# Construction and validation of the AI Ethics Scale in language research

**Bora Demir**
School of Foreign Languages, Çanakkale Onsekiz Mart University, Çanakkale, Turkey

**Selami Aydın**
Department of English Language Teaching, Istanbul Medeniyet University, Istanbul, Turkey

Despite the growing relevance of artificial intelligence (AI) in language research, there is a lack of validated instruments to assess researchers' ethical awareness regarding its use. Given the increasing integration of AI into technology-enhanced language teaching, assessment and learning environments, there is also a need to develop scales to provide a foundation for strengthening ethical AI practices across language research contexts. This study aimed to develop and validate the AI Ethical Awareness and Responsibility Scale (AI-EARS) to measure ethical awareness and responsibility in AI use among language researchers. The scale development process followed established validation standards and involved exploratory and confirmatory factor analyses. After item reduction, a two-factor structure was identified in the exploratory analysis, while a confirmatory factor analysis supported a refined one-factor, five-item model with strong model fit and excellent internal consistency. This final structure emerged because several items demonstrated weak loadings, high residual correlations or conceptual overlap during confirmatory analysis, which indicated that they did not sufficiently contribute to the latent construct. Evidence for concurrent validity and test-retest reliability further supported the psychometric robustness of the scale.

*Implications for practice or policy:*

- Language researchers should adopt the AI-EARS to self-assess and enhance their ethical awareness when integrating AI tools into their studies.
- Academic institutions could incorporate the AI-EARS into research ethics training programmes to promote responsible AI use among scholars.
- Research policymakers may use the AI-EARS as a benchmark to develop guidelines addressing ethical AI practices in language research.
- AI-EARS serves institutions as a diagnostic tool, enabling targeted training to ensure ethical and transparent AI integration in language research.

*Keywords:* language research, artificial intelligence (AI), ethics, validity, reliability, scale development

## Introduction

Research that contributes to understanding the way human beings communicate and interact socially and the language-education nexus is vital to applied linguistics (Żammit, 2024). Language is a vehicle for investigating how learners articulate their ideas, social beliefs and feelings (Canagarajah, 2012). It underlies all communication across linguistic boundaries, shaping social life and international progress (Byram, 2020), enabling learners in new language acquisition to acquire linguistic knowledge and intercultural awareness (Deardorff, 2020). Research, of course, is needed to explore the cognitive and emotional dimensions of technology and pedagogy (Qiao & Zhao, 2023), and it revitalises teaching processes and course production (Kern, 2006). Yet, the consequences of such contributions depend on ethical integrity; otherwise, all disrespectful and scheming research efforts add a footnote rather than a credible reference to the field.

Since the human aspect is fundamental to applied linguistics research, the issue of ethics is significant for several reasons. The whole of ethics, from broad issues of data integrity to participants' rights to privacy

and the protection of reported data, is vital given that the subjects of those data are undoubtedly humans capable of exploitation and personal harm (Dörnyei, 2020). Thus, ethical guidelines should mitigate any potential harm related to issues such as cultural and emotional sensitivity, misrepresentation and confidentiality (Zimmer, 2018). Thus, researchers are held accountable for trust, integrity and respect for cultural, emotional and social differences (Hultgren et al., 2016). Put simply, ethics is concerned with the respectful treatment of linguistic data and those involved with it, that is, usually, with subjects who are satellites of this sensitive and vulnerable core (De Costa, 2015). Furthermore, researchers who conduct their work in accordance with ethical procedures ensure the reliability and validity of their science, as well as the trustworthiness and permanence of their work. In this way, ethical procedures become necessary to protect those involved in research and raise the level of outcomes in practice, thereby generating findings that contribute theory in language education. In this context, as researchers increasingly integrate artificial intelligence (AI) into their work, ethical issues become more critical, because the use of AI tools in language research raises concerns that cannot be overlooked; on the one hand, AI tools provide a fantastic opportunity to the language research process; on the other hand, they pose serious concerns (Demir & Aydın, 2026, Chapter 16; Floridi et al., 2019).

While AI significantly contributes to language research and ethics, its rapid integration raises critical ethical concerns. AI can understand and process input information and provide reasoning and rationalisation (Gilson et al., 2023). Thus, researchers primarily use AI tools to write and edit documents, formulate specific research questions, compile literature reviews and identify discussion points (Dowling & Lucey, 2023). Moreover, AI tools can serve as collaborators, providing feedback on writing (Aydın, 2024) and supporting rewriting, proofreading and editing (Lin, 2023). Nevertheless, these benefits cannot be considered without ethical considerations and academic integrity. For instance, AI tools also pose ethical problems, including plagiarism, copyright infringement, fabrication, data manipulation and false authorship (Grimaldi & Ehrler, 2023; Lin, 2023). Moreover, other concerns include oversimplification, data privacy and misinterpretations (Yao et al., 2024). More importantly, overreliance on AI tools and environments during the research process raises questions about data ownership and consent (Holmes et al., 2019; Schembri & Jahić Jašić, 2022) and about cultural bias in research design and implementation (Pelzang & Hutchinson, 2018). To this end, it is evident that overreliance on AI in the research process may bring problems, including attrition in research. Thus, language researchers need awareness and knowledge when considering the abovementioned complexities and difficulties. From this perspective, ethical considerations regarding how language researchers perceive and use AI tools and environments in their research must be clarified within the research context. To achieve this aim, valid and reliable measurement tools must be used to explore how researchers perceive ethical issues in the research process. However, the literature lacks findings on the construction and validation of an AI ethics scale. To address this gap, the present study conceptualises AI ethics perceptions within a defined theoretical perspective that explains how individuals evaluate and regulate technology use in research contexts. Rather than treating ethical concerns as isolated attitudes, the study situates them within broader cognitive and behavioral processes that shape technology-related decision making. The following section therefore outlines the theoretical framework that underpins the construct of AI ethics perception and guides the development of a valid and reliable measurement tool for language researchers.

## Theoretical framework

Several clarifications are needed regarding the study variables. First, according to Nunan (1992, p. 3), *research* is a systematic inquiry process involving a question or hypothesis, data collection, analysis and interpretation. It typically follows steps such as identifying a problem, refining a general inquiry into a research question or hypothesis, reviewing relevant literature, designing an appropriate research method, selecting data collection and analysis tools and reporting results while addressing ethical issues (Denkci Akkaş et al., 2022; Elmas & Aydin, 2017; Seliger & Shohamy, 1989). Research is often characterised as systematic, empirical and objective in its examination of phenomena, acquisition of new knowledge and validation of theories across basic, applied and practical domains. Second, *ethics* establishes standards to ensure the integrity, transparency and fairness of the research process by preserving participants' rights, data integrity and overall credibility, emphasising principles such as honesty,

objectivity, confidentiality non-maleficence, responsible publication, fairness and adherence to legal guidelines (Comstock, 2012; Haneef & Agrawal, 2024; Petrova et al., 2016; Weinbaum et al., 2019). Third, *AI* is recognised for its capabilities in learning, adapting and creatively processing knowledge (Aydın, 2024; Korteling et al., 2021; Sarker, 2022) by demonstrating intelligence through understanding and processing information (Gilson et al., 2023), versatility in performing tasks such as generating texts and editing documents (Dowling & Lucey, 2023) and collaborative abilities through interactive knowledge sharing and responsiveness (Aydın, 2024; Lin, 2023), thereby highlighting its transformative potential in enhancing educational and research practices. Fourth, *perception* is the organisation and interpretation of sensory information to understand the environment (Schacter et al., 2009) and is shaped by an individual's attention, expectations, memory and learning (Bernstein et al., 2008). In line with social cognitive theory, these cognitive processes interact with environmental factors to influence human behaviours (Bruner, 1957). However, the constructive perspective defines perception as a product of the interaction between stimulus and information (Démuth, 2013). Moreover, from a computational perspective, perception involves simple mechanical rules governing unconscious entities (Gregory, 2015). Fifth and last, *scale development* is a systematic process used to design, construct and validate a measurement tool that quantifies abstract psychological, educational or social constructs (DeVellis, 2017). It involves clearly defining the construct, generating items that represent its dimensions and ensuring that the scale is reliable and valid (John & Benet-Martínez, 2000) to gather standardised data for interpretation, comparison and inference.

