International Journal for Research Trends in Social Science & Humanities
Volume 4, Issue 3
May-Jun 2026, PP 1-17

Assessment Literacy and Professional Development Needs Among Egyptian EFL Teachers: A Six-Domain Survey Study Including Smart Assessment and AI-Based Evaluation

Andrew Ghaly
Pharos University in Alexandria, Egypt

Abstract- Classroom-based assessment occupies a large share of an English-as-a-foreign-language (EFL) teacher's working week, yet many teachers receive little structured preparation for it. This study surveyed 140 in-service EFL teachers, all graduates of English department teaching divisions in Egyptian public university faculties of education, using a 16-item questionnaire covering six domains of assessment literacy: understanding assessment purposes and principles, validity, reliability, and fairness, test and task design, scoring, rubrics, and standard setting, data interpretation and use of results, and professional and ethical practices. Each item was rated on a five-point scale ranging from strongly disagree to strongly agree. The questionnaire also probed participants' familiarity with emerging concepts such as smart assessment and AI-based evaluation, and their perceptions of whether their pre-service methodology courses had adequately prepared them in these areas. Overall self-reported assessment literacy was moderate ($M = 3.03$, $SD = 1.31$), with the highest mean ratings in understanding assessment purposes and principles ($M = 3.30$) and the lowest in data interpretation and use of results ($M = 2.57$). A large majority of participants reported that their faculties had not embedded sufficient assessment literacy training in the pre-service curriculum, and nearly all indicated that their methodology coursework had contained no instruction on using AI tools for test design or data analysis. Familiarity with smart assessment concepts was similarly low. Teachers who had completed formal assessment training scored markedly higher on every domain (all $p < .001$, with mean differences of roughly two scale points), and years of teaching experience was positively correlated with assessment literacy ($r = .31$ to $.38$, all $p < .001$). Teaching level — primary, secondary, adult education, or university — showed no significant relationship with any domain. The findings suggest that Egyptian faculties of education need to substantially revise pre-service assessment curricula to address both foundational assessment literacy and the growing demands of technology-enhanced and AI-assisted assessment.

Keywords- language assessment literacy, EFL teachers, Egyptian faculties of education, smart assessment, AI-based evaluation, professional development, classroom assessment.

I. Introduction

Assessment takes up a large part of a language teacher's working life — writing tests, designing speaking and writing tasks, scoring student work, and using the results to make decisions about grades, placement, and what to teach next. Yet many teachers receive little structured preparation for any of this. Research across a range of EFL contexts has repeatedly found that in-service teachers rate their own assessment literacy as low to moderate and ask for more training (Isnawati, 2023; Lan et al., 2019; Sun et

al., 2022). In Egypt, where EFL teachers are predominantly trained through the English department teaching divisions of faculties of education in public universities, pre-service preparation in language assessment has received comparatively little scholarly attention, despite a national curriculum that increasingly emphasizes outcomes-based assessment and communicative testing approaches.

Language assessment literacy (LAL) has been defined in different ways over the past two decades, moving from a narrow focus on test construction and statistics toward a broader, more situated view that also includes teachers' beliefs, values, and capacity to reflect on their own practice (Coombe et al., 2020; Scarino, 2013). Coombe et al. (2020) described LAL as a combination of the knowledge teachers need, the skills required to apply that knowledge in their own classrooms, and the principles that guide how assessment information is interpreted and used in a given context. This broader view has implications for how LAL is measured: rather than producing a single score, researchers increasingly break the construct down into separate domains — such as understanding the purposes of assessment, designing valid and fair tasks, scoring consistently, and using results responsibly — each of which may show a different profile of strengths and needs (Bøhn et al., 2021; Lo et al., 2022). More recently, the emergence of smart assessment — broadly referring to technology-enhanced assessment systems that adapt to learner performance and generate diagnostic feedback automatically — and AI-based evaluation tools has introduced a further dimension of assessment literacy that conventional frameworks have yet to fully address. Teachers are increasingly expected to understand, critically evaluate, and potentially integrate these tools, yet there is little evidence about how well pre-service programs are preparing them to do so.

A growing number of survey-based needs analyses have applied this domain-based approach to EFL teachers in specific contexts, including Indonesia (Isnawati, 2023; Wiyaka et al., 2024), China (Lan et al., 2019; Sun et al., 2022), Taiwan (Chang et al., 2024), Yemen (Al-Akbari et al., 2024), and several European countries (Vogt et al., 2020). These studies converge on two broad findings. First, teachers' self-reported assessment literacy is rarely uniform across domains: conceptual knowledge of assessment purposes tends to be relatively stronger, while skills related to test specification, item writing, and especially the statistical interpretation of results tend to be weaker (Al-Akbari et al., 2024; Chang et al., 2024). Second, formal training and teaching experience are consistently associated with higher assessment literacy, although the size and pattern of these relationships vary across contexts (Sun et al., 2022). No comparable study has examined Egyptian EFL teachers specifically, nor has any study in the broader LAL literature explicitly investigated familiarity with smart assessment concepts or the presence of AI-related content in methodology courses.

