
RESEARCH ARTICLE

Beyond AI Familiarity: Explaining EFL Instructors' Professional Development Needs in AI-Mediated Assessment

Fawaz Al Mahmud¹, Hassan Saleem Alqurashi^{2*}

¹ Associate Professor, Department of English Language, College of Languages and Translation, University of Jeddah, Jeddah, Saudi Arabia; falaswad@uj.edu.sa

² Associate Professor, Department of Curriculum and Instruction, College of Education, University of Jeddah, Jeddah, Saudi Arabia; hsqurashi@uj.edu.sa

Corresponding Author: Hassan Saleem Alqurashi, **E-mail:** hsqurashi@uj.edu.sa

ABSTRACT

Professional-development need in generative-AI (GenAI) assessment is often treated as a deficit that should decline as teachers become more familiar and confident. This secondary analysis tests a different possibility: professional-development need may reflect a differentiated awareness of the demands of AI-mediated assessment rather than simple unfamiliarity. Using the verified anonymised responses of 216 GenAI-experienced EFL instructors in Saudi higher education, we modelled a three-item professional-development-need score from a previously validated instrument. Conditional quantile regression examined the 25th, 50th, and 75th percentiles of need, with 1000 case-bootstrap resamples. Familiarity/use alone explained 11.6%, 8.5%, and 3.3% of quantile variation, whereas the full competency model increased pseudo- R^2 by 12.6, 11.4, and 10.6 percentage points, respectively. The results revealed that familiarity/use was positively associated with need at the lower and median quantiles but not at the upper quantile. The results also showed that assessment literacy was negatively associated with need across all three quantiles, whereas ethical awareness and self-efficacy were consistently positive; assessment application contributed no independent association after adjustment. HC3 mean regression and item-specific ordinal-logit models reproduced the broad pattern. After demographic adjustment, ethical awareness and self-efficacy remained the most robust competency correlates. The findings challenge a simple deficit interpretation of training need and support differentiated professional development that combines assessment literacy, ethical judgement, and advanced evaluation of AI-generated assessment materials.

KEYWORDS

Generative AI; EFL; professional development; language assessment literacy; teacher self-efficacy; quantile regression

ARTICLE INFORMATION

ACCEPTED: 02 September 2026

PUBLISHED: 24 October 2026

DOI: 10.32996/jeltal.2026.8.11.2

1. INTRODUCTION

Generative artificial intelligence (GenAI) is increasingly considered as a component of the professional environment where EFL teachers create tasks, formulate feedback, revise rubrics, and assess student work. This development has increased an already demanding form of professional judgement: teachers must now evaluate the validity, appropriateness, transparency, and ethical acceptability of AI-mediated assessment artifacts, rather than focusing only on learner performance. The dataset used in the present study was originally collected to explore EFL instructors' GenAI assessment literacy in Saudi higher education. It indicated high perceived literacy, strong ethical concern, and considerable demand for training (Al Mahmud, 2026). However, what the need means and how variation is explained in it rather than whether professional-development need exists remained as unresolved questions.

Copyright: © 2026 the Author(s). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) 4.0 license (<https://creativecommons.org/licenses/by/4.0/>). Published by Al-Kindi Centre for Research and Development, London, United Kingdom.

Professional learning in the AI era has shifted from short tool demonstrations to the sustained development of pedagogical, evaluative, disciplinary, and ethical competence. The needs of the English language teacher have been highlighted in qualitative research published in 2026, which include understanding AI, integrating AI into the educational programme, addressing ethical considerations, aligning AI with pedagogy, and clarifying assessment methods (Zhang et al., 2026). Moreover, systematic reviews highlight the teacher professional development as a key but still underdeveloped element of AI integration (Tan et al., 2025a). Research on interventions shows that formal professional development with an AI focus can change both technical and pedagogical competence (Tan et al., 2025b). What can be inferred from the findings of previous research is that the need for professional development should not be understood as a lack of competence. Teachers with greater knowledge of AI, assessment or professional risk may be in a better position to see what they still need to learn.

This possibility is particularly reasonable in AI-mediated language assessment. Higher-education GenAI capability comprises basic understanding, teaching, research, and professional engagement (Xia et al., 2026), while EFL-specific GenAI literacy differentiates between knowledge, use, evaluation, design, and ethics (Wang et al., 2025). Research on language assessment additionally highlights that GenAI poses challenges related to validity, reliability, fairness, transparency, and human oversight (Chuang & Yan, 2025). Teachers assessing AI-assisted writing already report tensions around authenticity, rubric use, and defensible evaluative practice (Chen & Pi, 2026). Consequently, a teacher can be familiar with GenAI yet still need advanced development in evaluating AI-generated assessment materials, diagnosing validity threats, protecting student privacy, and deciding when human override is necessary.

The attempts of professional development have not always kept pace with these challenges. Faculty workshops increasingly focus on redesigning assessment to be AI-resistant or AI-aware (Alkouk & Khlaif, 2024). Moreover, Saudi teachers of EFL indicate concurrent pedagogical opportunities and ethical issues (Aljabr & Al-Ahdal, 2024). Analyses concerned with the use of ChatGPT in English as a foreign language education report potential benefits for teacher support besides concerns related to misinformation, privacy, and academic integrity (Lo et al., 2024). Beginning language teachers and teacher educators do not always feel equally prepared or confident about using AI (Moorhouse, 2024; Moorhouse & Kohnke, 2024). Recent research on future teachers emphasises that teacher education should enable them to integrate AI in their teaching rather than simply learning about how to use AI tools themselves (Kohnke et al., 2025). A 2026 study documents a number of gaps especially related to AI, such as awareness issues, application challenges in EFL settings, and the need for future professional development (Barghi et al., 2026). Likewise, in the Saudi context, research shows that university EFL teachers' confidence in using GenAI differs according to the area of application (Altalib, 2026).

These developments stand as a conceptual challenge for research based only on mean scores. If training need is assumed to be a single average outcome, then the same predictor is assumed to work similarly among instructors with relatively modest need and those already reporting very high need. But need can be distributionally varied. For example, basic familiarity may distinguish instructors at the lower end of the need distribution, while high-need instructors may already be familiar with GenAI and instead be responding to ethical complexity, evaluative responsibility, or the limits of their own assessment knowledge. A distribution-sensitive analysis is therefore appropriate when the outcome is bounded, highly endorsed, and potentially generated by different combinations of professional competence.