Building upon the conceptual framework drawn above, it is also necessary to establish a theoretical model that explains how AI ethical awareness functions as a multidimensional construct within the context of language research. To address the conceptual grounding of AI ethical awareness, the present study adopted a multidimensional perspective that views the construct as comprising interrelated cognitive, affective and behavioural components. Cognitively, AI ethical awareness involves researchers' understanding of ethical principles, recognition of risks such as plagiarism, bias and data misuse, and knowledge of responsibilities associated with AI-assisted research (Carobene et al., 2024). Affectively, it includes concerns, uncertainties and evaluative reactions that arise when they encounter issues related to privacy, consent, transparency or the reliability of AI-generated content (Khlaif, 2023). Behaviourally, the construct encompasses self-regulated actions such as verifying AI outputs, adhering to ethical procedures and taking responsibility for integrating AI into research processes (Yurt, 2025). The cognitive, affective and behavioural elements emphasised in the theory inform the initial item pool, including researchers' ethical understanding, their concerns about AI use and their self-regulated actions when working with AI tools. Considering this perspective, it should be noted that, in the current study, AI ethics was operationalised in a focused manner, referring specifically to language researchers' ethical awareness and sense of responsibility when using AI tools in the research processes. Thus, the scale is not intended to represent a comprehensive taxonomy of AI ethics principles; instead, it aims to provide a parsimonious measure of researchers' ethical use in AI-assisted language research.

## Literature review

Research has shown that specific instruments have been developed and validated to assess and ensure ethical practices in language research. One notable example was the Code of Ethics in Testing Inventory, designed by Hosseinnia and Kafi (2024) to evaluate English as a Foreign Language (EFL) instructors' perspectives on ethical issues, focusing only on language testing regarding fairness and honesty as values of integrity. However, the instrument was developed by ignoring ethical concerns such as data privacy, informed consent, data confidentiality, reliability, academic integrity and ethical risks related to AI. Ahmadi (2012) developed an EFL context-based questionnaire to collect information on students' cheating in the Iranian EFL context. The questionnaire, designed to understand students' attitudes towards cheating, did not include any items on the ethics of using AI to cheat. In a separate study, Ahmadi (2014) examined plagiarism among Iranian EFL learners in the academic context via a questionnaire. The first part consisted of eight items related to personal experiences of plagiarism during their academic life. The second part explored students' attitudes towards plagiarism with seven items, while the last part was designed to understand the reasons for plagiarism. Nevertheless, the questionnaire lacked items on the

use of AI for academic plagiarism. To this end, it should be noted that the current literature does not include scale development and validation regarding the ethical dimension of using AI in foreign language research. Moreover, the results of these studies indicate that existing language-oriented ethics instruments were developed before the widespread use of AI and therefore do not reflect the discipline-specific ethical concerns that arise when AI is integrated into language research or technology-enhanced language education. More importantly, none of these tools addressed ethical awareness and responsibility in the AI era, particularly in contexts where AI tools are used for corpus analysis, automated writing evaluation, data-driven feedback,or classroom-based research practices that place language researchers and educators in direct contact with potential ethical risks such as misinformation, biased outputs, confidentiality issues and a lack of transparency.

On the other hand, numerous scale development studies on the ethical integration of AI use in educational and research contexts have been advanced through a diverse array of validated instruments. Reflection-focused scales such as the AI Ethical Reflection Scale (Wang et al., 2025) and cross-cultural measures such as the AI and Ethics Perception Scale (Saatci, 2025) have provided researchers and practitioners with robust frameworks for assessing moral reasoning and cultural variation in ethical norms. In addition to these studies, several attitude scales have been developed, including the AI Attitude Scale for general and student populations (Grassini, 2023) and an instrument designed for undergraduates (Jang et al., 2022). These instruments focus on measuring affective and cognitive orientations underlying support for AI innovation. In educational assessment, Perkins et al. (2024) developed the Artificial Intelligence Assessment Scale to measure fairness, transparency and validity when integrating generative AI into the assessment process. In addition, Yurt and Kasarci (2024) explored motivational drivers of AI use through the Questionnaire of AI Use Motives, identifying convenience, social norms and integrity perceptions as key factors shaping AI engagement. Işık et al. (2024) emphasised enthusiasm for pedagogical innovation and caution regarding de-skilling and data privacy, underscoring the need for targeted support and policy development. In a separate study, Khalaf (2025) examined attitudes towards plagiarism and found that such attitudes predicted “aigiarism”.

According to the findings presented in Table 1, although several scales address ethical issues in general, no specific measurement tool currently focuses on ethical concerns in language research. For instance, the AI Ethical Reflection Scale by Wang et al. (2025) focused on awareness, critical evaluation and AI for social good among undergraduate students. The Questionnaire of AI Use Motives, developed by Yurt and Kaşarci (2024), assessed expectancy, attainment, utility value, intrinsic value and cost dimensions, although it does not directly address ethical concerns. Moreover, the Attitudes Toward Aigiarism Questionnaire by Khalaf (2025) explored the ethical implications of AI in academic misconduct and plagiarism. Jang et al. (2022) developed the Attitudes Toward the Ethics of Artificial Intelligence Scale, which included factors such as fairness, transparency, privacy, responsibility and non-maleficence. Işık et al. (2024) developed the Teachers’ Perception Scale Towards the Use of Artificial Intelligence Tools in Education, assessing competence, anxiety and perceived usefulness from a pedagogical perspective. Perkins et al. (2024) introduced the Artificial Intelligence Assessment Scale. The AI and Ethics Scale by Saatci (2025) was among the most comprehensive tools, as it addressed transparency, accountability, privacy, fairness and human oversight in a sector-agnostic, cross-cultural context. Lastly, Grassini (2023) developed the AI Attitude Scale, which includes a brief measure of general attitudes towards AI. All of these tools were developed in English.

Table 1
*Summary of the findings on scale development research*

| Scale | Authors | Items | Target groups | Factor solution | Cronbach's alpha | % of variance explained |
|---|---|---|---|---|---|---|
| AI Ethical Reflection Scale | Wang et al. (2025) | 12 | UG | AI ethical awareness | .77 | NR |
| | | | | AI critical evaluation | .73 | NR |
| | | | | AI for social good | .81 | NR |
| | | | | Overall | .77 | 51 |
| Questionnaire of AI Use Motives | Yurt & Kaşarci (2024) | 20 | | Expectancy | .88 | NR |
| | | | | Attainment | .92 | NR |
| | | | | Utility value | .91 | NR |
| | | | | Intrinsic and interest value | .93 | NR |
| | | | | Cost | .86 | NR |
| | | | | Overall | .82 | 82.85 |
| Attitudes Toward Aigiarism Questionnaire | Khalaf (2025) | 13 | | Attitudes towards plagiarism and "aigiarism" | .90 | 65.14 |
| Attitudes Toward the Ethics of Artificial Intelligence | Jang et al. (2022) | 17 | | Fairness | .86 | 28.99 |
| | | | | Transparency | .80 | 10.99 |
| | | | | Privacy | .75 | 8.60 |
| | | | | Responsibility | .76 | 7.73 |
| | | | | Non-maleficence | .81 | 6.18 |
| | | | | Overall | .84 | 62.49 |
| Teachers' Perception Scale Towards the Use of Artificial Intelligence Tools in Education | Işık et al. (2024) | 37 | Teachers | Competence | .97 | 34.42 |
| | | | | Anxiety | .86 | 15.09 |
| | | | | Usefulness | .87 | 14.78 |
| | | | | Overall | .97 | 64.29 |
| Artificial Intelligence Assessment Scale | Perkins et al. (2024) | 5 | | NR | NR | NR |
| AI and Ethics Perception Scale | Saatci (2025) | 30 | Sector-agnostic cross-cultural context | Transparency | .88 | 22 |
| | | | | Accountability | .85 | 18 |
| | | | | Privacy | .90 | 12 |
| | | | | Fairness | .91 | 9 |
| | | | | Human oversight | .94 | 6 |
| | | | | Overall | NR | 67.00 |
| AI Attitude Scale | Grassini (2023) | 4 | General population | AI attitudes | .83 | NR |

*Note*. UG = undergraduates. NR = not reported.

