully equivalent: Generalizability study. *Educational and Psychological Measurement, 55*(6), 991-997. https://doi.org/10.1177/0013164495055006007

Coates, H. (2016). Assessing student learning outcomes internationally: Insights and frontiers. *Assessment & Evaluation in Higher Education*, *41*(5), 662-676. https://doi.org/10.1080/02602938.2016.1160273

Cochran, W.G. (1977). *Sampling techniques*. John Wiley & Sons.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cole, K.L., Turner, R.C., & Gitchel, W.D. (2019). A study of polytomous IRT methods and item wording directionality effects on perceived stress items. *Personality and Individual Differences, 147*(6), 63-72. https://doi.org/10.1016/j.paid.2019.03.046

Conner, M., & Armitage, C.J. (1998). Extending the theory of planned behavior: A review and avenues for further research. *Journal of Applied Social Psychology*, *28*(15), 1429-1464. https://doi.org/10.1111/j.1559-1816.1998.tb01685.x

Creswell, J.W., & Clark, V.P. (2011). *Mixed methods research*. SAGE Publications.

Crick, R.D., Broadfoot, P., & Claxton, G. (2004). Developing an effective lifelong learning inventory: The ELLI project. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 247-272. https://doi.org/10.1080/0969594042000304582

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Dann, R. (2014). Assessment as learning: blurring the boundaries of assessment and learning for theory, policy and practice. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 149-166. https://doi.org/10.1080/0969594X.2014.898128

Dilek, H., & Akbaş, U. (2022). Investigation of education value perception scale's psychometric properties according to CTT and IRT. *International Journal of Assessment Tools in Education, 9*(3), 548-564. https://doi.org/10.21449/ijate.986530

Dill, D. (2007). *Quality assurance in higher education: Practices and issues.* The 3rd International Encyclopedia of Education.

Dunn, R., Hattie, J., & Bowles, T. (2018). Using the Theory of Planned Behavior to explore teachers' intentions to engage in ongoing teacher professional learning. *Studies in Educational Evaluation*, 59, 288-294. https://doi.org/10.1016/j.stueduc.2018.10.001

Eubanks, D. (2019). Reassessing the elephant, part 1. *Assessment Update*, *31*(2), 6-7. https://doi.org/10.1002/au.30166

Evans, C. (2016). *Enhancing assessment feedback practice in higher education: The EAT framework*. University of Southampton. https://www.southampton.ac.uk/assets/importe d/transforms/content-block/UsefulDownloads_Download/A0999D3AF2AF4C5AA24B5BEA08C61D8E/EAT%20Guide%20April%20FINAL1%20ALL.pdf

Field, A. (2003). *Discovering Statistics using IBM SPSS statistics*. Sage Publications.

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it [Editorial]. *European Journal of Psychological Assessment, 33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460

Fornell, C., & David, F.L. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research,* 18, 39-50. https://doi.org/10.2307/3151312

Henseler, J., Ringle, C.M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy Marketing Science,* 43, 115–135. https://doi.org/10.1007/s11747-014-0403-8

Hines, S.R. (2009). Investigating faculty development program assessment practices: What's being done and how can it be improved?. *The Journal of Faculty Development*, *23*(3), 5.

Holmboe, E.S., Ward, D.S., Reznick, R.K., Katsufrakis, P.J., Leslie, K.M., Patel, V.L., ... & Nelson, E.A. (2011). Faculty development in assessment: the missing link in competency-based medical education. *Academic Medicine, 86*(4), 460-467. https://doi.org/10.1097/acm.0b013e31820cb2a7

Hora, M.T., & Anderson, C. (2012). Perceived norms for interactive teaching and their relationship to instructional decision-making: A mixed methods study. *Higher Education*, *64*, 573-592. https://doi.org/10.1007/s10734-012-9513-8

Howard, M.C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve?. *International Journal of Human-Computer Interaction*, *32*(1), 51-62. https://doi.org/10.1080/10447318.2015.1087664

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

Jankowski, N.A., & Marshall, D.W. (2017). *Degrees that matter: Moving higher education to a learning systems paradigm*. Routledge. https://doi.org/10.4324/9781003444015

Kao, C.P., Lin, K.Y., & Chien, H.M. (2018). Predicting teachers' behavioral intentions regarding web-based professional development by the theory of planned behavior. *EURASIA Journal of Mathematics, Science and Technology Education*, *14*(5), 1887-1897. https://doi.org/10.29333/ejmste/85425

Kline, P. (1994). *An easy guide to factor analysis*. Routledge.

Knauder, H., & Koschmieder, C. (2019). Individualized student support in primary school teaching: A review of influencing factors using the Theory of Planned Behavior (TPB). *Teaching and Teacher Education*, *77*, 66-76. https://doi.org/10.1016/j.tate.2018.09.012

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

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Ligtvoet, R., Van der Ark, L.A., Marvelde, J.M. te, & Sijtsma, K. (2010). Investigating an Invariant Item Ordering for Polytomously Scored Items. *Educational and Psychological Measurement*, *70*(4), 578–595. https://doi.org/10.1177/0013164409355697

Liu, O.L., Bridgeman, B., & Adler, R.M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*(9), 352-362. https://doi.org/10.3102/0013189X12459679

Madigan, D.J., & Kim, L.E. (2021). Towards an understanding of teacher attrition: A meta-analysis of burnout, job satisfaction, and teachers' intentions to quit. *Teaching and Teacher Education*, 105, 103425. https://doi.org/10.1016/j.tate.2021.103425

Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for kernel smoothing in Item Response Theory. *Journal of Statistical Software*, *58*(6). https://doi.org/10.18637/jss.v058.i06

McDonald, R.P. (1999). *Test theory: A unified treatment*. Taylor & Francis.