This study seeks to find out whether EFL teachers' professional-development needs in AI-mediated assessment are connected to variables beyond their familiarity with GenAI. Instead of repeating the descriptive findings on training needs presented in the original publication (Al Mahmud, 2026), it uses the same verified dataset to address a different question: which competency areas are associated with professional-development needs after accounting for other factors, how do these relationships vary across different levels of need, and are similar patterns found across the three training needs measured in the study? Therefore, the contribution of the study is explanatory and distributional, not descriptive. The study provides a defined basis for establishing differentiated professional development. This is because it is not necessary that a training programme designed for unfamiliar users should be structured in the same way as one for confident instructors who are aware of the challenges related to ethics and assessment.

2. CONCEPTUAL BACKGROUND AND RESEARCH QUESTIONS

2.1. Professional Development as More Than a Deficit Signal

Teacher competence in AI should include understanding their practical, ethical, and pedagogical implications, rather than simply focusing on learning how to use the tools. This has been indicated in reviews of AI literacy in teacher education (Sperling et al., 2024). With the rapid advancement in GenAI, it became increasingly used to support teachers' own professional learning (Li et al.,

2026); controlled studies suggest that it can improve teachers' self-efficacy and higher-order thinking skills (Lu et al., 2024). However, research on responsible AI use indicates that professional development may strengthen teachers' critical awareness and focus on human-centred use, even when improvements in their confidence are less consistent (Huynh et al., 2025). This suggests that teachers can become more competent in using AI while still recognising areas where they need further training.

For AI-mediated assessment, this coexistence is theoretically meaningful. Professional development may be sought because a teacher lacks basic competence, but it may also be sought because the teacher understands enough to recognise advanced risks. Recent higher-education assessment reviews stress the importance of redesigning assessment and learning outcomes in response to GenAI (X. Weng et al., 2024). The advantages and challenges of implementation were perceived by the instructors in Saudi EFL settings (Alzubi & Alelaiwi, 2025). Thus, a positive relation between ethical awareness and training need would not necessarily suggest ethical weakness. Instead, it could indicate that instructors who recognise privacy, academic-integrity, validity, or transparency problems also recognise the need for structured support.

2.2. Language Assessment Literacy and Developmental Need

Language assessment literacy (LAL) provides the disciplinary foundation for interpreting professional-development need. Contemporary LAL is not limited to knowing testing terminology. It incorporates knowledge, skills, principles, contextual judgement, and the capacity to use assessment evidence responsibly (Coombe et al., 2020; F. Weng & Shen, 2022). Reviews of LAL-focused professional development show that effective programmes typically combine theory with practical assessment activity and reflection (Giraldo, 2021a, 2021b). Saudi and international studies likewise document continuing needs for assessment training among tertiary EFL practitioners (Latif, 2021; Salami & Alharthi, 2022; Sun & Zhang, 2022).

Language assessment literacy (LAL) may be enhanced by professional development. This can be performed through authentic test-design and evaluation work (Cui et al., 2022), and teachers report value in assessment-literacy courses when activities are directly related to classroom decisions (Giraldo et al., 2023). Empirical LAL models suggest that stakeholders do not have one general level of desired knowledge but rather differentiated needs profiles (Kremmel & Harding, 2020). Other conceptualisations focus on learners, context and process rather than fixed bodies of knowledge (Lee & Butler, 2020; Levi & Inbar-Lourie, 2020). The Teacher Assessment Literacy in Practice perspective also frames assessment literacy as situated professional practice (Xu & Brown, 2016), and foundational language-assessment work underscores practical use of assessment principles (Fulcher, 2012). In sum, this literature predicts professional-development need as a function of the intersection of teachers' knowledge, their capacity, and their perception of what is consequential in their context.

2.3. Distributional Rationale and Research Questions

Because the professional-development score is bounded and strongly endorsed, conditional-mean analysis alone may obscure meaningful heterogeneity. Quantile regression estimates the association between predictors and different conditional locations of an outcome distribution, rather than assuming a single mean effect (Koenker & Bassett, 1978; Koenker & Hallock, 2001). This is useful here because relatively low, median, and high training need may not be generated by the same combination of familiarity, assessment knowledge, application capability, ethical awareness and confidence. We therefore treat quantile analysis as the primary explanatory framework and use heteroskedasticity-consistent mean regression as a benchmark (Hayes & Cai, 2007). The present study poses the following research questions (RQs):

RQ1. To what extent does GenAI familiarity/use alone explain EFL instructors' professional-development need, and what additional explanatory value do assessment literacy, assessment application, ethical awareness, and self-efficacy contribute?

RQ2. Which competency domains are conditionally associated with professional-development need at the lower (25th), median (50th), and upper (75th) quantiles, and do these associations differ across quantiles?

RQ3. To what extent are these associations consistent across the three specific training needs measured: formal training in GenAI-based assessment, evaluation of AI-generated assessment materials, and ethical issues in AI-based assessment?

RQ4. How sensitive are the competency associations to adjustment for demographic and professional characteristics?

3. MATERIALS AND METHODS

3.1. Design, Data Source, and Analytic Distinctiveness

This study is a secondary cross-sectional analysis of an anonymised dataset previously collected from EFL instructors in Saudi higher education. The original publication used the dataset to validate the questionnaire and to report construct levels, ethical

concerns, demographic differences, and training needs (Al Mahmud, 2026). The present study addresses a distinct research question and unit of inference: professional-development need is modelled as an outcome and explained conditionally by other competency domains. No new participants or outcomes are introduced, and the present study does not reproduce the descriptive or demographic-comparison analyses of the original publication.

The original recruitment pool comprised 287 EFL instructors. Of 279 completed surveys, 47 respondents who had never used GenAI, 11 substantially incomplete records, and 5 records without consent were excluded, yielding 216 eligible participants (Al Mahmud, 2026). The verified analytic workbook used here contained exactly 216 rows, no missing questionnaire responses, and no duplicate rows. Its SHA-256 hash was 95cadb03e76dda99ba7a2f7b69c1c49cb72ff1a393051485d269137ba748deb7. The three professional-development item means replicated the published values to rounding precision as a data integrity check.