As illustrated in Table 1, the target groups used in the existing scales included undergraduates, teachers, the general population and sector professionals, whereas the factor solutions varied widely. On the other hand, the internal consistency values ranged from .73 to .97, indicating acceptable to excellent reliability, while several studies did not report percentage of variance values. Despite this diversity, none of the instruments capture the ethical awareness and responsibility specific to AI-mediated language research, in which researchers handle learner data, generate or verify AI-generated outputs, design AI-enhanced instructional materials and make pedagogical or methodological decisions informed by AI technologies. This gap highlights the need for a new, context-sensitive scale tailored to the ethical challenges faced by language researchers and educators who employ AI in language research.

## Overview of the current study

Several reasons guided this study. From the broadest perspective, applied linguistics research is essential to comprehend the nature of language teaching and learning. Thus, ethical considerations are vital in the process mentioned above. On the other hand, while AI significantly advances language research and ethical practices, its rapid integration raises critical concerns among researchers. Moreover, as AI-supported tools increasingly shape technology-enhanced language teaching, assessment and learning environments, ensuring that researchers demonstrate ethical awareness and responsibility is essential for supporting transparent, safe and pedagogically sound AI integration across educational contexts. Within this scope, there is a need to understand how language researchers perceive ethical issues concerning the use of AI in their research. However, while research has produced a variety of validated instruments addressing the ethical integration of AI in language education and research, they are not designed to reflect the ethical issues in language research. In other words, they do not focus on language-specific ethical concerns. Additionally, inconsistencies in factor structures and the lack of reporting variance further limit their applicability and adaptation to the language-teaching domain. Therefore, the present study aimed to construct and validate a measurement scale that uses language researchers' perceptions of the ethical use of AI in language research and addressing the following research question: "How valid and reliable is the AI Ethical Awareness and Responsibility Scale (AI-EARS) in language research?"

## Methodology

### Research context

This study, which aimed to develop and validate the AI-EARS, was guided by psychological and educational testing standards, which emphasise best practices in measurement and view validity as a unified concept centred on establishing construct validity (American Educational Research Association et al., 2014). According to this perspective, scale validation involves evidence from multiple sources, such as content relevance, response processes, internal structure, relationships with other variables and the implications of score interpretation (Downing, 2003). Accordingly, the study sought to gather evidence on the internal structure and associations with external variables at both the item and the overall scale levels. In doing so, it adhered to a rigorous process that included defining test specifications, generating and reviewing items, developing scoring and conducting iterative revisions (American Educational Research Association et al., 2014).

Scale development and validation are critical processes in the language research context, given the complexity of the psychological and pedagogical constructs involved (Aydın et al., 2024). Thus, the current study followed a three-phase model to ensure the robustness of the instrument. In Phase 1, substantive and content validity were established by reviewing relevant literature and conducting learner interviews to generate an initial item pool. Expert evaluations and target population feedback were then used to assess item clarity, relevance and alignment with defined constructs (Beck & Gable, 2001). Phase 2 focused on construct validity through exploratory factor analysis (EFA), which allowed for identifying latent dimensions and refining the scale by eliminating weak or redundant items. Internal consistency coefficients were also calculated to confirm the reliability of each factor (Kenyon & MacGregor, 2013). In Phase 3, confirmatory factor analysis (CFA) was used to validate the factor structure obtained in EFA. This was supported by concurrent validity testing with the Academic Integrity Scale – Short Version (Rettinger

& Kramer, 2009) and test-retest reliability to assess temporal stability (Boateng et al., 2018). Through these sequential steps (Figure 1), the study aimed to construct and validate the AI-EARS to ensure that it is both psychometrically sound and pedagogically meaningful.

| Phases | | Procedure | Item selection |
|---|---|---|---|
| 1 | Substantive validity | • Reviewing the literature<br>• Rationalising the need for the scale<br>• Defining constructs<br>• Interviews | Developing the initial pool |
| | Content validity | • Expert panel discussion<br>• Target population judgement | Item revision |
| 2 | Construct validity | EFA | Excluding irrelevant items |
| | Reliability | Internal consistency coefficient | |
| 3 | Construct validity | CFA | Finalising items |
| | Criterion validity | Concurrent validity | |
| | Reliability | • Internal consistency coefficient<br>• Test-retest reliability | |

*Figure 1.* Flowchart of the scale development process (Aydın et al., 2024, p. 662)

## Participants

As indicated in Table 2, data were collected from six independent samples corresponding to the interview, pilot, EFA, CFA, concurrent validity, and test–retest reliability phases. Each group was assigned to a specific stage of the scale development process to ensure methodological rigor across phases. The inclusion of both novice and senior researchers was intentional, as it enabled the examination of ethical perceptions of AI use in language research across different career stages and age groups. Across all phases, participants represented a broad range of ages and research experience levels, enhancing the diversity of perspectives captured in the study. The samples were predominantly female, reflecting the overall population distribution in the field, and academic qualifications ranged from bachelor's to doctoral degrees. Detailed demographic characteristics for each sample are presented in Table 2.

Table 2
*Demographic data of the participants*

| | Age | | Research experience in years | | Gender | | | | Degree | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Female | | Male | | Bachelor's | | Master's | | Doctorate | |
| | M | SD | M | SD | N | % | N | % | N | % | N | % | N | % |
| Sample 1: Interview (*n* = 20) | 39.5 (26–54) | 4.13 | 14.6 (3–32) | 8.46 | 11 | 55.0% | 9 | 45.0% | -- | -- | 5 | 25.0% | 15 | 75.0% |
| Sample 2: Pilot (*n* = 9) | 36.4 (25–56 | 9.33 | 7.6 (1–25) | 7.52 | 6 | 66.6% | 3 | 44.4% | 2 | 22.3% | 3 | 33.3% | 4 | 44.4% |
| Sample 3: EFA (*n* = 355) | 38.8 (20–72 | 19.4 | 11.7 (1–41) | 9.4 | 229 | 64.5% | 126 | 35.5% | 69 | 19.4% | 114 | 32.1% | 172 | 48.5% |
| Sample 4: CFA (*n* = 137) | 33.10 (20–60) | 11.50 | 8.91 (1–35) | 8.88 | 94 | 68.6% | 43 | 31.4% | 53 | 38.7% | 30 | 21.9% | 54 | 39.4% |
| Sample 5: Concurrent validity (*n* = 68) | 31.49 (20–56) | 11.17 | 7.72 (1–35) | 8.74 | 51 | 75.0% | 17 | 25.0% | 31 | 45.6% | 18 | 26.5% | 19 | 27.9v |
| Sample 6: Test-retest reliability (*n* = 99)6 | 37.26 (21–60) | 10.26 | 11.47 (1–35) | 8.88 | 70 | 70.7% | 29 | 29.3% | 18 | 18.2% | 30 | 30.3% | 51 | 51.5% |

## Tools

This study utilised three data collection tools: a background questionnaire, AI-EARS (the scale developed and validated in this study) and the Academic Integrity Scale – Short Version developed by Rettinger and Kramer (2009). First, the background questionnaire was used to interrogate participants' age, gender, years of research experience and academic background. Second, the items developed through the study were used to measure participants' perceptions of AI ethics in language research. The scale consisted of 35 items in the EFA process and five in the CFA process. The items in the scale were presented on a 5-point Likert scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree* and 5 = *strongly agree*). Third, the Academic Integrity Scale – Short Version was administered alongside the target scale. This 8-item instrument measured attitudes and behaviours related to academic integrity. The items in the scale were given on a 5-point Likert scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree* and 5 = *strongly agree*). The scale demonstrated acceptable reliability with Cronbach's alpha ranging from .61 to .83 and strong validity, with self-reported ethics behaviours significantly predicted by direct knowledge ($\beta = .50$, $p < .0001$), neutralising attitudes ($\beta = .41$, $p < .0001$), and extrinsic orientation ($\beta = .20$, $p = .003$), accounting for 57% of the variance ($R^2 = .57$) in a multiple regression model.

