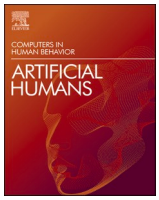
Computers in Human Behavior: Artificial Humans 5 (2025) 100176

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Development and validation of a short AI literacy test (AILIT-S) for university students

Marie Hornberger [a,*], Arne Bewersdorff [a], Daniel S. Schiff [b], Claudia Nerdel [a]

[a] Technical University of Munich, Germany
[b] Purdue University, USA

ARTICLE INFO

Keywords:
AI literacy
AI education
Higher education
Item response theory
Artificial intelligence

ABSTRACT

Fostering AI literacy is an important goal in higher education in many disciplines. Assessing AI literacy can inform researchers and educators on current AI literacy levels and provide insights into the effectiveness of learning and teaching in the field of AI. It can also inform decision-makers and policymakers about the successes and gaps with respect to AI literacy within certain institutions, populations, or countries, for example. However, most of the available AI literacy tests are quite long and time-consuming. A short test of AI literacy would instead enable efficient measurement and facilitate better research and understanding. In this study, we develop and validate a short version of an existing validated AI literacy test. Based on a sample of 1,465 university students across three Western countries (Germany, UK, US), we select a subset of items according to content validity, coverage of different difficulty levels, and ability to discriminate between participants. The resulting short version, AILIT-S, consists of 10 items and can be used to assess AI literacy in under 5 minutes. While the shortened test is less reliable than the long version, it maintains high construct validity and has high congruent validity. We offer recommendations for researchers and practitioners on when to use the long or short version.

1. Introduction

Even preceding the popularization of generative AI in the early 2020s, it has been generally accepted that basic AI competencies (known as *AI literacy*, Long & Magerko, 2020) are among the most important competencies for university students of many disciplines (Laupichler et al., 2022; Ng et al., 2021; Southworth et al., 2023), as these individuals constitute future workers, citizens, and designers of AI systems. However, research on how to foster these competencies effectively is still in its infancy. To evaluate different methods of learning and instruction, and progress made toward AI literacy policy goals expressed by governments and corporations around the world, effective assessment of AI literacy is crucial (Hornberger et al., 2023; Hornberger et al., 2025).

Several scales potentially appropriate for assessing AI literacy have been developed recently (Lintner, 2024). However, most of them are self-assessment scales rather than objective performance-based measures, and most of them are quite lengthy (Lintner, 2024). While thorough assessment instruments offer strong reliability and validity for high-stakes contexts, such as determining student outcomes or policy decisions, they are often too lengthy to be used in standard assessments or research studies that must be attentive to participant time or focus on additional concepts beyond AI literacy alone.

To reduce test-taking time and promote utilization, researchers thus typically develop short versions out of longer, more comprehensive tests (e.g., Botes et al., 2021; Schipolowski et al., 2014). This is a pragmatic way of reducing test-taking time, yet economic, objective, and validated tests for AI literacy have not been developed (Lintner, 2024). Tests with only a few items pose fewer time constraints on participants, consume fewer resources, and might increase participation rates (Schroeders et al., 2016). This is especially important in research settings when AI literacy is one variable of many to be measured, e.g., in large-scale assessments that might measure other student competencies, psychological traits, or various attitudes or behaviors (Schroeders et al., 2016; Ziegler et al., 2014). Plus, in teaching settings, it is beneficial to have access to a short test for quick evaluations, including pre-post studies of student learning. As AI literacy is gaining importance, a time-efficient literacy instrument can be a game changer both for research and practice. Therefore, the aim of this study is to develop a short AI literacy test that can measure AI literacy in an efficient way.

* Corresponding author. Technical University of Munich, Arcisstrasse 31, 80333, Munich, Germany.
E-mail address: marie.hornberger@tum.de (M. Hornberger).

https://doi.org/10.1016/j.chbah.2025.100176
Received 22 April 2025; Accepted 10 June 2025
Available online 16 June 2025

2. Theoretical framework

2.1. Current advancements in measuring AI literacy

AI literacy is commonly defined as "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" (Long & Magerko, 2020). With AI receiving more attention in the last years, since major AI advancements in the 2010s and the popularization of generative AI in the 2020s, there has been increased research on AI literacy, including its development and its effective, reliable, and valid assessment (for a review see Lintner, 2024). Most existing measures are self-assessment scales (Carolus et al., 2023; Laupichler et al., 2023; Lin et al., 2023; Ng et al., 2023; Pinski & Benlian, 2023; Wang et al., 2022), including short scales (e.g., Koch et al., 2024).