3.2. Participants and Measures

Participants included 110 men (50.9%) and 106 women (49.1%). Academic rank comprised 75 assistant professors, 65 associate professors, 35 lecturers, 26 instructors, and 15 professors. The largest age group was 41–50 years ($n = 95$), followed by 31–40 years ($n = 70$), 51–60 years ($n = 36$), and 20–30 years ($n = 15$). Teaching experience was distributed across 20 years or more ($n = 73$), 15–19 years ($n = 60$), 10–14 years ($n = 45$), 5–9 years ($n = 20$), and fewer than 5 years ($n = 18$).

The analysis used the 25 items retained in the published validation (Al Mahmud, 2026). Four items measured reported assessment-specific GenAI familiarity/use (FU), five perceived assessment literacy (AL), five reported assessment application (AA), four ethical-safeguard awareness (EF), four assessment self-efficacy (SE), and three professional-development need (PD). All items used five-point response options from 1 (strongly disagree) to 5 (strongly agree). Composite scores were the unweighted means of their retained items. Internal consistency in the verified file was acceptable to strong: $\alpha = .864$ (FU), $.875$ (AL), $.848$ (AA), $.786$ (EF), $.826$ (SE), and $.842$ (PD). Table 1 summarises the analytic variables before the multivariable models.

Table 1. Sample characteristics and composite-score descriptives.

Variable	Category / construct	n or M	% or SD
Gender	Male	110	50.9%
	Female	106	49.1%
Rank	Assistant Professor	75	34.7%
	Associate Professor	65	30.1%
	Lecturer	35	16.2%
	Instructor	26	12.0%
	Professor	15	6.9%
Age	20–30	15	6.9%
	31–40	70	32.4%
	41–50	95	44.0%
	51–60	36	16.7%
Composite	FU: familiarity/use	3.777	0.862
	AL: assessment literacy	4.256	0.634
	AA: assessment application	4.356	0.582
	EF: ethical awareness	4.556	0.523
	SE: self-efficacy	4.500	0.651
	PD: professional-development need	4.020	0.849

3.3. Outcome and Predictor Specification

The main outcome was the mean of three retained professional-development items: the need for formal training on GenAI in EFL assessment, the need for training on evaluating AI-generated assessment materials, and the need for training on ethical issues in AI-based assessment. Their means were 4.069 ($SD = 0.930$), 4.028 ($SD = 0.969$), and 3.963 ($SD = 1.020$), respectively. The five

predictor composites were FU, AL, AA, EF and SE. Predictors were standardised to mean 0 and standard deviation 1 prior to modelling. Thus, each coefficient represents the expected change in the PD score associated with a one-standard-deviation difference in the predictor, conditional on the other predictors.

3.4. Primary Distributional Analysis

Conditional quantile regression was estimated at $\tau = .25, .50$, and $.75$ (Koenker & Bassett, 1978; Koenker & Hallock, 2001). For each quantile two nested models were fitted. Model F consisted of FU only and operationalised a familiarity-centred explanation. Model C added AL, AA, EF, and SE and represented the full competency model. Explanatory gain beyond familiarity was summarised as change in quantile pseudo- R^2 from Model F to Model C. Uncertainty was assessed with 1000 nonparametric participant-level bootstrap resamples using fixed random seed 261016. Each resample refitted both nested models at all three quantiles, generating percentile intervals for coefficients and incremental pseudo- R^2 . No predictor, quantile, or model specification was changed after the final results were inspected.

3.5. Benchmark and Sensitivity Analyses

Three complementary checks were prespecified. First, ordinary least squares with HC3 heteroskedasticity-consistent standard errors provided a conditional-mean benchmark (Hayes & Cai, 2007). The same familiarity-only and full competency models were compared through R^2 . Second, each PD item was modelled separately as a proportional-odds logistic regression to test if the broad predictor pattern applies to formal-use, evaluation-focused, and ethics-focused training needs. Third, a demographic sensitivity model augmented the mean-regression specification with gender, academic rank, age group, and teaching-experience indicators. Only demographic variables were included as adjustments, and not as a replication of the group-comparison analyses already reported in the source article (Al Mahmud, 2026). Variance-inflation factors for the five competency predictors ranged from 1.43 to 2.79, suggesting no severe multicollinearity.

3.6. Ethical Considerations

The initial research was authorised by the Bioethics Committee of Scientific and Medical Research, University of Jeddah, Saudi Arabia (Registration No. HAP-02-J-094; Application No. UJ-REC-461). Participation was voluntary, electronic informed consent was obtained, and the final analytic file was anonymised (Al Mahmud, 2026). There was no new participant contact or new identifying information as part of this secondary analysis.

3.7. GenAI-Assisted Research Preparation

During the preparation of this manuscript, ChatGPT (OpenAI) was used to assist with language editing, manuscript organisation, and coding support. All analytic decisions, code execution, statistical outputs, source verification, interpretation and final conclusions were reviewed and approved by the authors, who are fully responsible for the integrity and accuracy of the work.

4. RESULTS

4.1. Professional-Development Need and Increment Beyond Familiarity

The professional-development need was high overall ($M = 4.020$, $SD = 0.849$), but there was sufficient dispersion in the score for conditional modelling. Familiarity/use was positively correlated with PD need ($r = .412$), but the other domains also had meaningful bivariate associations: AL $r = .224$, AA $r = .423$, EF $r = .465$, and SE $r = .377$. The conditional analysis was prompted by these correlations, which do not indicate which domains remain independently associated when modelled together.

Table 2 shows that familiarity alone gave a weak and declining explanation of professional-development need across the distribution. Model F pseudo- R^2 was $.116$ at Q25, $.085$ at Q50, and $.033$ at Q75. Adding AL, AA, EF, and SE increased pseudo- R^2 to $.241$, $.199$, and $.139$, respectively. The incremental gains were $.126$ at Q25, $.114$ at Q50, and $.106$ at Q75. In the 1000-resample bootstrap, all three incremental gains remained above zero: Q25 $[.061, .212]$, Q50 $[.063, .198]$, and Q75 $[.051, .189]$. Thus, competency information beyond familiarity contributed materially even among instructors at the upper end of the need distribution.

Table 2. Conditional quantile-regression results for professional-development need.