The present study extends this line of research by examining the self-reported assessment literacy of 140 in-service EFL teachers in Egypt across six domains — covering assessment purposes and principles, validity, reliability, and fairness, test and task design, scoring, rubrics, and standard setting, data interpretation and use of results, and professional and ethical practice — and by relating these profiles to teachers' formal training, years of experience, and teaching level. The study also investigates participants' perceptions of the adequacy of their faculty preparation, their familiarity

with smart assessment and AI-based evaluation, and whether their pre-service methodology courses included any content on using AI tools for test design or data analysis. Four research questions guided the study:

- Research Question 1. What is the overall level of self-reported assessment literacy among the surveyed Egyptian EFL teachers, and how does it vary across the six domains?
- Research Question 2. How do teachers' assessment literacy profiles differ according to formal assessment training, years of teaching experience, and the level at which they teach?
- Research Question 3. Which domains represent the greatest priorities for professional development?
- Research Question 4. How do participants perceive the adequacy of their faculty preparation for assessment, and what is their level of familiarity with smart assessment and AI-based evaluation concepts, including whether their methodology courses included instruction on AI tools for test design or data analysis?

II. Literature Review

Conceptualizing Language Assessment Literacy

Early definitions of assessment literacy focused mainly on the technical skills needed to construct and score tests — writing items, calculating reliability, and interpreting basic statistics. Over time, researchers argued that this technical view was too narrow for language teachers, who also need to understand how assessment relates to curriculum, how to give useful feedback, and how their own beliefs shape the decisions they make about students (Scarino, 2013). Scarino (2013) framed assessment literacy partly as a form of self-awareness: teachers need to recognize that their interpretations of student performance are not neutral, but are shaped by their own assumptions about language learning and their own assessment history.

Coombe et al. (2020) reviewed this shift and argued that language assessment literacy now needs to be understood as a combination of knowledge, skills, and principles — what teachers know about assessment concepts and procedures, what they can do with that knowledge in their own classrooms, and the values and context-sensitivity that guide how they use assessment in practice. Other researchers have proposed that different stakeholder groups, such as classroom teachers, teacher educators, and test developers, may need different combinations of these components (Bøhn et al., 2021; Lo et al., 2022). Lo et al. (2022), for example, argued that teachers working in content-based programs need an assessment literacy profile that links subject knowledge, language proficiency, and assessment skills in ways that a generic framework does not fully capture. Despite ongoing debate about how exactly to define the construct, most frameworks agree that assessment literacy is multidimensional, and that teachers are likely to be stronger in some areas, such as basic conceptual understanding, than in others, such as statistical analysis or rubric design (Coombe et al., 2020).

Survey Evidence from EFL Contexts

This multidimensional view is reflected in survey research on EFL teachers' assessment literacy. Isnawati (2023) administered an Assessment Literacy Inventory to secondary

school EFL teachers and found that their overall assessment literacy was relatively low, recommending more sustained training and professional development as a result. Sun et al. (2022) surveyed 272 college EFL teachers in China and reported similarly modest levels of involvement in assessment activities and self-perceived literacy, along with reported needs for training in nearly every aspect of language assessment; teachers' accumulated assessment experience and prior training were both significant predictors of their current literacy levels. Working with 344 middle school teachers, Lan et al. (2019) found that EFL teachers in their sample had reached only a functional level of classroom-based language assessment literacy and wanted further training to become more procedurally and conceptually literate, particularly around using assessment information in instruction.

Studies that break assessment literacy down by domain tend to find a consistent pattern: teachers report greater confidence in conceptual or principle-based domains than in domains that require statistical or technical skills. Chang et al. (2024), surveying 57 senior high school English teachers in Taiwan, found that teachers' greatest training needs centered on providing feedback, generating teaching content from assessment results, and assessing integrated skills — all of which depend on being able to interpret and act on assessment data. Al-Akbari et al. (2024), surveying 471 EFL teachers in Yemen, identified a ten-dimension model of assessment literacy and found that teachers' greatest training needs were in knowledge of assessment and in assessment administration and use, again pointing to the practical, results-oriented end of the construct. Similarly, Wiyaka et al. (2024) found that EFL teachers in Indonesian higher education institutions were comparatively stronger in assessing reading than listening, illustrating how assessment literacy needs can vary by skill area as well as by domain. At a larger scale, the Teachers' Assessment Literacy Enhancement project surveyed nearly 1,800 learners and 658 teachers across four European countries and found that classroom assessment practices remained heavily weighted toward traditional, discrete-point formats, with feedback often limited to marks and brief comments (Vogt et al., 2020).

Professional Development and Assessment Literacy

A separate line of research has examined whether and how professional development changes teachers' assessment literacy. Mertler (2009) found that a two-week workshop based on the Standards for Teacher Competence in Educational Assessment of Students produced clear gains on a pre-post assessment literacy inventory, with teachers' reflective journals describing the training as practically useful. Koh (2011) compared teachers who received ongoing, sustained professional development in designing authentic assessments and rubrics with teachers who received only short, one-off workshops, and found that assessment literacy gains were significantly larger for the sustained group. More recently, Mirsanjari (2025) examined a teacher development course for English-for-academic-purposes instructors and found significant pre-post gains in designing valid assessments and applying formative feedback, but also found that the size of these gains depended on teaching experience: less experienced instructors improved the most, while more experienced instructors, often working within established institutional routines, improved comparatively less.

