Are they lit? Developing, testing, and implementing an instrument to measure artificial intelligence literacy

- *Kumulative Arbeit* -

Inaugural-Dissertation

zur Erlangung der Doktorwürde

der

Philosophischen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität

zu Bonn

vorgelegt von

Matthias Carl Laupichler

aus

Neuwied

Bonn, 2024

Gedruckt mit Genehmigung der Philosophischen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Zusammensetzung der Prüfungskommission:

Prof. Dr. Ulrich Ettinger
(Vorsitzender)

Prof. Dr. Tobias Raupach
(Betreuer und Gutachter)

Prof. Dr. André Beauducel
(Gutachter)

Prof. Dr. Fani Lauermann
(weiteres prüfungsberechtigtes Mitglied)

Tag der mündlichen Prüfung: 26. August 2024

For Janosch

I could not have wished for a better deadline

Summary

This thesis presents the development and application of the "Scale for the assessment of non-experts' AI literacy" (SNAIL). Artificial intelligence (AI) is increasingly influencing various aspects of daily life and is being applied more and more frequently in professional contexts. To achieve a beneficial interaction between non-experts (i.e., individuals without specific AI education) and AI, a certain set of basic AI competencies is necessary. These basic competencies are commonly called "AI literacy" and have been the focus of intensive research in recent years. A particular branch of AI literacy research focuses on the reliable and valid measurement of AI literacy. Early AI literacy studies used unvalidated questionnaires, which were not suitable for reliably determining subjects' AI literacy. Some researchers, including myself[1], have therefore started to develop measurement instruments for AI literacy that meet psychometric quality criteria.

When this research project was registered and planned, there were no validated instruments for measuring AI competence. Therefore, I conducted a Delphi expert study in which an initial item set for assessing AI literacy was generated through three iterative Delphi rounds. In this initial study, particular emphasis was placed on the content validity of the items, aiming to create a set of questions that would cover the entire field of AI literacy without exceeding its scope. Afterwards, the item set was presented to a large sample of non-experts, who assessed their individual AI literacy using the items created in the first study. Based on the collected data, I conducted an exploratory factor analysis, which examined both the underlying factor structure and reduced the number of items. The result of this second study was the final SNAIL questionnaire. As a subsequent intermediate step, an investigation was conducted into the extent to which the adapted SNAIL could be suitable for evaluating AI courses. The adaptation involved two steps: firstly, the items were systematically translated from English to German. Secondly, all items were presented in

[1] Disclaimer: Throughout this paper, the first-person singular is consistently used to clarify that the research ideas as well as the organization of study implementation and analysis originate from me. However, this is by no means intended to diminish the invaluable support of my esteemed co-authors.

both a retrospective version (assessing AI literacy before the start of the AI course) and a post-version (assessment after the completion of the course). Specific statistical methods that are suitable for evaluating learning outcomes were employed to identify strengths and weaknesses of the evaluated course. Finally, a large-scale study was conducted, in which SNAIL was used for the first time to assess the AI literacy of a specific subgroup of non-experts. For this purpose, both SNAIL and a scale for assessing "attitudes towards AI" (ATAI) were distributed to medical students from two German medical schools. Conducting a confirmatory factor analysis revealed that the original three-factor model showed a good model fit for this new data set. At the conclusion of this thesis, avenues for further development and enhancement of SNAIL are presented. One potential area for improvement is to reduce the number of items in the final SNAIL questionnaire to increase the scale's efficiency while maintaining sufficient reliability of the subscales. Finally, I take another look at AI literacy research as a whole in order to identify potential research directions associated with SNAIL that extend beyond questionnaire development.

Keywords: Artificial intelligence; AI literacy; Questionnaire development; Self-assessment

Acknowledgements

First, I would like to thank Prof. Dr. Tobias Raupach for making it possible for me to conduct this doctoral project and whose valuable advice has had a big impact in making this thesis what it is. Additionally, I thank Prof. Dr. André Beauducel for helping to get my doctoral project rolling. Special thanks also go to Alexandra Aster, who co-authored every single study and with whom I have gone through the trials and tribulations of a doctoral thesis together. Thanks also to Johannes Schleiss, who was not only an excellent co-author, but also never tired of discussing my ideas in countless conversations. Furthermore, I thank my co-authors, my colleagues at the Institute of Medical Education of the University Hospital Bonn, and all individuals who were involved as advisors, experts, or participants in my research projects.

Special thanks go to my wife, Nina, who has supported me unconditionally and kept me sane whenever I struggled. Without her, I could not have made it. Additionally, I thank my parents, Mechtild and Frank, who supported my curiosity and always had an answer to my many *whys*. I also thank my brother, Moritz, who is on his way to becoming the next Dr. Laupichler, and my sister-in-law, Melissa, as well as the entire Thomas family. Last but not least, my gratitude extends to all companions, friends, colleagues, teachers, and role models who shaped my journey and believed in me.

Glossary

AI	Artificial Intelligence
AIEd	Artificial Intelligence in Education
AILQ	AI Literacy Questionnaire
AILS	AI Literacy Scale
AILT	AI Literacy Test
CA	Critical Appraisal (factor)
CFA	Confirmatory Factor Analysis
CSA	Comparative Self-Assessment (gain)
EFA	Exploratory Factor Analysis
GAAIS	General Attitudes towards Artificial Intelligence Scale
LLM	Large Language Model
MAILS	Meta AI Literacy Scale
MCQ	Multiple Choice Question
PA	Practical Application (factor)
SNAIL	Scale for the Assessment of Non-Experts' AI Literacy
TU	Technical Understanding (factor)
TUCAPA	3-Factor AI Literacy Model

List of Tables

List of Figures

List of references for Study 1 to 4

Study 1 - Laupichler, M. C., Aster, A., & Raupach, T. (2023a). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence*, *4*, 100126. https://doi.org/10.1016/j.caeai.2023.100126

Study 2 - Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023c). Development of the “Scale for the assessment of non-experts’ AI literacy”–An exploratory factor analysis. *Computers in Human Behavior Reports*, *12*, 100338. https://doi.org/10.1016/j.chbr.2023.100338

Study 3 - Laupichler, M. C., Aster, A., Perschewski, J. O., & Schleiss, J. (2023b). Evaluating AI Courses: A Valid and Reliable Instrument for Assessing Artificial Intelligence Learning through Comparative Self-Assessment. *Education Sciences*, *13*(10), 978. https://doi.org/10.3390/educsci13100978

Study 4 - Laupichler, M. C., Aster, A., Meyerheim, M., Raupach, T., & Mergen, M. (2024). Medical students’ AI literacy and attitudes towards AI: a cross-sectional two-center study

using pre-validated assessment instruments. *BMC Medical Education*, *24*(401).

https://doi.org/10.1186/s12909-024-05400-7

1 Artificial intelligence literacy

1.1 Artificial intelligence - 21st century’s most important technology

Since the advent of OpenAI's "ChatGPT" tool, the term artificial intelligence (AI) seems to be on everyone's lips. Still, most laypeople seem to think of AI as anthropomorphized robots ("Terminator", "I, Robot") or superintelligences that by far surpass the capabilities of mere mortals (HAL 9000 in "2001: A Space Odyssey"). However, the reality is much more nuanced, and true AI in the sense of human-like intelligence still seems a long way off[2].

1.1.1 The definition of AI

In fact, even on a scientific level, it is no easy task to provide a meaningful definition of AI (Wang, 2019). Many researchers either avoid the problem of providing a definition of AI altogether, or use a somewhat arbitrary, technical definition from a standard textbook on AI (Russell & Norvig, 2010, p. viii). Although both approaches may be justified in certain contexts, it is important to clarify how AI is understood in the context of this work. Most definitions seem to fall into one of two schools of thought. On the one hand, AI is often presented as "a branch of computer science dealing with the simulation of intelligent behavior in computers" (Merriam-Webster, 2023) and is thus treated as a subfield of the much larger field of computer science. However, this definition neglects the ever-increasing interprofessionalization of AI. Disciplines such as mathematics, linguistics, neuroscience, and others also have an influence on the development of AI systems. On the other hand, other definitions of AI relate artificial intelligence to human intelligence and capabilities. While a direct comparison of biological and machine intelligence is not very helpful, AI does, in essence, embody the concept of "computer programs that have some of the qualities of the human mind" (Cambridge Dictionary, 2023).

[2] However, some researchers and entrepreneurs may have opposing views (e.g., Bostrom, 1998).

1.1.2 The application of AI

While the definition of AI warrants its own doctoral thesis, the sheer number of application examples illustrates the relevance of AI in today's world. AI seems to permeate all areas of daily and professional life. Almost everyone interacts directly or indirectly with certain AI applications on a daily basis, be it movie recommendations (Bennett & Lanning, 2007), online shopping (Da'u & Salim, 2020) or facial recognition software (Kaur et al., 2020). Even technologies that were established decades ago have recently been influenced by AI. Examples of this are weather forecasting (McGovern et al., 2017) or traffic management (Boukerche et al., 2020), both of which are nowadays strongly supported by AI. In addition, the professional use of AI systems to solve domain-specific problems is steadily increasing as well (Furman & Seamans, 2019). For example, AI-supported computer vision can be found in the field of medical imaging (Pesapane et al., 2018), AI-supported predictive maintenance in the field of mechanical engineering (Lee et al., 2019), or AI-supported Intelligent Tutoring Systems in the field of education (Bond et al., 2024, Hobert, 2023). This development will have a significant impact on the labor market, as AI applications will partially or even completely replace certain jobs (Frey & Osborne, 2017).

When nearly all domains of private and professional life are permeated by AI, almost every individual will interact with AI on a daily basis. Consequently, specific fundamental AI skills become essential to facilitate deliberate and risk-aware engagement with AI. These basic skills are commonly referred to as *AI literacy*.

1.2 AI literacy - Essential today, indispensable tomorrow

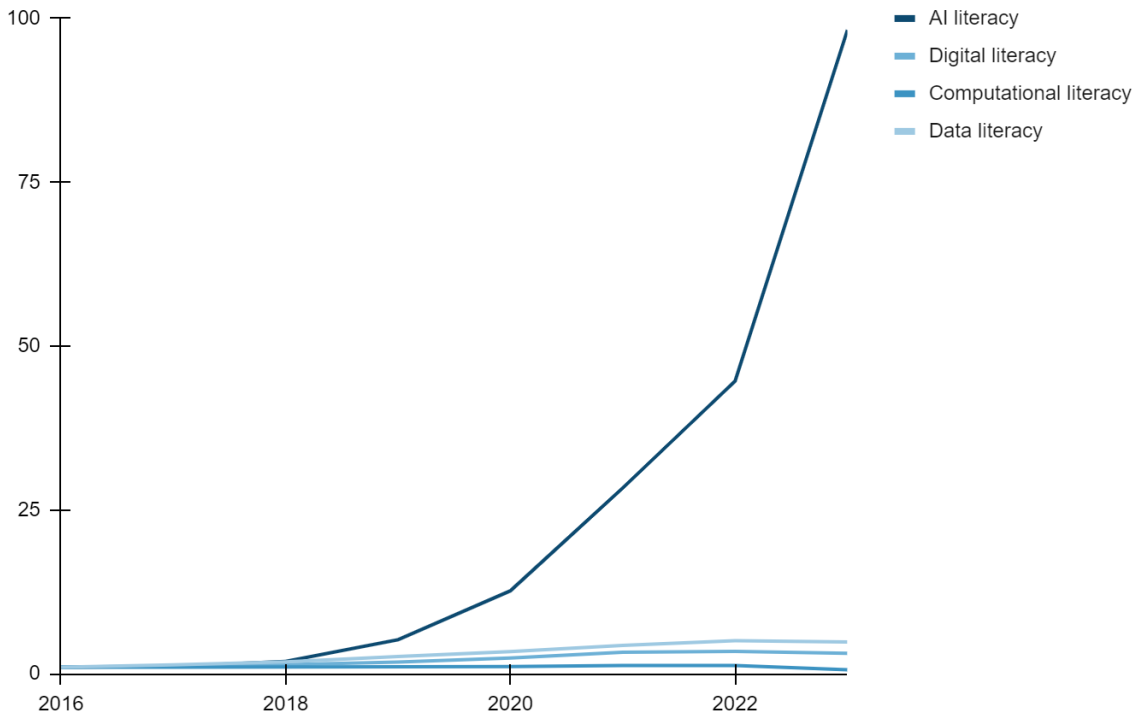
1.2.1 The "literacy" concept

The term AI literacy was first used in an online article by Konishi (2015) and "joins a long line of proposed literacies intended to symbolize the understanding of a particular technological construct" (Laupichler et al., 2022, p.1). These *literacies* combine the original concept of

literacy (i.e., alphabetical literacy; the ability to read and write) with an additional technical or cultural construct such as *media*. The combination of the two terms is intended to emphasize that the competencies in this area are basic skills that every educated citizen should possess, comparable to the ability to read and write (see Berkman et al., 2010). Examples of cultural or social literacies are media literacy (Livingstone, 2004), health literacy (Weiss, 2003), or financial literacy (Hastings et al., 2013). In addition, more technical constructs such as computational literacy (Magana et al., 2016), digital literacy (Gilster, 1997), and data literacy (Wolff et al., 2016) have also come to the fore, particularly in the last three decades. While research interest in technological literacies is either growing rather slowly (digital literacy, data literacy) or even seems to be stagnating (computational literacy), interest in AI literacy is growing rapidly (see Figure 1).

Figure 1

Standardized number of publications for different technological literacies



Note. The figure is based on the number of search results on Google Scholar. The diagram shows the relative number of publications. The number of publications in 2016 was set to 1

(as a starting point). The values for the following years were generated as the ratio of the number of publications in the interested year divided by the number of publications in 2016.

1.2.2 The definition of AI literacy

Even if there is no generally recognized definition of "AI literacy", the views of researchers are not as contradictory as they are when it comes to defining artificial intelligence. The most important definitions are presented below in order to identify commonalities and differences. The best known and most cited definition of AI literacy was published by Long and Magerko. The two authors define AI literacy 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, p.2). In the review by Ng et al. (2021b), the authors found that very few studies gave an explicit definition of AI literacy. However, the authors were able to identify four main constructs of AI literacy, which appeared repeatedly in various publications and postulated that AI literacy consists of "know & understand AI", "use & apply AI", "evaluate & create AI", and "AI ethics" (Ng et al., 2021b, p.4). Definitions by other authors go in a similar direction. Casal-Otero et al., for example, define AI literacy as "a set of skills that enable a solid understanding of AI through three priority axes: learning about AI, learning about how AI works, and learning for life with AI" (Casal-Otero et al., 2022, p.2). All definitions therefore seem to include a certain basic understanding of AI, which firstly encompasses a technical understanding of AI and secondly includes the capability to use or interact with AI tools to a certain extent. Furthermore, the aspect of AI application in daily life seems to be particularly important to the authors (Ng et al., 2021b; Casal-Otero et al., 2022). An important difference can be found in the consideration of AI ethics, a construct that is explicitly mentioned in Ng et al. (2021b), appears more implicitly in the definition by Long & Magerko (2020) ("critically evaluate") and is not mentioned in Casal-Otero et al.'s definition (2022). Nevertheless, all researchers seem to agree that programming skills and the ability to develop AI models

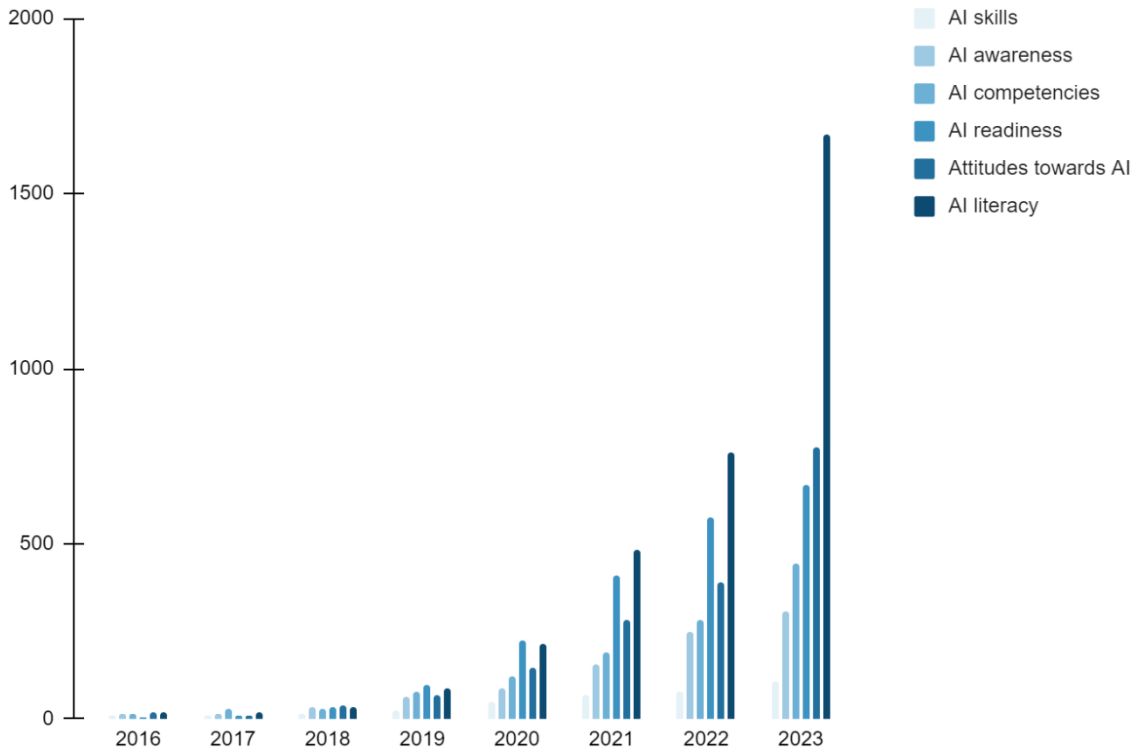
independently are not a part of AI literacy. Based on these findings, I have developed the following working definition of AI literacy, which should serve as a basis for understanding AI literacy in the context of this thesis: “The term AI literacy describes competencies that include basic knowledge and analytical evaluation of AI, as well as critical use of AI applications by non-experts.” (Laupichler et al., 2023c). In this definition, non-experts are all individuals who have not received a specific and extensive AI education. An example of non-experts would be medical students, physicians, or psychologists while most computer scientists or many mathematicians could potentially be called AI experts.

1.2.3 Related constructs

Although AI literacy seems to be the central concept when it comes to exploring the knowledge and understanding of AI (based on the number of publications, see Figure 2), some related constructs have also been developed alongside AI literacy.

Figure 2

Total number of publications containing the term AI literacy or related terms



Note. The figure is based on the number of search results on Google Scholar. I used different versions of the terms, for example: "AI litera*" OR "artificial intelligence litera*"

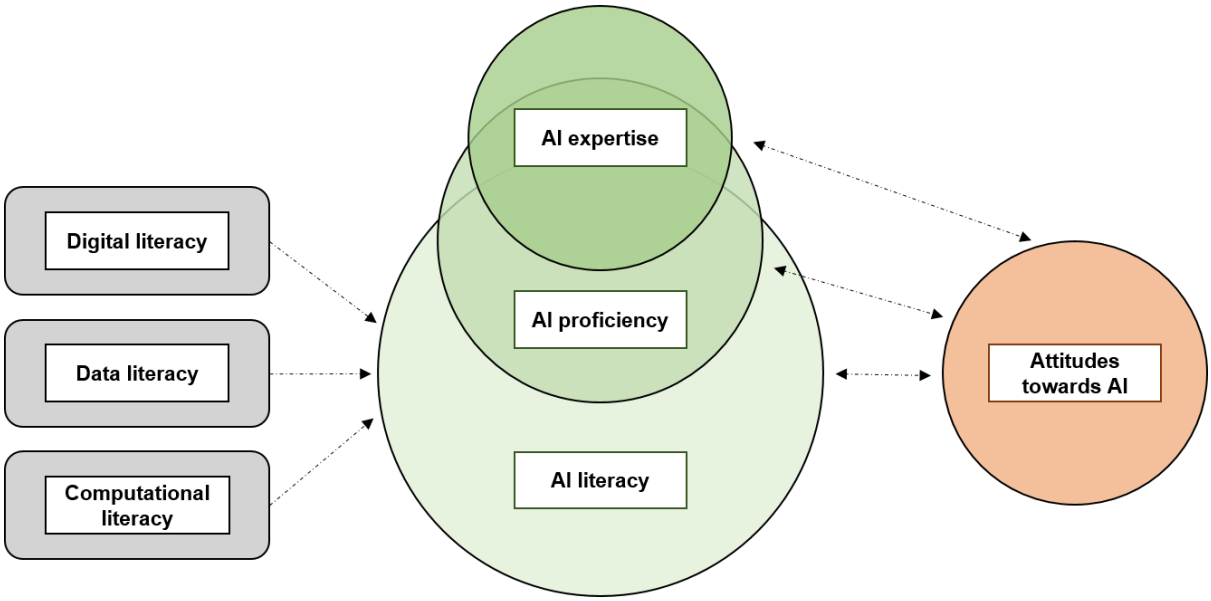
In comparison with the term AI literacy, other terms such as *AI skills*, *AI competencies*, and *AI awareness* have not caught on, which could be due to the fact that they are somewhat ambiguous. The term *AI readiness* is used more frequently. However, this term seems to have two different meanings: Firstly, Karaca et al. (2021) use it to describe the experienced preparedness of individuals to use AI, especially in the professional context (such as healthcare). Second, Holmström (2022) and others use the term "AI readiness" to describe an organization's readiness for the use of AI applications in its organizational processes. Accordingly, AI readiness must always be interpreted in the context of the corresponding research project, which could be counterproductive, especially in the interprofessional field of AI. The construct on which the second largest number of publications have appeared in recent years is "attitudes towards AI" (ATAI). Unfortunately, unlike AI literacy or AI readiness, it is not easy to delineate ATAI from other constructs, as no specific definition seems to exist. However, ATAI could be interpreted as an affective counterpart of the rather cognitive AI literacy. Schepman & Rodway (2020), for example, used an exploratory factor analysis and found that there are both positive attitudes towards AI (e.g., "Artificial intelligence is exciting") and negative views towards AI (e.g., "I find artificial intelligence sinister"). Sindermann et al. (2021) came to similar conclusions and identified two ATAI factors, which they called "acceptance" and "fear" of AI.