Quantile	Term	Estimate	Bootstrap 95% interval / model gain
Q25	FU	0.287	[0.074, 0.484]
Q25	AL	-0.346	[-0.532, -0.064]
Q25	AA	0.197	[-0.043, 0.370]
Q25	EF	0.258	[0.138, 0.472]
Q25	SE	0.298	[0.116, 0.423]
Q25	Pseudo-R ² : FU-only → full	0.116 → 0.241	$\Delta = 0.126$
Q50	FU	0.291	[0.086, 0.474]
Q50	AL	-0.238	[-0.434, -0.075]
Q50	AA	0.064	[-0.135, 0.277]
Q50	EF	0.329	[0.080, 0.443]
Q50	SE	0.293	[0.148, 0.468]
Q50	Pseudo-R ² : FU-only → full	0.085 → 0.199	$\Delta = 0.114$
Q75	FU	0.159	[-0.010, 0.340]
Q75	AL	-0.186	[-0.314, -0.070]
Q75	AA	0.055	[-0.111, 0.282]
Q75	EF	0.189	[0.055, 0.352]
Q75	SE	0.339	[0.043, 0.451]
Q75	Pseudo-R ² : FU-only → full	0.033 → 0.139	$\Delta = 0.106$

4.2. Which Competencies Matter at Different Levels of Need?

The pattern of coefficients in Table 2, illustrated in Figure 1, provided more information than familiarity alone. At Q25, FU was positive ($b = .287$, 95% bootstrap interval [.074, .484]), AL was negative ($-.346$ [-0.532, -0.064]), EF was positive (.258 [.138, .472]), and SE was positive (.298 [.116, .423]); AA was not independently associated. At Q50, the same four domains retained nonzero intervals: FU .291 [.086, .474], AL $-.238$ [-0.434, -0.075], EF .329 [.080, .443], and SE .293 [.148, .468]. At Q75, familiarity weakened to .159 [-0.010, .340], while AL remained negative ($-.186$ [-0.314, -0.070]) and both EF (.189 [.055, .352]) and SE (.339 [.043, .451]) remained positive. AA again had an interval crossing zero.

Although point estimates suggested that familiarity became less discriminating as need increased, bootstrap contrasts of Q75 minus Q25 did not establish statistically reliable between-quantile differences for the individual predictors. Thus, the right interpretation is not that the slopes of the coefficients definitively change across quantiles, but that the upper-need model can no longer be reduced to familiarity. Ethical awareness, self-efficacy and assessment literacy remain informative in areas where familiarity itself is no longer clearly associated. Figure 1 displays the standardised-predictor coefficients and bootstrap intervals before the discussion of their practical implications.

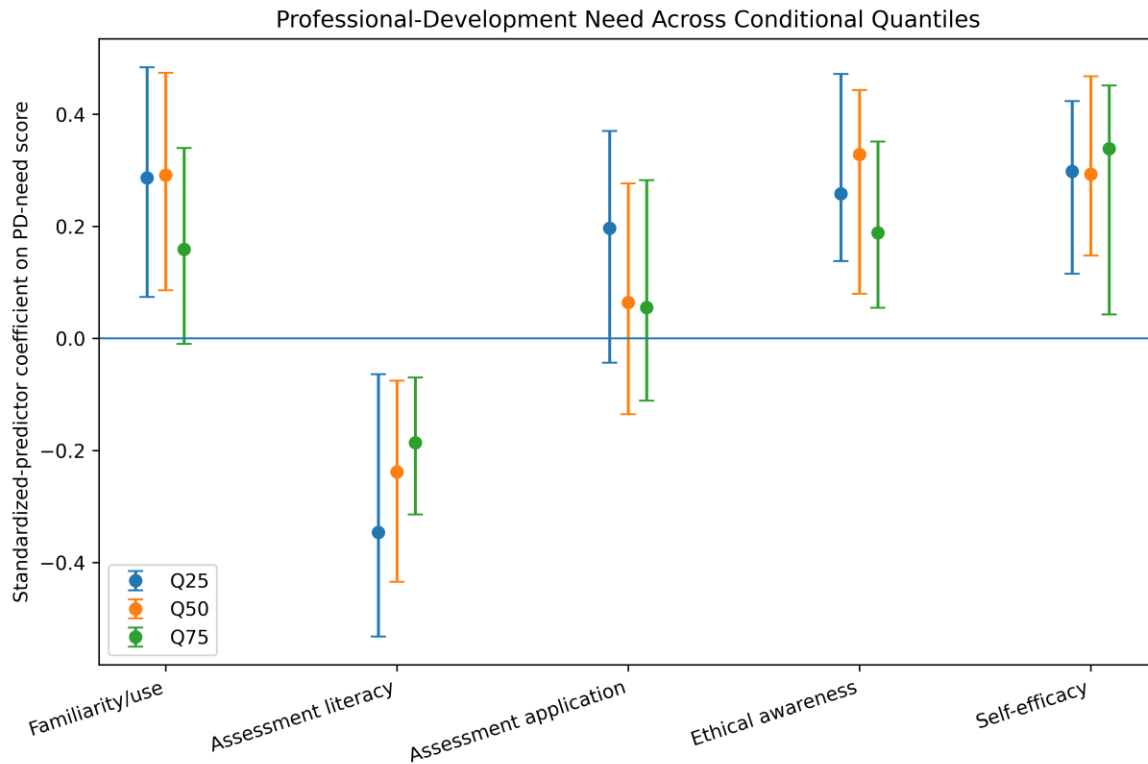


Figure 1. Conditional quantile-regression coefficients for professional-development need. Error bars are 95% participant-level bootstrap intervals.

4.3. Conditional-Mean Benchmark

The HC3 mean-regression benchmark supported the distributional results. Familiarity alone explained $R^2 = .169$. The full competency model explained $R^2 = .359$, an increase of .190. In the full model, FU was positive ($b = .220$, $p = .001$), AL negative ($b = -.252$, $p < .001$), EF positive ($b = .286$, $p < .001$), and SE positive ($b = .240$, $p < .001$); AA was not independently significant ($b = .111$, $p = .168$). The agreement of mean and quantile models reduces the chance that the main pattern is an artifact of one estimation framework.

Table 3. HC3 mean-regression benchmark.