Other studies have shown that professional development can build formative assessment literacy specifically, often through sustained, collaborative formats rather than single sessions. Li, J. et al. (2023) followed a group of secondary EFL teachers in China through a 12-week collaborative action research program and documented substantial growth in one teacher's knowledge, beliefs, and classroom practice around formative assessment. Li, Z. et al. (2023) similarly found that a professional development program improved primary teachers' formative assessment literacy in Hong Kong, with features such as professional learning communities contributing to its success. Taken together, this research suggests that the effects of training on assessment literacy are not uniform: they depend on how the training is structured and on the experience level of the teachers receiving it (Giraldo, 2021; Villa Larenas et al., 2022). Villa Larenas et al. (2022) add a further complication, showing that even the teacher educators responsible for training future EFL teachers in assessment often have gaps in their own assessment literacy, which may limit how well pre-service training translates into classroom practice.

The present study builds on this literature by providing a domain-level profile of assessment literacy for a sample of 140 in-service Egyptian EFL teachers and by examining how that profile relates to formal training, experience, and teaching level — factors that the literature suggests should matter, but whose combined effects have rarely been examined within a single six-domain instrument in the Egyptian context. Beyond the six established domains, the study responds to calls from researchers such as Coombe et al. (2020) and Giraldo (2021) to examine how well pre-service programs are preparing teachers for contemporary assessment demands, including the growing role of smart assessment systems and AI-based evaluation tools. Egyptian faculties of education represent a particularly important site for this investigation: virtually all state-sector EFL teachers in Egypt pass through these institutions, and if the assessment-related content of their methodology courses is inadequate or outdated, the effects on classroom practice are likely to be widespread.

II. Method

Participants

Participants were 140 in-service EFL teachers working across primary, secondary, adult/private language institute, and university settings in Egypt. All participants were graduates of the English department teaching divisions of faculties of education in Egyptian public universities, which constitute the primary pre-service training pathway for EFL teachers in the Egyptian state education system. Their ages ranged from 21 to 67 years (M = 39.84, SD = 10.10). In terms of teaching experience, 5 teachers (3.6%) had 0–2 years of experience, 25 (17.9%) had 3–5 years, 32 (22.9%) had 6–10 years, and 78 (55.7%) had more than 10 years. With regard to teaching level, 54 teachers (38.6%) taught at the university level, 38 (27.1%) taught secondary level, 34 (24.3%) taught in adult or private language institutes, and 14 (10.0%) taught at the primary level. Slightly more than half of the sample, 78 teachers (55.7%), reported having received formal training in language assessment since graduating, while 62 (44.3%) had not. Full demographic details are presented in Table 1.

Instrument

The questionnaire consisted of 16 statements organized into six assessment literacy domains, each rated on a five-point Likert scale (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree). The domains were (A) Understanding Assessment Purposes and Principles (3 items), (B) Validity, Reliability, and Fairness (4 items), (C) Test and Task Design (3 items), (D) Scoring, Rubrics, and Standard Setting (2 items), (E) Data Interpretation and Use of Results (2 items), and (F) Professional and Ethical Practices (2 items). Sample items included "I can explain the main purposes of formative, summative, and diagnostic assessment" (Domain A) and "I feel confident interpreting statistical indicators such as mean, item facility, and reliability coefficients" (Domain E).

Beyond the six domains, a supplementary section asked participants to respond to four additional statements using the same five-point scale: (G1) "My faculty of education provided sufficient training in language assessment to prepare me for classroom practice"; (G2) "I am familiar with the concept of smart assessment and how it differs from traditional testing approaches"; (G3) "I understand how AI-based evaluation tools can be used to support language assessment"; and (G4) "My pre-service methodology courses included at least one module or unit on using AI tools for test design or data analysis." These items were analyzed descriptively and were not included in the six-domain composite scores. The questionnaire also collected demographic information, including age, years of teaching experience, the level(s) currently taught, and whether the respondent had received formal training in language assessment since graduation.

Procedure

The questionnaire was distributed to in-service EFL teachers in Egypt via professional networks and faculty of education alumni contacts. Participation was voluntary and responses were collected anonymously. All participants confirmed that they had graduated from an English department teaching division of an Egyptian public university faculty of education. A total of 140 valid responses were obtained and retained for analysis. No responses contained missing values on the 16 assessment-literacy items or on the four supplementary faculty-and-technology items.