## Procedure

Following approval from the Scientific Research and Publication Ethics Board of Çanakkale Onsekiz Mart University (Approval No. 2025-YÖNP-249; Decision No. 45/143), all procedures involving human participants were conducted in accordance with institutional ethical standards. Participants in the interview, pilot, EFA, CFA, concurrent validity and test–retest reliability groups were informed about the purpose of the study and the voluntary nature of their participation and given assurance of anonymity and confidentiality. Prior to data collection, informed consent was obtained electronically from all participants in each sample group. Invitations were distributed via email, and weekly reminder messages were sent during the approximately 1-month data collection period to encourage participation. All responses were stored securely and used exclusively for research purposes in line with ethical guidelines.

As previously clarified, the study was conducted in three consecutive phases. In the first phase, item generation was based on interviews with participants, and the qualitative data were analysed using thematic analysis. For this purpose, all responses were transcribed and subjected to open coding, during which recurring expressions related to ethical responsibility, confidentiality, informed consent, data accuracy, transparency and researcher accountability were identified. Three independent coders compared codes and resolved discrepancies through discussion, ensuring triangulation and data saturation. Through coding, these initial codes were organised into broader thematic categories, each representing a key ethical dimension of AI use in language research. Then, these themes were directly transformed into item statements, ensuring that the initial pool reflected authentic researcher perceptions rather than assumptions. A literature review was also carried out to support item development. In the second phase, a pilot study was conducted to refine and reduce the independently developed items to 37. During this stage, nine participants reviewed the items for syntactic, lexical and semantic clarity and consistency, leaving 35 items. In the third phase, the refined items were administered to a new group of participants for EFA, resulting in a reduction to nine items based on statistical findings. The fourth phase involved CFA in assessing the factor structure and psychometric properties, resulting in five items based on statistical findings. In the fifth phase, concurrent validity was examined by administering the scale alongside an established measure. Finally, the scale was re-administered in the sixth phase to assess test-retest reliability. All procedures were carried out in English.

## Analysis

IBM SPSS (version 21) and IBM SPSS Amos (version 22) were utilised to conduct the statistical analyses. Before the main procedures, assumptions regarding sample size, univariate and multivariate normality, outlier detection, multicollinearity, linearity and homoscedasticity were examined (Tabachnick & Fidell, 2014). EFA was conducted using the maximum likelihood extraction method to identify the underlying structure of the scale. Items with factor loadings below .70 or communalities below .50 were eliminated,

and the remaining items were re analysed through EFA. Cronbach's alpha coefficients were calculated to assess internal consistency. Following EFA, CFA was employed to assess the fit of the proposed model. A range of fit indices was used, including the chi-square statistic ($\chi^2$), the *p* value, chi-square to degrees of freedom ratio (CMIN/df), root-mean-square error of approximation (RMSEA) with its 90% confidence interval, the *p* value for close fit (PCLOSE), comparative fit index (CFI), Tucker–Lewis index (TLI), incremental fit index (IFI), normed fit index (NFI), relative fit index (RFI), goodness of fit index (GFI), adjusted goodness of fit index (AGFI) and root-mean-square residual (RMR). Last, concurrent validity was examined by correlating the scale scores with those from the Academic Integrity Scale – Short Version developed by Rettinger and Kramer (2009). Item reduction was strictly statistically driven to prioritise psychometric robustness. Items demonstrating weak loadings, high residual correlations or conceptual overlap were eliminated to ensure unidimensionality. This concise structure reduces participant fatigue and enhances response rates in high-stakes research environments while maintaining exceptionally high reliability.

## Results

### EFA

Prior to the data analysis, the assumptions for EFA were checked for normality and sampling adequacy. For this purpose, the participant-to-item ratio was checked and found to be 10.14, which was established within 355 participants and 35 items (Hair et al., 1998). Then, the assumption of normality was checked by examining skewness and kurtosis, and was found to be within acceptable values (skewness > 3, kurtosis > 10) (Kim, 2013). Last, the Kaiser–Meyer–Olkin (KMO) was found to be .913, which is considered excellent and suggests that the sampling is adequate for factor analysis. In addition, Bartlett's test of sphericity yielded a significant result ($\chi^2 = 5680.642$, $df = 595$, $p < .001$), indicating that the variables are sufficiently correlated for factor extraction; thus, the data set was suitable for EFA. In conclusion, these findings showed that the data met the necessary assumptions for conducting factor analysis, as illustrated in Table 3.

Table 3
*EFA assumptions*

| Assumption checked | Result | Interpretation |
|---|---|---|
| Participant-to-item ratio | 10.14 (355 participants; 35 items) | Meets recommended minimum |
| Skewness and kurtosis | Skewness ranged from -2.919 to +0.925; kurtosis ranged from -0.931 to +9.082 | Within the acceptable range |
| KMO | .913 (excellent) | Sampling adequacy confirmed |
| Bartlett's test of sphericity | $\chi^2 = 5680.642$, $df = 595$, $p < .001$ | Suitable for factor extraction |

An item reduction process was performed after observing that the items performed best in the factor analysis. For this purpose, first, the communalities of all items were examined to ensure sufficient shared variance with the underlying factors. Based on the recommended threshold of .50 (Hair et al., 1998), items with rescaled communalities below .50 were considered for removal. As a result, six items (Items 8, 18, 19, 23, 27 and 29) were excluded from the analysis due to their low communality values, indicating they did not adequately contribute to the factor structure. Second, the item reduction process was performed by repeating the communalities and factor loadings analyses. Items with communalities below .50 and factor loadings below .70 were removed from the scale (Hair et al., 1998). After removing 26 items, the remaining nine items were re-evaluated for EFA. The KMO measure of sampling adequacy was .856, indicating that the sample was suitable for factor analysis, while Bartlett's test of sphericity was significant ($\chi^2(36) = 1527.642$, $p < .001$). In addition, the factor analysis revealed two factors, accounting for 64.84% of the variance. The first factor, comprising seven items, accounted for 45.91% of the total variance, while the second factor, consisting of two items, explained an additional 18.93%, as illustrated in Figure 2 and Table 4. Parallel analysis was also conducted to statistically simulate a random data set with the same number of participants and variables, to determine the appropriate number of factors (Watkins, 2018).

The eigenvalues of the random data for the root, mean and percentile were all 1.0, supporting the results indicated by the scree plot. Third and last, the reliability analysis showed a high level of internal consistency, with a Cronbach's alpha coefficient of .85 for the 9-item scale. Corrected item-total correlations ranged from .38 to .75, suggesting that most items were moderately to strongly correlated with the total scale score. The lowest correlations were observed for the last two items (.38 and .41). However, since these items contributed to the overall internal consistency and remained within acceptable psychometric ranges, no items were removed at this stage of the analysis. Although the second factor consisted of only two items and therefore represented a structurally weak factor, it was retained initially due to its high factor loadings and communalities, which indicated strong shared variance and theoretical coherence (Raubenheimer, 2004). This limitation is essential to note, as the restricted item count likely contributed to the instability of this factor and its subsequent removal during CFA. In addition, the privacy and consent dimension was not retained in later stages because it did not demonstrate sufficient statistical stability during CFA.