Finally, two additional tiers that extend beyond the scope of AI literacy could be added that might augment fundamental AI literacy. If one can assume that all people should be AI literate, it is reasonable to assume that there are competencies that go beyond basic AI literacy. “AI proficiency" can be located at the next hierarchical tier. AI proficiency describes the ability to use AI in certain domains in combination with one's own specialist expertise. An example of this would be radiologists who use AI applications for the acute diagnosis of strokes. The radiologists use their existing clinical knowledge and practical medical skills,

which are complemented by their ability to use AI applications safely and rationally and to recognize potential errors. The radiologists can take responsibility for the AI diagnosis because they understands how the AI application arrives at its results. In addition, AI proficient individuals are able to use their specialist skills (e.g., from the field of medicine) to contribute to the development of AI applications without having to program them themselves. The third and highest tier of this AI competence model, which I call “AI expertise”, is only achieved by very few people. AI experts have a deep technical and mathematical understanding of the processes behind AI. They are able to develop AI applications independently and collaborate with AI proficient individuals to develop AI applications for the relevant domains. Figure 3 attempts to illustrate the theoretical constructs surrounding AI literacy and their interrelationships.

Figure 3

Theoretical model of concepts related to AI literacy



Note. All of the relationships shown in the figure are based on purely theoretical considerations (i.e., not on empirical analyzes), as illustrated by the dashed arrows. In the case of bidirectional arrows, it is assumed that the two constructs influence each other, whereas in the case of unidirectional arrows, it is assumed that just one construct has an effect on the other.

2 Assessing AI literacy

2.1 AI literacy assessment - An important endeavor

2.1.1 The need for AI literacy assessment

As described in the first chapter, AI literacy is an indispensable skill of the 21st century. It has been shown that researchers across various research domains are aware of the relevance of this competence, which is reflected in the rapidly growing number of AI literacy publications (see Figure 1 and 2). This development raises a number of questions: How AI literate are individuals today, in general? In which AI literacy areas does an adequate AI knowledge base already exist? Can AI literacy aspects be identified in which further training is necessary? Answering these general questions is crucial, as societies should become as AI literate as possible to ensure a critically reflective and ethically acceptable use of AI applications. In addition to general questions, there are also specific research questions that require the existence of a validated AI literacy measurement tool. For example, it would be relevant to assess AI literacy in certain subgroups (e.g., doctors, engineers, teachers), as well as any potential differences between them. This information could be used to develop profession-specific AI courses precisely addressing the strengths and weaknesses of the respective discipline. In addition, the learning effectiveness of existing AI courses could be evaluated with the help of a reliable and valid AI literacy assessment instrument, which would enable quality assurance and continuous improvement of the course. Another interesting research question that can only be answered with the help of an AI literacy assessment instrument addresses the relationship between AI literacy and other constructs. For instance, the correlation between AI literacy and ATAI (see Section 1.2.3) must be investigated, as it could be assumed that a high level of AI literacy is associated with more positive attitudes towards AI (comparable to the relationship between scientific literacy and attitudes towards science; Einsiedel, 1994). Besides, it would also be interesting to compare AI literacy with personality traits, since people who are more open to new experiences

(Costa & McCrae, 2008) might be more AI literate because they could be intrinsically motivated to educate themselves on the new and exciting topic of AI.

In addition to the predominantly research-oriented benefits of accurately assessing AI literacy, corresponding tools could also be utilized in practical AI education. In the context of test-enhanced learning (Roediger & Karpicke, 2006), items from objective AI literacy scales (such as multiple-choice questions, MCQ) could particularly be employed in formative quizzes, as the repeated retrieval of knowledge has been shown to result in improved learning outcomes (Pan & Rickard, 2018).

Ultimately, reliable and valid AI literacy assessment instruments could even support political decision-making. While there used to be no legal frameworks for AI and AI education, many governments and organizations are currently in the process of adopting corresponding laws or recommendations (e.g., European AI Act, Madiega, 2023). These efforts are to be expected in the area of AI literacy, as there are already corresponding guidelines for related constructs such as data literacy (Schüller et al., 2021). To ensure that political decision-makers do not have to cast their votes solely on the basis of opinions and information from non-scientific media, a representative assessment of citizens' AI literacy would be desirable.

2.1.2 Requirements for AI literacy assessment instruments

The main problem underlying AI literacy assessment is that it is not enough to simply *ask* people how AI literate they think they are. Firstly, most laypeople simply do not know what the term "AI literate" means. Secondly, it is conceivable that person A considers themself to be rather AI literate because they have already worked with an AI-supported chatbot such as ChatGPT. Person B, on the other hand, could apply a different cognitive framing and rate their individual AI literacy as rather low because they have the feeling that they do not understand the structural and mathematical processes underlying the large language model on which ChatGPT is based. Accordingly, the validity of a direct question about people's AI literacy would be questionable at best, as it cannot cover the full scope of the definitions

listed in Section 1.2.2. Therefore, the development of a psychological measurement instrument that meets psychometric quality criteria and contains multiple items is necessary. While there are several standards that an AI literacy assessment instrument would need to fulfill, the two most important psychometric quality criteria are reliability and validity (Kaplan & Saccuzzo, 2005)[3]. Reliability is often described as the consistency of results: “For a psychometric test to be reliable, its results should be consistent across time (test-retest reliability), across items (internal reliability), and across raters (inter-rater reliability).” (White et al., 2022, p.16). In an AI literacy assessment questionnaire, the first two aspects are particularly important. Firstly, subjects should receive a similar "AI literacy score" if they take the test on consecutive occasions and have not educated themselves in the topic of AI in the meantime. Secondly, all items of the assessment instrument should yield similar results. This means that if subjects rate their competencies as high on item A, they should also achieve a higher score (on average) on item B, as long as the two items measure the same construct (i.e., AI literacy). The reliability of the questionnaire is also important because it has a strong influence on the second important criteria, validity. Put simply, “validity is the extent to which a test or examination assesses what it purports to assess” (Rust, 2007, p.25). Accordingly, an AI literacy assessment instrument should on the one hand cover the entire breadth of AI literacy and on the other hand not assess aspects that fall within the scope of other constructs (such as digital literacy, ATAI, etc.). Thus, the items of an instrument for assessing AI literacy should be mutually exclusive and collectively exhaustive. It should be noted that there are different subtypes of validity. Content validity, criterion validity, construct validity (including discriminant validity), external and internal validity, face validity, and concurrent validity (Fitzner, 2007) should all be taken into account when creating psychological measurement instruments.

[3] Although some authors also count objectivity and other quality criteria such as standardization or freedom of bias among the most important psychometric quality criteria (Rust, 2007).

In addition to these quality criteria, which are particularly important in psychometrics, there are other factors that influence the applicability of an AI literacy assessment instrument. One example of this is the *generalizability* of AI literacy assessment instruments, which is a criterion that evaluates how effectively the instrument can be used in different contexts. While generalizability is important for many measurement instruments, it is crucial for AI literacy instruments, as it can be assumed that all individuals should have a certain level of AI literacy. Accordingly, the AI literacy of groups as diverse as high school students, PhD students, employees, and pensioners as well as nurses, military personnel, accountants and teachers would have to be assessed using the same questionnaire. In addition, it is crucial how *efficiently* the measurement tool can capture an individual's AI literacy. This concerns aspects such as the length of the questionnaire and the number of items, but also the comprehensibility of the individual questions. The more people are asked to complete a questionnaire, the more important it is that this questionnaire does not take up too much time and cognitive resources, which would waste valuable resources (e.g., participants' time or researchers' money). Accordingly, the rule of thumb is that a questionnaire should be as long as necessary and as short as possible. Of course, an AI literacy instrument should also meet other psychometric quality criteria (e.g., fairness, usefulness; see Kubinger, 2019). Nevertheless, reliability and validity seem to be particularly important in the novel field of AI literacy measurement, and generalizability and efficiency should also be taken into account when developing AI literacy questionnaires.

2.2 AI literacy assessment - The status quo

For concepts related to AI literacy, such as digital literacy (Covello, 2010; Nguyen & Habók, 2023), data literacy (Cui et al., 2023; Kim et al., 2023), and scientific literacy (Istyadji & Sauqina, 2023; Atta et al., 2020), a large number of assessment instruments have been developed and validated in recent decades. In addition, for some years now, many instruments have been generated that are designed to evaluate ATAI, which have been

developed for different target groups and examine various attitudes of individuals towards AI. Probably best-known is the “General Attitudes towards Artificial Intelligence Scale” (GAAIS) developed by Schepman & Rodway (Schepman & Rodway, 2020; Schepman & Rodway, 2023), which uses 16 items to assess perceived opportunities, benefits and positive emotions and 16 items to assess concerns and negative emotions related to AI (Schepman & Rodway, 2020, p.3). Examples of other ATAI scales that differ from the GAAIS, for example in the number of items, are Sindermann et al. (2021), Suh et al. (2022), and Grassini (2023). In addition, instruments already exist that assess attitudes towards AI in specific contexts (Hadlington et al., 2023).

In contrast to the instruments listed above, no instrument for measuring AI literacy existed at the beginning of the research project described in this thesis. At the same time, however, various research groups already understood the relevance of such measurement instruments at this time and called on the research community to develop appropriate tools (Ng et al., 2021a). Since then, however, several researchers have heeded this call and developed, validated, evaluated and published AI literacy assessment instruments. Even if the research described in this thesis was unknowingly carried out in parallel with this other work, a thorough analysis of the other scales offers important insights. In order to compare all published AI literacy assessment instruments, I conducted a structured literature search on Web of Science (Clarivate), PubMed (National Library of Medicine) and PsycINFO (American Psychological Association).[4] I used the following search terms: ("AI literacy") AND ("assessment" OR "scale" OR "questionnaire" OR "instrument" OR "test" OR "item*"). Moreover, I extended the search to Google Scholar, while limiting my focus to the titles of the articles (i.e., only articles containing the search terms in their titles were analyzed).

The literature search led to the identification of eight instruments focused on measuring AI literacy (see Table 1). Five scales were designed as self-assessment scales, allowing participants to self-evaluate their AI literacy using Likert-type items (Carolus et al., 2023;

[4] The search was conducted on March 27th, 2024.

Laupichler et al., 2023c; Ng et al., 2023, Pinski et al., 2023; Wang et al., 2022).[5] The remaining three instruments were aimed at objectively testing participants' AI knowledge using MCQs and similar exercises (Hornberger et al., 2023; Tully et al., 2023; Weber et al., 2023). The first research group to develop and validate an AI literacy assessment instrument using exploratory (EFA) and confirmatory factor analyses (CFA) was Wang et al. (2022). Their "Artificial Intelligence Literacy Scale" (AILS) consists, in its final form, of 12 items loading onto four factors: awareness, usage, evaluation, and ethics. Following this publication in 2022, the subsequent year saw the release of the other seven instruments. This underscores the increasing significance of reliable and valid AI literacy measurement methods. Notably, despite the development of the scales occurring independently of each other, discernible overlaps are evident. Without delving excessively into specifics, a distinct emphasis on the "knowledge aspect" of AI literacy can be found in all self-assessment scales. While the knowledge factors bear different names such as "Understand AI," "Technical Understanding," "Cognitive," "AI Technology Knowledge," "AI Steps Knowledge," and "Awareness," all self-assessment scales include items that can unequivocally be traced back to the cognitive, knowledge-focused aspect of AI literacy. Even though factor names are somewhat arbitrarily determined by the researchers and may not necessarily allow to draw conclusions about the actual content of the items, these overlaps are nonetheless noteworthy. The same focus on understanding can be found in the objective AI literacy tests, which exclusively emphasize the more quantifiable and testable knowledge component. While the other factors may not be quite as similar, certain themes can also be identified. For instance, Carolus et al. (2023), Ng et al. (2023), and Wang et al. (2022) all list a factor addressing ethical aspects of AI literacy. Moreover, many of the instruments appear to emphasize the utilization of AI (Carolus et al., 2023; Pinski et al., 2023; Wang et al., 2022; Weber et al., 2023) as well as the evaluation of AI outcomes (Laupichler et al., 2023c; Wang

[5] It should be noted that the scale described in this thesis is already included as one of the five self-assessment scales in the table (Laupichler et al., 2023c).

et al., 2023; Weber et al., 2023). A more detailed examination of the individual scales would exceed the scope of this thesis. However, it is crucial to emphasize that despite the diversity of the scales in terms of theoretical background, number of items, content focus, and target audience, all scales appear to measure the same construct (see Table 1).

Table 1

Features of various AI literacy assessment instruments

Authors	Year	Scale/ instrument acronym	Dimension or factor names	Number of items	AI literacy self-assessment or test	Primary method and distinctive methodological features	Primary target audience
Carolus et al.	2023	MAILS	1. Use & Apply AI 2. Understand AI 3. Detect AI 4. AI Ethics	34	Self-assessment	Used hypothetical model derived from literature; tested model fit by conducting a CFA. Includes further psychological competencies in addition to AI literacy.	Not specified
Hornberger et al.	2023	/	/	31	Test (MCQs and sorting item)	Used CFA to test the assumption of unidimensionality before analyzing the data by using item response theory.	University students
Laupichler et al.	2023	SNAIL	1. Technical Understanding 2. Critical Appraisal 3. Practical Application	31	Self-assessment	Used Delphi method to create initial itemset, EFA to develop the model and refine the self-assessment questionnaire, and CFA to validate the model.	“Non-experts”, i.e., individuals without AI education
Ng et al.	2023	AILQ	1. Affective 2. Behavioral 3. Cognitive 4. Ethical	32	Self-assessment	Used expert- and layperson interviews for content validation and EFA & CFA for model development and verification.	Secondary students

Pinski et al.	2023	/	1. AI Technology Knowledge 2. Human actors in AI knowledge 3. AI Steps Knowledge 4. AI Usage Experience 5. AI Design Experience	13	Self-assessment	Used expert interviews, card sorting and SEM for instrument development and validation. “Pre-test study” with 50 participants.	Employees in AI-related positions
Tully et al.*	2023	AILT	/	25	Test (MCQs)	Tested individuals’ AI literacy with MCQs. Authors conducted several validation studies using different samples and methods.	Not specified
Wang et al.	2022	AILS	1. Awareness 2. Usage 3. Evaluation 4. Ethics	12	Self-assessment	Based on research on digital literacy. Conducted an EFA and a CFA for model development and verification.	Ordinary users
Weber et al.	2023	/	1. Socio & Technical AI literacy 2. User & Creator/ Evaluator	16	Test (MCQs)	Used expert interviews, card sorting, and validation study for test development. Tests AI literacy of individuals with MCQs.	Human stakeholders (separated in three classes: evaluators, creators, users)

Note. Entries are structured alphabetically. The table only contains instruments that have been validated using structural equation models, factor analyses, or other high-quality analysis methods. Publications marked with an asterisk are published as preprints (at the time the literature search was conducted, i.e., March 27, 2024) and have therefore not yet been peer-reviewed. MCQs = Multiple-choice questions.

3 Research framework

The research project described in this thesis adhered to a well-defined research framework, which is described below. The findings of each step in the framework were published as peer-reviewed research articles in scientific journals (see Section “List of references for Study 1 to 4”). Consequently, the research framework is intended to provide an overview of the project's progression and serve as a guideline for this thesis (see Table 2). A more in-depth examination of the results can be found in Chapters 4 to 7.

Table 2

The key features of each step in the research framework

Step in research framework	Goal	Primary method	Participants
1	Generating content valid item set	Delphi expert method	AI literacy experts from various institutions
2	Analyzing latent factor structure and finalizing AI literacy self-assessment questionnaire	Exploratory factor analysis (EFA)	Online sample of non-experts
3	Investigating whether the AI literacy questionnaire is suitable for evaluation of AI courses	Retrospective vs. post-course self-assessment	Participants of a university-level AI course
4	Using the questionnaire to assess AI literacy of specific subgroups and examining its relation to ATAI	Confirmatory factor analysis (CFA)	Medical students from two German medical schools

The objective of this project was to develop a reliable and valid AI literacy measurement instrument that could efficiently assess the AI literacy of diverse groups of non-experts.

3.1 Generating the item set

When the research framework was developed, there were no publicly available items to measure AI literacy. Therefore, the initial first step of the plan involved generating an item

set that, firstly, encompasses the entire breadth of the AI literacy concept (see Section 1.2.2) and, secondly, excludes items focused on related constructs such as data or computational literacy (see Section 1.2.3). Fulfilling these two fundamental requirements could be considered a reasonable proxy for content validity, which "addresses the degree to which items of an instrument sufficiently represent the content domain" (Zamanzadeh et al., 2014, p.164). Many scale developers generate the items somewhat unsystematically, either by developing the questions with the help of a few colleagues who are familiar with the topic, or by coming up with the questions themselves. While this approach is not illegitimate per se, it has the disadvantage that prior knowledge or preferences of individuals have a major influence on item generation. Therefore, a risk remains that the field of AI literacy is not covered to its full extent, or that certain aspects of AI literacy are given disproportionate importance. To counter this problem, I used the Delphi technique, in which a large number of experts with different professional backgrounds take a position on a topic in several study rounds in order to ultimately reach a common consensus (Rowe & Wright, 1999). While the Delphi method was originally developed to make predictions about the future (Sackman, 1974), it has been used to develop measuring instruments numerous times (e.g., Antcliff et al., 2013; Gagnon et al., 2014; Mengual-Andrés et al., 2016). A detailed description of the procedure and an assessment of the strengths and weaknesses of this project step can be found in Chapter 4 and in Laupichler et al. (2023a).

3.2 Developing the scale

In a second step, the itemset generated in step 1 was distributed to a sufficiently large and representative sample of non-experts. In this step, an EFA was conducted to examine if latent factors could be identified that influence the participants' response behavior to the manifest items (Watkins, 2021). The second goal of this project step was to finalize the AI literacy questionnaire, which has since been referred to as the "Scale for the assessment of non-experts' AI literacy" (SNAIL, see Figure 4). For this purpose, the communalities and

intercorrelations of the individual items in the final model were examined. If values were unsatisfactory, items were eliminated in order to increase the efficiency of the SNAIL and avoid redundancies. Further information on this project step can be found in Chapter 5 and in Laupichler et al. (2023c).

Figure 4

Official logo of the "Scale for the assessment of non-experts' AI literacy" (SNAIL)

3.3 Adapting the scale for evaluation

The third step can be interpreted as an intermediate step within the research framework, as it was designed as a "proof of concept" study. Its aim was to investigate the extent to which the newly developed AI literacy questionnaire is suitable for evaluating AI courses and thereby enabling quality assurance and course improvement. A slight adaptation was

necessary for this purpose, as the original questionnaire was designed as a "status quo" measurement, while the focus of the course evaluations should lie on examining the learning outcomes (i.e., a change in AI literacy from before to after attending the course). The modification involved assessing all items twice: once as a retrospective self-assessment of AI literacy before the start of the course and once at the time of data collection itself (i.e., after the completion of the course). Furthermore, it was investigated whether the comparative self-assessment gain (CSA gain) calculation, as proposed by Schiekirka et al. (2013), facilitated a more detailed examination of the AI literacy change compared to a traditional t-test. A more comprehensive description of this project step can be found in Chapter 6 and in Laupichler et al. (2023b).

3.4 Using the scale

In the fourth and final step of the research framework, the SNAIL was employed for the assessment of AI literacy within a specific subgroup of non-experts, namely medical students in Germany. In addition to the SNAIL, participants were presented with the ATAI scale developed by Sindermann et al. (2021). This approach allowed drawing preliminary conclusions regarding the relationship between the predominantly cognitive AI literacy and the more affectively oriented ATAI. Furthermore, a CFA was conducted to come full circle by evaluating the fit of the model developed in step 2 based on the data of the new sample. A detailed discussion of the results can be found in Chapter 7 of this text and in Laupichler et al. (2024).