Data Analysis

Descriptive statistics (means and standard deviations) were computed for each of the 16 items and for each of the six domains, with domain scores calculated as the mean of their constituent items. Internal consistency was assessed using Cronbach's alpha for each domain and for the full 16-item scale. To address Research Question 2, Welch's independent-samples t tests compared domain scores between teachers who had and had not received formal assessment training, one-way ANOVAs compared domain scores across four experience groups (0–2, 3–5, 6–10, and more than 10 years) and across four teaching-level groups (primary, secondary, adult/private institute, and university), and Pearson correlations examined the relationship between domain scores and both age and years of experience, the latter coded as the approximate midpoint of each category. Intercorrelations among the six domain scores were also examined to address Research Question 3. To address Research Question 4, the four supplementary items (G1–G4) were analyzed descriptively, with mean ratings, standard deviations,

and the proportion of responses in the disagree (1–2), neutral (3), and agree (4–5) categories reported for each item. All inferential analyses used an alpha level of .05.

III. Results

Overall Level and Domain Profile

Across the full 16-item scale, teachers' mean self-reported assessment literacy was 3.03 (SD = 1.31) on the five-point scale — roughly midway between neutral and agree. Internal consistency was excellent for the full scale (α = .98) and good to excellent for each of the six domains (α = .88–.97; see Table 2).

The six domains were not rated equally. Domain A (Understanding Assessment Purposes and Principles) received the highest mean rating (M = 3.30, SD = 1.30), followed by Domain B (Validity, Reliability, and Fairness; M = 3.18, SD = 1.31) and Domain F (Professional and Ethical Practices; M = 3.10, SD = 1.54). Domains C (Test and Task Design; M = 2.90, SD = 1.44) and D (Scoring, Rubrics, and Standard Setting; M = 2.86, SD = 1.39) fell slightly below the scale midpoint, and Domain E (Data Interpretation and Use of Results; M = 2.57, SD = 1.40) received the lowest mean rating of all six domains (Table 2).

Item-level results (Table 3) show that the highest-rated item overall was "I understand how classroom assessment supports learning, not just grading" (M = 3.65, SD = 1.37), and the lowest-rated item was "I feel confident interpreting statistical indicators such as mean, item facility, and reliability coefficients" (M = 2.44, SD = 1.33).

Table 4 reports the proportion of responses falling into the disagree (1–2), neutral (3), and agree (4–5) categories for each domain. Domain E had the highest proportion of disagree responses (53.6%) and the lowest proportion of agree responses (32.5%), confirming its position as the domain with the greatest reported need. Domains C and D followed closely, with disagree rates of 48.3% and 49.6%, respectively.

Differences by Training, Experience, and Teaching Level

Formal training in language assessment was strongly associated with self-reported assessment literacy in every domain (Table 5). Teachers who reported having received formal training scored, on average, between 2.05 and 2.69 scale points higher than untrained teachers across the six domains. Welch's t tests confirmed that all six differences were highly significant (all $t > 15.0$, all $p < .001$), with the largest absolute gaps observed in Domain F (Professional and Ethical Practices), Domain C (Test and Task Design), and Domain E (Data Interpretation and Use of Results).

Years of teaching experience was also significantly related to assessment literacy. One-way ANOVAs (Table 6) showed significant differences across the four experience groups for all six domains (all $F(3, 136) > 5.1$, all $p < .01$), with mean scores generally increasing from the least experienced group (0–2 years) to the most experienced group (more than 10 years). Pearson correlations between years of experience (coded as group midpoints) and domain scores ranged from r = .31 (Domain C) to r = .38 (Domain A), all $p < .001$. Age, by contrast, was not significantly correlated with any domain (all |r|

< .04, all p > .70), suggesting that the experience effect reflects accumulated teaching and assessment exposure rather than age per se.

Teaching level showed no significant association with any domain. One-way ANOVAs comparing primary, secondary, adult/private institute, and university teachers were non-significant for all six domains (all $F(3, 136) < 2.1$, all $p > .11$), although teachers at the primary level had numerically the highest mean scores on five of the six domains.

Relationships Among the Six Domains

The six domain scores were strongly and positively intercorrelated (Table 7), with correlations ranging from r = .80 (Domains A and E) to r = .94 (Domains C and F). This pattern suggests that, at least in this sample, the six domains tap a common underlying dimension of assessment literacy as much as they tap distinct skill sets: teachers who rated themselves highly in one domain tended to do so across all domains. Combined with the training and experience results reported above, Domain E (Data Interpretation and Use of Results), together with Domains C (Test and Task Design) and D (Scoring, Rubrics, and Standard Setting), stand out as the clearest priorities for professional development based on their low absolute ratings, while Domains F, C, and E showed the largest gaps between trained and untrained teachers.

Faculty Preparation, Smart Assessment, and AI-Based Evaluation

Responses to the four supplementary items revealed a striking and consistent picture of perceived under-preparation. On item G1 ("My faculty of education provided sufficient training in language assessment to prepare me for classroom practice"), the mean rating was 2.11 (SD = 1.03), with 71.4% of participants selecting a disagree response (1–2) and only 12.9% selecting an agree response (4–5). This represents one of the most negative response distributions in the entire questionnaire and indicates that, regardless of their current assessment literacy level, most participants felt that their faculties had failed to embed adequate assessment training in the pre-service curriculum.