As self-assessment scales can be biased (Dunning et al., 2004), it is recommended to use objective performance-based assessments (Claro et al., 2024). This is especially important for a complex and evolving topic like AI, where relevant definitions, concepts, and public familiarity are changing rapidly. A reliance on self-reported measurements, the most commonly used form of knowledge assessment and research in the field (Dreksler et al., 2025), is also problematic given the low correlation between self-assessments and objective knowledge tests (Sitarenios, 2022). To our knowledge, all existing performance-based assessments are multiple-choice tests consisting of 20 items or more (Hornberger et al., 2023; Soto-Sanfiel et al., 2024; Zhang et al., 2024). This makes it quite time-consuming to implement these tests. Test duration is a common concern in test development and a major barrier for researchers, who must undergo extensive efforts to recruit and often pay for samples. Long test durations consume more resources, reduce participation rates, and can have a negative effect on data quality (Botes et al., 2021), for example resulting from fatigue (Ackerman & Kanfer, 2009) and satisficing behavior (Barge & Gehlbach, 2012). There is further a growing need for more parsimonious tests generally, as research designs become more complex and multivariate (Rammstedt & Beierlein, 2014; Schroeders et al., 2016; Ziegler et al., 2014). This includes, for example, large-scale-assessments and longitudinal studies which might simultaneously assess multiple constructs.

2.2. The long version of the AI literacy test

This short version development is based on a longer AI literacy test that was first developed and validated by Hornberger et al. (2023), adopting an Item Response Theory (IRT) approach. A second, updated version of the test has been validated as part of an international study on students' AI literacy (Hornberger et al., 2025). The test was validated in both English and German versions, with responses from 1,465 students from Germany, the UK, and the US, three countries that are leaders in AI development, regulation, education, and adoption conversations. This long version (28 questions) of the test has been tested for country- and gender-specific differential item functioning and is additionally fair with regard to gender and across the tested countries and languages, meaning that it can be used to compare groups with different languages or genders fairly. Due to the number of items, the full-length test takes, on average, approximately 12 minutes to complete. This ensures good measurement accuracy but presents a hurdle for implementation when resources are constrained. Therefore, the goal of this study is to select a subset of items for a short test which can nevertheless assess AI literacy in a reliable and valid manner. We aimed to find a set of 10 items, as this number is a balance between the need to cover the breadth of the construct of AI literacy and the desire to be as short as possible for pragmatic reasons. This corresponds to under 5 min test-taking time.[1]

[1] Three-quarters of the sample required less than 30 s, on average, to complete each item in the long version.

2.3. Short scale development

There are several benefits of short scales: they pose fewer time constraints on participants, consume fewer resources (such as financial costs), and might increase participation rates, data quality, and improve participant retention and reduce dropout in longitudinal studies (Botes et al., 2021; Schroeders et al., 2016; Ziegler et al., 2014). This is especially important in research settings, when AI literacy is one variable besides many that may need to be measured (Koch et al., 2024). Plus, in teaching settings, it is beneficial to have a short scale for quick evaluations.

However, short scales have their own drawbacks. A main issue is a loss of reliability, which has been documented in various studies (Schipolowski et al., 2014; Ziegler et al., 2014). The reduction of items, while trying to retain the breadth of the construct, often leads to a decrease in internal consistency (Rammstedt & Beierlein, 2014; Ziegler et al., 2014). The goal of short scales is to be efficient, which means balancing measurement accuracy and required resources when attempting to realize the overall aims of the test (Ziegler et al., 2014). Considerations related to efficiency thus include the application context, for example, research or screening, and the target group for which the short version is to be developed. In the present study, the objective is to develop a short version for assessing AI literacy among students. The aim is to create a test that allows for efficient estimation of AI literacy at the group level (e.g., enabling statistically sound comparisons of group means). Consequently, some reduction in reliability is accepted in exchange for shorter administration time, thereby enabling rapid assessment of literacy levels.

3. Method

3.1. Sample

This study used a sample already described by Hornberger et al. (2025) and used in Bewersdorff et al. (2025). The sample consists of 1,465 university students from three Western countries (US: $N = 494$, UK: $N = 499$, and Germany: $N = 472$). The mean age of the participants is $M = 28.4$ ($SD = 10.3$). Among the participants, 747 students (51.0 %) were male, 680 (46.4 %) were female, 33 (2.3 %) were non-binary, and 5 (0.3 %) preferred not to disclose their gender. Regarding educational level, 1010 students were pursuing a bachelor's or similar degree (68.9 %), 393 (26.8 %) were enrolled in a master's or equivalent program, and 61 (4.2 %) were participating in other programs, such as graduate or Ph.D. studies. The distribution of disciplines can be viewed in Fig. 1.