4 Study 1 - Developing an initial item set to assess AI literacy with a focus on content validity

Laupichler, M. C., Aster, A., & Raupach, T. (2023a). Delphi study for the development and preliminary validation of an item set for the assessment of non-

experts' AI literacy. *Computers and Education: Artificial Intelligence, 4*, 100126. https://doi.org/10.1016/j.caeai.2023.100126

4.1 Summary of Study 1

As described above, no validated scales for measuring AI literacy were available at the beginning of this research project. A systematic literature search in April 2022 found only a small number of articles reporting the use of an AI literacy scale. However, none of these scales had been psychometrically validated. A second search in October 2022 yielded a new paper by Wang et al. (2022), which turned out to be the first publication that used EFAs and CFAs to develop an AI literacy assessment scale (see Section 2.2 and Table 1). While this scale was rightly praised as pioneering work, some areas for improvement were identified. One of the limitations of Wang et al.'s (2022) publication was that the authors of the study generated the items themselves, and that only five experts contributed to content validation, which could have caused some form of selection bias (Blackwell & Hodges, 1957). Study 1 therefore focused exclusively and in detail on the content validation of the items, using a much larger and more diverse group of experts. Thus, the aim of the study was to conduct preliminary groundwork for the development of the final scale and to formulate a set of items to facilitate a content-valid measurement of AI literacy among individuals without expertise in the field (i.e., non-experts). I wanted to find out which "items are relevant for and representative of AI literacy", and how these items could "be rephrased to most accurately represent the construct of AI literacy" (Laupichler et al., 2023a, p.2). When choosing the method, I opted for a Delphi expert survey. As explained in Chapter 3, this method has the advantage that it enabled a very elaborate testing and re-testing of content validity and "that it is not the opinions of individual persons that count, but the assessments of a large group that is very well versed in the field" (Laupichler et al., 2023a, p.2). Since the people who participated in the Delphi study were all experts in the field of AI or education (or both), it could be assumed that they are capable of assessing the validity of AI literacy items. The

study was divided into three Delphi rounds. In round 1, the participants were asked in an open question about topics or items that they considered important for assessing AI literacy. They were also presented with 40 items that had been generated in advance by interviewing AI and education experts and consulting books on AI, machine learning, and AI ethics (see method section of Study 1, Laupichler et al., 2023a) and were asked to rate their relevance. In rounds 2 and 3, some items were excluded while new items were generated and evaluated from the experts' open answers and possible reformulations were evaluated. Of a total of 47 possible items (40 generated in advance, 7 derived from expert suggestions) that were evaluated by the experts, 39 were ultimately included in the final item set. Furthermore, the experts made improvement or rewording suggestions for 66% of all items. In the subsequent rounds, all experts were able to select the item wording that they found to be most likely to validly assess a facet of AI literacy. “This approach further increased the scale’s content validity by ensuring that no important item content was omitted or that the inclusion of unnecessary item content negatively affected the relevance of the item.” (Laupichler et al., 2023a, p.6).

4.2 Strengths and limitations of Study 1

As with every research project, internal and external factors led to methodological and content-related limitations in Study 1. The first limitation lay in the selection of subject matter experts. Although the sample of 53 experts enabled a much more representative evaluation of the items than was the case in comparable studies that deployed only a few experts, the selection of participants in this study may also have been subject to selection bias. As the experts were mainly recruited through contact lists from national network meetings, the experts had fairly similar positions and were mostly academic staff at German universities. Accordingly, it would have been desirable to have also invited subject matter experts who came from a different background (i.e., non-university institutions; industry) or who had a different perspective on AI. The second limitation could best be elucidated by referring to

Figure 2 in Laupichler et al. (2023a). Typically, Delphi studies involve fixed consensus criteria that must be met to arrive at a conclusion for a specific research question. However, in the case of Study 1, this proved unattainable due to significant divergence of opinions, making it challenging to interpret the absence of response variance as a consensus. Consequently, contrary to the original idea of using standard deviations as a consensus criterion, the research team had to make decisions about consensus rules, which may also have been subject to biases. Since these decisions were not arbitrary and continued to be based on statistical measures (rather than the researchers' intuition), it can be presumed that the decisions were nonetheless valid and reliable.

I trust that throughout the summary and in the publication (Laupichler et al., 2023a), I have effectively conveyed the strengths of the paper in contrast to its limitations. Beyond the provision of a freely available and content-valid item set for constructing an AI literacy assessment scale, a notable strength of this work lay in the combination of the recruitment of a substantial number of subject matter experts and the three-stage Delphi process. Particularly, the iterative evaluation of content validity by the experts could be construed as a clear advantage of this study compared to similar endeavors by other researchers.

4.3 Integration of Study 1 into the research framework and subsequent steps

Study 1 was an essential first step in the development of a reliable and valid AI literacy assessment instrument. Without the availability of a set of items that was as representative and clearly delineable as possible, it would not have been possible to carry out the subsequent steps in the research framework. However, as Carolus et al. (2023) correctly pointed out, no factor analysis was carried out to validate and further develop the item set.[6] I addressed this criticism in Study 2, in which the underlying latent factor structure was to be investigated by conducting an exploratory factor analysis. The advantage of conducting an

[6] Carolus et al. (2023) refered to Study 1 in their critique, unaware that studies 2 through 4 of the research framework were already planned and specifically designed to address this very criticism.

EFA was that no hypothetical considerations needed to be made regarding the assignment of certain items to previously defined factors. In contrast to many other scale development projects, this project did not follow a deductive procedure (deriving and grouping items based on theoretical considerations), but an inductive procedure (developing theories on the basis of empirical findings). The reason for this lay in the somewhat weak theoretical landscape surrounding AI literacy. There were some theoretical considerations regarding competency dimensions and skill areas related to AI literacy (Touretzky et al., 2019; Long & Magerko, 2020; Ng et al., 2021a; Ng et al., 2021b). However, these were often based on considerations of individual authors or on the results of literature reviews, which in turn relied on publications that did not base their assessments on empirical evidence. Since I believed that especially with novel constructs like AI literacy, an empirical foundation should be established from which theoretical considerations can be derived (i.e., inductive reasoning), I chose the sequential process described in Study 1 and Study 2.

5 Study 2 - Conducting an exploratory factor analysis to finalize the AI literacy assessment scale

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

5.1 Summary of Study 2

As mentioned above, defining AI literacy is not as straightforward as one might think. This is partly due to the abundance of different definitions of AI literacy, some of which even contradict each other (see Section 1.2.2). Therefore, in the context of Study 2, I have developed my own AI literacy working definition, which emerged from the empirical results of Studies 1 and 2 and places a special focus on “non-experts”. My AI literacy definition read:

"The term AI literacy describes competencies that include basic knowledge and analytical evaluation of AI, as well as critical use of AI applications by non-experts." (Laupichler et al., 2023c, p.1).

It was easy to see the level of attention AI literacy assessment received when comparing the number of published AI literacy scales at the time of the publication of the first and second study. While at the time of the first study's publication, only the scale by Wang et al. (2022) existed, Study 2 identified two more published scales (Wang et al., 2022; Pinski & Benlian, 2023; Carolus et al., 2023). In Study 2, unlike as in Wang et al. (2022), no hypotheses were defined, since a purely exploratory approach was pursued. Nonetheless, three open research questions were formulated to structure the research process, loosely following the steps of EFAs. Firstly, the number of latent factors and the factor loadings of the items were to be examined. Secondly, it was to be evaluated to what extent the items of a factor followed a thematic content, to which a descriptive and fitting name could be assigned. Thirdly, it was to be examined which items could potentially be excluded from the final SNAIL based on statistical considerations to increase the efficiency of the questionnaire. To address these research questions, the items developed in Study 1 were presented to a total of 415 non-expert online participants, who rated their abilities regarding each item on a Likert scale from 1 to 7. The choice of sample size was not arbitrary but based on methodological considerations and recommendations (Mundfrom et al., 2005; Comrey & Lee, 1992; Benson & Nasser, 1998). In conducting the EFA, I followed the recommendations of Watkins (2021). For the data analysis, I used R (R Core Team, 2021) and RStudio (RStudio Team, 2020). A detailed overview of the underlying R code, which includes all R packages used, can be found in the Markdown document in the Supplementary Material of the article on the journal's website (Link: https://ars.els-cdn.com/content/image/1-s2.0-S2451958823000714-mmc1.pdf). For the sake of brevity, it can only be mentioned here that the `fa` function from the `psych` package (version 2.4.6.26) was used for factor rotation. The final results are based on the promax rotation method, which is an oblique modification of the orthogonal

varimax rotation. Apart from the number of factors (`nfactors`), the rotation method (`rotate = "promax"`), the sample size (`n.obs = 415`, as I used a correlation matrix as the basis for analysis), and the factoring method (`fm = "ml"`, maximum likelihood), all other `fa`-arguments have been set to the default version.[7] The considerations relating to the implementation of the EFA were listed in Section 2.3 of Study 2. (Laupichler et al., 2023c). To increase efficiency, I eliminated items by identifying items with "salient pattern coefficients on more than one factor on the one hand, and a particularly low communality on the other." (Laupichler et al., 2023c, p.4). The results showed that a one-factor model would be susceptible to underfactoring, and a four-factor model would have too few items loading on the fourth factor. In contrast, both a two- and a three-factor model would have been acceptable based on the parameters found. However, since the correct number of factors depends not only on absolute numbers, but rather on "not missing any factor of more than trivial size" (Cattell, 1978, p.61), I opted for the three-factor model. A more detailed justification would exceed the scope of this thesis, so I would like to refer to the results section of Study 2 (Laupichler et al., 2023c). With respect to the third research question, eight items were excluded, resulting in the final SNAIL comprising 31 items, which loaded onto three factors. Based on the content of the most salient items of each factor, I labelled the factors "Technical Understanding" (TU), "Critical Appraisal" (CA), and "Practical Application" (PA). Accordingly, the final three-factor model was abbreviated as the TUCAPA model.

5.2 Strengths and limitations of Study 2

One of the main drawbacks of assessing AI literacy through a self-assessment instrument was the possibility of responding untruthfully (consciously or unconsciously). Reasons for this included response biases such as social desirability (Nederhof, 1985), acquiescence

[7] Note that the default κ in the `fa` function of the R `psych` package is not specified and cannot be adapted through existing arguments. However, promax rotation was found to be relatively insensitive to the κ setting, as long as $\kappa \leq 5$ (Gorsuch, 2003; Tataryn et al., 1999).

(Messick, 1966; Hinz et al., 2007), or simply lacking knowledge about one's own skills (Kruger & Dunning, 1999; McDonald, 2008). Accordingly, I welcomed the development of objective AI literacy performance tests (Hornberger et al., 2023; Weber et al., 2023, Tully et al., 2023), which could potentially counteract the influence of response biases when used in conjunction with self-report measures like SNAIL. Furthermore, the results presented here all come from a single sample. If the sample happened to consist of individuals particularly knowledgeable about certain facets of AI, this could have introduced a sample bias that could significantly influence the EFA results. It would therefore be essential to reevaluate the factor structure identified through EFA through a CFA based on data from another sample. On the other hand, a major strength of Study 2 was the methodological rigor with which the EFA was conducted. Conducting an EFA requires a mutlitude of decisions on the part of the researcher, including choosing a factor model, deciding on the number of factors to retain, and choosing a rotation method. In the past, it has often been found that these decisions were made arbitrarily or simply incorrectly (Ford et al., 1986). Therefore, it was especially important that these decisions and the underlying rationale be reported (Watkins, 2018). Finally, the clearly formulated working definition could also be interpreted as a positive result of Study 2. In particular, the focus on the non-expert target group was advantageous, as it underlined the assumption that higher-level AI skills are not a part of AI literacy (but rather of AI proficiency or AI expertise, see Figure 3), which makes it easier to conceptualize AI literacy.

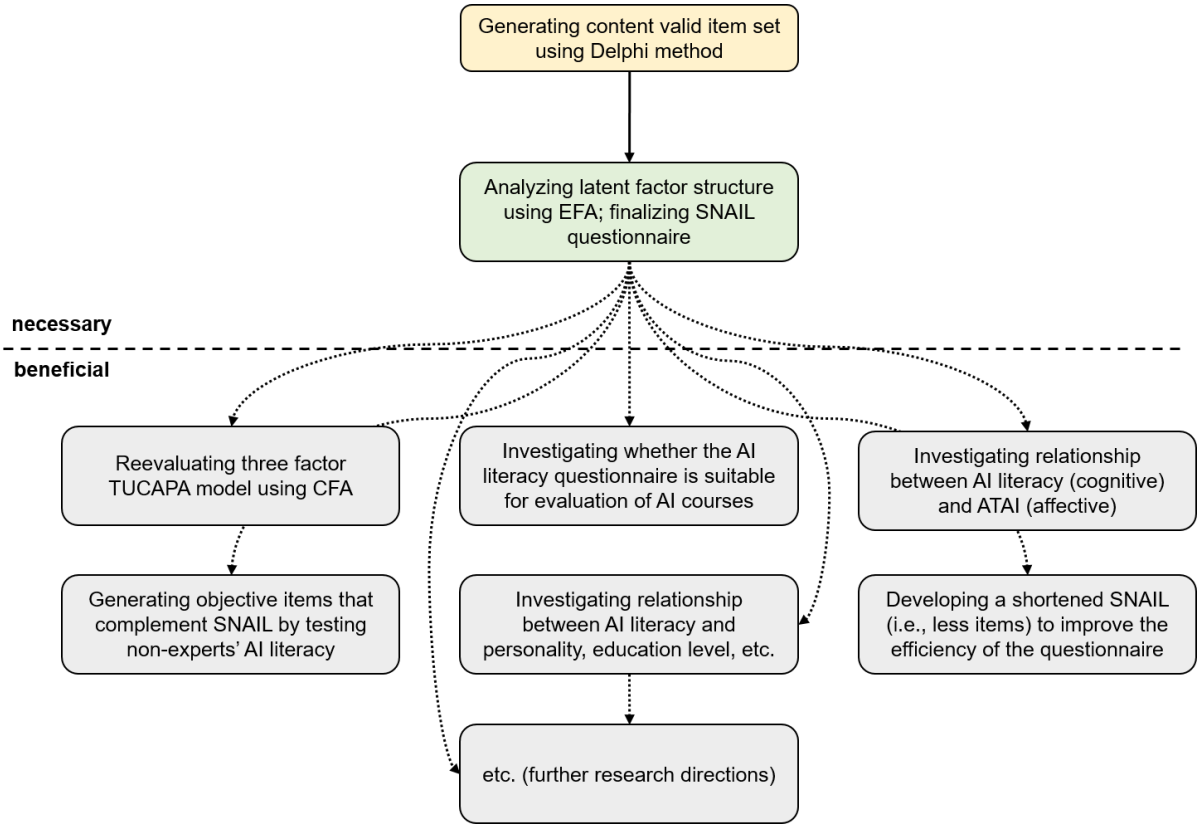
5.3 Integration of Study 2 into the research framework and subsequent steps

Study 2 can be understood in many respects as the centerpiece of the research framework, as within this work, SNAIL was developed and finalized, enabling its use in future research projects. As mentioned in Section 5.2, the subsequent step was to examine whether the model fit of the three-factor TUCAPA model remained adequate when analyzed based on data from another sample. However, SNAIL was essentially ready for use with the

completion of Study 2 and could be employed for various research endeavors. Of course, the questionnaire was not set in stone, and further improvement or adaptation of the scale may be beneficial in the future (see Figure 5 and Section 8.1).

Figure 5

Division of potential research projects into necessary and beneficial steps



Note. The yellow and green boxes at the top of the figure summarize the steps that necessarily had to be taken in order to generate and use SNAIL, while the gray boxes at the bottom represent interesting and useful, but not necessary, further research directions.

6 Study 3 - Translating the AI literacy scale and evaluating whether it is useful for AI course evaluation

Laupichler, M. C., Aster, A., Perschewski, J. O., & Schleiss, J. (2023b). Evaluating AI Courses: A Valid and Reliable Instrument for Assessing Artificial Intelligence Learning through Comparative Self-Assessment. *Education Sciences, 13*(10), 978. https://doi.org/10.3390/educsci13100978

6.1 Summary of Study 3

Study 3 could be understood as a proof of concept paper since “we found *preliminary* evidence that the adapted SNAIL questionnaire enables a valid evaluation of AI-learning gains” (Laupichler et al., 2023b, p.1, emphasis mine). Unlike the other three studies presented within this thesis, Study 3 constituted the report of an attempt to conduct course evaluations using SNAIL, contributing to quality assurance and improvement of AI courses. While there were already some publications presenting the evaluation results of AI courses aimed at increasing AI literacy, they did not utilize validated scales, often resorting to rather basic evaluation instruments that capture only the lowest level of the Kirkpatrick Model (Kirkpatrick & Kirkpatrick, 2006). Within the scope of Study 3, I addressed three research questions. Firstly, the extent to which SNAIL could be used for the reliable and valid evaluation of AI courses was investigated. Secondly, I analyzed the potential correlative relationship between AI literacy and ATAI for the first time, as such a relationship could also influence the success of AI courses (e.g., AI courses needing to address fears about AI). Lastly, it was examined whether prior AI education (beyond the evaluated course) had an impact on participants' self-assessment of AI literacy.

The evaluated program was a 150-hour interdisciplinary AI course, mainly focusing on artificial neural networks. In order for SNAIL to be used as an evaluation instrument for this course, it had to be slightly modified in two steps. First, the scale initially published in English was translated into German, following recognized guidelines for translating psychological questionnaires (Harkness et al., 2004). The reason for this was that most of the course participants were German native speakers, and I wanted to avoid potential language barriers that could distort the responses. Furthermore, each item was presented once in the “retrospective” version (assessment of AI literacy looking back on the time before the course began) as well as in a "post" version (assessment of AI literacy at the time of evaluation, i.e., after attending the course). In addition to SNAIL, the very short (and thus very efficient) ATAI scale by Sindermann et al. (2021) was presented, which was also available in a validated

German version. In addition to the two main questionnaires, several items were added that captured "to what extent the participants had already educated themselves on the topic of AI prior to the course, in other courses or with other methods." (Laupichler et al., 2023b, p.4). For the analysis of learning outcomes (i.e., changes in AI literacy after participation in the course), in addition to the classical one-tailed t-tests, the so-called comparative self-assessment gain was calculated (CSA gain, Raupach et al., 2011; Schiekirka et al., 2013). The reason for this was that participants usually rate themselves significantly better after attending the course than before, which greatly limits the informative value of t-tests. The CSA gain value (expressed as a percentage, range -100% to +100%) incorporates the learners' prior knowledge, enabling a differentiated assessment of learning outcomes. Somewhat unsurprisingly, almost every t-test mean comparison was significant, and for more than 80% of the items, at least a moderate positive effect (expressed as Cohen's d > .50) could be found. As previously described, the use of CSA gain allowed for a more nuanced examination of the learning outcomes. It was observed that course participants' CSA gain was higher on items of the Technical Understanding (TU) and Practical Application (PA) factors than of the Critical Appraisal (CA) factor. However, this difference represented, on average, only a 7% lower CSA gain. No significant relationship was found between AI literacy and ATAI, providing preliminary evidence that cognitive AI literacy and affective attitudes towards AI might not influence each other. However, further research in this direction is necessary. For a detailed discussion of Study 3's results, I would like to refer to Sections 3.2 to 3.4 in Study 3 (Laupichler et al., 2023b).

6.2 Strengths and limitations of Study 3

Even though the discussion section of Study 3 explains why the sample described in the study is adequate for the conducted analyses, one could criticize the composition of the sample. Since SNAIL is, by definition, an instrument for assessing AI literacy in non-experts, it should not be used to assess competencies in groups proficient in AI. However, the

sample of Study 3 included, among others, computer science students who could potentially be considered AI experts (or at least AI proficient). Future research should evaluate to what extent SNAIL maintains its validity when the analyzed samples consist of individuals educated in AI. Furthermore, the course itself had a strong technical focus. Aspects such as the ethical implications of AI or AI applications in daily life were discussed only marginally, if at all. This may have explained why the TU and PA factors showed a higher CSA gain than the CA factor. However, this provided further evidence that SNAIL has good construct validity. Nevertheless, further research projects should investigate if SNAIL leads to similar or different results in comprehensive AI courses. Lastly, the sample was quite small and originated from a single AI course, which could have favored the emergence of sample effects.

Despite these limitations, conducting Study 3 provided added value for the development and improvement of SNAIL. The results constituted preliminary evidence that even small adjustments to SNAIL could be sufficient to use the scale as a course assessment tool, although they need to be re-evaluated in full-scale AI courses with more learners. Furthermore, the publication (Laupichler et al., 2023b) offered a structured and clearly outlined guide for future research projects (potentially conducted by other research groups) on how these adaptations could be implemented. Additionally, researchers and educators in German-speaking regions benefited from the systematic translation of the scale into German, which was published as supplementary material and is openly accessible on the journal's website.

6.3 Integration of Study 3 into the research framework and subsequent steps

Study 3 can be regarded as a first step for further research in the field of AI course evaluation. If future studies prove that SNAIL can be used to evaluate AI courses, practitioners could use the adapted questionnaire (in its German and English versions) to

evaluate AI courses that are mainly focused on promoting the AI literacy of their participants (see Section 8.1).

As I attempted to illustrate in Figure 5 (lower part), besides the "evaluation direction," additional research directions could be pursued, each of which would advance the field of AI literacy or AI literacy assessments. Therefore, within this thesis, I chose not to further pursue the evaluation direction but explored other possible directions to provide as broad a picture as possible of SNAIL's potential.

7 Study 4 - Using the AI literacy scale and examining its relationship to other constructs

> Laupichler, M. C., Aster, A., Meyerheim, M., Raupach, T., & Mergen, M. (2024). Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. *BMC Medical Education*, *24*(401). https://doi.org/10.1186/s12909-024-05400-7

7.1 Summary of Study 4

Study 4 represents the first research project in which SNAIL was utilized to assess the average AI literacy of a specific group of individuals (in this case, medical students). Alongside SNAIL, the 5-item ATAI scale by Sindermann et al. (2021) was once again employed. While Study 3 did not find a statistically significant correlational relationship between AI literacy and ATAI, it had the aforementioned methodological limitations (see Section 6.2), which could have obscured a potential correlational effect. Since Study 4 benefited from a significantly larger sample size that was also closer to the target population of non-experts than the sample in Study 3, I decided to examine both constructs simultaneously in Study 4. Although there were existing publications that attempted to measure the AI literacy (or AI knowledge) and ATAI of medical students (see the literature review by Mousavi Baigi et al., 2023), in most cases, short, unvalidated questionnaires with

questionable reliability and validity were employed, and their results must be interpreted with caution. The use of the two validated scales (SNAIL and ATAI-scale) should therefore ensure that reliable and valid statements about the general AI literacy of medical students could be made.