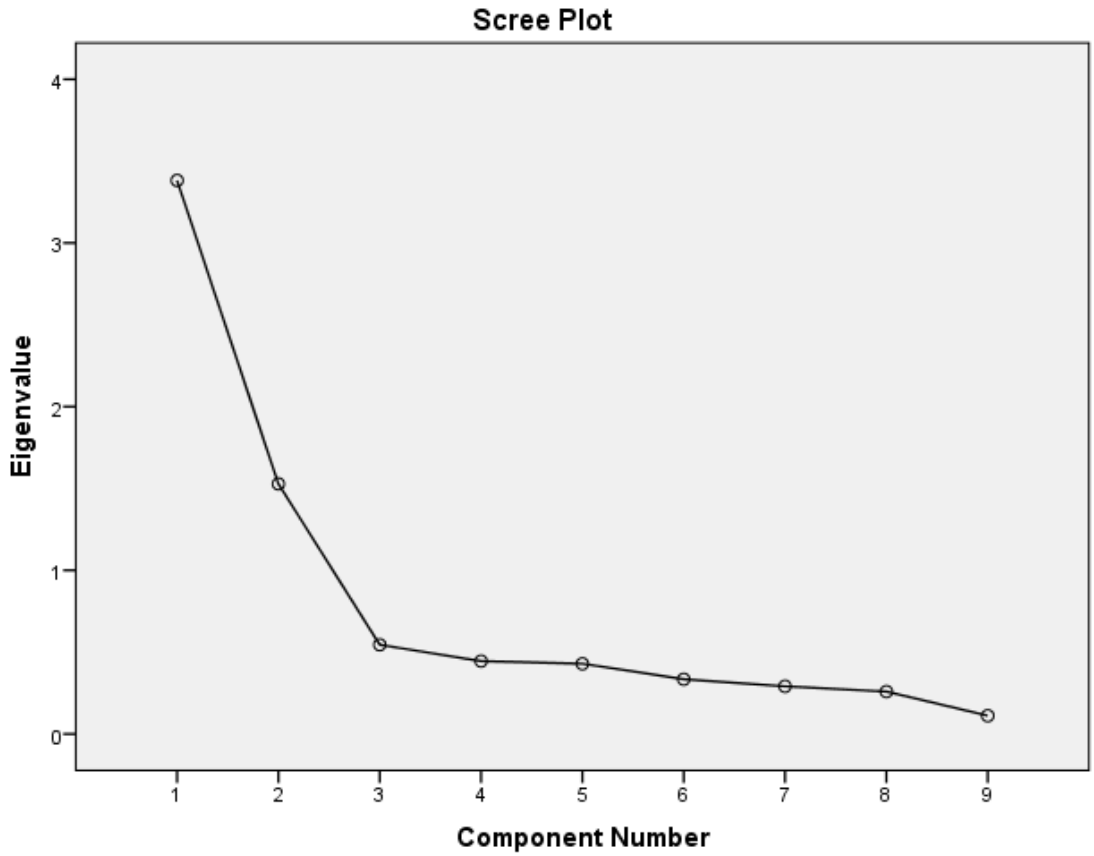


Figure 2. Scree plot

Table 4
*Communalities, factor loadings and item-total statistics of the items (N = 355)*

| Factors | Items | Factor loadings | Total variance | Communalities | Corrected item-total correlation | Cronbach's alpha if item deleted |
|---|---|---|---|---|---|---|
| Ethical Awareness and Responsibility in AI Use | 2. I think it is important to follow ethical rules when using AI in my studies. | .87 | 45.91 | .69 | .75 | .82 |
| | 1. I know that using AI tools in language research comes with ethical responsibilities. | .85 | | .73 | .71 | .82 |
| | 11. Open discussion among colleagues about AI's ethical implications can improve research quality. | .74 | | .82 | .60 | .83 |
| | 6. I believe AI can be useful in language research if it is used ethically. | .73 | | .60 | .56 | .84 |
| | 22. I try to follow ethical rules whenever I include AI in my research process. | .72 | | .79 | .66 | .83 |
| | 9. I feel researchers should always check AI-generated content carefully before using it. | .72 | | .60 | .64 | .83 |
| | 31. I believe language researchers are responsible for how they use AI and the content it creates. | .71 | | .73 | .59 | .83 |
| Privacy and Consent Concerns in AI Use | 15. I am concerned about how AI tools handle the privacy of research participants. | .91 | 18.93 | 1.26 | .38 | .86 |
| | 16. I find it unclear how most AI platforms deal with privacy and consent. | .86 | | 1.12 | .41 | .86 |

## CFA

To clarify the transition from the EFA findings to the final CFA model, the two-factor, nine-item solution obtained in EFA was first tested in CFA. In this process, the initial two-factor model did not demonstrate acceptable fit ($\chi^2/df > 5$, RMSEA > .10, CFI < .90), indicating that the factor structure identified through EFA did not generalise to the new sample. Thus, the two items constituting the second factor, Privacy and Consent Concerns, showed high residual correlations, weak standardised loadings (< .60) and substantial cross-loadings. These indicators suggested that the second factor was not sufficiently distinct and that its two items lacked the psychometric strength required for retention in a confirmatory framework. As a result, the privacy dimension identified in EFA was removed from the final model because it did not meet the required statistical criteria for factor retention**.** The exclusion of this dimension does not dismiss the ethical importance of privacy and informed consent; rather, the CFA results indicate that privacy- and consent-related concerns did not demonstrate sufficient structural stability in the final model. These findings suggest that such concerns may constitute a related but conceptually distinct construct that requires independent measurement and validation beyond the scope of the current scale.

In the CFA process, the participant-to-item ratio was computed, yielding 15.22 (137:9), indicating an adequate sample size for the analysis (Hair et al., 1998). Then, skewness and kurtosis were computed to see the normality of the data set. The values for the items ranged from -.079 to -2.52 for skewness and from -.22 to 5.71 for kurtosis, indicating acceptable skewness and kurtosis values (skewness > 3, kurtosis > 3). Regarding the appropriateness of the data for the analysis, Bartlett's test of sphericity, the KMO test and approximate chi-square were computed. The values were .86 for the KMO test, 746.51 for approximate chi-square and .00 for significance. These values demonstrate that the data could be used for further analysis. In addition, Mahalanobis distances were used to identify outliers and to assess whether the CFA assumptions were met (Tabachnick & Fidell, 2014). Since *p* values were .00, the assumption was met. Last, the variance inflation factor values ranged from 1.42 to 4.53, remaining well below the critical threshold of 10. In contrast, tolerance values ranged from .22 to .70, exceeding the minimum acceptable value of .10. These results indicate that the assumption of multicollinearity was satisfactorily met, since all tolerance values exceeded the acceptable threshold of .20. All variance inflation factor values were below 5, indicating no serious multicollinearity concerns among the variables (Marcoulides & Raykov, 2019).

CFA was processed to observe the structure of the scale. First, the analysis demonstrated an acceptable to excellent model fit across all indices. Four items were removed during the model refinement process due to low factor loadings and high residual correlations, resulting in a more parsimonious five-item, one-factor structure. First, the chi-square value was significant ($\chi^2 = 13.07$, $p = .02$), as expected given the sensitivity to sample size. The relative chi-square (CMIN/df=2.61) fell within the acceptable range (< 3). Second, the RMSEA value of .06 and its 90% confidence interval (.02–.09) indicated a good fit, further supported by the non-significant PCLOSE value of .31 (> .05). Third, incremental fit indices confirmed a strong model (CFI = .99, TLI = .98, IFI = .99, NFI = .98 and RFI = .96), all exceeding the recommended threshold of .95. Last, absolute fit indices were also satisfactory, with a GFI of .99 and an AGFI of .97, both above the .90 benchmark. The RMR was low at .04, indicating acceptable residuals. As illustrated in Table 5, these results provided robust evidence for the validity of the final five-item model.