Predictor	b	HC3 SE	95% CI	p
FU	0.220	0.068	[0.086, 0.354]	0.0013
AL	-0.252	0.073	[-0.396, -0.109]	0.0006
AA	0.111	0.081	[-0.047, 0.270]	0.1683
EF	0.286	0.065	[0.158, 0.414]	0.0000
SE	0.240	0.067	[0.109, 0.372]	0.0003
Model R^2	0.169 → 0.359			$\Delta R^2 = 0.190$

4.4. Specific Training Needs

Item-specific proportional-odds models revealed a remarkably consistent pattern across the three professional-development needs. Figure 2 summarises these results as odds ratios. A one-standard-deviation increase in FU was associated with higher odds of stronger reported need for formal training ($OR = 1.62$), evaluation training (1.80), and ethics training (1.46). Higher AL showed the opposite pattern across all three outcomes ($OR = 0.63$, 0.44 , and 0.64). EF was consistently positive ($OR = 1.82$, 1.81 , and 1.69), as was SE ($OR = 1.64$, 2.06 , and 1.69). AA did not reach conventional significance for any of the three items. The strongest item-specific association was between self-efficacy and need for training in evaluating AI-generated assessment materials ($OR = 2.06$, 95% CI [1.48, 2.86]).

Figure 2 is therefore important for interpretation: the main composite-level pattern is not driven by only one professional-development item.

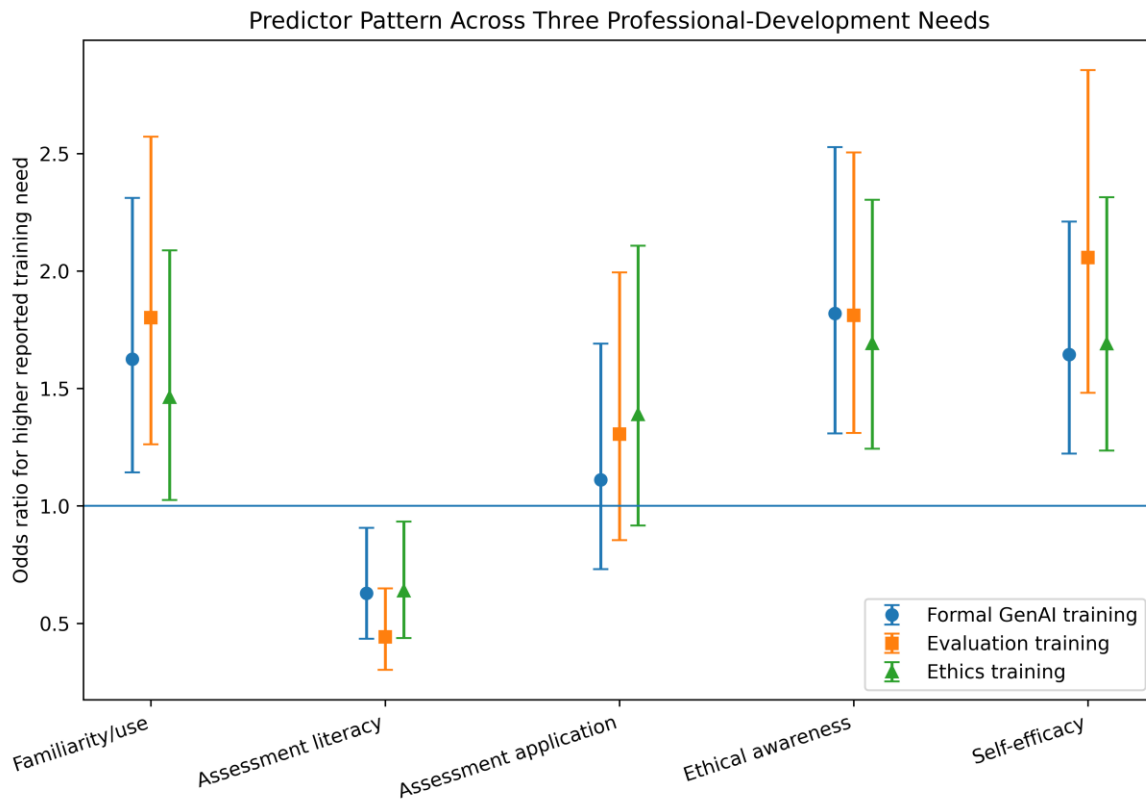


Figure 2. Odds ratios for higher response categories on the three professional-development-need items. Error bars are Wald 95% confidence intervals.

4.5. Demographic Sensitivity

The mean-model R^2 increased to .726 with the addition of gender, academic rank, age group, and teaching-experience indicators. Under this strong adjustment, the FU coefficient attenuated to .034 ($p = .487$) and AL to $-.104$ ($p = .185$), while EF remained positive (.193, $p < .001$) and SE remained positive (.155, $p = .017$); AA became borderline (.108, $p = .062$). This sensitivity result implies that familiarity and perceived assessment literacy partially covary with instructors' professional backgrounds, while ethical awareness and self-efficacy have more robust conditional associations with professional-development need. Because the source article already analysed demographic group differences (Al Mahmud, 2026), the demographic coefficients themselves are not interpreted substantively here.

5. DISCUSSION

The study started with an intentionally simple question: does professional-development need disappear once EFL teachers are familiar with GenAI? The answer is no. Familiarity/use alone explained a small part of need and its explanatory power was weakest at the upper conditional quantile. Adding assessment literacy, application capability, ethical awareness and self-efficacy provided consistent incremental explanatory gains. The outcome provides direct support for the key proposition of the title: professional-development need in AI-mediated assessment is not reducible to lack of familiarity.

The most important finding is the positive association between self-efficacy and professional-development need. According to conventional deficit reasoning, confident teachers would be expected to report less need for development. The data did not support this expectation. Instructors who reported higher self-efficacy also reported greater need for professional development, and this pattern held across the lower, median, and upper quantiles. It was also consistent across the robustness checks, persisting in the HC3 mean regression, after controlling for demographic characteristics, and in each of the three item-specific models. This pattern is consistent with a developmental-awareness interpretation. Teachers who feel empowered to use GenAI may engage with the technology at a deeper level to identify more advanced limitations, responsibilities, and learning opportunities. Recent

research (Huynh et al., 2025; Lu et al., 2024; Tan et al., 2025a, 2025b) indicates that AI competence can be developed through reflective recalibration and technical learning. Although this study does not establish that self-efficacy leads to greater demand for training, the findings challenge the assumption that reporting a need for further development necessarily reflects low confidence. Similarly, awareness of ethical issues indicated a positive connection with professional-development needs. This is theoretically consistent in AI-mediated assessment since issues such as privacy, academic integrity, human oversight, fairness, and validity are central to responsible practice. Studies investigating the use of GenAI in language have emphasised these concerns (Chuang & Yan, 2025), whereas the findings of the other studies focusing on the teacher have revealed the necessity of training for integration and modification of the existing practices rather than simply providing access to AI tools (Aljabr & Al-Ahdal, 2024; Alkouk & Khlaif, 2024; X. Weng et al., 2024). The findings of this study showed that teachers who emphasised the significance of ethical safeguards also highlighted the need for professional development including training on ethical issues themselves. In other words, greater awareness of the risks of AI use may lead teachers to acknowledge gaps in their knowledge and to seek further training and support.