Within the scope of Study 4, a total of five research questions were addressed. The investigation aimed to explore how medical students assessed their AI literacy, whether there were differences among old and young, female and male, as well as experienced and inexperienced medical students, and how medical students perceived their ATAI. Additionally, the analysis examined whether the two constructs of AI literacy and ATAI correlated significantly with each other and whether a higher level of AI education or interest in AI is associated with higher AI literacy. In addition to the items of the two scales (SNAIL and ATAI), an attention check, and a bogus item, the previous AI education of medical students and their interest in the topic of AI were surveyed. To minimize sampling effects and achieve an adequate sample size, the study was conducted at two German medical schools (University of Bonn and Saarland University). Besides typical inferential statistical methods such as t-tests (or Welch tests), Mann-Whitney-Wilcoxon tests, Fisher's test, etc. (see Section 2.3 in Study 4, Laupichler et al., 2024), a confirmatory factor analysis (CFA) was conducted to assess the model fit of the TUCAPA model based on the new sample. Additionally, Cronbach's α was calculated for each subscale of SNAIL and ATAI to examine the internal consistency of the questionnaires.

Since the data were multivariate non-normal and the variables were ordinal (due to Likert-type representation), a polychoric correlation matrix was used for conducting the CFA. An acceptable to good model fit of the three-factor SNAIL model was found. ATAI exhibited an excellent model fit. However, this was somewhat unsurprising, as the ATAI scale consists of the two factors "fear" and "acceptance," which could be considered polar opposites. Regarding Cronbach's α, the effect was reversed, with the SNAIL subscales achieving good to excellent α values, while the internal consistency of the ATAI subscales was relatively low.

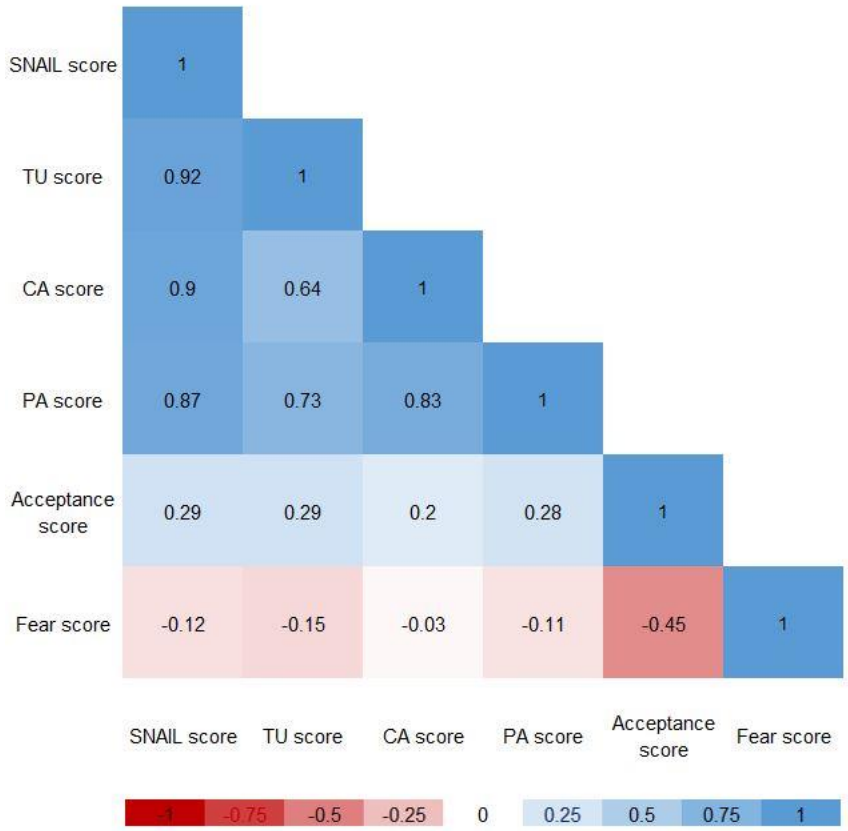
However, this may have been due to the number of items comprising the scales. Since Cronbach's α also increases with a greater number of items (and vice versa; Kopalle & Lehmann, 1997), it was not surprising that the two ATAI subscales (3 and 2 items, respectively) did not exhibit good internal consistency. At the same time, with α values >.9, the question arises whether the number of items in the TU subscale of SNAIL could be reduced, as some redundancy might be suspected. However, this exceeded the scope of Study 4.

Students from both medical schools rated their abilities highest on items of the CA subscale, closely followed by the PA subscale. Both CA and PA were rated significantly higher than TU skills. Participants' gender seemed to be the most important third variable influencing an individual's overall SNAIL score.[8] Across both medical schools, male medical students rated their AI literacy higher than females. Furthermore, all SNAIL factors correlated significantly positively with each other, which is unsurprising as they all measure the same overarching construct (AI literacy). Interestingly, all AI literacy subscores as well as the overall score correlated significantly positively with the "acceptance" score of the ATAI scale. While the picture was somewhat more subtle for the "fear" factor, a clear negative trend could still be observed, as most of the AI literacy subscales correlated significantly negatively with "fear" (see Figure 6). Finally, it was found that both the students' previous AI education and their interest in AI correlated significantly positively with the overall SNAIL score.

Figure 6

Heatmap depicting the correlations between SNAIL and ATAI subscales

[8] It is possible that there are other mediator or moderator variables that have an effect on AI literacy that were not investigated in this study. For example, the general level of education, openness to new experiences or the use of digital media could have an influence on the SNAIL score.



Note. Darker shading indicates higher correlations between the constructs. Red stands for negative correlations, blue for positive correlations.

7.2 Strengths and limitations of Study 4

As in all cross-sectional studies that capture the association or correlation between constructs, Study 4 also only examined a correlational, rather than a causal relationship. Therefore, it would be desirable to investigate the relationship between SNAIL and ATAI (as well as other constructs) trough psychological experiments, also known as "randomized controlled trials", in the future.

However, a clear strength of Study 4 is the analysis of two independent samples that, despite their independence, belonged to a common population, namely medical students. The use of two independent samples reduced the likelihood of sample effects and increased the validity of the CFAs, which were conducted separately for each medical school. Furthermore, Study 4 was the first study to not only advance the understanding of SNAIL

and the field of AI literacy assessments but also to provide important insights for another research field (i.e., medical education) through its focus on medical students. Thus, Study 4 demonstrated that SNAIL can also contribute to knowledge acquisition in other domains.

7.3 Integration of Study 4 into the research framework and subsequent steps

Study 4 was the final study conducted within the research framework outlined in this thesis. Although all the necessary and most of the beneficial aspects in Figure 5 have been investigated, there are now more questions after completing this research framework than before. However, this is not because the research presented here has led to ambiguous or conflicting results, but because the completion of the SNAIL development enables a plethora of new questions to be explored. These potential research directions, which future research projects can explore, are further elucidated in the concluding Chapter 8.

8 The future of AI literacy assessment

8.1 Quo vadis, SNAIL? - The future of the “Scale for the assessment of non-experts’ AI literacy”

The following chapter builds on Figure 5 and explores further possible research directions that can improve and advance SNAIL. One initial and relatively tangible aspect that could be targeted in future research projects is the reduction of SNAIL’s length (see Figure 5). With 31 items, SNAIL is relatively long in its final version, which affects the efficiency of its use in a questionnaire. Since particularly in Study 4, an extraordinarily high Cronbach's α of >.9 per subscale was found, it is suspected that redundancies exist. Therefore, the effect of eliminating certain items on the scale’s internal consistency should be investigated, since this could increase efficiency. This would require careful balancing to evaluate which items have a low informative value and can therefore be excluded. Additionally, different lengths of the subscales should be examined to determine at what number of items (per subscale) the internal consistency drops significantly. Following the development of the shortened SNAIL

(e.g., "ShortSNAIL"), a confirmatory factor analysis (CFA) should be conducted using another sample to examine whether the TUCAPA model is valid based on the reduced item set.

SNAIL has the disadvantage of only capturing the subjective self-assessment of AI literacy. Self-assessments can be influenced by external incentives or internal biases, both explicitly and implicitly. Therefore, it would be interesting to create an objective version of SNAIL, which supplements the self-assessment items with performance test questions. For each SNAIL item, one or more objective items could be developed, which, for example, objectively test the respondents' self-assessment through MCQs or higher-order (i.e., production test) questions. While Hornberger et al. (2023) and other researchers have already published tests with a similar goal, these were mostly developed independently of self-assessment instruments, which might render a comparison between self-perception and objective measurements difficult. Nevertheless, it would be interesting to compare existing AI competency tests with respondents' self-assessment using SNAIL, as this could shed light on areas where the test or the self-assessment scale could be improved.

In addition to the parallelism of the two measurement methods, it would also be helpful to develop a larger set of objective items. For instance, if five to ten MCQs were generated for each SNAIL item, an adaptive test could be developed using item response theory, based on a large question set, yet providing a very precise assessment of AI literacy with only a few questions for each participant. Since each person is presented with a slightly different set of questions in the adaptive test, this test could even be used in fields such as job assessment centers, as cheating is made significantly more difficult. Furthermore, this would have the advantage that objective AI literacy assessment items could also enhance learning outcomes in accordance with test-enhanced learning principles (see Section 2.1.1, Roediger & Karpicke, 2006). Lastly, one could go even further by automating the creation of objective items using language models (LLMs) such as ChatGPT. As I have demonstrated in another context, LLMs are capable of generating a large number of targeted and high-quality MCQs

(Laupichler et al., 2023d). This automation would lead to an even larger set of questions, further enhancing the aforementioned benefits. However, the exclusive use of MCQs also entails some problems, for which alternative solutions are proposed in Section 8.2. Furthermore, preliminary evidence was collected in Study 3 suggesting that SNAIL is also suitable for evaluating AI courses. However, since Study 3 should be considered a proof-of-concept study, future research projects could explore whether SNAIL is also suitable for evaluating larger-scale introductory AI courses. Particularly, the combination of "then-/post-" tests (retrospective assessment of skills before the course begins and assessment after completing the course) and calculating CSA gains that was introduced in Study 3 could be promising. During the course of this research project, several researchers and educators from national and international contexts have already inquired about using SNAIL to evaluate their courses. Therefore, it is possible that publications in this area will soon follow. Additionally, it would be interesting to evaluate large-scale AI education programs such as "Elements of AI" (University of Helsinki and MinnaLearn, elementsofai.com) or "AI-Campus" (Stifterverband für die Deutsche Wissenschaft, ki-campus.org) using SNAIL as they represent a large and representative sample of non-experts who come from different educational backgrounds. Beyond these target groups, it would also be essential to investigate the extent to which SNAIL is also suitable for use with children, for example in K-12 education. While much of the current AI literacy research deals with adult non-experts such as university students (Laupichler et al., 2022), there is an increasing focus on the AI literacy of schoolchildren (Casal-Otero et al., 2023). Corresponding educational programs in schools should also be evaluated using measurement instruments that are as reliable and valid as possible.

Another step that would advance SNAIL is the combined distribution of SNAIL and other AI literacy self-assessment scales (see Table 1). It would be theoretically possible to present the items of all five self-assessment scales in one questionnaire, so that every participant completes all AI literacy items. Although practical problems such as participant dropouts due

to the very long and redundant questionnaire might arise, the advantage of the project would be immense. On the one hand, it could be examined how the individual scales, especially the subscales, relate to each other, and whether there are central or outlier items. On the other hand, an exploratory factor analysis (EFA) could be conducted based on all items to arrive at a "super-scale" (in the literal and in the figurative sense) that encompasses the most important factors measured only with the best items from each questionnaire.
Lastly, a theoretical examination of the TUCAPA model would be highly beneficial. Within this work, the TUCAPA model has been derived solely from empirical results. The names of the factors were derived from the content of the most salient items, which critics could interpret as being somewhat arbitrary.[9] Therefore, a conceptual discussion of TUCAPA with other models, which have either originated purely theoretically (Long & Magerko, 2020, Ng et al., 2021b) or have been empirically developed (Wang et al., 2022), would be advisable.

8.2 Will they be (AI) lit (erate)? - The future of AI literacy (assessment) research

The following section takes another look at the entirety of AI literacy research. While the focus remains on AI literacy assessment, SNAIL itself is no longer in the foreground.
An important insight, which can serve as a basis for further research, is that all AI literacy assessments currently rely on one of two possible methodological modalities. The majority of measurement instruments, including SNAIL, consist of self-assessment questionnaires, which typically contain a specific number of statements, to which respondents indicate their agreement on a Likert-type scale. The second modality used to assess individuals' AI literacy lies in performance tests, which currently primarily use MCQs (specifically, single best answer questions). The results of these tests are subsequently analyzed using either classical test theory methods (Weber et al., 2023) or item response theory (Hornberger et al., 2023). While the application of these two measurement modalities is justified as they are

[9] My current approach still is a legitimate and common method for labelling the factors in the context of EFAs (Yong & Pearce, 2013).

straightforward, intuitive, efficient, and economical, additional modalities for assessing AI literacy are conceivable. Further possible modalities can be classified into two different groups: firstly, text-based methods, including the two aforementioned existing modalities, and secondly, behavior-based modalities. Possible text-based modalities include situational judgment tests, which are “measures of trainees’ awareness about what is effective behavior in work-relevant contexts in important interpersonal domains" (Patterson et al., 2012, p.858). More extensive questions based on case vignettes, for example "key feature questions" (Hrynchak et al., 2014; Berens et al., 2022) could also be employed. The latter would be particularly important for capturing the sub-construct "Practical Application," as PA aspects are more challenging to assess using self-report items (or MCQs) compared to aspects of the TU and CA factors. On the other hand, behavior-based modalities include methods such as behavioral observation or the analysis of so-called "log files", which are obtained by analyzing user data of computer programs. By analyzing the behavioral interaction between humans and AI, implicit or unexpected reactions can also be captured, which cannot be represented by text-based methods. Behavior-based methods could be particularly useful for the valid assessment of PA skills, as the efficient practical application of AI-powered software is reflected in the behavior of users.

Another aspect that may influence future research on AI literacy has been problematized several times throughout this thesis. The majority of currently existing studies that have examined the relationship between individuals' AI literacy and other constructs such as ATAI or interest in AI have overwhelmingly been restricted to exploring *correlative* relationships. The primary reason for this is that they were conducted as cross-sectional studies, which assessed the constructs at one point in time using a single group of individuals. Since correlations do not necessarily allow inferences on causality, future AI literacy research should be supplemented by entirely different empirical methods. Foremost among these is the (psychological) experiment, also known in some disciplines as a "randomized controlled trial". Following this method, study participants are randomly assigned to one of at least two

groups. Subsequently, a variable (or a construct such as ATAI) is manipulated to examine the effect of this independent variable on a dependent variable (for example, AI literacy).[10] There are also other methods that are aimed at analyzing causal relationships, which largely rely on data from longitudinal or cross-over design studies. These methods primarily involve statistical techniques such as propensity score matching (Caliendo & Kopeinig, 2008) or fixed effect models (Firebaugh et al., 2013). Nevertheless, conducting experiments remains the preferred choice for establishing causality, and it would be desirable if future AI literacy research projects employ this method.

In studies 3 and 4, relationships between AI literacy, ATAI, prior AI education, interest in AI, as well as personal variables such as gender and age were examined. Another interesting research idea would be to investigate the relationship between AI literacy and other, broader constructs. One of the most well-known personality theories is the so-called Big Five personality model (McCrae & Costa, 1987; Zuckerman et al., 1993). This model comprises five overarching personality factors that are more or less pronounced in all individuals. Especially the factor of "openness" could have an influence on AI literacy, as it could be assumed that more open-minded individuals are more likely to engage with novel AI technology and thus possess more knowledge and skills in this area. Of course, there are countless other constructs that could directly influence AI literacy or serve as moderator/mediator variables.

Lastly, I postulate that a large-scale survey will be necessary in the medium- to long-term to assess the AI literacy of a sample that can be considered representative of European citizens. The reason for this is that the European Union's "Artificial Intelligence Act" makes it very clear that AI literacy is an essential issue, and that "the Union and the Member States shall promote measures for the development of a sufficient level of AI literacy, across

[10] To provide a more illustrative example: It would be possible for an experimenter to randomly assign participants to one of two groups: Group 1 is shown a video highlighting the dangers and risks of AI. Group 2, on the other hand, would watch a video explaining that AI has a positive impact on humanity and can solve many problems. Subsequently, the difference between the AI literacy scores of the two groups and those of a control group that received no intervention could be analyzed.

sectors and taking into account the different needs of groups of providers, deployers and affected persons concerned, including thorough education and training, skilling and reskilling programmes and while ensuring proper gender and age balance, in view of allowing a democratic control of AI systems" (European Commission, 2023, Act. 4b). However, at the current time, we do not yet know which knowledge or skill gaps exist among the "average European". Consequently, one would need to measure all facets of AI literacy in a representative comparative sample in order to record the current state. Based on these results, focused AI education endeavors could then be developed to strengthen the AI literacy of all European citizens. It is even conceivable to plan an entire "AI curriculum" that links AI education at various educational levels (from kindergarten to university, Kandlhofer et al., 2016). In the context of curriculum development, the assessment of pan-European AI literacy would need to be repeated at regular intervals to evaluate the success and quality of the AI curriculum (Thomas et al., 2022).

Even though these ideas may currently sound like visions of the future, the influence of AI on daily life is undeniable. Extensive AI education for all citizens, comparable to education in other subjects such as mathematics or politics, will sooner or later become a necessity.

8.3 Conclusion

Within the scope of this dissertation project, a psychological measurement instrument was developed, validated, and applied to assess the AI literacy of non-experts. The measurement instrument was called the "Scale for the assessment of non-experts' AI literacy" (SNAIL).

In the first step, a content-valid item set was created through a Delphi study involving a large group of AI and education experts. Subsequently, the item set was examined based on data from a non-expert sample using exploratory factor analysis. The analysis revealed that the 31 items of the scale loaded onto three factors, which were termed "Technical Understanding," "Critical Appraisal," and "Practical Application."

Next, an interim study examined the suitability of SNAIL for evaluating AI courses. Finally, SNAIL was used for the first time to assess the AI literacy of a specific subgroup of non-experts, namely medical students. Additional results indicated that AI literacy is positively correlated with acceptance of AI and negatively correlated with fear of AI.

The development of a reliable and valid instrument for assessing AI literacy enables researchers to accurately measure the AI literacy of non-experts in the future, thereby enriching AI literacy research in general. Several possibilities for future research projects utilizing SNAIL have been illustrated within this work. For instance, SNAIL allows for the exploration of the relationship between AI literacy and other constructs such as ATAI. Ultimately, SNAIL could even advance AI literacy theory development, as the TUCAPA model is not solely based on theoretical assumptions but has a strong empirical foundation. In addition to the manifold theoretical and empirical implications, there are also practical opportunities afforded by the existence of SNAIL. Educators, for example, can use SNAIL to assess the AI literacy of their students or evaluate their own AI courses. As explained in Section 8.2, one could even go so far as to employ SNAIL for the political context, ensuring that political decisions can be made based on data obtained through SNAIL.

Disclosures

During the course of this doctoral project, I was employed as a research assistant at the Institute of Medical Education of the University Hospital Bonn. Besides my salary, I did not receive any financial support or other assistance for conducting the studies listed in this work. The execution of the four research projects received approval from the Ethics Committee of the University of Bonn. The respective file numbers can be found in the publications.

Furthermore, I assure that I have conducted this dissertation independently. Whenever I have used or referred to external material, I have clearly indicated this through a citation and an entry in the reference list. All aids were named either in the main text of the work or in the

individual papers. I have used the programs DeepL (DeepL SE, https://www.deepl.com/de/translator) and ChatGPT (OpenAI, Version 3.5, https://chat.openai.com) for the purpose of translating text passages from German to English. These applications were used exclusively for the translation of text passages. At no point did I use generative AI to develop new ideas or generate text that I did not write myself.

References

1) Antcliff, D., Keeley, P., Campbell, M., Oldham, J., & Woby, S. (2013). The development of an activity pacing questionnaire for chronic pain and/or fatigue: a Delphi technique. *Physiotherapy, 99*(3), 241–246. https://doi.org/10.1016/j.physio.2012.12.003
2) Atta, H. B., Vlorensius, Aras, I., & Ikhsanudin. (2020). Developing an instrument for students' scientific literacy. *Journal of Physics: Conference Series, 1422*(1), 012019. https://doi.org/10.1088/1742-6596/1422/1/012019
3) Bennett, J., & Lanning, S. (2007). The Netflix Prize. *Proceedings of KDD Cup and Workshop*, 35–38.
4) Benson, J., & Nasser, F. (1998). On the Use of Factor Analysis as a Research Tool. *Journal of Vocational Education Research, 23*(1), 13–33.
5) Berens, M., Becker, T., Anders, S., Sam, A. H., & Raupach, T. (2022). Effects of Elaboration and Instructor Feedback on Retention of Clinical Reasoning Competence Among Undergraduate Medical Students: A Randomized Crossover Trial. *JAMA Network Open*, *5*(12), e2245491. https://doi.org/10.1001/jamanetworkopen.2022.45491
6) Berkman, N. D., Davis, T. C., & McCormack, L. (2010). Health literacy: what is it?. *Journal of health communication*, *15*(S2), 9-19.
7) Blackwell, D., & Hodges, J. L. (1957). Design for the Control of Selection Bias. *The Annals of Mathematical Statistics, 28*(2), 449–460. https://doi.org/10.1214/aoms/1177706973
8) Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., ... & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(1), 4.

9) Bostrom, N. (1998). How long before superintelligence? *International Journal of Future Studies, 2*(1), 11–30. http://www.nickbostrom.com

10) Boukerche, A., Tao, Y., & Sun, P. (2020). Artificial intelligence-based vehicular traffic flow prediction methods for supporting intelligent transportation systems. *Computer Networks, 182,* 107484. https://doi.org/10.1016/j.comnet.2020.107484

11) Caliendo, M., & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys, 22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

12) Cambridge Dictionary. (2023). Artificial Intelligence. In *Cambridge Dictionary*. Cambridge University Press, Cambridge.