Table 5
*CFA model fit indices*

| Fit index | Value | Reference values |
|---|---|---|
| Chi-square ($\chi^2$) | 13.07 | Significant ($p < .05$); sensitive to sample size |
| *p* value | .02 | < .05 indicates significance |
| CMIN/df | 2.61 | < 3 = acceptable |
| RMSEA | .06 | < .06 = good |
| 90% CI for RMSEA | .02 – .09 | < .10 = acceptable |
| PCLOSE | .31 | > .05 = good fit |
| CFI | .99 | > .95 = excellent |
| TLI | .98 | > .95 = excellent |
| IFI | .99 | > .95 = excellent |
| NFI | .98 | > .95 = excellent |
| RFI | .96 | > .95 = excellent |
| GFI | .99 | > .90 = good |
| AGFI | .97 | > .90 = good |
| RMR | .04 | < .05 = acceptable |

The standardised factor loadings confirmed a one-factor structure, with values ranging from .84 to 1.00, all exceeding the .50 threshold typically considered acceptable (see Figure 3). Four items were removed during the model refinement process due to low factor loadings and high modification indices, which indicated limited contribution to the construct and potential disruption to model fit. The residual variances ranged from .18 to .48 and were all statistically significant ($p < .001$), indicating acceptable levels of unexplained variance. Model comparison indices also supported the adequacy of the final model, with the Akaike information criterion (36.99), Browne-Cudeck criterion (37.92) and expected cross-validation index (.27) all lower than those of the independence model, suggesting better fit and generalisability. Hoelter's critical N values were found to be 89 at the .05 level and 121 at the .01 level, indicating a sufficient sample size for stable model estimation. Finally, internal consistency was excellent, with a Cronbach's alpha of .92, confirming the reliability of the five-item scale within the range of .89–.91, as indicated in Table 6.

Table 6
*Summary of model estimates and reliability statistics*

| Statistic | Value | Interpretation / Criterion |
|---|---|---|
| Standardised factor loadings | 1.00 – .84 | > .50, acceptable |
| Residual variances | .18 – .48 | $p < .001$, significant |
| Number of items removed | 4 | Low loadings |
| Model Akaike information criterion / Browne-Cudeck criterion / expected cross-validation index | 36.99 / 37.92 / .27 | Better fit |
| Hoelter's critical N (.05 / .01) | 89 / 121 | ≥ 50, sufficient sample size |
| Cronbach's alpha | .92 | > .90, excellent internal consistency |

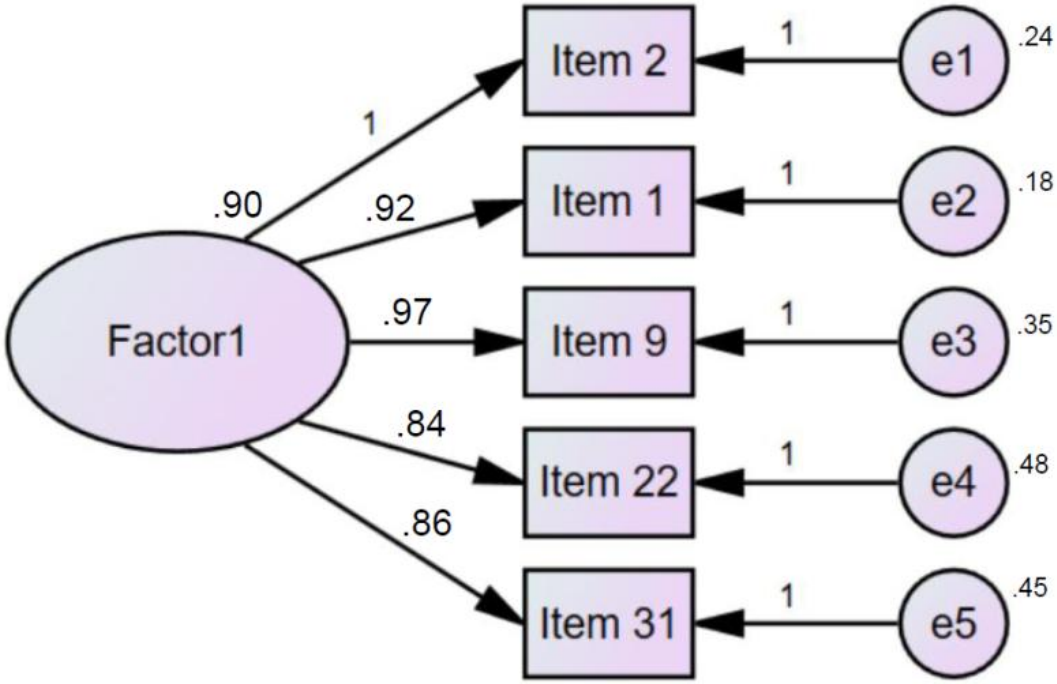


*Figure 3.* The CFA diagram

### Criterion validity and test-retest reliability

Hypothesising that the AI-EARS would be associated with the Academic Integrity Scale – Short Version, developed by Rettinger and Kramer (2009), a Pearson correlation analysis was conducted to evaluate concurrent validity. Rettinger and Kramer's scale was selected because academic integrity encompasses ethical values such as honesty, responsibility and accountability, which align with the cognitive and behavioural components of ethical awareness in research contexts. Results showed a statistically significant, moderate, positive correlation between the target scale and the criterion measure, $r(68) = .339$, $p = .03$, indicating that although the two scales share common ethical foundations, the AI-EARS measures a related yet more specific construct. This outcome supported concurrent validity and suggested that AI ethical awareness encompassed additional dimensions not fully captured by general academic integrity measures.

After a 15-day interval, the test-retest reliability of the scale was checked. For this purpose, several statistical analyses were performed. First, a Pearson correlation was computed between the scores obtained at two different time intervals. The results indicated a moderate, but significantly positive, correlation between the two scores ($r = .41$, $p < .001$), suggesting acceptable scale stability over time. The scale's internal consistency was evaluated using Cronbach's alpha, which yielded a coefficient of .72, indicating acceptable reliability. Descriptive analysis of the scale scores showed a total mean of 23.62 (SD=2.23), with a 95% CI ranging from .61 to .79, indicating acceptable reliability over time. The single-measures ICC was .33, suggesting moderate agreement among individual measurements. The ICC values were statistically significant ($F = 3.460$, $p < .001$), supporting the scale's temporal stability. Lastly, the paired-samples $t$ test revealed no significant difference between the mean scores of the two administrations ($M = 0.08$, $SD = 0.45$), $t(1) = 1.79$, $p = .08$. This result suggested that participants' scores remained stable across the two testing occasions, as summarised in Table 7.

Table 7
*Test-retest evidence*

| Analysis method | Result | Interpretation |
|---|---|---|
| Pearson correlation | r = .41, p < .001 | Moderate, statistically significant correlation |
| ICC[b] (Single measures) | .33, CI[a] (.79, .91) | Moderate agreement for individual scores |
| Cronbach's alpha | .72 | Acceptable internal consistency |
| ICC (average) | .72, CI (.61, .79) | Acceptable reliability over time |
| Paired *t* test | p = .08 | No significant difference between the two test scores |

[a]Confidence interval
[b] Intraclass correlation coefficient

## Conclusion and discussion

Based on the results of the current study, which aimed to develop a valid and reliable scale to measure ethical awareness and responsibility in the use of AI in language research, several conclusions are drawn. First, the scale (see appendix) demonstrates strong psychometric properties in EFA. The final version, consisting of nine items, is supported by adequate sampling procedures and the assumption of normality. The factor analysis reveals a two-factor structure, showing high internal consistency and conceptual clarity. Second, CFA supports a refined one-factor structure consisting of five items. In other words, the privacy dimension that appeared in EFA is excluded because it does not demonstrate adequate stability and model fit in CFA, resulting in a final one-factor structure that focuses on ethical awareness and responsibility. The model fit indices meet the recommended thresholds, and the standardised factor loadings are statistically acceptable and theoretically coherent. It should be pointed out that although the final items are intentionally phrased in general terms, the scale's discipline-specific nature is established through its theoretical grounding, target population and validation context. In other words, the instrument, developed specifically for language researchers, reflects ethical awareness as it arises in AI-assisted language research practices. General item wording was also preferred to ensure applicability across diverse research designs within the field while maintaining strong psychometric properties. Third, the scale shows concurrent validity through a significant correlation with an established measure of academic integrity, supporting its relevance in the broader context of ethical research practices. Fourth and last, the test-retest reliability analysis confirms the temporal stability of the scale. The results also indicate that the final five items collectively represent a coherent construct of ethical awareness and responsibility in AI-mediated language research. These items indicate researchers' perceived obligations to critically examine AI-generated content, verify the accuracy and transparency of AI outputs, acknowledge the role of AI in research reporting, and remain attentive to potential ethical risks such as bias, misinformation and inappropriate reliance on automated systems. The items also reflect a form of ethical competence that is increasingly necessary in both research and educational contexts, where AI tools actively shape data analysis, assessment practices and pedagogical decision-making. The scale's concise nature demonstrates that ethical awareness in AI use can be captured by core behaviours and dispositions rather than broad or diffuse constructs. In light of these findings, it can be concluded that the scale developed in this study is a valid, reliable and psychometrically sound instrument for assessing ethical awareness and responsibility in AI use among language researchers. The final five-item structure effectively captures core ethical behaviours, specifically verification and accountability. Despite its brevity, the scale demonstrates excellent internal consistency ($\alpha$ = .92). This unidimensional design ensures psychometric stability and practical efficiency, providing an ideal diagnostic tool for assessing ethical AI responsibility.