3.2. The long version AI literacy test

The long version of the AI literacy test was first developed by Hornberger et al. (2023) based on the AI literacy framework by Long and Magerko (2020). After the release of ChatGPT in late 2022, the test was updated and validated in both a German and English version (Hornberger et al., 2025). The long version consists of 28 items, 27 of which are multiple choice items (one out of four answers is correct), while one is a sorting task. All items are based on the AI literacy competencies proposed by Long and Magerko (2020), which are organized around five overarching themes derived from a literature review. Table 1 provides an overview of the items and their distribution across these themes and associated competencies.

3.3. Analysis

3.3.1. Short test development strategy

The recommended approach to creating a short version based on a longer version is to select items according to two different types of criteria. First, content validity must be preserved, ensuring that the short version adequately represents the full breadth of the intended construct

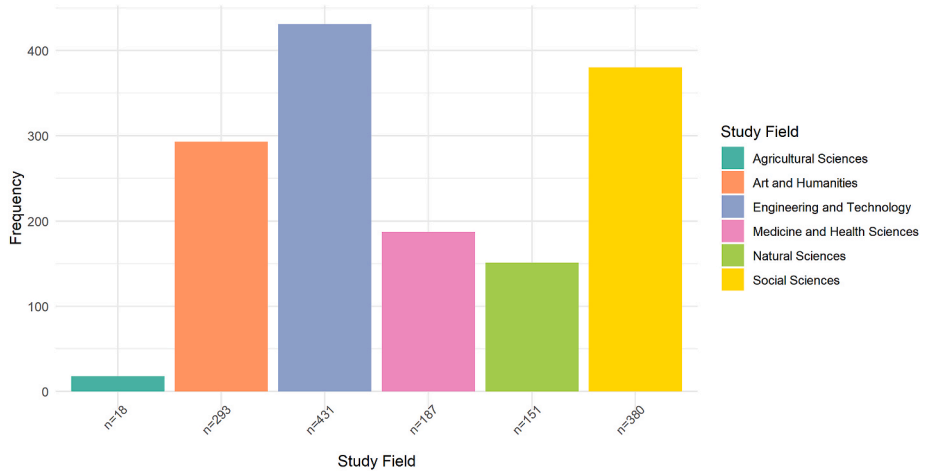


Fig. 1. Proportion of students from different disciplines.

Table 1
AI literacy item overview according to Long and Magerko's (2020) themes and competencies.

Theme	Competency	Item labels (no.)
What is AI?	Recognizing AI	Typical applications (01), Recognizing a chatbot (02)
	Interdisciplinarity	AI systems (03), Interdisciplinary research fields (04)
	Understanding Intelligence	Intelligence of AI (05), Intelligence of AI 2 (06)
	General vs. Narrow	Weak and strong AI (07), Capabilities of weak AI (08)
What can AI do?	AI's Strengths & Weaknesses	Superiority of AI (09), Superiority of humans (10)
How does AI work?	Representations	Knowledge representations 1 (11),
	Decision-Making (of AI)	Decision-making (13), Optimization (14), Supervised and unsupervised learning (15),
	ML Steps	Iterative process (16), Steps in supervised learning (17)[a], Training and test data (18)
	Human Role in AI	Human influence (19), Human influence 2 (20)
	Data Literacy	Visualization of data (22)
	Learning from Data	Learning from data (23), Learning form user data (24)
	Critically Interpreting Data	Representativeness of data (25)
How do people perceive AI?	Programmability	Programmability (21)
How should AI be used?	Ethics	Black box (27), Societal challenges (28), Risks of AI (29), Legal challenges (30)

[a] Sorting task.

(Schipolowski et al., 2014). Second, statistical criteria, such as item difficulty and item discrimination, should guide the selection of well-performing items (Schipolowski et al., 2014). The item selection process should thus take into account both content validity and empirical measures of item quality. For the validation process, we concentrate on reliability, construct validity, and congruent validity in this study. Importantly, the selection of items and the validation of the short version need to be conducted with independent datasets to ensure the meaningfulness of the validation (Schipolowski et al., 2014). However, as new data collections are often not possible due to resource constraints, a pragmatic solution is to randomly split a given dataset into two independent samples, and use one sample for item selection and the other for validation (Schipolowski et al., 2014).

3.3.2. Data split

We randomly split our dataset into two subsamples of equal size. One half was used for the item selection procedure ('Item Selection Sample'), and the other half served to validate the resulting short test ('Validation Sample', see Fig. 2).

3.3.3. Item selection

We selected 10 items according to three criteria. The first criterion is content validity: Our goal was to retain the coverage of the full breadth of the AI literacy framework by Long and Magerko (2020) because AI literacy is a broad construct which risks losing relevance if it is overly abbreviated. We decided to orient item selection at the level of themes instead of competencies because they form broader and fewer categories. Table 2 shows the number of items in the long version per theme and the number of items that we aimed for in the short version. The goal was to reduce the number of items to approximately one-third to one-fourth of the original set while ensuring that each theme was represented by at least one item. We decided to put an emphasis on the theme *How does AI work?* as the associated competencies related to this theme represent the core of the construct. Table 2 shows how many items per question should be chosen.