13) Carolus, A., Koch, M., 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, 2*(1), 100014. https://doi.org/10.1016/j.chbah.2023.100014

14) Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: a systematic literature review. *International Journal of STEM Education, 10*(1), 29. https://doi.org/10.1186/s40594-023-00418-7

15) Cattell, R. B. (1978). *Use of factor analysis in behavioral and life sciences*. Plenum Press, New York, NY.

16) Comrey, A., & Lee, H. (1992). Interpretation and application of factor analytic results. In *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates.

17) Costa, P. T., & McCrae, R. R. (2008). The Revised Neo Personality Inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment* (2nd ed., Vol. 2, pp. 179–198).

18) Covello, S. (2010). A Review of Digital Literacy Assessment Instruments. *Front-End Analysis Research* (IDE-712).

19) Cui, Y., Chen, F., Lutsyk, A., Leighton, J. P., & Cutumisu, M. (2023). Data literacy assessments: a systematic literature review. *Assessment in Education: Principles, Policy & Practice, 30*(1), 76–96. https://doi.org/10.1080/0969594X.2023.2182737

20) Da'u, A., & Salim, N. (2020). Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review, 53*(4), 2709–2748. https://doi.org/10.1007/s10462-019-09744-1

21) Einsiedel, E. F. (1994). Mental Maps of Science: Knowledge and Attitudes among Canadian Adults. *International Journal of Public Opinion Research, 6*(1), 35–44. https://doi.org/10.1093/ijpor/6.1.35

22) European Commission (2023). *Artificial Intelligence Act*.

23) Firebaugh, G., Warner, C., & Massoglia, M. (2013). Fixed Effects, Random Effects, and Hybrid Models for Causal Analysis. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 113–132). Springer, Heidelberg.

24) Fitzner, K. (2007). Reliability and Validity: A Quick Review. *The Diabetes Educator, 33*(5), 775–780. https://doi.org/10.1177/0145721707308172

25) Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The Application of Exploratory Factor Analysis in Applied Psychology: A Critical Review and Analysis. *Personnel Psychology, 39*(2), 291–314. https://doi.org/10.1111/j.1744-6570.1986.tb00583.x

26) Frey, C. & Osborne, M. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254-280. https://doi.org/10.1016/j.techfore.2016.08.019

27) Furman, J., & Seamans, R. (2018). AI and the Economy. *Innovation Policy and the Economy, 19*, 161–191. https://doi.org/10.1086/699936

28) Gagnon, A. J., DeBruyn, R., Essén, B., Gissler, M., Heaman, M., Jeambey, Z., Korfker, D., McCourt, C., Roth, C., Zeitlin, J., & Small, R. (2014). Development of the Migrant Friendly Maternity Care Questionnaire (MFMCQ) for migrants to Western societies: an international Delphi consensus process. *BMC Pregnancy and Childbirth, 14*(1), 200. https://doi.org/10.1186/1471-2393-14-200

29) Gilster, P. (1997). *Digital literacy*. John Wiley & Sons, Hoboken, NJ.

30) Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2., pp. 143-164). Wiley.

31) 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

32) Hadlington, L., Binder, J., Gardner, S., Karanika-Murray, M., & Knight, S. (2023). The use of artificial intelligence in a military context: development of the attitudes toward AI in defense (AAID) scale. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1164810

33) Harkness, J., Pennell, B., & Schoua-Glusberg, A. (2004). Survey Questionnaire Translation and Assessment. In *Methods for Testing and Evaluating Survey Questionnaires* (pp. 453–473). Wiley. https://doi.org/10.1002/0471654728.ch22

34) Hastings, J. S., Madrian, B. C., & Skimmyhorn, W. L. (2013). Financial Literacy, Financial Education, and Economic Outcomes. *Annual Review of Economics, 5*(1), 347–373. https://doi.org/10.1146/annurev-economics-082312-125807

35) Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *GMS Psycho-Social Medicine, 20*(4).

36) Hobert, S. (2023). Fostering skills with chatbot-based digital tutors–training programming skills in a field study. *i-com*, *22*(2), 143-159. https://doi.org/10.1515/icom-2022-0044

37) Holmström, J. (2022). From AI to digital transformation: The AI readiness framework. *Business Horizons, 65*(3), 329–339. https://doi.org/10.1016/j.bushor.2021.03.006

38) 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*, 100165. https://doi.org/10.1016/j.caeai.2023.100165

39) Hrynchak, P., Glover Takahashi, S., & Nayer, M. (2014). Key-feature questions for assessment of clinical reasoning: a literature review. *Medical Education, 48*(9), 870–883. https://doi.org/10.1111/medu.12509

40) Istyadji, M., & Sauqina, S. (2023). Conception of scientific literacy in the development of scientific literacy assessment tools: a systematic theoretical review. *Journal of Turkish Science Education*. https://doi.org/10.36681/tused.2023.016

41) Kandlhofer, M., Hirschmugl-Gaisch, S., & Huber, P. (2016). Artificial Intelligence and Computer Science in Education: From Kindergarten to University. *2016 IEEE Frontiers in Education Conference (FIE)*, 1–9.

42) Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications and issues* (6th ed.). Lincoln: University of Nebraska Press.

43) Karaca, O., Çalışkan, S. A., & Demir, K. (2021). Medical artificial intelligence readiness scale for medical students (MAIRS-MS) – development, validity and reliability study. *BMC Medical Education, 21*(1). https://doi.org/10.1186/s12909-021-02546-6

44) Kaur, P., Krishan, K., Sharma, S. K., & Kanchan, T. (2020). Facial-recognition algorithms: A literature review. *Medicine, Science and the Law, 60*(2), 131–139. https://doi.org/10.1177/0025802419893168

45) Kim, J., Hong, L., Evans, S., Oyler-Rice, E., & Ali, I. (2023). Development and Validation of a Data Literacy Assessment Scale. *Proceedings of the Association for Information Science and Technology, 60*(1), 620–624. https://doi.org/10.1002/pra2.827

46) Kirkpatrick, D., & Kirkpatrick, J. (2006). *Evaluating training programs: The four levels*. Berrett-Koehler Publishers, Oakland, CA.

47) Konishi, Y. (2015, December 25). *What is Needed for AI Literacy? Priorities for the Japanese Economy in 2016*. https://www.rieti.go.jp/en/columns/s16_0014.html

48) Kopalle, P. K., & Lehmann, D. R. (1997). Alpha Inflation? The Impact of Eliminating Scale Items on Cronbach's Alpha. *Organizational Behavior and Human Decision Processes, 70*(3), 189-197. https://doi.org/10.1006/obhd.1997.2702

49) Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

50) Kubinger, K. D. (2019). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens*. Hogrefe Verlag GmbH & Company KG, Göttingen.

51) 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*, 100101. https://doi.org/10.1016/j.caeai.2022.100101

52) Laupichler, M. C., Aster, A., & Raupach, T. (2023a). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence*, *4*, 100126.

53) Laupichler, M. C., Aster, A., Perschewski, J. O., & Schleiss, J. (2023b). Evaluating AI Courses: A Valid and Reliable Instrument for Assessing Artificial-Intelligence Learning through Comparative Self-Assessment. *Education Sciences*, *13*(10), 978.

54) Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023c). Development of the "Scale for the assessment of non-experts' AI literacy"–An exploratory factor analysis. *Computers in Human Behavior Reports*, *12*, 100338.

55) Laupichler, M. C., Rother, J. F., Kadow, I. C. G., Ahmadi, S., & Raupach, T. (2023d). Large Language Models in Medical Education: Comparing ChatGPT-to Human-Generated Exam Questions. *Academic Medicine*, published ahead of print. DOI: 10.1097/ACM.0000000000005626

56) Laupichler, M. C., Aster, A., Meyerheim, M., Raupach, T., & Mergen, M. (2024). Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. *BMC Medical Education, 24*(401). https://doi.org/10.1186/s12909-024-05400-7

57) Lee, W. J., Wu, H., Yun, H., Kim, H., Jun, M. B. G., & Sutherland, J. W. (2019). Predictive maintenance of machine tool systems using artificial intelligence techniques applied to machine condition data. *Procedia CIRP, 80*, 506–511. https://doi.org/10.1016/j.procir.2018.12.019

58) Livingstone, S. (2004). What is media literacy? *Intermedia, 32*(3), 18–20.

59) Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

60) Madiega, T. (2023). Artificial intelligence act. In *Briefing on the Artificial Intelligence Act*. European Union.

61) Magana, A. J., Falk, M. L., Vieira, C., & Reese, M. J. (2016). A case study of undergraduate engineering students' computational literacy and self-beliefs about computing in the context of authentic practices. *Computers in Human Behavior, 61*, 427–442. https://doi.org/10.1016/j.chb.2016.03.025

62) McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

63) McDonald, J. D. (2008). Measuring Personality Constructs: The Advantages and Disadvantages of Self-Reports, Informant Reports and Behavioural Assessments. *Enquire, 1*(1), 75–94.

64) McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., & Williams, J. K. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society, 98*(10), 2073–2090. https://doi.org/10.1175/BAMS-D-16-0123.1

65) Mengual-Andrés, S., Roig-Vila, R., & Mira, J. B. (2016). Delphi study for the design and validation of a questionnaire about digital competences in higher education. *International Journal of Educational Technology in Higher Education, 13*(1), 12. https://doi.org/10.1186/s41239-016-0009-y

66) Merriam-Webster. (2023). Artificial Intelligence. In *Merriam-Webster*. Merriam-Webster, Springfield, MA.

67) Messick, S. (1966). The Pschology of Acquiescence: An Interpretation of Research Evidence. *ETS Research Bulletin Series, 1*. https://doi.org/10.1002/j.2333-8504.1966.tb00357.x

68) Mousavi Baigi, S. F., Sarbaz, M., Ghaddaripouri, K., Ghaddaripouri, M., Mousavi, A. S., & Kimiafar, K. (2023). Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: A systematic review. *Health Science Reports, 6*(3). https://doi.org/10.1002/hsr2.1138

69) Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing, 5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

70) Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

71) Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021a). AI Literacy: Definition, Teaching, Evaluation and Ethical Issues . *Proceedings of the Association for Information Science and Technology, 58*(1), 504–509. https://doi.org/10.1002/pra2.487

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

73) Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2023). Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13411

74) Nguyen, L. A. T., & Habók, A. (2023). Tools for assessing teacher digital literacy: a review. *Journal of Computers in Education*. https://doi.org/10.1007/s40692-022-00257-5

75) Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education, 46*(9), 850–868. https://doi.org/10.1111/j.1365-2923.2012.04336.x

76) Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological bulletin*, *144*(7), 710. https://doi/10.1037/bul0000151

77) Pesapane, F., Codari, M., & Sardanelli, F. (2018). Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. In *European Radiology Experimental* (Vol. 2, Issue 1). Springer, Heidelberg. https://doi.org/10.1186/s41747-018-0061-6

78) Pinski, M., & Benlian, A. (2023). AI Literacy - Towards Measuring Human Competency in Artificial Intelligence. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 165–174.

79) Raupach, T., Münscher, C., Beißbarth, T., Burckhardt, G., & Pukrop, T. (2011). Towards outcome-based programme evaluation: Using student comparative self-assessments to determine teaching effectiveness. *Medical Teacher, 33*(8), e446–e453. https://doi.org/10.3109/0142159X.2011.586751

80) Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

81) Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting, 15*(4), 353–375. https://doi.org/10.1016/S0169-2070(99)00018-7

82) Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

83) Rust, J. (2007). Discussion piece: The psychometric principles of assessment. *Res Matters, 3*, 25–27.

84) Sackman, H. (1974). *Delphi Assessment: Expert Opinion, Forecasting, and Group Process*. RAND Corporation, Santa Monica, CA.

85) Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports, 1*, 100014. https://doi.org/10.1016/j.chbr.2020.100014

86) Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human–Computer Interaction, 39*(13), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

87) Schiekirka, S., Reinhardt, D., Beibarth, T., Anders, S., Pukrop, T., & Raupach, T. (2013). Estimating Learning Outcomes From Pre- and Posttest Student Self-Assessments. *Academic Medicine, 88*(3), 369–375. https://doi.org/10.1097/ACM.0b013e318280a6f6

88) Schüller, K., Koch, H., & Rampelt, F. (2021). *Data Literacy Charter.* Stifterverband, Berlin.

89) Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language. *KI - Künstliche Intelligenz, 35*(1), 109–118. https://doi.org/10.1007/s13218-020-00689-0

90) Suh, W., & Ahn, S. (2022). Development and Validation of a Scale Measuring Student Attitudes Toward Artificial Intelligence. *SAGE Open, 12*(2), 215824402211004. https://doi.org/10.1177/21582440221100463

91) Tataryn, D. J., Wood, J. M., & Gorsuch, R. L. (1999). Setting the value of k in promax: A Monte Carlo study. *Educational and Psychological Measurement, 59*(3), 384-391. https://doi.org/10.1177/00131649921969938

92) Thomas, P. A., Kern, D. E., Hughes, M. T., Tackett, S. A., & Chen, B. Y. (2022). *Curriculum development for medical education: a six-step approach* (4th ed.). Johns Hopkins University Press.

93) Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What Should Every Child Know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 9795–9799. https://doi.org/10.1609/aaai.v33i01.33019795

94) Tully, S., Longoni, C., & Appel, G. (2023). Knowledge of Artificial Intelligence Predicts Lower AI Receptivity. *Psyarxiv*. https://doi.org/10.31234/osf.io/t9u8g

95) 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, 42*(9), 1324–1337. https://doi.org/10.1080/0144929X.2022.2072768

96) Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence, 10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

97) Watkins, M. R. (2021). *A Step-by-Step Guide to Exploratory Factor Analysis with R and RStudio*. Routledge.

98) Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology, 44*(3), 219–246. https://doi.org/10.1177/0095798418771807

99) Weber, P., Pinski, M., & Baum, L. (2023). Toward an Objective Measurement of AI Literacy. *Pacific Asia Conference on Information Systems (PACIS) 2023 Proceedings*, 60.

100) Weiss, B. D. (2003). *Health literacy - A manual for clinicians*. American Medical Association Foundation, Chicago, IL.

101) White, R. F., Braun, J. M., & Kopylev, L. (2022). *NIEHS Report on Evaluating Features and Application of Neurodevelopmental Tests in Epidemiological Studies*. Research Triangle Park, NC.

102) Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2016). Creating an Understanding of Data Literacy for a Data-driven Society. *The Journal of Community Informatics, 12*(3), 9–26. www.ci-journal.net/index.php/ciej/article/view/1286.

103) Yong, A. G., & Pearce, S. (2013). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79–94. https://doi.org/10.20982/tqmp.09.2.p079

104) Zamanzadeh, V., Rassouli, M., Abbaszadeh, A., Majd, H. A., Nikanfar, A., & Ghahramanian, A. (2015). Details of content validity and objectifying it in instrument development. *Nursing Practice Today, 1*(3), 163–171.

105) Zuckerman, M., Kuhlman, D. M., Joireman, J., Teta, P., & Kraft, M. (1993). A comparison of three structural models for personality: The Big Three, the Big Five, and the Alternative Five. *Journal of Personality and Social Psychology, 65*(4), 757–768. https://doi.org/10.1037/0022-3514.65.4.757

Unaltered original publications

The four original publications discussed in this dissertation can be found below. The publications were downloaded from the website of the respective journal and have not been adapted or otherwise altered. To find further information on the respective publications or to access the supplementary materials, the sources of all publications are listed in the section "List of references for Study 1 to 4".

Computers and Education: Artificial Intelligence 4 (2023) 100126

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy

Matthias Carl Laupichler *, Alexandra Aster, Tobias Raupach

Institute of Medical Education, University Hospital Bonn, Bonn, Germany

ARTICLE INFO

Keywords:
AI literacy
Assessment
Scale
Questionnaire
Delphi study
Validity

ABSTRACT

Artificial intelligence literacy is a concept that has been the focus of exhaustive research recently. However, there are very few psychometrically sound and thoroughly evaluated instruments that attempt to assess AI literacy in a valid way. Therefore, this study presents an item set to assess the AI literacy of non-experts. In the context of a Delphi expert study, 53 subject matter experts participated in three iterative questionnaire rounds to generate potential AI literacy items and assess their content validity. In addition, the experts made suggestions on how the items' wording accuracy could be improved and evaluated the wording suggestions of the other experts. Of 47 potential items, 38 were judged relevant for inclusion in a final AI literacy questionnaire. The result is one of the first freely available AI literacy item sets and represents an important first step in assessing AI literacy and its subconstructs. Finally, the development of the items through the execution of an iterative Delphi study and the strong focus on content validity contribute to the advancement of AI literacy theory.

1. Introduction

1.1. AI literacy and its relevance

Although the importance of AI literacy research has increased in recent years, there is still no commonly accepted definition of AI literacy. However, one of the most commonly cited definitions is that of Long & Magerko, who describe AI literacy 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, p. 2).

The importance of an AI literate population is growing as more and more aspects of personal and professional life are permeated by AI. On the one hand, individuals engage (consciously or unconsciously) with AI-based applications in their spare time, such as smart speakers (Bentley et al., 2018), face recognition (Adjabi et al., 2020), or recommender systems for web-applications (Zhang et al., 2019). On the other hand, AI applications are also increasingly finding their way into the workplace, and employees have to learn how to deal with these novel systems (Chowdhury et al., 2022) to ensure they can continue to work in decent work environments. (Braganza, Chen, Canhoto, & Sap, 2021).

1.2. Assessing AI literacy and related concepts

Several efforts have been made to develop scales to capture constructs related to AI literacy. However, they mostly deal with the affective component of AI collaboration and cannot be used to capture AI literacy itself. Examples include the "Attitudes Towards Artificial Intelligence" Scale (Sindermann et al., 2021), the "General Attitudes Towards Artificial Intelligence" Scale (Schepman & Rodway, 2022), as well as the "Artificial Intelligence Anxiety Scale" (Wang & Wang, 2022).

In contrast to the assessment of attitudes towards AI, there are few research projects that seek to advance the psychometrically valid measurement of AI literacy. In order to capture the current state of research on methods of AI literacy assessment, we conducted a brief literature review with five search terms synonymous with "ai literacy scale" in five different databases. We ran a first search in April 2022, and a second search with the same terms in the same databases in October 2022. The initial search yielded ten results, whereof two were published in another language than English and two called for the creation of AI literacy scales (Ng et al., 2021a, 2021b). In the remaining six papers, an AI literacy scale was developed as a means to evaluate the learning outcomes of specific educational programs. However, these scales have not been psychometrically evaluated (de Souza, 2021; Kong et al., 2021) and were often developed specifically for particular courses (Dai et al.,

* Corresponding author. Institute of Medical Education, University Hospital Bonn, Venusberg-Campus 1, 53127, Bonn, Germany.
E-mail addresses: matthias.laupichler@ukbonn.de (M.C. Laupichler), alexandra.aster@ukbonn.de (A. Aster), tobias.raupach@ukbonn.de (T. Raupach).

https://doi.org/10.1016/j.caeai.2023.100126
Received 21 November 2022; Received in revised form 10 January 2023; Accepted 10 January 2023
Available online 12 January 2023

2020). In addition, some authors seemed to have a different understanding of AI literacy. For example, Lin et al. (2021) and Shih et al. (2021) reported an AI literacy scale that contains the two factors "teamwork" and "attitudes toward AI" and thus does not reflect the central aspects of the AI literacy definitions reported above. Finally, although some authors provide examples of the items in the scale (Kong et al., 2021), most papers do not include the entire scale, making it difficult for other researchers to replicate the results.

The second search yielded one additional result. Wang et al. (2022) introduced the first psychometrically evaluated "Artificial Intelligence Literacy Scale", consisting of twelve items on four dimensions (i.e., "awareness", "usage", "evaluation", and "ethics"). In many ways, this scale represents a significant improvement over the previously developed scales in that it approaches the development of an instrument to measure AI literacy in a structured and methodically sound manner. It must be mentioned, however, that the "Artificial Intelligence Literacy Scale" was developed as "a valid and reliable scale to measure people's AI literacy for future [human-AI interaction] research" (Wang et al., 2022, p. 5). Human-AI interaction (HAII) research has traditionally viewed the use of AI from the perspective of program design rather than user capabilities (Amershi et al., 2019). While focusing on HAII is a legitimate approach, the extent to which the "Artificial Intelligence Literacy Scale" is valid outside the HAII research domain is debatable. A general AI literacy scale that can be used universally should be applicable in all research areas related to AI literacy. As an illustration, a general method for measuring AI could be used to assess AI literacy before and after attending an introductory AI-course.

The underlying definition, which was formulated by the authors themselves, differs in some respects from established definitions such as that of Long and Magerko (2020). Most importantly, the strong focus on AI awareness is certainly relevant, but somewhat neglects AI-knowledge and -understanding, which constitutes a main aspect of most AI literacy definitions (Ng et al., 2021b). Thus, we see the need for a universally applicable AI literacy scale.

1.3. Non-experts as target group

Similar to other technological literacies like data literacy (Wolff et al., 2016) or computational literacy (Jacob & Warschauer, 2018), AI literacy is commonly used to describe the competencies of non-experts rather than (AI-) professionals. Non-experts are defined as individuals who have not received formal training in AI and are using AI applications rather than developing them. Thus, in general, all adults interacting with sophisticated and modern digital applications can be considered non-experts, as it can be assumed that most of today's digital applications are at least partially based on AI algorithms. In our interpretation, a non-expert is on one of the two lower levels of the framework proposed by Faruqe et al. (2021). Therefore, he or she is either a "consumer […] who uses the outputs of AI to improve their work or life" or a "co-worker [who] knows the basics of how the AI systems work and uses AI outputs in the work" (Faruqe et al., 2021, p. 1). Although (younger) children are no AI experts, they are not included in the target group of the items either, as they are not yet consumers or co-workers.