The study is significant for several reasons. First, the AI-EARS is one of the first instruments developed specifically to measure ethical awareness and responsibility in the use of AI within the context of language research. Unlike existing tools that broadly assess digital ethics or general academic integrity, the AI-EARS focuses on the challenges and ethical concerns arising from integrating AI technologies in research settings. Second, developing and validating the scale in English is particularly significant, as it enhances its applicability in diverse international research settings. Given that the scale specifically measures

awareness and responsibility, the findings also suggest that institutions can strengthen researchers' ethical practices by providing training in critically evaluating AI-generated content, verifying its accuracy, documenting AI use transparently and taking responsibility for ethical decision-making when employing AI tools. In addition, given that the scale specifically measures ethical awareness and responsibility, the findings indicate that institutions can improve educational practice by offering targeted training that helps researchers and educators critically evaluate AI-generated content, verify its accuracy, document AI use transparently and take responsibility for ethical decision-making when integrating AI into language teaching, learning and assessment processes. In conclusion, the study contributes to the related literature by promoting ethical awareness, guiding responsible AI integration and supporting future research.

This study has certain limitations. First, the statistical analyses are confined to EFA, CFA, concurrent validity and test-retest reliability. Within this scope, criterion validity is examined using a single external instrument. Second, in accordance with the Standards for Psychological and Educational Testing (American Educational Research Association et al., 2014), the study primarily focuses on internal structure, without addressing other sources of validity evidence. Third, although the initial factor solution includes two dimensions, the confirmatory procedures validate a one-factor structure. While these are retained on theoretical grounds and for statistical acceptability, they may indicate a need to refine the factor structure further. Fourth, the participant groups consisted primarily of language researchers, which may limit applicability to broader educational contexts and to other populations, such as teachers, teacher candidates, or students who also use AI tools in instructional settings. Fifth, although the final five-item structure demonstrates strong psychometric properties, the concise format may not capture all dimensions of ethical AI use that might emerge in different instructional, institutional or cross-cultural contexts. Sixth, the study relies solely on self-reported perceptions, which may be influenced by social desirability or participants' prior experiences with AI technologies. Seventh, it should be noted that although model fit indices are within acceptable to excellent ranges, the findings reflect the characteristics of the current sample groups. As a final limitation, while the five-item scale demonstrates strong psychometric properties, it captures a narrow aspect of AI ethics: ethical awareness and responsibility in AI use. In other words, in the final model, other core ethical dimensions, such as research ethics and data privacy, remain outside the scope of the instrument developed. Thus, further research may benefit from developing complementary or modular scales that address these dimensions in greater depth, thereby enabling a more comprehensive assessment of AI ethics in language research using the CFA version presented in Table 4.

Several recommendations for further research can be made. First, further research should examine the validity and reliability of the AI-EARS across multiple institutional and cultural settings to enhance its cross-contextual applicability. Second, researchers should test alternative factor structures and more complex models, including hierarchical or bifactor models, to explore the multidimensional nature of ethical awareness and observe the robustness of the factor solution obtained in this study. Third, further investigation is needed to determine whether the scale functions equivalently across certain variables such as academic rank, gender or years of research experience. Finally, longitudinal research could explore how researchers' ethical awareness evolves, particularly in response to advances in AI tools and emerging ethical guidelines in academic practice.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing (4th ed.)*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Ahmadi, A. (2012). Cheating on exams in the Iranian EFL context. *Journal of Academic Ethics*, *10*(2), 151–170. https://doi.org/10.1007/s10805-012-9156-5

Ahmadi, A. (2014). Plagiarism in the academic context: A study of Iranian EFL learners. *Research Ethics*, *10*(3), 151–168. https://doi.org/10.1177/1747016113488859

Aydın, S. (2024). Using ChatGPT in foreign language research: An overview. *Innovational Research in ELT*, *5*(1), 20–26. https://doi.org/10.29329/irelt.2024.1045.2

Aydın, S., Tekin, I., & Akkaş, F. (2024). Construction and validation of the foreign language learning enjoyment scale. *Psychology in the Schools*, *61*(2), 657–670. https://doi.org/10.1002/pits.23076

Beck, C., & Gable, R. (2001). Ensuring content validity: An illustration of the process. *Journal of Nursing Measurement*, *9*(2), 201–215. https://doi.org/10.1891/1061-3749.9.2.201

Bernstein, D., Nash, P., Clarke-Stewart, A., Penner, L., & Roy, E. (2008). *Essentials of psychology* (4th ed.). Houghton Mifflin Company.

Boateng, G. O., Neilands, T. B, Frongillo, E., Frongillo, E. A., Melgar-Quiñonez, H., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*. https://doi.org/10.3389/fpubh.2018.00149

Bruner, J. (1957). On perceptual readiness. *Psychological Review*, *64*(2), 123–152.

Byram, M. (2020). *Teaching and assessing intercultural communicative competence*. Multilingual Matters. https://doi.org/10.21832/BYRAM0244

Canagarajah, S. (2012). *Translingual practice: Global Englishes and cosmopolitan relations*. Routledge. https://doi.org/10.4324/9780203073889

Carobene, A., Padoan, A., Cabitza, F., Banfi, G., & Plebani, M. (2024). Rising adoption of artificial intelligence in scientific publishing: Evaluating the role, risks, and ethical implications in paper drafting and review process. *Clinical Chemistry and Laboratory Medicine*, *62*(5), 835–843. https://doi.org/10.1515/cclm-2023-1136

Comstock, G. (2012). *Research ethics: A philosophical guide to the responsible conduct of research*. Cambridge University Press. https://doi.org/10.1017/CBO9780511902703

De Costa, P. (2015). Ethics and applied linguistics research. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 190–199). Bloomsbury Academic.

Deardorff, D. (2020). *Manual for developing intercultural competencies: Story circles*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000370336

Demir, B., & Aydin, S. (2026). The ethics of using artificial intelligence in language research. In K. Carlson (Ed.), *Ethical, legal, and pedagogical perspectives on AI in education* (pp. 429–452). IGI Global. https://doi.org/10.4018/979-8-3373-3000-6.ch016

Démuth, A. (2013). *Perception theories*. University of Trnava. https://ff.truni.sk/sites/default/files/publikacie/demuth_perception_theories_1.1.pdf

Denkci Akkaş, F., Tekin, I., & Aydın, S. (2022). Does developing research skills increase academic motivation among foreign language learners? *The Literacy Trek*, *8*(2), 142–164. https://doi.org/10.47216/literacytrek.1124192

DeVellis, R. (2017). *Scale development: Theory and applications*. Sage Publications Ltd.