The next two criteria involved key item parameters. As in the validation of the long version (Hornberger et al., 2025), we fitted a 3-PL model with a fixed guessing parameter ($g = .25$) to the data of the item selection sample (one-quarter guessing rate given multiple choice questions). We retrieved parameter estimations for item difficulty and discrimination from this model. First, the variance of person ability in our sample (from -2 to 2) should be covered by items with different levels of *difficulty* to enhance measurement at different points of the latent ability dimension (Schipolowski et al., 2014). Second, items with the highest *discrimination* should be chosen. Fig. 3 shows the IRT estimations of difficulty and discrimination of all items that are included in the long version.

3.3.4. Validation

We evaluated reliability, construct validity, and congruent validity of the short version based on the validation sample. To assess reliability, we not only examined Cronbach's alpha, but also utilized McDonalds' omega, as it is less affected by scale shortening (Schipolowski et al., 2014). Acceptable values are higher than .70 (Tavakol & Dennick, 2011). Concerning construct validity, we examined dimensionality, model fit, and item fit with the goal of replicating the results of the long version. We evaluated unidimensionality by fitting a one-factor model using the package *lavaan* (Rosseel, 2012) for *R* (R Core Team, 2022). We used the following fit indices and cutoffs: $X^2/df < 2$–3; RMSEA <0.05; SRMR <0.1 (Backhaus et al., 2015; Brown, 2015). We fitted a 3-PL

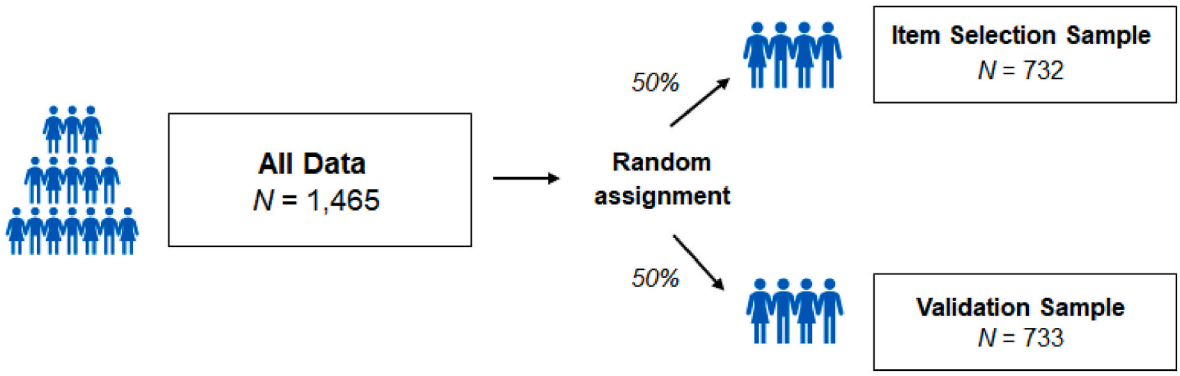


Fig. 2. Date split procedure.

Table 2
AI literacy item overview according to Long and Magerko's (2020) questions and competencies.

Themes from Long & Magerko	No. items long version	No. items short version	Selected items
What is AI?	8	2	(1, 4)
What can AI do?	2	1	(9)
How does AI work?	13	4	(15, 17, 19, 25)
How do people perceive AI?	1	1	(21)
How should AI be used?	4	2	(27, 29)

model with a fixed guessing parameter ($g = .25$) to the short version based on the validation sample. We determined the model fit using the M_2 statistic according to the following criteria: RMSEA, SRMR≤.05; TLI, CFI≥.95 (Maydeu-Olivares, 2013). For item fit, we examined the signed chi-squared ($S - X^2$) index (Orlando & Thissen, 2003).

Regarding congruent validity, we aimed to test if the short version correlates highly with the long version. Importantly, this correlation is calculated based on the validation sample, which was not involved in the item selection process. A high correlation (>.80) with the long version would indicate high congruent validity.

4. Results

4.1. Item selection

We fitted a 3-PL model with a fixed guessing parameter ($g = .25$) to the data of the item selection sample. The parameter estimations can be viewed in the appendix. We chose the items according to the following criteria, in this order: content validity, item difficulty, and item discrimination. Table 2 shows which items have been selected for each question. Fig. 4 shows the difficulty and discrimination of items that have been selected vs. not selected. As shown in the figure, the chosen items cover the breadth of person ability (from -2 to 2). Furthermore, we chose those items with highest discrimination, when possible. The full text version of these items can be viewed in the appendix.