1.4. Aim of this study

The aim of this study was to generate a face and content valid AI literacy item set that can be used to develop a scale to assess the AI literacy of non-experts. To evaluate the relevance of the different items for the assessment of AI literacy and to determine its content validity, we conducted an expert Delphi study with subject matter experts (SMEs). Content validity can be defined as "the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose." (Haynes et al., 1995, p. 238).

We developed a primary research question, which was divided into two subquestions (1a and 1b). The first question was related to the content validity of potential items on AI competence. Therefore, research question 1a was.

RQ1a. Which items are relevant for and representative of AI literacy?

The second subquestion related to the wording of the items. Since the items on AI literacy must contain all necessary information without including superfluous or irrelevant aspects, research question was 1b.

RQ1b. How can the items be rephrased to most accurately represent the construct of AI literacy?

2. Method

An iterative, three-round Delphi expert study (Hsu & Sandford, 2007) was conducted to develop a face and content valid AI literacy item set. Using a Delphi study allowed for very elaborate validity testing. This is primarily due to the fact that the Delphi methodology is a multistage, iterative procedure. This has the advantage that the experts involved can take into account each other's opinion and thus a consensus can be reached. In addition, the participant pool in Delphi surveys consists of several experts. This means that it is not the opinions of individual persons that count, but the assessments of a large group that is very well versed in the field.

Ethical approval to perform this study was granted by the Ethics Committee of the University of Bonn, Germany (Reference 194/22).

2.1. Expert panel

2.1.1. Expert panel selection

We contacted a total of 471 potential SMEs in the field of AI literacy and AI education by email (see Fig. 1). The SMEs' contact details were retrieved from three sources. 400 contacts originated from the "Networking Event: AI in University Education 2022" and 50 from the "AI Networking Event North Rhine-Westphalia 2022", both of which were organized by the German Federal Ministry of Education and Research (BMBF). The other 21 people were members of smaller AI working groups in which the first author participated. Since the participants of these events were engaged in AI education on a professional basis (mostly as researchers or lecturers), they could be considered AI experts. Nevertheless, the actual expertise was assessed at a later stage (see section 3.1). In addition to AI expertise, it can be assumed that the experts had good pedagogical and didactic skills, since most of them were either lecturers from the university sector or worked at the intersection of AI and education. Although some participants may have been more knowledgeable or skilled in one of the two areas (i.e., AI or pedagogy), this population still provided the best opportunity to reach a reasonably large sample of participants. Of all those contacted, 85 prospective participants (18%) completed a brief registration survey. In the actual Delphi study rounds, 59 (Round 1), 55 (Round 2), and 53 (Round 3) SMEs participated. Thus, the dropout rate was 5% between rounds 1 and 2 and 4% between rounds 2 and 3.

2.1.2. Experts' characteristics

Most participants (N = 53) answered all or part of the sociodemographic questions asked at the beginning of Round 3. About two in five (41%, N = 22) of the SMEs identified as being female, 53% (N = 28) as being male, 2% (N = 1) as "other", and two did not wish to disclose their gender. Most SMEs (43%, N = 23) were between 30 and 39 years old, with the youngest participants being between 18 and 29 years old and the oldest being between 50 and 65 years old. The majority of study participants had a Master's degree (55%, N = 29), while 26% (N = 14) had doctorate degrees and 15% (N = 8) were professors. Nearly two thirds of the SMEs worked at a university (60%, N = 32) and 23% (N = 12) worked at a university of applied sciences. In addition, some SMEs did not work at a research or educational institution (13%, N = 7) or worked at another form of research or educational institution (4%, N =

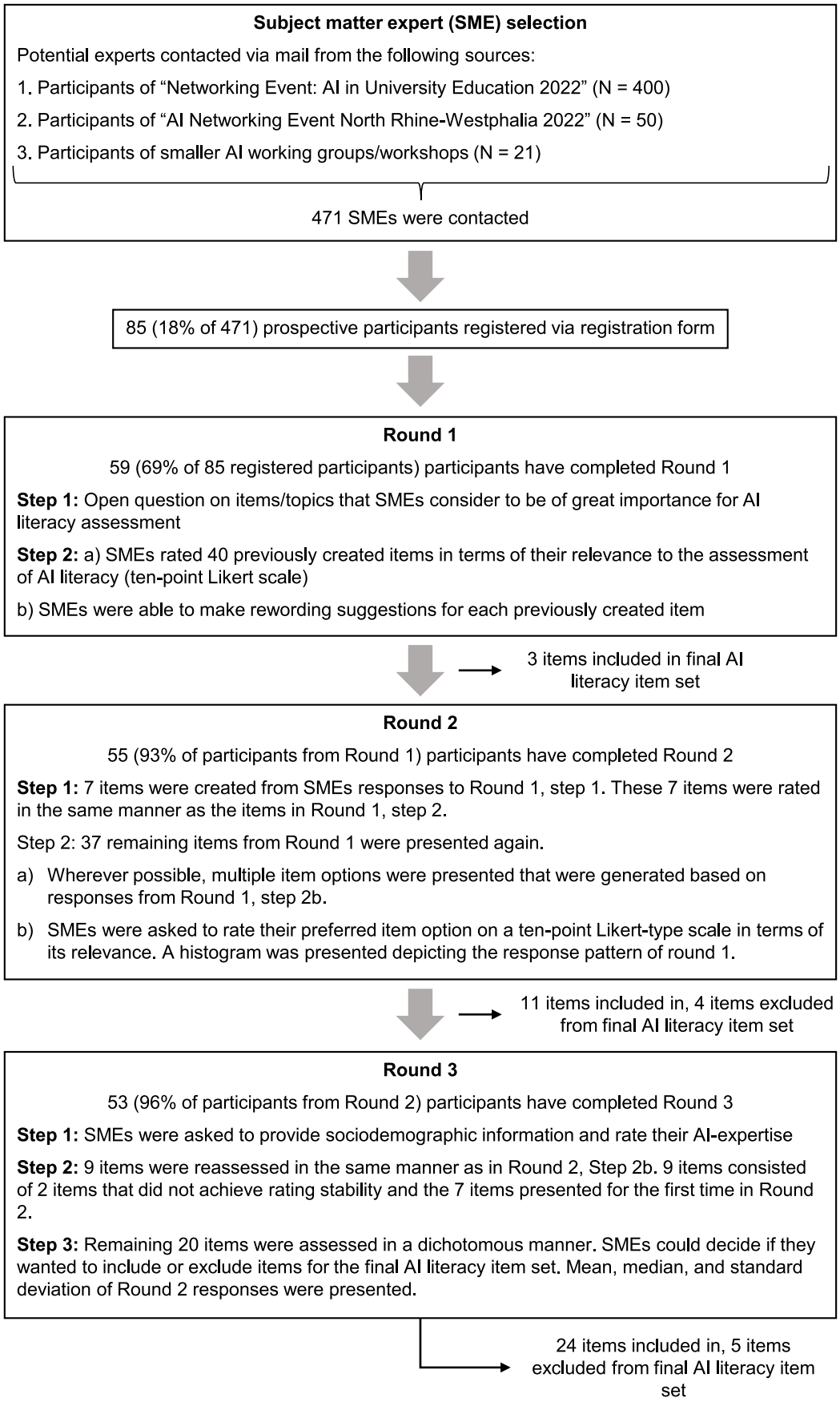


Fig. 1. Delphi procedure across three iterative rounds, including subject matter expert selection.

2).

Most participants stated that they had either a “good understanding of how AI works and where it is used” (43%, N = 23) or a “deep understanding of AI” (36%, N = 19) (see Table 1 for a detailed breakdown of the response frequencies). Participants reported dealing with AI once a week (28%, N = 15), almost every workday (36%, N = 19), or on a daily basis (23%, N = 12). The majority of the expert panel members had been working in the field of AI for “1–3 years” (45%, N = 24), followed by “3–10 years” (26%, N = 14). Nearly all of the participants were German native speakers, with one participant rating his or her German language skills at C2-level.[1] Moreover, 68% (N = 36) rated their English language skills to be on the C-level (36%, N = 19 for C1, and 32%, N = 17 for C2, respectively). Some participants self-assessed their English language skills to be on the B2-level (25%, N = 13). All participants lived and worked in Germany.

Table 1
Subject matter experts‘ characteristics (N = 53).

Response options	N	%
Own AI-Expertise		
No AI knowledge/experience at all	0	0
Basic idea of what AI is	7	13.2
Good understanding of how AI works, where it is used, etc.	23	43.4
Deep understanding of AI (conducted initial AI research/ development/knowledge accumulation)	19	35.8
Very deep understanding of AI (several years of intensive AI research/development/knowledge accumulation)	4	7.5
Frequency of engagement with AI		
Almost never	1	1.9
Less than once a week	6	11.3
Approximately once a week	15	28.3
Almost every (working) day	19	35.8
Every day	12	22.6
Duration of engagement with AI		
0 to 1 year	10	18.9
1 to 3 years	24	45.3
3 to 10 years	14	26.4
More than 10 years	5	9.4
Age		
18 to 29	12	22.6
30 to 39	23	43.4
40 to 49	9	17.0
50 to 65	9	17.0
Older than 65	0	0.0
Gender		
Female	22	41.5
Male	28	52.8
Other	1	1.9
Not specified	2	3.8
Highest level of education		
Secondary school leaving certificate	0	0.0
High school diploma	0	0.0
Bachelor’s degree	1	1.9
Master’s degree	29	54.7
Doctorate/PhD	14	26.4
Professorship	8	15.1
Other	1	1.9
Type of employment		
No employment at research or educational institution	7	13.2
University	32	60.4
University of Applied Sciences	12	22.6
Other type of educational/research institution	2	3.8

Note: N = Number of SMEs that chose this response option. % = Percentage of this response option in the total sample.

[1] A1 stands for the lowest language proficiency level, C2 for the highest language proficiency level. Individuals at the "A" level are considered basic users, individuals at the "B" level are considered independent users, and individuals at the "C" level are considered proficient users. The meaning of each level was explained to the SMEs in the questionnaire.

2.2. Procedure

In a first step preceding the Delphi study, an initial set of 40 AI literacy items was created. For this purpose, well-known and relevant AI literacy courses such as "Elements of AI" (University of Helsinki & MinnaLearn, 2018; www.elementsofai.com), "AI for Everyone" (Ng & DeepLearning.AI, 2022; www.coursera.org/learn/ai-for-everyone), "Introduction to AI" (Waldmann et al., 2020, www.ki-campus.org/courses/einfuehrungki2020) and books such as “Human + Machine” (Daugherty & Wilson, 2018) and “Artificial Intelligence: The Insights You Need from Harvard Business Review” (Davenport et al., 2019) were reviewed in an unsystematic manner to identify recurring content. Key terms from the various sources were collected and compared. Terms that appeared in at least two independent sources were transformed into items. In addition, Long & Magerko’s (2020) AI literacy framework with its 16 AI competency domains was used as a further basis for item generation. To avoid a rigid classification of each item into one of the 16 competencies, the framework was only used as an implicit decision support tool. Although “it is both a common and an acceptable modification of the Delphi process format to use a structured questionnaire in Round 1 that is based upon an extensive review of the literature” (Hsu, 2007, p. 2), we wanted to ensure that the preliminary themes identified reflected the most important AI constructs. Therefore, the topics were discussed with a small convenience sample of AI experts (N = 5) to generate the items presented in round 1.

The actual Delphi study was conducted online via the questionnaire tool “evasys” (evasys Giannarou & Zervas, 2014).

In the first round, participants were given a common definition of AI literacy (definition by Long & Magerko, 2020) in order for all participants to be able to share one definition of the underlying construct (please find the questionnaires for all three rounds in the original German version as well as in the English translation in Supplementary Material 1). In addition, it was explained to the SMEs for which target group the questionnaire will be designed and what exactly the term "non-experts" means (see section 1.3). Subsequently, participants were asked to enter their own ideas regarding topics and items that would be highly relevant for an AI literacy scale in a text box (see Fig. 1). This question was presented before the evaluation of the initially generated items in order to avoid possible influencing effects. Accordingly, participants were then asked to rate the items in terms of their relevance to an AI literacy scale. Relevance was rated on a ten-point Likert-type scale from 1 ("not relevant at all") to 10 ("very relevant"). There was an option to abstain ("no answer"). After each item, participants could indicate whether they wanted to suggest a rewording ("Would you reword, change, add, or shorten the preceding item?"). If they clicked "Yes," a text box appeared in which they could enter their suggestions.

At the beginning of Round 2, items that were generated from the free-text responses at the beginning of Round 1 were presented. The rating procedure was the same as for the items in Round 1. Afterwards, all items from round 1 were presented again for which no final decision regarding inclusion/exclusion could be made (see Fig. 2 for an overview of the inclusion/exclusion decision process). The procedure was structured as follows. Wherever possible, multiple item options (i.e., slightly different versions of the same item) were created based on the rewording suggestions from Round 1. Participants could first select their preferred item wording and then rate the preferred version on a ten-point Likert scale. As additional information, the rating results from Round 1 were presented as a histogram.

In Round 3, the first step was to gather some information about the SMEs themselves. This information included age, gender, highest level of education, country of main affiliation, and if they worked in a research or education facility. In addition, the experts were asked to rate

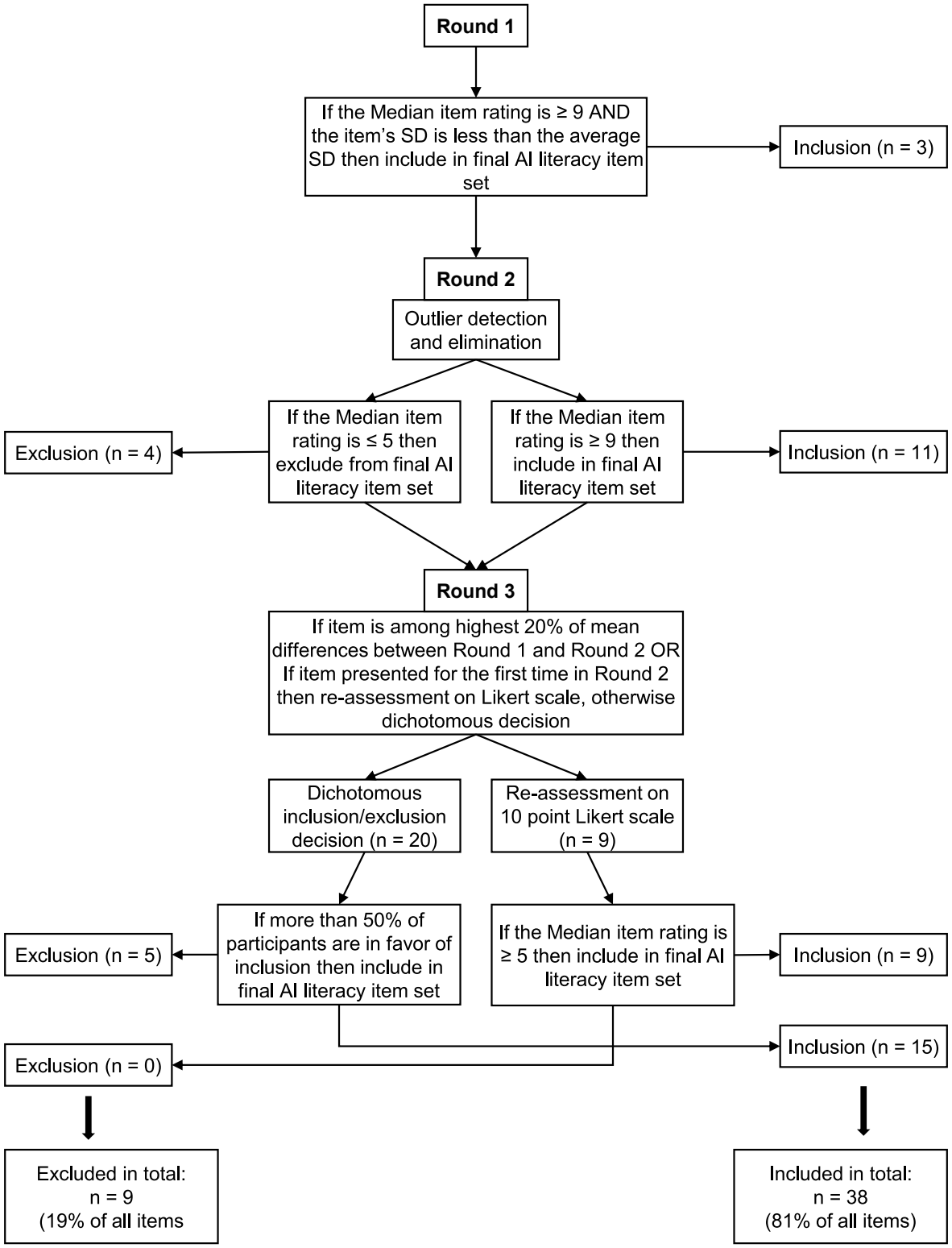


Fig. 2. Rules for including items in/excluding items from the final AI literacy item set and number of items included/excluded due to these rules.
Note. "Inclusion" means inclusion in the final AI literacy item set, "exclusion" means exclusion from the final AI literacy item set.

their English and German proficiency on a single question (from A1 to C2, according to the levels of the "Common European Framework of Reference", Council of Europe, 2022) since the instructions were presented in German but the items themselves in English. Finally, the participants were asked to rate their own AI expertise and to indicate since when and how regularly they have been involved with AI.

Subsequently, items were reassessed whose response patterns had not achieved stability over the first two rounds. In addition, the items that were presented for the first time in Round 2 were reassessed. The remaining items, which could not yet be included or excluded from the final item set, but already showed a stable response pattern, were assessed in a dichotomous manner. This means that the SMEs were able

to make a final decision on whether the items should be included or excluded. To support the decision, the mean, median, and standard deviation of the Round 2 ratings were provided (as text) next to the item text.

2.3. Consensus criteria

Initially, we had planned to use a fixed consensus criterion comparable to that proposed by Giannarou and Zervas (2014). In our case, this would have been a standard deviation of less than 1 (please refer to OSF-preregistration https://doi.org/10.17605/OSF.IO/B7R4H). However, during the evaluation of the first round, it turned out that the experts' opinions differed too much to apply a fixed consensus criterion. Therefore, we generated a set of hierarchically organized rules that determined whether an item should be presented again in the next round or not. In summary, an item was presented again if it could not yet be definitively decided whether the respective item should be included in or excluded from the final item set (see Fig. 2 for a detailed description of the consensus rules).

2.4. Data analysis

The mean, median, and standard deviation for each item were calculated after each round. In addition, histograms were created to summarize the response pattern of the previous round for the participants in a structured way. To calculate the stability of the response pattern, the mean difference between round 1 and round 2 was calculated (as an absolute value). This calculation was performed to determine if there was a stable dissent of expert responses (i.e., SMEs stick to their assessments despite differing opinions). If this was the case, further presentation of the items was considered unnecessary and they were included or excluded based on the criteria described above. After Round 2, a boxplot was created for each item to identify and subsequently eliminate potential outliers. All values that fell outside of the whiskers (max. 1.5 x interquartile range) were treated as outliers. Data analysis was performed using SPSS Statistics (IBM, 2022).

3. Results

3.1. Pool of potentially useful items

Forty items were generated by analyzing AI introductory courses and books. The items included were chosen to assess core competencies of AI that appeared repeatedly in the various popular science courses and books. Examples that occurred in at least two of the sources mentioned in section 2.2 and were therefore included as items in the preliminary item pool included "I can tell if the things I use frequently are supported by artificial intelligence." or "I can describe what a Turing test is supposed to find out." (both initial item wording prior to rewording suggestions). In addition to these 40 items, another sample of items was generated by analyzing the SMEs' responses to an open question posed at the very beginning of Round 1. Somewhat unsurprisingly, most topics or items that were suggested by the SMEs were already covered by items generated in advance. However, seven items that were not included in the initial item set were added to the pool of items which could be potentially relevant for assessing AI literacy. Thus, a total of 47 items were evaluated by the SMEs regarding their relevance throughout the three rounds.

3.2. Relevance of potential items and decision on inclusion in the final scale

44 of the 47 items were rated at least twice on a ten-point Likert scale. The remaining three items were rated only once, as the evaluation resulted in a median of 10 in the first round, while showing low variability. Thus, in the case of the three items, the experts agreed very early on that they were important for the assessment of AI literacy. The three items were: "I can describe risks that may arise when using artificial intelligence systems.", "I can explain why data plays an important role in the development and application of artificial intelligence.", and "I can identify ethical issues surrounding artificial intelligence."

After Round 2 and the elimination of outlier values, 11 items whose median was $\geq$9 were included in the final item set, whereas four items whose median was $\leq$5 were excluded from the final item set.

In Round 3, two items had to be presented again due to a lack of response stability (i.e., high difference between rounds). Those items were "I can describe the potential impact of artificial intelligence on the future." and "I can explain how sensors are used by computers to collect data that can be used for AI purposes." Both were subsequently included in the final item set. Of the 20 items assessed in a dichotomous manner (i.e., include or exclude), 15 items were included in the final item set. Finally, all of the seven items presented for the first time in Round 2 were included in the item set, since all of them had a median of $\geq$6.

In summary, a total of 47 items were evaluated regarding their relevance for inclusion in an AI literacy scale. Of these, 38 items were rated as relevant and representative for AI literacy, while nine items were not included in the final item set due to lack of relevance (see Table 2).