Dörnyei, Z. (2020). *Innovations and challenges in language learning motivation*. Routledge. https://doi.org/10.4324/9780429485893

Dowling, M., & Lucey, B. (2023). ChatGPT for (Finance) research: The Bananarama Conjecture. *Finance Research Letters*, *53*, Article 103662. https://doi.org/10.1016/j.frl.2023.103662

Downing, S. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Elmas, E., & Aydin, S. (2017). Pre-service foreign language teachers' perceptions of research skills: A qualitative study. *The Qualitative Report*, *22*(12), 3088–3101. https://doi.org/10.46743/2160-3715/2017.3194

Floridi, L., Cowls, J., King, T. C., Taddeo, M., Floridi, L. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics, 26*, 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Gilson, A., Safranek, C., Huang, T., Socrates, V., Chi, L., Taylor, R., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, *9*. https://doi.org/10.2196/45312

Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): A brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1191628

Gregory, R. (2015). *Eye and brain: The psychology of seeing*. Princeton University Press.
Grimaldi, G., & Ehrler, B. (2023). AI et al. : Machines are about to change scientific publishing forever. *ACS Energy Letters*, *8*(1), 878–880. https://doi.org/10.1021/acsenergylett.2c02828
Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (1998). *Multivariate data analysis*. Pearson.
Haneef, I., & Agrawal, M. (2024). Ethical issues in educational research. *Asian Research Journal of Arts & Social Sciences*, *22*(5), 29–38. https://doi.org/10.9734/arjass/2024/v22i5535
Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promise and implications for teaching and learning*. Center for Curriculum Redesign.
Hosseinnia, M., & Kafi, Z. (2024). Constructing and validating a code of ethics in testing inventory: Investigating EFL instructors' perspectives. *Language Testing in Asia*, *14*, Article 56. https://doi.org/10.1186/s40468-024-00331-y
Hultgren, A. K., Erling, E. J., & Chowdhury, Q. (2016). Ethics in language and identity research. In S. Preece (Ed.), *The Routledge handbook of language and identity* (pp. 15–30). Routledge. https://doi.org/10.1080/00437956.2017.1347315
Işık, S., Çakır, R., & Korkmaz, Ö. (2024). Teachers' Perception Scale towards the use of artificial intelligence tools in education. *Participatory Educational Research*, *11*, 80–94. https://doi.org/10.17275/per.24.95.11.6
Jang, Y., Choi, S., & Kim, H. (2022). Development and validation of an instrument to measure undergraduate students' attitudes toward the ethics of artificial intelligence (AT-EAI) and analysis of its difference by gender and experience of AI education. *Education and Information Technologies*, *27*(8), 11635–11667. https://doi.org/10.1007/s10639-022-11086-5
John, O. T, & Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H. Reis & C. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 339–369). Cambridge University Press.
Kenyon, D., & MacGregor, D. (2013). Pre-operational testing. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 309–320). Routledge. https://doi.org/10.4324/9781003220756
Kern, R. (2006). Perspectives on technology in learning and teaching languages. *TESOL Quarterly*, *40*(1), 183–120. https://doi.org/10.2307/40264516
Khalaf, M. (2025). Does attitude towards plagiarism predict aigiarism using ChatGPT? *AI and Ethics*, *5*(1), 677–688. https://doi.org/10.1007/s43681-024-00426-5
Khlaif, Z. (2023). Ethical concerns about using AI-generated text in scientific research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4387984
Kim, H. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative Dentistry & Endodontics*, *38*(1), 52–54. https://doi.org/10.5395/rde.2013.38.1.52
Korteling, J., van de Boer-Visschedijk, G., Blankendaal, R., Boonekamp, R., & Eikelboom, A. (2021). Human- versus artificial intelligence. *Frontiers in Artificial Intelligence*, *4*, Article 622364. https://doi.org/10.3389/frai.2021.622364
Lin, Z. (2023). Why and how to embrace AI such as ChatGPT in your academic life. *Royal Society Open Science, 10*(8), Article 230658. https://doi.org/10.1098/rsos.230658
Marcoulides, K., & Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable modeling methods. *Educational and Psychological Measurement*, *79*(5), 874–882. https://doi.org/10.1177/0013164418817803
Nunan, D. (1992). *Research methods in language learning*. Cambridge University Press.
Pelzang, R., & Hutchinson, A. (2018). Establishing cultural integrity in qualitative research. *International Journal of Qualitative Methods*, *17*(1). https://doi.org/10.1177/1609406917749702
Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). The Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment. *Journal of University Teaching and Learning Practice*, *21*(6), 49–66. https://doi.org/10.53761/q3azde36
Petrova, E., Dewing, J., & Camilleri, M. (2016). Confidentiality in participatory research. *Nursing Ethics*, *23*(4), 442–454. https://doi.org/10.1177/0969733014564909

Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology*, *14*, Article 1255594. https://doi.org/10.3389/fpsyg.2023.1255594

Raubenheimer, J. (2004). An item selection procedure to maximise scale reliability and validity. *SA Journal of Industrial Psychology*, *30*(4), Article a168. https://doi.org/10.4102/sajip.v30i4.168

Rettinger, D., & Kramer, Y. (2009). Situational and personal causes of student cheating. *Research in Higher Education*, *50*(3), 293–313. https://doi.org/10.1007/s11162-008-9116-5

Saatci, E. (2025). AI and ethics: Scale development for measuring ethical perceptions of artificial intelligence across sectors and countries. *International Journal of Economic Behavior and Organization*, *13*(1), 35–50. https://doi.org/10.11648/j.ijebo.20251301.14

Sarker, I. (2022). AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, *3*, Article 158. https://doi.org/10.1007/s42979-022-01043-x

Schacter, D., Gilbert, D., & Wegner, D. (2009). *Introducing psychology*. Macmillan.

Schembri, N., & Jahić Jašić, A. (2022). Ethical issues in multilingual research situations: A focus on interview-based research. *Research Ethics*, *18*(3), 210–225. https://doi.org/10.1177/17470161221085857

Seliger, H. W., & Shohamy, E. (1989). *Second language research methods*. Oxford University Press.

Tabachnick, B., & Fidell, L. (2014). *Using multivariate statistics*. Pearson.

Wang, Z., Chai, C., Li, J., & Lee, V. (2025). Assessment of AI ethical reflection: The development and validation of the AI ethical reflection scale (AIERS) for university students. *International Journal of Educational Technology in Higher Education*, *22*, Article 19. https://doi.org/10.1186/s41239-025-00519-z

Watkins, M. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

Weinbaum, C., Landree, E., Blumenthal, M., Piquado, T., & Gutierrez, C. (2019). *Ethics in scientific research*. RAND Corporation. https://doi.org/10.7249/RR2912

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, *4*(2), Article 100211. https://doi.org/10.1016/j.hcc.2024.100211

Yurt, E. (2025). The Self-Regulation for AI-Based Learning Scale: Psychometric properties and validation. *International Journal of Current Educational Studies*, *4*(1), 95–118. https://doi.org/10.46328/ijces.176

Yurt, E., & Kasarci, I. (2024). A questionnaire of artificial intelligence use motives: A contribution to investigating the connection between AI and motivation. *International Journal of Technology in Education*, *7*(2), 308–325. https://doi.org/10.46328/ijte.725

Żammit, J. (2024). Capturing the full potential of Maltese language learning through ChatGPT. *Technology in Language Teaching & Learning*, *6*(1), 1–22. https://doi.org/10.29140/tltl.v6n1.1082

Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media + Society*, *4*(2). https://doi.org/10.1177/2056305118768300

---

**Corresponding author:** Bora Demir, boradem ir@comu.edu.tr

**Copyright**: Articles published in the *Australasian Journal of Educational Technology* (AJET) are available under Creative Commons Attribution Non-Commercial No Derivatives Licence (CC BY-NC-ND 4.0). Authors retain copyright in their work and grant AJET right of first publication under CC BY-NC-ND 4.0.

**Please cite as**: Demir, B., & Aydın, S. (2026). Construction and validation of the AI Ethics Scale in language research. *Australasian Journal of Educational Technology*, *42*(2), 18–38. https://doi.org/10.14742/ajet.11034

## Appendix

*AI Ethical Awareness and Responsibility Scale (AI-EARS)*

| Items | *Strongly disagree* (1) | *Disagree* (2) | *Neutral* (3) | *Agree* (4) | *Strongly agree* (5) |
|---|---|---|---|---|---|
| 2. I think it is important to follow ethical rules when using AI in my studies. | | | | | |
| 1. I know that using AI tools in language research comes with ethical responsibilities. | | | | | |
| 22. I try to follow ethical rules whenever I include AI in my research process. | | | | | |
| 9. I feel researchers should always check AI-generated content carefully before using it. | | | | | |
| 31. I believe language researchers are responsible for how they use AI and the content it creates. | | | | | |