McKown, C., Kharitonova, M., Russo-Ponsaran, N.M., & Aksu-Dunya, B. (2023). Development and Validation of a shortened form of SELweb EE, a Web-Based Assessment of Children's Social and Emotional Competence. *Assessment, 30*(1), 171-189. https://doi.org/10.1177/10731911211046044

Medland, E. (2019). 'I'm an assessment illiterate': Towards a shared discourse of assessment literacy for external examiners. *Assessment & Evaluation in Higher Education*, *44*(4), 565-580. https://doi.org/10.1080/02602938.2018.1523363

Meijer, R.R., & Baneke, J.J. (2004). Analyzing psychopathology items: A case for Nonparametric Item Response Theory Modeling. *Psychological Methods*, *9*(3), 354–368. https://doi.org/10.1037/1082-989X.9.3.354

Meijer, R.R., Tendeiro, J.N., & Wanders, R.B.K. (2015). The use of nonparametric item response theory to explore data quality. In S.P. Reise & D.A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to typical performance assessment* (pp. 85–110). Routledge.

Mokken, R.J. (1971). *A theory and procedure of scale analysis*. De Gruyter.

Morell, L., & Tan, R.J.B. (2009). Validating for use and interpretation: A mixed methods contribution illustrated. *Journal of Mixed Methods Research, 3*(3), 242-264. https://doi.org/10.1177/1558689809335079

Muthén, B.O. (1993). Goodness of fit with categorical and other nonnormal variables. In K.A. Bollen, & J.S. Long (Eds.), *Testing structural equation models* (pp. 205-234). Sage Publishing.

O'Neill, G., McEvoy, E., & Maguire, T. (2023). Supporting assessment literacy in changing times. In C. Evans and M. Waring (Eds.), *Research handbook on innovations in assessment and feedback in higher education*. Elgar Publishing.

Padilla, J.L., & Leighton, J.P. (2017). Cognitive interviewing and think aloud methods. In B. Zumbo & A. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 211-228). Springer.

Pastore, S. (2022). Assessment Literacy in the higher education context: A critical review. *Intersection: A Journal at the Intersection of Assessment and Learning, 4*(1). https://doi.org/10.61669/001c.39702

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

Pett, M.A., Lackey, N.R., & Sullivan, J.J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage Publications.

Price, M., Rust, C., ODonovan, B., Handley, K., & Bryant, R. (2012). *Assessment literacy: The foundation for improving student learning*. ASKe, Oxford Centre for Staff and Learning Development.

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramsay, J.O., & Silverman, B.W. (2005). *Functional data analysis* (2nd ed.). Springer.

Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University, Evanston, Illinois. R package version 2.3.9, https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rimal, R.N., & Real, K. (2003). Understanding the influence of perceived norms on behaviors. *Communication Theory*, *13*(2), 184-203. https://doi.org/10.1111/j.1468-2885.2003.tb00288.x

Sadler, D.R. (2017). Academic achievement standards and quality assurance. *Quality in Higher Education*, *23*(2), 81-99. https://doi.org/10.1080/13538322.2017.1356614

Scholl, K., & Olsen, H.M. (2014). Measuring student learning outcomes using the SALG instrument. *SCHOLE: A Journal of Leisure Studies and Recreation Education*, *29*(1), 37-50. https://doi.org/10.1080/1937156X.2014.11949710

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Sage Publications.

Sijtsma, K., & van der Ark, L.A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158. https://doi.org/10.1111/bmsp.12078

Singh, M., & Ramya, K.R. (2011). Outcome based education. *International Journal of Nursing Education*, *3*(2), 87-91.

Sliter, K.A., & Zickar, M.J. (2014). An IRT examination of the psychometric functioning of negatively worded personality items. *Educational and Psychological Measurement, 74*(2), 214-226. https://doi.org/10.1177/0013164413504584

Tabachnick, B., & Fidell, L.S. (1996). *Using multivariate statistics*. Harper Collins.

Van der Ark, L.A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20, 1-19. https://doi.org/10.18637/jss.v020.i11

Van der Ark, L.A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software*, 48, 1-27. https://doi.org/10.18637/jss.v048.i05

Velicer, W.F., Eaton, C.A., & Fava, J.L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R.D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment* (pp. 41-71). Kluwer.

Williams, J. (2016). Quality assurance and quality enhancement: Is there a relationship?. *Quality in Higher Education*, *22*(2), 97-102. https://doi.org/10.1080/13538322.2016.1227207

Wolf, R., Zahner, D., & Benjamin, R. (2015). Methodological challenges in international comparative post-secondary assessment programs: Lessons learned and the road ahead. *Studies in Higher Education*, *40*(3), 471-481. https://doi.org/10.1080/03075079.2015.1004239

Wright, B.D., & Masters, G.N. (1982). *Rating Scale Analysis*. MESA Press.

Wright, S.L., & Jenkins-Guarnieri, M.A. (2023). Further validation of the social efficacy and social outcome expectations scale. *Journal of Psychoeducational Assessment*, *42(1), 74-88.* https://doi.org/10.1177/07342829231198277

Xu, Y., & Brown, G.T. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149-162. https://doi.org/10.1016/j.tate.2016.05.010

Zhou, Y. (2019). A mixed methods model of scale development and validation analysis. *Measurement: Interdisciplinary Research and Perspectives, 17*(1), 38-47. https://doi.org/10.1080/15366367.2018.1479088

Zhu, X., & Evans, C. (2022). Enhancing the development and understanding of assessment literacy in higher education. *European Journal of Higher Education*, 1-21. https://doi.org/10.1080/21568235.2022.2118149

Zoom Video Communications, Inc. (2023). *ZOOM cloud meetings* (Version 5.15.5). https://zoom.com