4.2. Validation of the short version

To validate the short version, we again fitted a 3-PL model with a fixed guessing parameter ($g = .25$) on both the short and long collection of items using the validation sample.

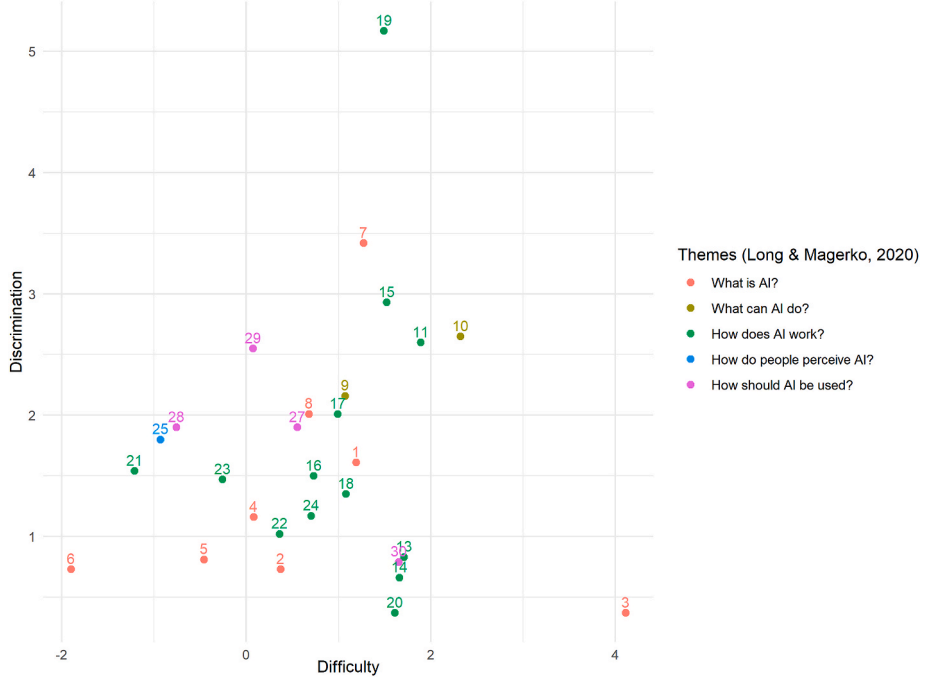


Fig. 3. Difficulty and discrimination parameters of long version items.

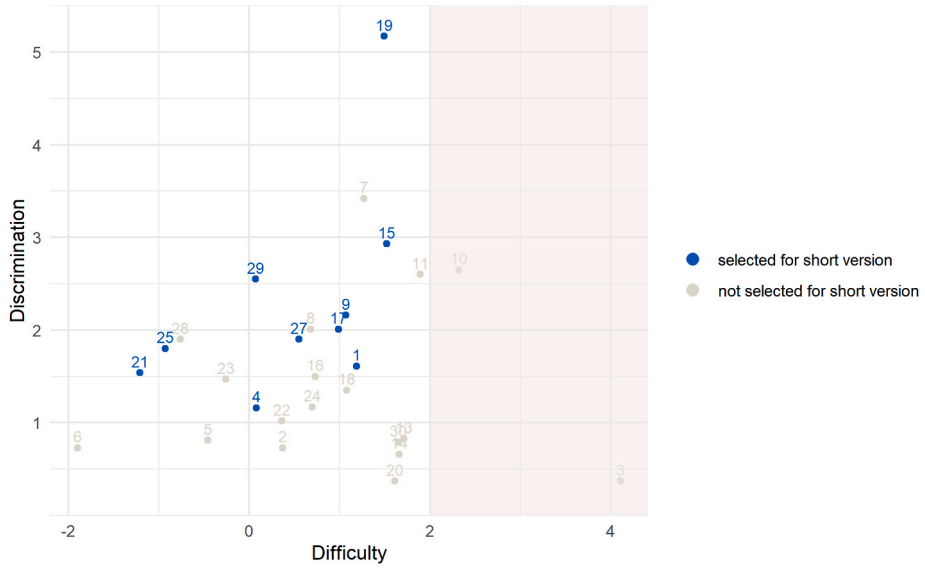


Fig. 4. Difficulty and discrimination parameters of items selected vs. not selected for short version.

4.2.1. Reliability

First, we assessed reliability using different measures from classical test theory (Cronbach's α and McDonald's ω) and IRT (EAP reliability). As shown in Table 3, all measures are smaller in the short version than in the long version, which points towards some loss in measurement accuracy.