3.3. Validity of item wording

In addition to generating potentially relevant AI literacy items and evaluating them, the Delphi study had a third purpose. To verify that the items proposed by the research team or SMEs were worded as clearly, concisely, and validly as possible, the SMEs were given the opportunity to make rewording suggestions for each item. The rewording suggestions were evaluated by members of the research team. The first step was to check whether the entry was an actual rewording or improvement proposal. An example of a comment that was interesting but did not contain a rewording suggestion was "I think many experts don't know this term". In a second step, it was examined whether the suggestion met the purpose of the proposed assessment. For example, some SMEs suggested changing certain items to an item that would test respondents' knowledge of AI. While this would also be an interesting research project, it goes beyond the scope of the work presented here. Finally, the number of times a particular rewording suggestion was made was counted. Frequent mentions were presented as alternative item options. As an example for the different item options presented to the SMEs, one can look at item #6, which originally said: "I can distinguish media representations of AI (e.g., in movies or video games) from realistic AI.". The alternative item options generated based on the SMEs' responses were "I can distinguish science fiction representations of AI (e.g., in movies or video games) from real AI." and "I can evaluate whether media representations of AI (e.g., in movies or video games) go beyond the current capabilities of AI technologies." Alternative item options could not be generated for each item. Nevertheless, at least one to a maximum of three alternative options were created for 66% of all items ($N_{item} = 31$). Of these alternatives, the preferred version was selected by each participant in the following rounds.

This approach further increased the scale's content validity by ensuring that no important item content was omitted or that the inclusion of unnecessary item content negatively affected the relevance of the item.

4. Discussion

We conducted a Delphi expert study to generate an item set that supports the development of a scale for the assessment of non-experts' AI literacy.

Table 2
Mean (M), median (Mdn), and standard deviation (SD) for all items across all 3 rounds.

	Item	Round 1			Round 2			Round 3			Inclusion/ Exclusion,
	I can …	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*	Round
1	tell if the technologies I use are supported by artificial intelligence.	8.9	9	1.49	9.0	9	1.26	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
2	name examples of technical applications that are supported by artificial intelligence.	8.2	9	2.31	9.1	9	0.99	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
3	explain the differences between human and artificial intelligence.	7.6	8	2.52	8.6	9	1.22	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
4	describe how artificial intelligence consists of an interplay of complex algorithms and mathematical formulas.	5.4	6	2.65	4.6	4	2.07	n.a., f.	n.a., f.	n.a., f.	Excluded, Round 2
5	explain the difference between general (or strong) and narrow (or weak) artificial intelligence	7.2	8	2.23	7.1	7	1.69	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
6	evaluate whether media representations of AI (e.g., in movies or video games) go beyond the current capabilities of AI technologies.	7.2	8	2.32	7.8	8	1.32	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
7	explain what is meant by the term singularity in the context of artificial intelligence.	4.8	4	2.60	3.4	3	1.54	n.a., f.	n.a., f.	n.a., f.	Excluded, Round 2
8	name weaknesses of artificial intelligence.	8.6	9	1.94	8.9	9	1.04	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
9	name strengths of artificial intelligence.	8.2	9	1.88	8.6	9	1.23	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
10	describe risks that may arise when using artificial intelligence systems.	9.2	10	1.25	n. a., f.	n.a., f.	n.a., f.	n.a., f.	n.a., f.	n.a., f.	Included, Round 1
11	describe advantages that can come from using artificial intelligence systems.	8.0	8	1.88	8.1	8	1.22	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
12	describe the potential impact of artificial intelligence on the future.	6.2	6	2.77	7.1	7	1.82	7.8	8	1.46	Included, Round 3
13	distinguish AI applications that already exist from AI applications that are still in the future.	7.3	8	2.23	7.7	8	0.94	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
14	describe what knowledge representation means.	6.0	7	2.71	6.3	7	2.15	n.a., d.	n.a., d.	n.a., d.	Excluded, Round 3
15	explain how AI applications make decisions.	7.3	8	2.15	7.8	8	1.33	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
16	explain how AI-expert systems work.	6.3	6	2.48	5.7	6	2.24	n.a., d.	n.a., d.	n.a., d.	Excluded, Round 3
17	explain how machine learning works at a general level.	7.2	8	2.26	7.8	8	1.44	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
18	describe how machine learning models are trained, validated, and tested.	6.8	7	2.46	7.3	7	1.73	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
19	explain the difference between 'supervised learning' and 'unsupervised learning' (in the context of machine learning).	7.0	7	2.25	7.3	8	1.96	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
20	explain how 'reinforcement learning' works on a basic level (in the context of machine learning).	6.2	7	2.51	6.5	7	1.82	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
21	explain how deep learning relates to machine learning.	6.3	7	2.67	7.2	7	1.59	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
22	explain what the term 'artificial neural network' means.	6.9	7	2.24	7.6	8	1.01	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
23	critically evaluate the implications of artificial intelligence applications in at least one subject area.	8.1	8	1.93	8.6	9	1.02	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
24	explain why data plays an important role in the development and application of artificial intelligence.	9.2	10	1.32	n. a., f.	n.a., f.	n.a., f.	n.a., f.	n.a., f.	n.a., f.	Included, Round 1
25	describe why humans play an important role in the development of artificial intelligence systems.	8.4	9	2.12	9.0	9	1.13	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
26	describe how some artificial intelligence systems can act in their environment and react to their environment.	7.1	7	2.08	6.8	7	1.71	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
27	explain how sensors are used by computers to collect data that can be used for AI purposes.	6.2	7	2.72	7.3	8	1.8	6.4	6	2.06	Included, Round 3
28	name applications in which AI-assisted computer vision is used.	6.6	7	2.61	6.5	7	1.81	n.a., d.	n.a., d.	n.a., d.	Excluded, Round 3
29	name applications in which AI-assisted natural language processing/ understanding is used.	6.9	7.5	2.66	7.3	7	1.72	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
30	identify ethical issues surrounding artificial intelligence.	9.1	10	1.36	n. a., f.	n.a., f.	n.a., f.	n.a., f.	n.a., f.	n.a., f.	Included, Round 1
31	explain what the term 'black box' means in relation to artificial intelligence systems.	7.9	9	2.29	8.7	9	1.23	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
32	describe how biases arise in AI systems.	8.4	9	1.86	9.3	10	1.00	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
33	critically reflect on the potential impact of artificial intelligence on individuals and society.	7.6	8	2.37	8.7	9	1.23	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
34	give a short overview about the history of artificial intelligence.	4.2	4	2.44	2.4	2	1.22	n.a., f.	n.a., f.	n.a., f.	Excluded, Round 2
35	explain what the term 'artificial intelligence winter' means.	3.8	3	2.46	1.7	2	0.65	n.a., f.	n.a., f.	n.a., f.	Excluded, Round 2
36	explain why AI has recently become increasingly important.	7.1	7	1.94	7.3	8	1.57	n.a., d.	n.a., d.	n.a., d.	Included, Round 3

(*continued on next page*)

Table 2 (*continued*)

	Item	Round 1			Round 2			Round 3			Inclusion/ Exclusion,
	I can …	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*	Round
37	describe what a Turing test is supposed to find out.	5.9	6	2.73	5.2	6	2.36	n.a., d.	n.a., d.	n.a., d.	Excluded, Round 3
38	explain how rule-based systems differ from machine learning systems.	7.1	7	2.47	7.6	7.5	1.61	n.a., d.	n.a., d.	n.a., d.	Included, Round 3
39	explain how decision tree systems work.	6.4	7	2.35	7.1	7	1.57	n.a., d.	n.a., d.	n.a., d.	Excluded, Round 3
40	assess if a problem in my field can and should be solved with artificial intelligence methods.	7.5	8	2.05	8.4	9	1.55	n.a., f.	n.a., f.	n.a., f.	Included, Round 2
41	describe what artificial intelligence is.	n.a., 2nd	n.a., 2nd	n.a., 2nd	8.3	9	1.94	9.5	10	0.64	Included, Round 3
42	describe the concept of explainable AI.	n.a., 2nd	n.a., 2nd	n.a., 2nd	7.8	8	1.69	8	8	1.44	Included, Round 3
43	* explain why data security must be considered when developing and using artificial intelligence applications. & explain why data privacy must be considered when developing and using artificial intelligence applications.	n.a., 2nd	n.a., 2nd	n.a., 2nd	8.4	9	1.94	8.75	9	1.59	Included, Round 3
44	describe the concept of big data.	n.a., 2nd	n.a., 2nd	n.a., 2nd	7.8	8	2.04	8.2	8	1.26	Included, Round 3
45	give examples from my daily life (personal or professional) where I might be in contact with artificial intelligence.	n.a., 2nd	n.a., 2nd	n.a., 2nd	8.8	9	1.58	9	9	0.98	Included, Round 3
46	explain what an algorithm is.	n.a., 2nd	n.a., 2nd	n.a., 2nd	7.5	8	2.15	8	8	1.6	Included, Round 3
47	describe potential legal problems that may arise when using artificial intelligence.	n.a., 2nd	n.a., 2nd	n.a., 2nd	7.1	7.5	2.40	7.4	8	1.65	Included, Round 3

n.a., f. = Not applicable, final decision was made. n.a., d. = Not applicable, dichotomous decision. n.a., 2nd = Not applicable, item created after first round.
Note: The items presented here represent the final item options which were selected by the participants. The final seven items were generated from the initial responses from Round 1, which is why the statistical characteristics for these items are reported starting with Round 2. After Round 1, three items were included in the final questionnaire version because the median score was ten. After Round 2, 11 items were included in the final questionnaire version because the median score was at least nine. In addition, after Round 2, four items were excluded from the final questionnaire version because the median score was five or lower. *At the beginning, "data security" and "data privacy" were combined in one item. However, the SMEs decided that this item should be divided into two items.

4.1. Significance of the findings

As described earlier, a strong increase in scientific AI literacy publications and popular scientific AI literacy courses, books, etc. has been observed in recent years (Laupichler et al., 2022; Long & Magerko, 2020; Ng et al., 2021a). While there are various efforts to improve AI literacy of non-experts, there is no way to assess individuals' AI literacy, which has a detrimental effect on AI literacy research. Therefore, this research project was conducted to support the development of one of the first measurement tools to assess AI literacy in non-experts.

The primary concern in this study was to assess the content validity of the items in a reliable manner. While content validity is the basic prerequisite for the existence of a meaningful questionnaire and should accordingly be given the highest priority (Zamanzadeh et al., 2014), it is often only evaluated through methodologically problematic procedures, or disregarded completely. Especially for a topic as new and complex as AI literacy, simply assessing content validity by a small sample of SMEs would be problematic. This is especially true when the experts are not selected from a large pool of potential participants, but are personally chosen by the researchers, which can, for example, lead to selection bias (Blackwell & Hodges, 1957). To circumvent this problem, we contacted over 450 potential experts, of whom 53 contributed their heterogeneous opinions.

It must be mentioned in this regard that the experts rated 81% of all items (38 out of 47) as relevant for capturing AI literacy. On the one hand, this could mean that the expert evaluation or the exclusion criteria were too insensitive. On the other hand, it could also be that this large number of items is necessary to validly capture the rather complex model of AI literacy.

Another interesting finding is that attitudes or affective components toward AI do not appear in the item set generated in this study. This is true for both the initial 40 items and the items suggested by the SMEs. Thus, the item set differs from the AI scales presented in the theory section. While some of these were developed specifically to assess AI attitudes (Sindermann et al., 2021; Schepman & Rodway, 2022), even the scales primarily developed to assess AI literacy often contain some items covering affective components. For example, Lin et al. (2021) and Shih et al. (2021) reported an AI literacy scale with two factors, "teamwork" and "attitudes toward AI." The item set presented here contains some items that could be loosely connected to the "teamwork" factor, for example "I can describe why humans play an important role in the development of artificial intelligence systems." or "I can assess if a problem in my field can and should be solved with artificial intelligence methods." However, no items from this item set seem fit to the proposed "attitudes toward AI"-factor, although this has to be examined further by conducting factor analyses. This is consistent with the content of most AI literacy definitions, which are more concerned with knowledge and understanding of AI, its application, evaluation, and creation, and AI ethics (Ng et al., 2021b).

The most recent scale, which is also the only one that has been psychometrically studied ("Artificial Intelligence Literacy Scale", Wang et al., 2022), does not include attitude items. However, as already described, it does not contain many items that test understanding or knowledge about AI. The item set presented here has several items that can be interpreted as enabling individuals to assess their knowledge about AI and its most important subfields (e.g., machine learning). Exemplary items that are concerned with AI understanding would be "I can describe the concept of explainable AI." or "I can explain how deep learning relates to machine learning.". Since most researchers include a knowledge and understanding component in their definitions of AI literacy, it should be included in an AI literacy scale.

4.2. Strengths

This research project developed the first freely available item set for assessing the AI literacy of non-experts. This work thus forms the basis for the development of a psychometrically evaluated, generally applicable AI literacy assessment scale.

The primary strength of the research presented here is the elaborate face and content validation of the item set. While measures of external validity (i.e., construct validity, criterion validity) are usually evaluated in relative detail, too little attention is paid to content validity in the development of tests and questionnaires. By involving more than 50 experts and repeatedly evaluating the relevance, we achieved a high content validity of the item set, ensuring the representativeness of the items for AI literacy.

Another advantage of the item set presented here is that all of the items are listed in this article and can therefore be considered "open access". This is not the case with other AI literacy scales, as they describe the use of the scales but do not report the content (i.e., the items). Thus, other researchers cannot use the items for their own research or replicate the corresponding studies. Moreover, in our case, both the included and excluded items were reported, so that readers can evaluate whether they agree or disagree about the correctness of the SMEs' decisions.

4.3. Limitations

Even though the main objective of this study was to develop and validate an AI literacy item set, it can be considered a limitation that no factor analysis was performed using a test sample. Conducting an exploratory factor analysis would have the advantage of identifying the common factors underlying AI literacy (Mulaik, 2010). In addition to different benefits for questionnaire development and presentation, this could even support the development of AI literacy theories, as most proposed AI literacy subfields are currently based on purely theoretical considerations. Furthermore, with the help of factor analysis it would be possible to reduce the total number of variables (Wirtz & Nachtigall, 2004), which in turn would have a positive effect on participants' commitment and reduce respondent fatigue (Schatz et al., 2012). Therefore, it must be reiterated that the item set presented here is not a definitive AI literacy scale, but an item set whose applicability as an AI literacy scale in real-world settings can only be evaluated through future research.

Another issue that all AI literacy questionnaires encounter is the selection of the most appropriate AI literacy definition. Since a valid AI literacy item set is, by its nature, intended to "measure a representative sample of the subject matter" (APA, 2022), the definition of the item must be as precise and unchallengeable as possible. However, due to the plethora of different AI literacy definitions (e.g., Kandlhofer et al., 2016; Long & Magerko, 2020; Ng et al., 2021b), it is impossible to use a single universally valid definition as a basis. Theoretically, instead of using Long & Magerko's (2020) definition, we could have presented the SMEs with any other definition as a starting point. However, the choice of this particular AI literacy representation was not arbitrary. Rather, we used it because it is the most widely accepted and most frequently cited definition. This does not necessarily mean that it could not be improved, but at least it provides a commonly accepted foundation.

Finally, two methodological limitations have to be considered. First, the SME selection method resulted in a sample that was predominantly from academia and higher education. However, the opinions of representatives from other subpopulations, such as industry or secondary education, might reveal slightly different AI literacy items. Future research projects should therefore investigate the extent to which the item set can be usefully applied in areas outside of higher education. Second, the choice of the consensus criteria is rather uncommon when compared to other Delphi studies (see Table 1 in Giannarou & Zervas, 2014). Although the rules described in Fig. 2 reflect empirically based decisions, they nevertheless have the disadvantage of being based, at least in part, on decisions made by the research team. This, however, is due to the fact that the initially planned measure of consensus turned out to be infeasible in the context of this study (as described in section 2.3), which is why the alternative had to be deployed.

4.4. Future research directions

The next major step should be to distribute the item set to a larger normative sample. The data obtained from this can be used to further test the psychometric properties of item set and to develop a final (i.e. non-preliminary) AI literacy scale. This would entail conducting a factor analysis and reliability testing. In addition to psychometric evaluations, other questions arise. For example, it could be examined whether the finalized scale could also be used as a pre/post or then/post assessment for the evaluation of AI literacy courses. In addition, it would be useful to examine the extent to which AI literacy and attitudes toward AI or trust in AI are related. It can be hypothesized that an increase in AI literacy correlates with higher trust in AI, a relationship that has been found for scientific literacy as well (Einsiedel, 1994). Last but not least, the participating SMEs in this study themselves pointed out an interesting research direction, namely the development of a knowledge or skills test (as opposed to a psychological questionnaire). The item set presented in this work has the goal to enable the development of a scale for the assessment of the AI literacy of non-experts. In the future, however, it may become equally important to *test* the AI literacy of individuals, for example in the sense of a classic multiple choice knowledge test. Companies, among others, could use this knowledge or skills test to assess the AI literacy of applicants without bias, avoiding social desirable responses.

5. Conclusion

With the generation of the AI literacy item set, we responded to the call for ways to assess AI literacy, which was expressed by several researchers. The purpose of this study was to generate a set of potential items for assessing AI literacy and to test its representativeness for the AI literacy construct. Future research will examine the further psychometric properties of the item set. This concerns both an additional evaluation of validity by distributing the questionnaire to a sample population, as well as the testing of reliability and objectivity. We therefore want to encourage other research teams to use the item set as an preliminary assessment tool to further evaluate the questionnaire in an iterative manner.

Funding

This work was supported by the Open Access Publication Fund of the University of Bonn. The fund is paying the costs of the article processing charge and had no influence, substantive or otherwise, on the scientific process or content of this work.

Ethical approval

The study was conducted between July 2022 and October 2022 by the University Hospital Bonn and the University of Bonn, Bonn, Germany. Participation in the study was voluntary and participants gave their written informed consent to participate in the study. The study was approved by the Research Ethics Committee of the University of Bonn (Reference 194/22).

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

First and foremost, we would like to thank the subject matter experts for sharing their valuable and thoughtful insights with us. Furthermore, we would like to thank Mike Bernd, Elena Trunz, Marko Jovanovic and

Stephan Jonas for their support in the development of the initial item set. Last but not least, we would like to thank all colleagues who helped with the organization and evaluation of the study as well as with feedback regarding the manuscript: Manuel Müller, Jana Schirch, Saskia Zimmer, and all others.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.caeai.2023.100126.

References

Adjabi, I., Ouahabi, A., Benzaoui, A., & Taleb-Ahmed, A. (2020). Past, present, and future of face recognition: A review. *Electronics, 9*(8), 1–53. https://doi.org/10.3390/electronics9081188

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Conference on human factors in computing systems - proceedings*. https://doi.org/10.1145/3290605.3300233. May 2.

APA. (2022). Content validity. *APA Dictionary of Psychology*. https://dictionary.apa.org/content-validity.

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the Long-term Use of smart speaker Assistants. In *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* (pp. 1–24). https://doi.org/10.1145/3264901, 2(3).

Blackwell, D., & Hodges, J. L. (1957). Design for the Control of selection bias. *The Annals of Mathematical Statistics, 28*(2), 449–460.

Braganza, A., Chen, W., Canhoto, A., & Sap, S. (2021). Productive employment and decent work: The impact of AI adoption on psychological contracts, job engagement and employee trust. *Journal of Business Research, 131*, 485–494. https://doi.org/10.1016/j.jbusres.2020.08.018

Chowdhury, S., Budhwar, P., Dey, P. K., Joel-Edgar, S., & Abadie, A. (2022). AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework. *Journal of Business Research, 144*, 31–49. https://doi.org/10.1016/j.jbusres.2022.01.069

Council of Europe. (2022). Global scale - table 1 (CEFR 3.3): Common reference levels. *Common European Framework of Reference for Languages (CEFR)*. https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale.

Dai, Y., Chai, C. S., Lin, P. Y., Jong, M. S. Y., Guo, Y., & Qin, J. (2020). Promoting students'well-being by developing their readiness for the artificial intelligence age. *Sustainability, 12*(16). https://doi.org/10.3390/su12166597

Daugherty, P. R., & Wilson, H. J. (2018). *Human+ machine: Reimagining work in the age of AI*. Harvard Business Press.

Davenport, T. H., Brynjolfsson, E., McAfee, A., & Wilson, H. J. (2019). *Artificial intelligence: The insights you need from Harvard business review*. Harvard Business Press.

Einsiedel, E. F. (1994). Mental Maps of science: Knowledge and attitudes among Canadian adults. *International Journal of Public Opinion Research, 6*(1), 35–44. https://doi.org/10.1093/ijpor/6.1.35

Faruqe, F., Watkins, R., & Medsker, L. (2021). *Competency model approach to AI literacy: Research-based path from initial framework to model*. arXiv preprint arXiv:2108.05809.

Giannarou, L., & Zervas, E. (2014). Using Delphi technique to build consensus in practice. *Journal of Business Science and Applied Management, 9*(2).

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A Functional approach to concepts and methods. *Psychological Assessment, 7*(3).

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research and Evaluation, 12*, 10. https://doi.org/10.7275/pdz9-th90

Jacob, S. R., & Warschauer, M. (2018). Computational thinking and literacy. *Journal of Computer Science Integration, 1*(1). https://doi.org/10.26716/jcsi.2018.01.1.1

Kandlhofer, M., Hirschmugl-Gaisch, S., & Huber, P. (2016). Artificial intelligence and computer science in education: From Kindergarten to university. In *2016 IEEE frontiers in education conference (FIE)* (pp. 1–9).

Kong, S. C., Man-Yin Cheung, W., & Zhang, G. (2021). Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. *Computers and Education: Artificial Intelligence, 2*. https://doi.org/10.1016/j.caeai.2021.100026

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, C.-H., Yu, C.-C., Shih, P.-K., & Wu, L. Y. (2021). International Forum of educational Technology & Society STEM based artificial intelligence learning in general education for non-engineering Undergraduate students. *Technology in Society, 24*(3), 224–237. https://doi.org/10.2307/27032867

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Conference on human factors in computing systems - proceedings*. https://doi.org/10.1145/3313831.3376727. April 21.

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Taylor & Francis.

Ng, A., & DeepLearning.AI. (2022). In *AI for Everyone*. www.coursera.org/learn/ai-for-everyone.