Familiarity with smart assessment and AI-based evaluation was likewise very low. On item G2 ("I am familiar with the concept of smart assessment and how it differs from traditional testing approaches"), the mean rating was 1.93 (SD = 0.98), with 76.4% of participants in the disagree category. On item G3 ("I understand how AI-based evaluation tools can be used to support language assessment"), the mean rating was 1.87 (SD = 0.94), with 78.6% in the disagree category. These figures indicate that the large majority of participants had not encountered these concepts in a meaningful way during or since their pre-service preparation.

Responses to item G4 ("My pre-service methodology courses included at least one module or unit on using AI tools for test design or data analysis") were the most uniformly negative of the four supplementary items (M = 1.54, SD = 0.72). An overwhelming 91.4% of participants selected a disagree response, and fewer than 4% selected agree. This near-unanimous pattern suggests that the integration of AI tools into assessment methodology coursework was effectively absent from the pre-service programs completed by this sample, regardless of the institution attended or the year of graduation.

Table 1
Demographic Characteristics of Participants (N = 140)

Characteristic	**n**	**%**
Age (years)		
M = 39.84, SD = 10.10, range = 21–67		
Teaching experience		
0–2 years	5	3.6
3–5 years	25	17.9
6–10 years	32	22.9
More than 10 years	78	55.7
Teaching level		
Primary	14	10.0
Secondary	38	27.1
Adult / private language institute	34	24.3
University	54	38.6
Formal training in language assessment		
Yes	78	55.7
No	62	44.3

Note. Percentages may not sum to exactly 100 due to rounding.

Table 2
Descriptive Statistics and Internal Consistency Reliability for the Six Assessment Literacy Domains

Domain	**k**	**M**	**SD**	**α**
A. Understanding Assessment Purposes and Principles	3	3.30	1.30	.92
B. Validity, Reliability, and Fairness	4	3.18	1.31	.95
C. Test and Task Design	3	2.90	1.44	.97
D. Scoring, Rubrics, and Standard Setting	2	2.86	1.39	.95
E. Data Interpretation and Use of Results	2	2.57	1.40	.88
F. Professional and Ethical Practices	2	3.10	1.54	.95
Overall scale	16	3.03	1.31	.98

Note. k = number of items; M and SD are based on the domain composite score (mean of constituent items), N = 140.

Table 3
Item-Level Descriptive Statistics

Domain	**Item**	**M**	**SD**
A	I can explain the main purposes of formative, summative, and diagnostic assessment.	3.17	1.50
A	I understand how classroom assessment supports learning, not just grading.	3.65	1.37
A	I can describe the difference between assessment of, for, and as learning.	3.07	1.36

B	I understand what validity means in test design and why it matters.	3.11	1.42
B	I know how to improve reliability in scoring students' work.	3.03	1.40
B	I consider fairness and bias when developing or selecting test materials.	3.24	1.40
B	I can judge whether a test measures what it is supposed to measure.	3.34	1.37
C	I can write test specifications describing the skills, formats, and scoring methods to be used.	2.76	1.48
C	I can select appropriate item types (e.g., multiple choice, gap-fill, essay) for different skills.	3.00	1.52
C	I can balance test length, timing, and difficulty when designing assessments.	2.95	1.47
D	I can create analytic and holistic rubrics for productive skills (speaking, writing).	2.89	1.37
D	I can apply scoring rubrics consistently when rating students' work.	2.84	1.47
E	I can analyze test results to identify patterns of student strengths and weaknesses.	2.70	1.63
E	I feel confident interpreting statistical indicators such as mean, item facility, and reliability coefficients.	2.44	1.33
F	I understand the ethical issues related to test security, confidentiality, and grading.	3.10	1.60
F	I continuously reflect on my own assessment practices and seek ways to improve them.	3.11	1.57

Note. Items are presented in the order in which they appeared in the questionnaire. N = 140.

Table 4
Response Distribution by Domain (Percentage of Responses)

Domain	**Disagree (1–2)**	**Neutral (3)**	**Agree (4–5)**
A. Understanding Assessment Purposes and Principles	32.6	20.2	47.1
B. Validity, Reliability, and Fairness	35.9	20.2	43.9
C. Test and Task Design	48.3	11.0	40.7
D. Scoring, Rubrics, and Standard Setting	49.6	11.8	38.6
E. Data Interpretation and Use of Results	53.6	13.9	32.5
F. Professional and Ethical Practices	48.6	6.1	45.4

Note. Percentages are based on all individual item responses within each domain and may not sum to exactly 100 due to rounding.