4.2.2. Construct validity

Next, we verified the assumption of unidimensionality by performing a confirmatory factor analysis with a one-factor model. All indices indicated a good model fit: $X^2/df = 1.59$, RMSEA = .028, SRMR = .032. Therefore, the assumption of unidimensionality is fulfilled and an unidimensional IRT model can be used. Next, we evaluate the model and item fit of the 3-PL model of the short version. Table 4 shows that all cutoff values for the model fit have been reached. No index for item fit was significant (see appendix). This supports the construct validity of the test.

4.2.3. Congruent validity

To evaluate congruent validity, we calculated the Pearson correlation between the AI literacy scores based on the short vs. long version. The correlation is $r = .91$, indicating high congruent validity. Fig. 5 shows the person parameter scores based on the short version as a function of the long version.

5. Discussion

5.1. Discussion of the short version development

The aim of this study was to develop and validate a short test that can be used to measure students' AI literacy in an efficient way. With AILIT-S we created a 10-item version out of a validated 28-item long version (Hornberger et al., 2025) based on both content validity (covering the breadth of the AI literacy framework by Long & Magerko, 2020) and statistical criteria (item difficulty and discrimination). In our validation we find that AILIT-S shows lower reliability than the long version.

Table 3
Reliability measures for the short version vs. long version.

	Cronbach's α	McDonald's ω	EAP reliability
Long version	.74	.75	.78
Short version	.61	.64	.62

Table 4
Model fit indices.

	M_2	RMSEA	SRMSR	TLI	CFI
Long version	505	.025	.042	.945	.949
Short version	46	.021	.040	.977	.982

Note. RMSEA = root mean square error of approximation; SRMSR = standardized root mean squared residuals, TLI = Tucker–Lewis index; CFI = comparative fit index.

Nevertheless, it demonstrates good model and item fit indices, and, like the long version, a unidimensional structure could be confirmed. Further, AILIT-S displays high congruent validity, indicated by the high correlation with the long version.

The short version developed in this study is affected by decreasing internal consistency due to scale-shortening, which is a known issue in short scale development (Rammstedt & Beierlein, 2014; Schipolowski et al., 2014). As the aim of scale-shortening is to reduce the number of items while maintaining the breadth of the construct, this typically leads to a more heterogeneous set of items, resulting in less internal consistency (Rammstedt & Beierlein, 2014; Ziegler et al., 2014). Therefore, the usefulness of AILIT-S for individual diagnostics is limited due to poor measurement accuracy (Ziegler et al., 2014). However, if it is used for group statistics, it is acceptable to prioritize measurement efficiency over internal consistency (Rammstedt & Beierlein, 2014; Ziegler et al., 2014). Moreover, as Sitarenios (2022) highlight, there is no firm consensus on what constitutes an acceptable level of internal consistency, and some scholars argue that reliabilities in the range of .60 - .70 may be sufficient for short versions (Clark & Watson, 1995). Concerning construct and congruent validity, our analyses yield very good results, implicating that AILIT-S can be used instead of the long version for valid assessments at the group level.

The brevity of the 10-item version allows easier integration in complex research settings (e.g., large scale, multivariate assessments) and in practice (e.g., evaluation of courses). Due to the reduced effort for participants, fewer resources are needed, and higher participation rates may be achievable (Schroeders et al., 2016). Therefore, this short AI literacy has the potential to facilitate research in the field and advance our understanding of AI literacy.

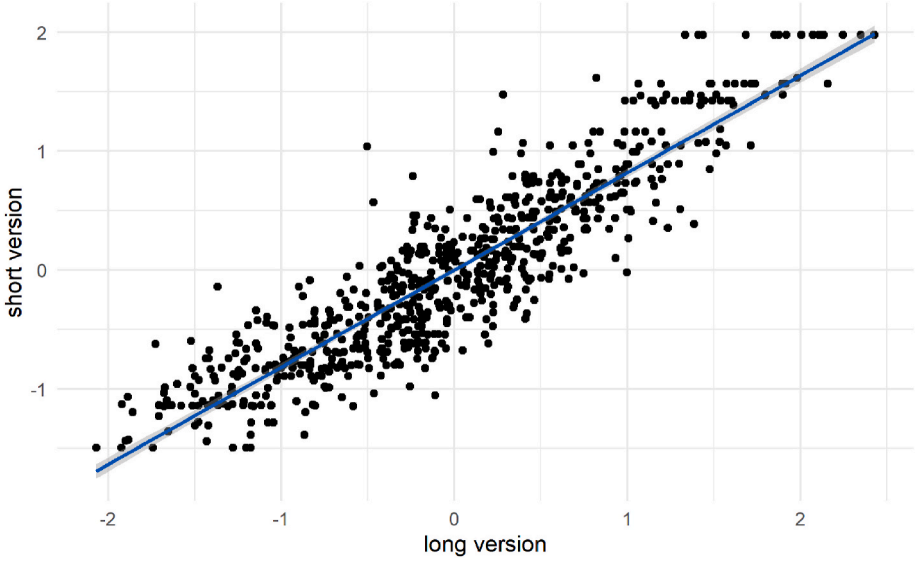


Fig. 5. Person ability estimations based on long version vs. short version.