Assessment literacy showed a different pattern. Although it was positively correlated with professional-development needs, its association became negative once the other domains were taken into account, across Q25, Q50, Q75, the three item-specific models, and the mean benchmark. The change from positive to negative correlation implies that the relationship is partly explained by the overlap between assessment literacy and other factors, such as GenAI familiarity, practical application, ethical awareness, and self-efficacy. Considering these accounts, those teachers who felt more confident in their assessment literacy reported less need for professional development. This finding supports the findings of previous studies on language assessment literacy (LAL) which highlight that stronger assessment knowledge and practice can decrease the needs for perceived training (Coombe et al., 2020; Cui et al., 2022; Giraldo, 2021a, 2021b; Giraldo et al., 2023; Latif, 2021; Salami & Alharthi, 2022; Sun & Zhang, 2022; F. Weng & Shen, 2022). However, this association weakened and was no longer significant after demographic adjustment, suggesting that assessment literacy is partly tied to rank, age, or experience.

Assessment application did not show a consistent independent relationship with professional-development needs. This does not mean, however, that the ability to apply AI in assessment is unimportant. Assessment application had a substantial correlation with professional-development needs when examined on its own, and it also had a high mean score. Yet it overlapped with other domains, particularly GenAI familiarity and assessment literacy, and once these factors were taken into account, its independent contribution was minimal. This has implications for professional-development programmes: a demonstrated or perceived ability to use GenAI to create or adapt assessment tasks should not be seen as a replacement for sound assessment knowledge or an understanding of ethical issues. For example, teachers may know how to use GenAI to develop formative assessment activities, but they may still need support in evaluating assessment validity, checking AI-generated scoring suggestions, identifying bias, or protecting student data.

The quantile perspective is an extra layer. Familiarity was clearly positive at Q25 and Q50, but the bootstrap interval at Q75 contained zero. We did not find significant formal interquantile contrasts so we do not claim a statistically proved slope change. Nevertheless, the pattern is educationally useful. For teachers already at the high end of professional-development need, familiarity is not very discriminating, while ethics, self-efficacy, and assessment literacy continue to be informative. This suggests that advanced professional development should not be designed around introductory familiarity modules alone.

Additionally, the findings related to the individual training-need models provide support for this recommendation. The findings showed that training-in-evaluation needs was strongly connected to self-efficacy, whereas its association with assessment literacy was strongly negative. The significance of this lies in that AI-assisted assessment requires teachers to judge whether AI-generated tasks, rubrics, scoring suggestions, and feedback are appropriate and reliable. As for teachers who are confident in using GenAI, professional development should focus on evaluating AI-generated materials, assessing validity, verifying outputs, and knowing when to reject AI suggestions and rely on human judgement, rather than being limited to basic skills. The original study indicated that teachers needed more training (Al Mahmud, 2026). The present analysis supports those findings and determines the areas of competency related to greater professional-development needs.

The context of Saudi higher education also matters. Regionally, there is evidence of growing use of AI-generative tools by EFL faculty, as well as concerns of implementation and ethics (Aljabr & Al-Ahdal, 2024; Alzubi & Alelaiwi, 2025). Persistent assessment-development needs are observed even outside GenAI in prior Saudi LAL research (Latif, 2021; Salami & Alharthi, 2022). AI professional development for EFL instructors should combine language-assessment literacy, responsible AI use, and disciplinary task design rather than operating as separate technology and assessment workshops.

5.1. Implications for Differentiated Professional Development

The predictor pattern is a poor match with a one-size-fits-all workshop. Institutions should at least distinguish three developmental layers. First, basic GenAI fluency for educators who have not yet achieved basic operational proficiency. The second is development of assessment literacy, which is focused on learning outcomes, construct representation, scoring logic, quality of feedback and interpretation of evidence. Third, advanced responsible use development with a focus on privacy, bias, academic integrity, oversight by humans, verification and governance. These layers need to be trained jointly, not individually.

The positive relationships between ethical awareness, self-efficacy, and professional-development needs also suggest that the need for training should not be viewed simply as a sign of limited competence. Even instructors who are highly confident in using AI, or highly aware of its ethical issues, may need more training because they recognise the challenges involved in using AI for assessment. Therefore, it is necessary for professional-development programmes take these needs seriously instead of considering them weaknesses. Recent research on AI-related professional learning stressed the importance of continuous development that takes into consideration instructors' existing knowledge, confidence and attitudes (Huynh et al., 2025; Li et al., 2026; Lu et al., 2024; Tan et al., 2025a, 2025b; Zhang et al., 2026). Similarly, the findings of the present study support this approach.

The three types of training needs identified in the study suggest some practical ways to structure professional-development programmes. Training in formal use could focus on writing effective prompts, designing assessment tasks, and setting clear boundaries for AI use. Evaluation training could help teachers check AI-generated questions, rubrics, scores, and feedback for accuracy and quality. Ethics training could address privacy, fairness, transparency, academic integrity, and situations in which human review is necessary. The findings showed broadly similar patterns across the three training needs, suggesting that they could be addressed through a shared competency framework. However, each module should also reflect the specific decisions teachers need to make when using AI for assessment.

5.2. Theoretical Implications

The findings suggest that professional-development needs should be understood as part of teachers' ongoing professional growth, rather than simply as an indication of low competence. Teachers who recognise the limits of their knowledge may be more aware of the complexities involved in using AI for assessment and more willing to seek further training. This interpretation is consistent with process-oriented models of language assessment literacy (LAL) (Kremmel & Harding, 2020; Lee & Butler, 2020; Levi & Inbar-Lourie, 2020; Xu & Brown, 2016) and recent research on AI competence, which treats ethics, professional engagement, self-efficacy, and reflective judgement as distinct but related areas (Altalib, 2026; Sperling et al., 2024; Wang et al., 2025; Xia et al., 2026). Future studies of teacher readiness to use AI should therefore distinguish between what teachers can do, how confident they feel about doing it, and how much further training they believe they need. Treating these as a single measure of readiness could overlook important differences between competence, confidence, and the desire to improve.