Table 5
Comparison of Domain Means by Formal Assessment Training

Domain	Trained M (SD)	Untrained M (SD)	t	df	p
A. Understanding Assessment Purposes and Principles	4.21 (0.87)	2.16 (0.74)	15.08	137.5	< .001
B. Validity, Reliability, and Fairness	4.10 (0.91)	2.03 (0.64)	15.76	136.1	< .001
C. Test and Task Design	4.01 (0.90)	1.51 (0.39)	22.09	110.6	< .001
D. Scoring, Rubrics, and Standard Setting	3.88 (1.00)	1.59 (0.41)	18.31	106.8	< .001
E. Data Interpretation and Use of Results	3.60 (1.05)	1.28 (0.28)	18.68	90.6	< .001
F. Professional and Ethical Practices	4.29 (0.95)	1.60 (0.44)	22.15	112.8	< .001

Note. Trained n = 78; Untrained n = 62. Degrees of freedom (df) are Welch-adjusted to account for unequal variances.

Table 6
Domain Means by Years of Teaching Experience and One-Way ANOVA Results

Domain	0–2 yrs (n = 5)	3–5 yrs (n = 25)	6–10 yrs (n = 32)	>10 yrs (n = 78)	F(3, 136)	p
A. Purposes & Principles	2.53	2.43	3.12	3.70	7.94	< .001
B. Validity, Reliability, Fairness	2.35	2.44	2.98	3.56	6.46	< .001
C. Test & Task Design	2.20	2.05	2.85	3.24	5.14	.002
D. Scoring, Rubrics, Standard Setting	1.90	2.00	2.70	3.27	7.22	< .001
E. Data Interpretation & Use	1.70	1.78	2.48	2.92	5.36	.002
F. Professional & Ethical Practices	2.40	2.20	2.92	3.51	5.69	.001

Note. Cell entries are domain composite means for each experience group.

Table 7
Intercorrelations Among the Six Assessment Literacy Domains

Domain	A	B	C	D	E	F
A. Purposes & Principles	—	.89	.85	.85	.80	.89
B. Validity, Reliability, Fairness		—	.89	.83	.89	.91
C. Test & Task Design			—	.87	.90	.94
D. Scoring, Rubrics, Standard Setting				—	.85	.89

E. Data Interpretation & Use					—	.89
F. Professional & Ethical Practices						—

Note. All correlations significant at $p < .001$.

IV. Discussion

Overall Level of Assessment Literacy

The overall mean of 3.03 places this Egyptian sample close to the midpoint of the scale — neither strongly confident nor strongly lacking. This is somewhat higher than the low levels reported in some EFL contexts (Isnawati, 2023; Lan et al., 2019), but broadly consistent with the more mixed picture reported by Sun et al. (2022), who likewise found that prior training and experience were strong predictors of current literacy. One plausible explanation lies in the composition of the present sample: more than half of the participants (55.7%) reported having received formal training in language assessment since graduation, and most (78.6%) had more than five years of teaching experience. Given how strongly both factors were related to assessment literacy in this study, a sample drawn mainly from novice or untrained teachers would likely report a lower overall mean than the one observed here. It is also worth noting that the moderate overall mean masks a substantial gap between what teachers felt their faculties had prepared them for and what they actually know: despite achieving midpoint scores on the six-domain scale, 71.4% of participants reported that their faculty preparation in assessment had been insufficient, suggesting that whatever assessment literacy they now possess was largely acquired through experience or post-graduation training rather than through their pre-service program.

Domain Profile

The domain-level pattern — stronger in Domain A (assessment purposes and principles) and weaker in Domain E (data interpretation and use of results) — echoes findings from several other contexts. Chang et al. (2024) found that Taiwanese teachers' greatest training needs centered on using assessment results to provide feedback and inform teaching, and Al-Akbari et al. (2024) similarly found that Yemeni teachers' greatest needs were in assessment administration and use rather than in foundational conceptual knowledge. A plausible interpretation is that conceptual knowledge about why assessment matters is more likely to be covered, at least briefly, in general teacher education or to be picked up through everyday teaching experience, whereas the more technical skills involved in writing test specifications, constructing rubrics, and interpreting statistics such as item facility or reliability coefficients require more deliberate, often specialist, instruction that many teachers in this sample appear not to have received.

The strong intercorrelations among domains ($r = .80–.94$) are also worth noting. Rather than six clearly separable skill sets, the pattern observed here looks more like a single underlying factor: teachers who feel confident in one area tend to feel confident across the board, and vice versa. This is broadly consistent with views of assessment literacy as a holistic professional orientation rather than a checklist of unrelated competencies (Coombe et al., 2020; Scarino, 2013). Teachers who have engaged seriously with

assessment, whether through training or experience, appear to develop a generally more assessment-literate stance overall, even though some specific domains — particularly Domain E — remain comparatively weak for nearly everyone.

Training, Experience, and Teaching Level

The very large differences between trained and untrained teachers (Table 5) are consistent with a substantial body of evidence that formal training measurably raises assessment literacy (Koh, 2011; Mertler, 2009; Mirsanjari, 2025). What is striking here is the size of the gap — on the order of two scale points across every domain — suggesting that, in this sample, having received formal training (or not) may be the single most important factor differentiating teachers' assessment literacy profiles, more so than experience or teaching level.