5.2. Recommendations

The short version of the AI literacy test (AILIT-S) offers a time-efficient way to assess AI literacy in under 5 min, making it especially useful in research and educational settings where time and resources are limited. However, this comes with a trade-off in measurement accuracy compared to the long version. Therefore, we emphasize that, if possible, the long version should always be preferred, as it provides higher reliability and a more comprehensive assessment of AI literacy. The short version should only be considered when resources, time, or participant capacity are limited, and when a quick estimation of AI literacy at the group level is sufficient.

We recommend using AILIT-S only for group-level assessments, such as comparing AI literacy across different student groups, evaluating the impact of educational interventions, or conducting large-scale surveys where AI literacy is only one of several constructs being measured. Other application examples include course evaluations or exploratory research aiming to identify general trends and gaps in AI knowledge. For any contexts requiring high accuracy – such as individual diagnostics, formal certification, or selection processes – we strongly advise using the long version, as it meets higher psychometric standards, especially acceptable reliability (Tavakol & Dennick, 2011). Researchers and practitioners should thus carefully consider their measurement goals and available resources when choosing between the short and long versions.

5.3. Limitations and future research

This study is limited by several factors. First, the sample is not representative of the population of students in the respective countries. Second, the three countries are all Western countries. Third, we did not have an independent sample to validate the short version but relied on a random sample split to create independence between our item selection and validation procedures. Future research should address these limitations by using independent samples to examine the validity and reliability of the short version, and by providing additional evidence for the validity of the test (e.g., criterion validity).

Furthermore, future research could adapt and validate AILIT-S for other adult populations beyond university students. For instance, examining AI literacy amongst working professionals or the general adult population would provide insights into the level of AI literacy in key segments of society and inform different stakeholders like educators and policymakers. In addition, testing AILIT-S in non-Western contexts would be valuable for understanding its cross-cultural applicability and ensuring that it captures AI literacy across more diverse cultural settings.

6. Conclusion

The goal of this study was to develop and validate a short test of AI literacy for university students. The 10-item AILIT-S can be administered in under 5 minutes and demonstrates high validity. This test provides several benefits for research and practice. First, it saves resources and is easier to integrate into complex research designs. Second, it may increase participation rates and data quality. Third, it is easier to implement in education settings, as it needs less time from teaching time and fewer constraints on students and educators.

CRediT authorship contribution statement

Marie Hornberger: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft. **Arne Bewersdorff:** Conceptualization, Investigation, Project administration, Writing – review & editing. **Daniel S. Schiff:** Conceptualization, Funding acquisition, Investigation, Project administration, Writing – review & editing. **Claudia Nerdel:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (www.chat.openai.com) as well as Grammarly (www.grammarly.com) in order to improve the readability and language of single sentences. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding statement

This work was supported in part by funding from Google Research. The project on which this paper is based was partly funded by the German Federal Ministry of Education and Research under the funding code 16DHBKI051. The responsibility for the content of this publication lies with the authors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to express our gratitude to Sarah-Alina Günzer for her support with the data analysis for this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbah.2025.100176.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*(2), 163–181. https://doi.org/10.1037/a0015719

Backhaus, K., Erichson, B., & Weiber, R. (2015). *Fortgeschrittene Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*. Springer-Verlag.

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53*(2), 182–200. https://doi.org/10.1007/s11162-011-9251-2

Bewersdorff, A., Hornberger, M., Nerdel, C., & Schiff, D. S. (2025). AI advocates and cautious critics: How AI attitudes, AI interest, use of AI, and AI literacy build university students' AI self-efficacy. *Computers and Education: Artificial Intelligence, 8*, Article 100340. https://doi.org/10.1016/j.caeai.2024.100340

Botes, E., Dewaele, J.-M., & Greiff, S. (2021). The development and validation of the short form of the foreign language enjoyment scale. *The Modern Language Journal, 105*(4), 858–876. https://doi.org/10.1111/modl.12741

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford publications.