6. LIMITATIONS

Several limitations constrained the conclusions. First, the study is a secondary analysis of a cross-sectional self-report dataset and cannot determine the direction of causality. Higher ethical awareness may cause a greater perceived need for training, but training-oriented instructors may also have higher ethical awareness. Second, the sample consisted of only GenAI-experienced EFL instructors recruited through a voluntary accessible pool in Saudi higher education and so the results should not be generalised to non-users or to all EFL instructors. Third, the professional-development outcome is operationalised by three highly endorsed items. Internal consistency was high, and item-specific models captured the general pattern, but future work should include a wider range of development needs, such as institutional policy, advanced validation, data governance, and discipline-specific AI evaluation. Fourth, all predictor domains are perceived rather than demonstrated competencies. Thus, the positive association with self-efficacy is not evidence that instructors who perform better objectively require more training. Research should examine if instructors perceiving greater need indeed make more calibrated decisions when deciding on flawed AI-generated assessments. Fifth, we estimated quantile models on a moderate sample size. The bootstrap with 1000 resamples offers a more robust way to report uncertainty but the exact quantile-specific differences need to be replicated in larger, independent samples. Finally, demographic adjustment greatly increased explained variance and attenuated some of the competency coefficients. This indicates that a professional-development need is contextualised in career and institutional context, not just individual AI competence.

7. CONCLUSION

The professional-development need in AI-mediated EFL assessment goes beyond a lack of familiarity with AI. Beyond familiarity, competency information added considerable explanatory value across conditional quantiles. After controlling for the other

domains, perceived assessment literacy was linked to less training need. Ethical awareness and self-efficacy were linked to greater need, and were the strongest competency correlates after controlling for demographics. At lower and median need familiarity was informative, but did not clearly discriminate instructors at the upper quantile. The practical implication is that universities should do more than introductory AI workshops. Professional development should be differentiated by developmental need and should include operational fluency, language-assessment literacy, evaluation of AI-generated materials, ethics, and supervision by humans. Future work should validate these patterns using independent samples and performance-based measures of teachers' actual human-AI assessment judgement.

Author Contributions: Conceptualization, F.A.M. and H.S.A.; methodology, F.A.M. and H.S.A.; software, F.A.M. and H.S.A.; validation, F.A.M. and H.S.A.; formal analysis, F.A.M. and H.S.A.; investigation, F.A.M.; resources, F.A.M.; data curation, F.A.M. and H.S.A.; writing—original draft preparation, F.A.M. and H.S.A.; writing—review and editing, F.A.M. and H.S.A.; visualization, F.A.M. and H.S.A.; supervision, F.A.M. and H.S.A.; project administration, F.A.M. and H.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The original study was approved by the Bioethics Committee of Scientific and Medical Research, University of Jeddah, Saudi Arabia (Registration No. HAP-02-J-094; Application No. UJ-REC-461).

Informed Consent Statement: Electronic informed consent was obtained from all participants in the original data collection.

Data Availability Statement: The data sets generated during and/or analysed during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank the participants in the original survey study.

Conflicts of Interest: The authors have no conflict of interest to declare.

ORCID iD: Fawaz Al Mahmud, <https://orcid.org/0000-0002-1029-1556>; Hassan Saleem Alqurashi, <https://orcid.org/0000-0002-5190-3088>

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

REFERENCES

- [1] jabr, F. S., & Al-Ahdal, A. A. M. H. (2024). Ethical and pedagogical implications of AI in language education: An empirical study Ha'il University. *Acta Psychologica*, 251, 104605. <https://doi.org/10.1016/j.actpsy.2024.104605>
- [2] kouk, W. A., & Khlaif, Z. N. (2024). AI-resistant assessments in higher education: Practical insights from faculty training workshops. *Frontiers in Education*, 9, 1499495. <https://doi.org/10.3389/educ.2024.1499495>
- [3] Mahmud, F. (2026). EFL instructors' assessment literacy for generative AI in Saudi higher education: Readiness, ethical concerns, and policy needs. *Education Sciences*, 16, 1081. <https://doi.org/10.3390/educsci16071081>
- [4] talib, A. (2026). Domain-differentiated generative AI competency self-efficacy among university EFL teachers: Patterns and predictors. *Behavioral Sciences*, 16, 1252. <https://doi.org/10.3390/bs16071252>
- [5] zubi, A. A. F., & Alelaiwi, A. S. (2025). Exploring EFL university teachers' perceptions of AI-generative tools in Saudi Arabia: A mixed-methods study. *Discover Computing*, 28, 282. <https://doi.org/10.1007/s10791-025-09793-x>
- [6] irghi, A. H., Zengin, R., Dahal, N., Hasan, M. K., & Shrestha, S. (2026). Examining EFL teachers' awareness, use and challenges of integration in ELT context. *Frontiers in Education*, 11, 1779680. <https://doi.org/10.3389/educ.2026.1779680>
- [7] ren, X., & Pi, X. (2026). Assessing EFL students' GenAI-assisted writing: Teachers' pains, perceptions and practices. *Assessing Writing*, 67, 100996. <https://doi.org/10.1016/j.asw.2025.100996>
- [8] ruang, P.-L., & Yan, X. (2025). Language assessment in the era of generative artificial intelligence: Opportunities, challenges, and future directions. *System*, 134, 103846. <https://doi.org/10.1016/j.system.2025.103846>
- [9] rombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia*, 10, 3. <https://doi.org/10.1186/s40468-020-00101-6>
- [10] ji, Y., Liu, Y., Yu, H., & Gao, Y. (2022). Developing English teachers' language assessment literacy in an EAP reform context through test design: A case study. *System*, 109, 102866. <https://doi.org/10.1016/j.system.2022.102866>
- [11] ilcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9, 113–132. <https://doi.org/10.1080/15434303.2011.642041>
- [12] raldo, F. (2021a). Language assessment literacy and teachers' professional development: A review of the literature. *Profile: Issues in Teachers' Professional Development*, 23(2), 265–279. <https://doi.org/10.15446/profile.v23n2.90533>
- [13] raldo, F. (2021b). A reflection on initiatives for teachers' professional development through language assessment literacy. *Profile: Issues in Teachers' Professional Development*, 23(1), 197–213. <https://doi.org/10.15446/profile.v23n1.83094>
- [14] raldo, F., Escalante-Villa, D., & Isaza-Palacio, D. (2023). English teachers' perceptions of a language assessment literacy course. *OW*, 30, 2. <https://doi.org/10.19183/how.30.2.750>
- [15] ayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction to software implementation. *Behavior Research Methods*, 39(4), 709–722. <https://doi.org/10.3758/BF03192961>