The positive relationship with years of experience (r = .31–.38) suggests that assessment literacy also develops, at least to some extent, through accumulated classroom practice, independent of formal training. However, the fact that age itself was unrelated to any domain indicates that this association is specifically about time spent teaching and assessing, not simply about being older. This is broadly compatible with Mirsanjari's (2025) finding that less experienced instructors gained the most from a teacher development course, while more experienced instructors, who may already have built up some assessment literacy through years of practice and who may also be more entrenched in existing routines, gained comparatively less.

The absence of any significant difference by teaching level was somewhat unexpected, given that primary, secondary, adult, and university contexts differ considerably in curriculum, assessment formats, and stakes. One interpretation is that the six domains assessed here are general enough — covering principles, validity, design, scoring, interpretation, and ethics — that they apply similarly across teaching contexts, and that what matters more is whether and how much training and experience a teacher has had, regardless of where they teach. This is consistent with the Teachers' Assessment Literacy Enhancement project's finding that classroom assessment challenges were broadly shared across different national and institutional contexts in Europe (Vogt et al., 2020).

Faculty Preparation and the AI and Smart Assessment Gap

Perhaps the most striking findings in this study are those from the four supplementary items. The near-unanimous view that pre-service methodology courses contained no instruction on AI tools for test design or data analysis (G4: M = 1.54; 91.4% disagreeing) is not easily explained by the recency of these technologies alone. AI-assisted language assessment tools have been available and discussed in the applied linguistics literature for at least a decade, covering applications from automated essay scoring to adaptive testing systems and natural language processing-based feedback tools. The finding that virtually none of the participants — whose graduation dates span a wide range of years — encountered such content in their methodology courses suggests a systemic gap in Egyptian faculty of education curricula, not simply a lag behind the latest developments.

The low familiarity with smart assessment (G2: M = 1.93) and AI-based evaluation (G3: M = 1.87) reinforces this interpretation. Smart assessment, understood as assessment systems that adapt in real time to learner responses, generate personalized feedback, and support diagnostic decision-making, is increasingly present in language learning platforms used in Egyptian schools and private institutes. If in-service teachers are unfamiliar with the conceptual basis of these tools, they are poorly positioned either to use them effectively or to evaluate their outputs critically. This aligns with the broader argument made by Coombe et al. (2020) that assessment literacy needs to be understood as a living, evolving construct that expands as the assessment landscape changes.

The perception that faculties did not do enough to embed assessment literacy training more broadly (G1: M = 2.11; 71.4% disagreeing) also deserves attention. This is not simply a complaint about a missing AI module: the majority of participants felt that their preparation was inadequate across assessment literacy as a whole, including foundational areas such as test design, scoring, and the interpretation of results. Taken alongside the very large differences observed between trained and untrained teachers on the six domains, this finding suggests that the formal assessment content of pre-service programs was insufficient to bring graduates to a functional level of readiness, and that the gains in assessment literacy observed among more experienced and post-graduation-trained teachers represent — in part — a compensatory effect: teachers filling gaps left by their initial education.

Implications for Professional Development

Taken together, these findings point to several practical implications for designing assessment literacy training for in-service Egyptian EFL teachers and for reforming pre-service curricula in faculties of education.

First, Domain E (data interpretation and use of results), along with Domains C (test and task design) and D (scoring, rubrics, and standard setting), should be prioritized in professional development content. These are the domains where teachers in this sample reported the least confidence, suggesting both that there is room for improvement and that targeted training in these areas is likely to be useful.

Second, the size of the training effect, combined with prior evidence that sustained professional development produces larger and more durable gains than one-off workshops (Koh, 2011), suggests that brief, single-session training is unlikely to close the gaps identified here. Programs structured around ongoing cycles of practice, feedback, and reflection, such as those described by Li, J. et al. (2023) and Li, Z. et al. (2023) for formative assessment, appear better suited to developing the more technical domains, particularly data interpretation.

Third, the experience-related findings suggest that professional development may need to be differentiated rather than designed as a single program for all teachers. Mirsanjari (2025) found that newer teachers benefited most from a development course, while more experienced teachers showed smaller gains, often because of entrenched institutional habits. Program designers might therefore consider distinct tracks: foundational training for less experienced or untrained teachers, and more targeted,

practice-focused modules for veteran teachers, perhaps focused on translating existing experience into more systematic use of assessment data.

Fourth, the supplementary findings point to a clear and urgent need for Egyptian faculties of education to revisit their pre-service assessment curricula. The near-universal perception among participants that their methodology courses provided no preparation in the use of AI tools for test design or data analysis, combined with very low familiarity with smart assessment concepts, indicates that pre-service programs are currently producing graduates who are not equipped to engage with the technological dimensions of contemporary assessment practice. Faculties should consider integrating dedicated modules on assessment literacy into methodology courses, including content on AI-based evaluation tools, adaptive assessment systems, and the critical evaluation of automated scoring and feedback platforms. Given the pace of development in this area, such modules would need to be reviewed and updated regularly, rather than treated as fixed additions to a stable curriculum.