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). Mails - meta AI literacy scale: development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans, 1*(2), Article 100014. https://doi.org/10.1016/j.chbah.2023.100014

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Claro, M., Castro-Grau, C., Ochoa, J. M., Hinostroza, J. E., & Cabello, P. (2024). Systematic review of quantitative research on digital competences of in-service school teachers. *Computers & Education, 215*, Article 105030. https://doi.org/10.1016/j.compedu.2024.105030

Dreksler, N., Law, H., Ahn, C., Schiff, D., Schiff, K. J., & Peskowitz, Z. (2025). What does the public think about AI? An overview of the public's attitudes towards AI and a resource for future research. https://doi.org/10.2139/ssrn.5108572.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest: A Journal of the American Psychological Society, 5*(3), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Hornberger, M., Bewersdorff, A., & Nerdel, C. (2023). What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence, 5*, Article 100165. https://doi.org/10.1016/j.caeai.2023.100165

Hornberger, M., Bewersdorff, A., Schiff, D. S., & Nerdel, C. (2025). A multinational assessment of AI literacy among university students in Germany, the UK, and the US. *Computers in Human Behavior: Artificial Humans, 4*, Article 100132. https://doi.org/10.1016/j.chbah.2025.100132

Koch, M. J., Carolus, A., Wienrich, C., & Latoschik, M. E. (2024). Meta AI literacy scale: Further validation and development of a short version. *Heliyon, 10*(21), Article e39686. https://doi.org/10.1016/j.heliyon.2024.e39686

Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023). Development of the "Scale for the assessment of non-experts' AI literacy" – An exploratory factor analysis. *Computers in Human Behavior Reports, 12*, Article 100338. https://doi.org/10.1016/j.chbr.2023.100338

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence, 3*, Article 100101. https://doi.org/10.1016/j.caeai.2022.100101

Lin, X.-F., Zhou, Y., Shen, W., Luo, G., Xian, X., & Pang, B. (2023). Modeling the structural relationships among Chinese secondary school students' computational thinking efficacy in learning AI, AI literacy, and approaches to learning AI. *Education and Information Technologies*, 1–27. https://doi.org/10.1007/s10639-023-12029-4

Lintner, T. (2024). A systematic review of AI literacy scales. *Npj Science of Learning, 9*(1), 50. https://doi.org/10.1038/s41539-024-00264-4

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*.

Maydeu-Olivares, A. (2013). Goodness-of-Fit assessment of item response theory models. *Measurement: Interdisciplinary Research & Perspective, 11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ng, D. T. K., Wu, W., Leung, J. K. L., & Chu, S. K. W. (2023). Artificial intelligence (AI) literacy questionnaire with confirmatory factor analysis. In *2023 IEEE international conference on advanced learning technologies (ICALT)*. https://www.researchgate.net/profile/tsz-kit-ng/publication/372284548_artificial_intelligence_ai_literacy_questionnaire_with_confirmatory_factor_analysis.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Pinski, M., & Benlian, A. (2023). Ai literacy - Towards measuring human competency in artificial intelligence. In *56th annual Hawaii international conference on system sciences*. https://scholarspace.manoa.hawaii.edu/handle/10125/102649.

R Core Team. (2022). *R: A language and environment for statistical computing [computer software]*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? *Journal of Individual Differences, 35*(4), 212–220. https://doi.org/10.1027/1614-0001/a000141

Rosseel, Y. (2012). Lavaan : An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schipolowski, S., Schroeders, U., & Wilhelm, O. (2014). Pitfalls and challenges in constructing short forms of cognitive ability measures. *Journal of Individual Differences, 35*(4), 190–200. https://doi.org/10.1027/1614-0001/a000134

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLoS One, 11*(11), Article e0167110. https://doi.org/10.1371/journal.pone.0167110

Sitarenios, G. (2022). Short versions of tests: Best practices and potential pitfalls. *Journal of Pediatric Neuropsychology, 8*(3), 101–115. https://doi.org/10.1007/s40817-022-00126-0

Soto-Sanfiel, M. T., Angulo-Brunet, A., & Lutz, C. (2024). The scale of artificial intelligence literacy for all (SAIL4ALL): A tool for assessing knowledge on artificial intelligence in all adult populations and settings. https://doi.org/10.31235/osf.io/bvyku.

Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023). Developing a model for AI across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Computers and Education: Artificial Intelligence, 4*, Article 100127. https://doi.org/10.1016/j.caeai.2023.100127

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Wang, B., Rau, P.-L. P., & Yuan, T. (2022). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, 1–14. https://doi.org/10.1080/0144929X.2022.2072768

Zhang, H., Perry, A., & Lee, I. (2024). Developing and validating the artificial intelligence literacy concept inventory: An instrument to assess artificial intelligence literacy among middle school students. *International Journal of Artificial Intelligence in Education. Advance online publication*. https://doi.org/10.1007/s40593-024-00398-x

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – Five misunderstandings and ways to overcome them. *Journal of Individual Differences, 35*(4), 185–189. https://doi.org/10.1027/1614-0001/a000148