- [16] Jynh, L., Le, T. H., Dang, B., et al. (2025). Mindset matters: Fostering teachers' responsible AI use through professional development. *Journal of New Approaches in Educational Research*, 14, 23. <https://doi.org/10.1007/s44322-025-00043-y>
- [17] Jenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50. <https://doi.org/10.2307/1913643>
- [18] Jenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4), 143–156. <https://doi.org/10.1257/jep.15.4.143>
- [19] Jhnke, L., Zou, D., Ou, A. W., & Gu, M. M. (2025). Preparing future educators for AI-enhanced classrooms: Insights into AI literacy and integration. *Computers and Education: Artificial Intelligence*, 8, 100398. <https://doi.org/10.1016/j.caeai.2025.100398>
- [20] Kemmel, 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, 100–120. <https://doi.org/10.1080/15434303.2019.1674855>
- [21] Kif, M. W. (2021). Exploring tertiary EFL practitioners' knowledge base component of assessment literacy: Implications for teacher professional development. *Language Testing in Asia*, 11, 19. <https://doi.org/10.1186/s40468-021-00130-9>
- [22] Lee, J., & Butler, Y. G. (2020). Reconceptualizing language assessment literacy: Where are language learners? *TESOL Quarterly*, 54, 1098–1111. <https://doi.org/10.1002/tesq.576>
- [23] Levi, T., & Inbar-Lourie, O. (2020). Assessment literacy or language assessment literacy: Learning from the teachers. *Language Assessment Quarterly*, 17, 168–182. <https://doi.org/10.1080/15434303.2019.1692347>
- [24] Li, Wang, C., & Bonk, C. J. (2026). Generative AI for teachers' self-directed professional development: A mixed-methods study. *edTechTrends*, 70, 23–38. <https://doi.org/10.1007/s11528-025-01123-8>
- [25] Li, C. K., Yu, P. L. H., Xu, S., Ng, D. T. K., & Jong, M. S.-Y. (2024). Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies. *Smart Learning Environments*, 11, 50. <https://doi.org/10.1186/s40561-024-00342-5>
- [26] Li, J., Zheng, R., Gong, Z., & Xu, H. (2024). Supporting teachers' professional development with generative AI: The effects on higher order thinking and self-efficacy. *IEEE Transactions on Learning Technologies*, 17, 1279–1289. <https://doi.org/10.1109/TLT.2024.3369690>
- [27] Liorhouse, B. L. (2024). Beginning and first-year language teachers' readiness for the generative AI age. *Computers and Education: Artificial Intelligence*, 6, 100201. <https://doi.org/10.1016/j.caeai.2024.100201>
- [28] Liorhouse, B. L., & Kohnke, L. (2024). The effects of generative AI on initial language teacher education: The perceptions of teacher educators. *System*, 122, 103290. <https://doi.org/10.1016/j.system.2024.103290>
- [29] Mami, F. A., & Alharthi, R. M. (2022). Improving language assessment literacy for in-service Saudi EFL teachers. *Arab World English Journal*, 13(3), 536–554. <https://doi.org/10.24093/awej/vol13no3.35>
- [30] Merling, K., Stenberg, C.-J., McGrath, C., Åkerfeldt, A., Heintz, F., & Stenliden, L. (2024). In search of artificial intelligence (AI) literacy in teacher education: A scoping review. *Computers and Education Open*, 6, 100169. <https://doi.org/10.1016/j.caeo.2024.100169>
- [31] Min, H., & Zhang, J. (2022). Assessment literacy of college EFL teachers in China: Status quo and mediating factors. *Studies in Educational Evaluation*, 74, 101157. <https://doi.org/10.1016/j.stueduc.2022.101157>
- [32] Min, X., Cheng, G., & Ling, M. H. (2025a). Artificial intelligence in teaching and teacher professional development: A systematic view. *Computers and Education: Artificial Intelligence*, 8, 100355. <https://doi.org/10.1016/j.caeai.2024.100355>
- [33] Min, X., Cheng, G., & Ling, M. H. (2025b). Enhancing teachers' AI competency: A professional development intervention study based on intelligent-TPACK framework. *Computers and Education: Artificial Intelligence*, 9, 100521. <https://doi.org/10.1016/j.caeai.2025.100521>
- [34] Ming, Y., Derakhshan, A., & Ghiasvand, F. (2025). EFL teachers' generative artificial intelligence (GenAI) literacy: A scale development and validation study. *System*, 133, 103791. <https://doi.org/10.1016/j.system.2025.103791>
- [35] Ming, F., & Shen, B. (2022). Language assessment literacy of teachers. *Frontiers in Psychology*, 13, 864582. <https://doi.org/10.3389/fpsyg.2022.864582>
- [36] Ming, X., Xia, Q., Gu, M., Rajaram, K., & Chiu, T. K. F. (2024). Assessment and learning outcomes for generative AI in higher education: A scoping review on current research status and trends. *Australasian Journal of Educational Technology*, 40, 37–55. <https://doi.org/10.14742/ajet.9540>
- [37] Ming, Q., Yang, Y., Weng, X., Cilsalar-Sagnak, H., Cheng, W. K., & Chiu, T. K. F. (2026). Generative artificial intelligence competency self-efficacy scale for teachers and researchers (GAICS-TR) in higher education. *Journal of Computers in Education*, 13, 769–793. <https://doi.org/10.1007/s40692-025-00373-y>
- [38] Ming, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 53, 149–162. <https://doi.org/10.1016/j.tate.2016.05.010>
- [39] Ming, Y., Derakhshan, A., & Park, Y. (2026). Unpacking English language teachers' professional learning needs in the age of AI: descriptive qualitative analysis of perceptions, expectations and suggested supports. *European Journal of Education*, 61, 1–17. <https://doi.org/10.1111/ejed.70856>