Finally, because teaching level was not a significant factor, institutions serving multiple sectors — as in this sample, which included primary, secondary, adult, and university teachers — may be able to design a shared core curriculum addressing the six domains and the technology-enhanced assessment dimension, with optional context-specific modules layered on top. Given that even teacher educators themselves sometimes report gaps in their own assessment literacy (Villa Larenas et al., 2022), faculties of education in Egypt should also ensure that those responsible for designing and delivering methodology courses are themselves adequately prepared in both conventional and technology-enhanced assessment, not only the teachers they train.

V. Limitations

Several limitations should be considered when interpreting these results. First, the data are based entirely on self-report; teachers' perceptions of their own assessment literacy may not correspond closely to their actual knowledge or classroom practice, and may be subject to social desirability or to differing personal standards of self-assessment. This concern applies equally to the supplementary items: teachers' reports of what their pre-service methodology courses did or did not contain are retrospective and may be subject to recall bias, particularly for those who graduated many years ago. Second, the design is cross-sectional, so the relationships observed between training, experience, and assessment literacy cannot be interpreted causally; teachers who seek out training may differ in other ways from those who do not.

Third, the sample, while reasonably diverse in terms of teaching level and experience, was a convenience sample of 140 teachers drawn through professional networks, which limits how confidently the findings can be generalized to all Egyptian EFL teachers or to other Arab EFL contexts. Fourth, the familiarity items for smart assessment and AI-based evaluation (G2 and G3) were single-item measures with no established validity evidence, and the concepts themselves — particularly "smart assessment" — may not have been interpreted identically by all respondents. Future studies should use more fully developed and validated instruments to assess these dimensions. Finally, three of the six assessment literacy domains (D, E, and F) were measured with only two items

each, which, while showing acceptable to high internal consistency in this sample, provides a narrower basis for assessing those domains than the three- or four-item domains.

VI. Conclusion

This study examined the self-reported assessment literacy of 140 in-service EFL teachers across six domains, from understanding assessment purposes to professional and ethical practice. Teachers reported moderate overall assessment literacy, with clear strengths in foundational conceptual knowledge and a clear weakness in interpreting and using assessment data. Formal training in language assessment was associated with substantially higher scores across every domain, and years of teaching experience showed a smaller but still significant positive relationship, while the level at which teachers taught made little difference. These findings reinforce calls from research in other EFL contexts for assessment literacy training that goes beyond foundational concepts to address the more technical, results-oriented aspects of assessment — particularly data interpretation, test and task design, and rubric development — and that is structured as sustained, differentiated professional development rather than a single generic workshop.

References

1. Al-Akbari, S., et al. (2024). EFL teachers' language assessment literacy training needs. Social Sciences & Humanities Open.
2. Bøhn, H., et al. (2021). Teacher educators' conceptions of language assessment literacy in Norway. Journal of Language Teaching and Research.
3. Chang, D. Y.-S., et al. (2024). Exploring language assessment literacy and needs of English teachers at senior high school level. Asia Pacific Journal of Education.
4. Coombe, C., et al. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? Language Testing in Asia.
5. Farmasari, S., et al. (2023). Pre-service EFL teachers' language assessment literacy satisfaction and assessment preparedness. International Journal of Language Education.
6. Giraldo, F. (2021). Language assessment literacy and teachers' professional development: A review of the literature. Profile: Issues in Teachers' Professional Development.
7. Isnawati, I. (2023). EFL teachers' assessment literacy. LLT Journal: A Journal on Language and Language Teaching.
8. Koh, K. (2011). Improving teachers' assessment literacy through professional development. Teaching Education.
9. Lan, C.-C., et al. (2019). Developing classroom-based language assessment literacy for in-service EFL teachers: The gaps. Studies in Educational Evaluation.
10. Li, J., et al. (2023). Developing classroom-based formative assessment literacy: An EFL teacher's journey. Chinese Journal of Applied Linguistics.
11. Li, Z., et al. (2023). The role of a professional development program in improving primary teachers' formative assessment literacy. Teacher Development.

12. Lo, Y., et al. (2022). Conceptualising assessment literacy of teachers in Content and Language Integrated Learning programmes. International Journal of Bilingual Education and Bilingualism.
13. Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. Improving Schools.
14. Mirsanjari, Z. (2025). Enhancing assessment literacy in EAP instruction: The role of teacher development courses in overcoming systemic barriers. Language Testing in Asia.
15. Park, E. (2024). The exploration of EFL preservice teachers' self-perceived importance of assessment literacy. Language Teaching Research Quarterly.
16. Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. Language Testing.
17. Sun, H., et al. (2022). Assessment literacy of college EFL teachers in China: Status quo and mediating factors. Studies in Educational Evaluation.
18. Villa Larenas, S., et al. (2022). But who trains the language teacher educator who trains the language teacher? An empirical investigation of Chilean EFL teacher educators' language assessment literacy. Language Testing.
19. Vogt, K., et al. (2020). Linking learners' perspectives on language assessment practices to teachers' assessment literacy enhancement (TALE): Insights from four European countries. Language Assessment Quarterly.
20. Wiyaka, et al. (2024). Investigating skills-based language assessment literacy of EFL teachers in Indonesia. KnE Social Sciences.