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021a). AI literacy: Definition, teaching, evaluation and ethical issues. In *Proceedings of the association for information science and technology* (pp. 504–509). https://doi.org/10.1002/pra2.487, 58(1).

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

Schatz, R., Egger, S., & Masuch, K. (2012). The impact of test Duration on user fatigue and reliability of subjective Quality ratings. *Journal of the Audio Engineering Society, 60*(1), 63–73.

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards artificial intelligence scale. *Computers in Human Behavior Reports, 1*, Article 100014. https://doi.org/10.1016/j.chbr.2020.100014

Shih, P. K., Lin, C. H., Wu, L. Y., & Yu, C. C. (2021). Learning ethics in AI-teaching non-engineering undergraduates through situated learning. *Sustainability, 13*(7). https://doi.org/10.3390/su13073718

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., & Montag, C. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English Language. *KI-Künstliche Intelligenz, 35*(1), 109–118.

de Souza, C. E. C. (2021). *What if AI is not that fair? - understanding the impact of fear of algorithmic bias and AI literacy on information disclosure*. BI Norwegian Business School. Master thesis.

University of Helsinki, & MinnaLearn. (2018). In *Elements of AI*. https://www.elementsofai.com/.

Waldmann, A., Liebl, A., & Gerbert, P. (2020). In *Einführung in die KI*. www.ki-campus.org/courses/einfuehrungki2020.

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

Wang, Y. Y., & Wang, Y. S. (2022). Development and validation of an artificial intelligence anxiety scale: An initial application in predicting motivated learning behavior. *Interactive Learning Environments, 30*(4), 619–634. https://doi.org/10.1080/10494820.2019.1674887

Wirtz, M., & Nachtigall, C. (2004). *Deskriptive statistik* (3rd ed.). Juventa Verlag.

Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2016). Creating an understanding of data literacy for a data-driven Society. *Journal of Community Informatics, 12*(3), 9–26. www.ci-journal.net/index.php/ciej/article/view/1286.

Zamanzadeh, V., Rassouli, M., Abbaszadeh, A., Majd, H. A., Nikanfar, A., & Ghahramanian, A. (2014). Details of content validity and objectifying it in instrument development. *Nursing Practice Today, 1*(3), 163–171. http://npt.tums.ac.ir.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. In *ACM computing surveys*. Association for Computing Machinery. https://doi.org/10.1145/3285029. Vol. 52, Issue 1.

Computers in Human Behavior Reports 12 (2023) 100338

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Development of the "Scale for the assessment of non-experts' AI literacy" – An exploratory factor analysis

Matthias Carl Laupichler [a,*], Alexandra Aster [a], Nicolas Haverkamp [b], Tobias Raupach [a]

[a] Institute of Medical Education, University Hospital Bonn, Bonn, Germany
[b] Department of Psychology, University of Bonn, Bonn, Germany

ARTICLE INFO

Keywords:
AI literacy
AI competencies
AI literacy scale
AI literacy questionnaire
Assessment
Exploratory factor analysis

ABSTRACT

Artificial Intelligence competencies will become increasingly important in the near future. Therefore, it is essential that the AI literacy of individuals can be assessed in a valid and reliable way. This study presents the development of the "Scale for the assessment of non-experts' AI literacy" (SNAIL). An existing AI literacy item set was distributed as an online questionnaire to a heterogeneous group of non-experts (i.e., individuals without a formal AI or computer science education). Based on the data collected, an exploratory factor analysis was conducted to investigate the underlying latent factor structure. The results indicated that a three-factor model had the best model fit. The individual factors reflected AI competencies in the areas of "Technical Understanding", "Critical Appraisal", and "Practical Application". In addition, eight items from the original questionnaire were deleted based on high intercorrelations and low communalities to reduce the length of the questionnaire. The final SNAIL-questionnaire consists of 31 items that can be used to assess the AI literacy of individual non-experts or specific groups and is also designed to enable the evaluation of AI literacy courses' teaching effectiveness.

1. Introduction

Artificial intelligence (AI) is having an increasing impact on various aspects of daily life. These effects are evident in areas such as education (Zhai et al., 2021), healthcare (Reddy et al., 2019), or politics (König & Wenzelburger, 2020). However, AI is not only used in niche areas that require a high degree of specialization, but it is also integrated into everyday life applications. Programs like ChatGPT (OpenAI, 2023) provide free and low-threshold access to powerful AI applications for everyone. It is already becoming apparent that the use of these AI applications requires a certain level of AI competence that enables a critical appraisal of the programs' capabilities and limitations.

1.1. Defining AI literacy

These competencies are often referred to in the literature as *AI literacy*. There are several definitions of AI literacy, but one of the most commonly used can be found in a paper by Long and Magerko (2020), which lists 16 core AI literacy competencies. They define AI literacy 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" (p. 2). Furthermore, Ng et al. (2021a) state in their literature review that „instead of merely knowing how to use AI applications, learners should be inculcated with the underlying AI concepts for their future career, as well as the ethical concerns of AI applications to become a responsible citizen" (p. 507). Despite these and other attempts to define AI literacy, there is still no clear consensus on which specific skills fall under the umbrella term AI literacy. However, researchers seem to agree that AI literacy is aimed at *non-experts*, which are laymen who have not had specific AI or computer science training. These may be individuals who could be classified as consumers of AI, or individuals who interact with AI in a professional manner (Faruqe et al., 2021). Because of this somewhat ambiguous definitional situation, we propose the following AI literacy working definition: *The term AI literacy describes competencies that include basic knowledge and analytical evaluation of AI, as well as critical use of AI applications by non-experts.* It should be emphasized that programming skills are explicitly not included in AI literacy in this definition, since in our view they represent a separate set of competencies and go beyond AI literacy.

* Corresponding author. Institute of Medical Education, University Hospital Bonn, Venusberg-Campus 1, 53127, Bonn, Germany.
E-mail addresses: matthias.laupichler@ukbonn.de (M.C. Laupichler), alexandra.aster@ukbonn.de (A. Aster), nicolas.haverkamp@ukbonn.de (N. Haverkamp), tobias.raupach@ukbonn.de (T. Raupach).

https://doi.org/10.1016/j.chbr.2023.100338
Received 18 April 2023; Received in revised form 7 August 2023; Accepted 25 September 2023
Available online 27 September 2023

1.2. Assessing AI literacy

Although comparatively young, the field of AI literacy and AI education has been the subject of increasing research for several years (Kandlhofer et al., 2016, Cetindamar et al., 2022). In addition, there are many examples in the literature of courses and classes that strive to increase AI literacy of individuals at different levels of education, e.g., kindergarten (Su & Ng, 2023), high school (Ng et al., 2022), or university (Laupichler et al., 2022). However, few attempts have been made to develop instruments for assessing individuals' AI literacy. However, the existence of such instruments would be essential, for example, to evaluate the teaching effectiveness of the courses described above. Another advantage of AI literacy assessment tools would be the ability to compare the AI literacy of different subgroups (e.g., high school or medical students), identify their strengths and weaknesses, and develop learning opportunities based on these findings. In addition, a scale reliably assessing AI literacy could be used to characterize study populations in AI-related research. It is important that such assessment instruments meet psychometric quality criteria. In particular, the reliability and validity of the instruments are vitally important and should be tested extensively (Verma, 2019).

To our knowledge, there are currently three publications dealing with the development of psychometrically validated scales for AI literacy which allow a general and cross-sample assessment of AI literacy. The first published scale by Wang et al. (2022) found four factors that constitute AI literacy: "awareness", "usage", "evaluation", and "ethics". This scale was developed primarily to "measure people's AI literacy for future [Human-AI interaction] research" (p. 5). The authors developed their questionnaire based on digital literacy research and found that digital literacy and AI literacy overlap to some extent. Another study was published by Pinski and Benlian (2023). This study primarily presents the development of a set of content-valid questions and supplements this with a pre-test of the item set with 50 participants. Based on the preliminary sample, structural equation modelling was used to examine whether their notion of a general model of AI capabilities was accurate. While the study is well designed overall, the results of the pre-test based on only 50 subjects can indeed only be considered preliminary. It is also interesting to note that the questionnaire is intended to be used to assess general AI literacy, but in the pre-selection of participants, a certain level of programming ability was required. The most recent study in this area was published as a preprint by Carolus et al. (2023) and is still in the peer-review process at this time. The authors generated a set of potential AI literacy items derived from the categories listed in the review by Ng et al. (2021b). Afterwards, the "items were discussed, rephrased, rejected, and finalised by [their] team of researchers" (Carolus et al., 2023, p.6). They then tested the fit of the items to the theoretical categories using confirmatory factor analysis. Of note, this procedure corresponds to the top-down process of deduction, as the authors derive practical conclusions (i.e., items) from theory.

1.3. Developing the „scale for the assessment of non-experts' AI literacy"

The main objective of this paper is to present the development of the "Scale for the assessment of non-experts' AI literacy" (SNAIL), which aims to expand the AI literacy assessment landscape. It differs from existing AI literacy assessment tools in several essential ways. First, the focus of the scale is clearly on non-experts, i.e., individuals who have not had formal AI training themselves and who interact or collaborate with AI rather than create or develop it (in contrast to Carolus et al. (2023)). Second, we focused exclusively on AI literacy items, as the assessment of AI literacy must be detached from related constructs such as digital literacy (in contrast to Wang et al., 2022). Third, we take an inductive, exploratory, bottom-up research approach by moving from specific items to generalized latent factors. The main reason for this approach is the prelusive theoretical basis for AI literacy (as described in section 1.1). Since this inductive research approach derives theoretical assumptions from practical observations (i.e., participants' responses to the AI literacy items), we deliberately refrained from formulating hypotheses.

However, three research questions can still be formulated that can structure the development of the scale. First, we are interested in whether hidden (or latent) factors influence item responses. These could be subconstructs that map different capabilities in the field of AI literacy. For example, it would be possible that AI literacy consists of the specific subcategories of "awareness," "usage," "evaluation," and "ethics," as postulated by Wang et al. (2022). As a first step, it would therefore be interesting to determine how many factors there are and which items of the item set can be assigned to each individual factor. Thus, research question (RQ) 1 is:

RQ1. How many factors should be extracted from the available data, and which items of the SNAIL-questionnaire load on which factor?

While RQ1 can be answered mainly with statistical methods (more on this in the section 2), RQ2 is more concerned with the meaning of the factors in terms of factor content. Often, multiple items loading on a single factor follow a specific content theme. This theme can be identified and named, and the name can be used as the "title" for the respective factor.

RQ2. Can the items loading on a factor be subsumed under a particular theme that can be used as a factor name?

Lastly, in most item sets there are certain items whose added value is rather low. This could be due, to the fact that an item is worded ambiguously or measures something other than what it is supposed to measure. Such items should be excluded from the final scale because they can negatively influence the psychometric quality criteria. In addition, a scale is more efficient if it requires fewer items while maintaining the same quality.

RQ3. Do items exist in the original item set that can be excluded to increase the efficiency of the final SNAIL-questionnaire?

2. Material and methods

This study was approved by the local Ethics Committee (application number 194/22), and all participants gave informed consent.

2.1. Variable selection and study design

Laupichler et al. (2023) developed a preliminary item set for assessing individuals' AI literacy in a Delphi expert study. In this study, 53 experts in the field of AI education were asked to evaluate pre-generated items in terms of their relevance to an AI literacy questionnaire. In addition, the experts were asked to contribute their own item suggestions as well as to improve the wording of the pre-generated items. The relevance and the wording of 47 items were evaluated in three iterative Delphi rounds (for more information on the Delphi process, see Laupichler et al., 2023). This resulted in a preliminary set of 39 content-valid items designed to cover the entire domain of AI literacy. The authors argued that the item set is preliminary because their psychometric properties were not assessed in the study. The items were formulated as "I can…" statements, e.g. "I can tell if the technologies I use are supported by artificial intelligence".

We used an analytical, observational, cross-sectional study design. All 39 items created by Laupichler et al. (2023) were presented to the participants in an online questionnaire. Participants rated the corresponding competency on a seven-point Likert scale from "strongly disagree" (one) to "strongly agree" (seven), as recommended by Lozano et al. (2008). The items were presented in random order, and the online questionnaire system ensured that the items were presented in a different (randomized) order for each participant. In addition to the actual AI literacy items, some sociodemographic questions were asked

about age, gender, country of origin, etc. In addition, two bogus items were used to control the participants' attention (see next section).

2.2. Participants

2.2.1. Participant selection and sampling method

The final "Scale for the assessment of non-experts' AI literacy" (SNAIL) is intended to be used by non-experts and can be applied in a variety of educational (i.e., high school and beyond) and professional settings. Thus, we did not survey a specific (sub-) population but rather attempted to obtain a sample that is as heterogenous as possible. We recruited 479 participants through Prolific (www.prolific.com) to take part in our study. Prolific is an incentive-based platform and participants received 1.80£ for answering the questionnaire. Participants had to speak English as their primary language and be over 18 years old. Therefore, our sampling procedure can be defined as non-probabilistic and consecutive (total enumerative), since we included every Prolific participant who met the inclusion criteria until our required sample size was achieved. The only limitation of the consecutive sampling procedure in our study was that exactly 50% of the participants (n = 240) should identify as male and 50% (n = 240) as female. Thus, once the 240 participants of one gender were reached, no further participants of that gender were allowed to participate in the study. Compliance with this sampling procedure was ensured by Prolific's automated participant sampling feature, which randomly sends study invitations to eligible participants and allows them to participate in the study until the required number of participants is reached. Since the total population of all AI non-experts is very large, difficult to delineate, and poorly studied, we refrained from attempting to achieve a probabilistic and representative sample.

Since careless responses have a significant influence on the reliability of factor analyses (Woods, 2006), we used three identification criteria for careless or inattentive response patterns and excluded those cases before analysis of the data (see Fig. 1). First, we used an attention check item "Please check "Somewhat disagree" (3) for this item (third box from the left).", which was randomly placed between the actual items. Participants who failed to choose the correct response option were excluded from the data set (n = 9). Second, we used a bogus item which was meant to identify nonsensical or intentionally wrong response patterns (Meade & Craig, 2012), "I count myself among the top 10 AI researchers in the world." Participants who at least partly agreed (five to seven on a seven-point Likert scale) to the statement were excluded from data analysis (n = 16). Finally, we excluded all participants whose questionnaire completion time was one standard deviation (2:59 min) below the mean completion time (5:23 min) of all participants (n = 39). Since our questionnaire consisted of a total of 39 AI literacy questions, 10 additional questions and some introductory, explanatory and concluding text elements, it can be assumed that the probability of careless responses increased strongly with completion times of less than 3 min.

Mundfrom et al. (2005) suggest calculating the number of participants needed to conduct an EFA based on communality, variables per factor, and the number of factors found in comparable studies. Because our study is one of the first studies to develop an AI literacy questionnaire, these parameters were not available in our case. Nevertheless, we believe that the final sample of n = 415 participants is adequate for EFA. This is in line with recommendations made by different research groups. For example, Comrey and Lee (1992) found that 300 to 500 participants is "good" to "very good", and Benson and Nasser (1998) found a participant to variable ratio of 10:1 to be adequate for EFA (10.6:1 in our study).

2.2.2. Sample characteristics

Most participants were from the United Kingdom (n = 316 or 76.1%), South Africa (n = 32 or 7.7%), the United States (m = 27 or 6.5%), Australia (n = 9 or 2.2%), and Canada (n = 9 or 2.2%). The average age of the participants was 39.5 years (SD = 13.6), and 208 (50.1%) identified as female. On average, the participants included in the final data set (i.e., after exclusion) took 5:39 min (SD = 2:19 min) to complete the questionnaire.

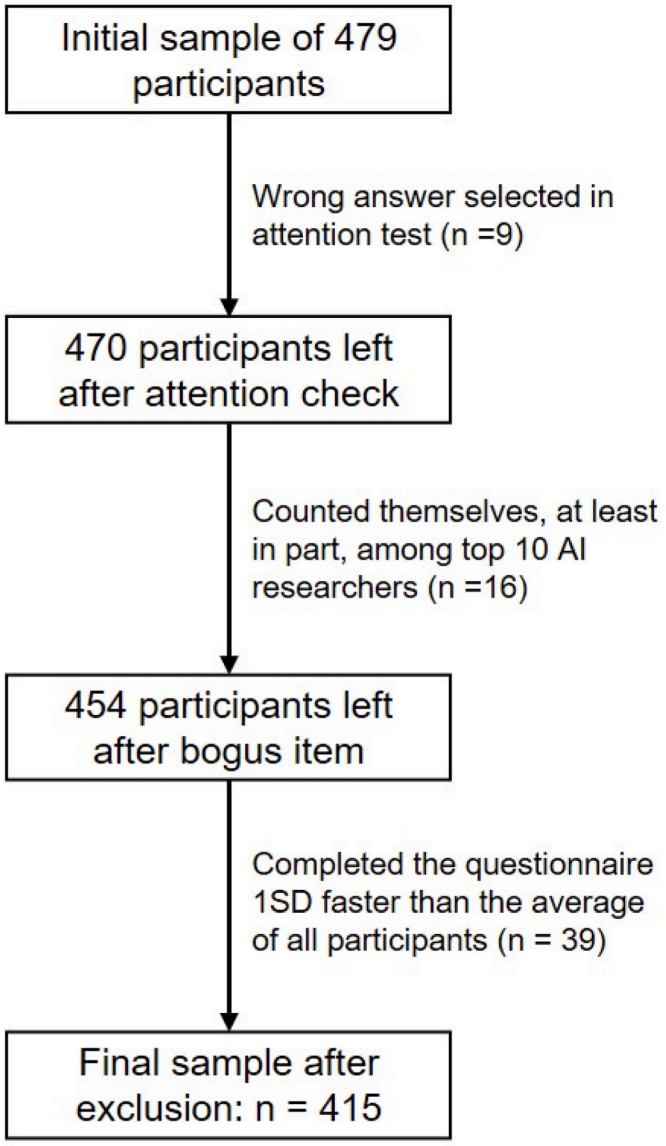


Fig. 1. Number of participants excluded from data analysis based on three exclusion criteria.

2.3. Data analysis

To conduct a methodologically sound data analysis, we followed the recommendations of Watkins (2021) in conducting the EFA, as appropriate. In a first step, the data set was analysed for various univariate descriptive statistical parameters such as skew, kurtosis, the presence of outliers, and the number and distribution of missing values. In addition, Mardia's test of multivariate skew and kurtosis (Mardia, 1970) and Mahalanobi's distance (Mahalanobis, 1936) were calculated to test the multivariate distribution of the data. Afterwards, the appropriateness of the data for conducting an EFA was examined. For this purpose, Bartlett's test of sphericity (Bartlett, 1950) and the Kaiser-Meyer-Olkin criterion (Kaiser, 1974) were calculated and a visual inspection of the correlation matrix was performed to determine whether a sufficient number of correlations ≥.30 was present.

Since our goal was to "understand and represent the latent structure

of a domain" (Widaman, 2018, p. 829), we chose common factor analysis over principal component analysis. However, since we used a relatively high number of variables (39), both techniques would likely produce fairly similar results (Watkins, 2021).

Although different factor extraction methods generally yield similar results (Tabachnik et al., 2019), we compared the results of maximum likelihood extraction and iterated principal axis extraction due to the multivariate non-normality of our data. The differences between the two extraction methods were negligible, so we applied the more commonly used maximum likelihood extraction. We used squared multiple correlations for the initial estimation of communalities. Since our variables were in principle ordinal at least, we based the analysis on the polychoric correlation matrix instead of the more commonly used Pearson correlation matrix. We used parallel analysis by Horn (1965) and the minimum average partial (MAP) method of Velicer (1976) to decide how many factors to retain. A scree-plot (Catell, 1966) was used for visual representation, but not as a decisive method, since it was found to be rather subjective and researcher-dependent (Streiner, 1998). Since we expected the various factors in the model to be at least somewhat correlated, we used an oblique rotation method. We used the promax rotation method as a basis for interpretation, but compared the results with the oblimin rotation method. Norman and Streiner (2014) suggested to set the threshold at which pattern coefficients (factor loadings) will be considered meaningful (i.e., salient) to $\frac{5.152}{\sqrt{N-2}}$ (for $p = .01$). However, due to the large number of participants in our study, this would imply a relatively low salience threshold of 0.25, which is why we followed the more conservative suggestion made by Comrey and Lee (1992), who considered a minimum loading of 0.32 as salient.

After the EFA was conducted, the SNAIL-questionnaire was shortened to improve questionnaire economy and thereby increase the acceptability of using SNAIL as an assessment tool. As a basis for deciding whether to exclude variables, we looked at salient pattern coefficients on more than one factor on the one hand, and a particularly low communality on the other.

Data pre-processing was done partially in Microsoft Excel (Microsoft Corporation, 2018) or R (R Core Team, 2021) and RStudio (RStudio Team, 2020), respectively. Data analysis and data visualization was conducted entirely in R and RStudio.

3. Results

3.1. Data screening and appropriateness of data for EFA

The univariate distribution of all variables was acceptable, with skewness values ranging from −1.18 to 0.87, which is in the acceptable range of −2.0 to 2.0. Similar results were found for univariate kurtosis, with values ranging from −1.26 to 1.85, which is in the acceptable range of −7.0 to 7.0 (see supplementary material 1). Because Mardia's test of multivariate skew and kurtosis became significant ($p < .001$), multivariate non-normality had to be assumed. Bentler (2005) found that increased multivariate kurtosis values of ≥5.0 can influence the results of EFA when working with Pearson correlation matrices, which is another reason to base calculations on the polychoric correlation matrix. Using the Mahalanobis distance (D^2), some outliers were identified, but these were still within the normal range and showed no signs of systematic error. Data entry errors or other third-party influences are highly unlikely because we used automated questionnaire programs.

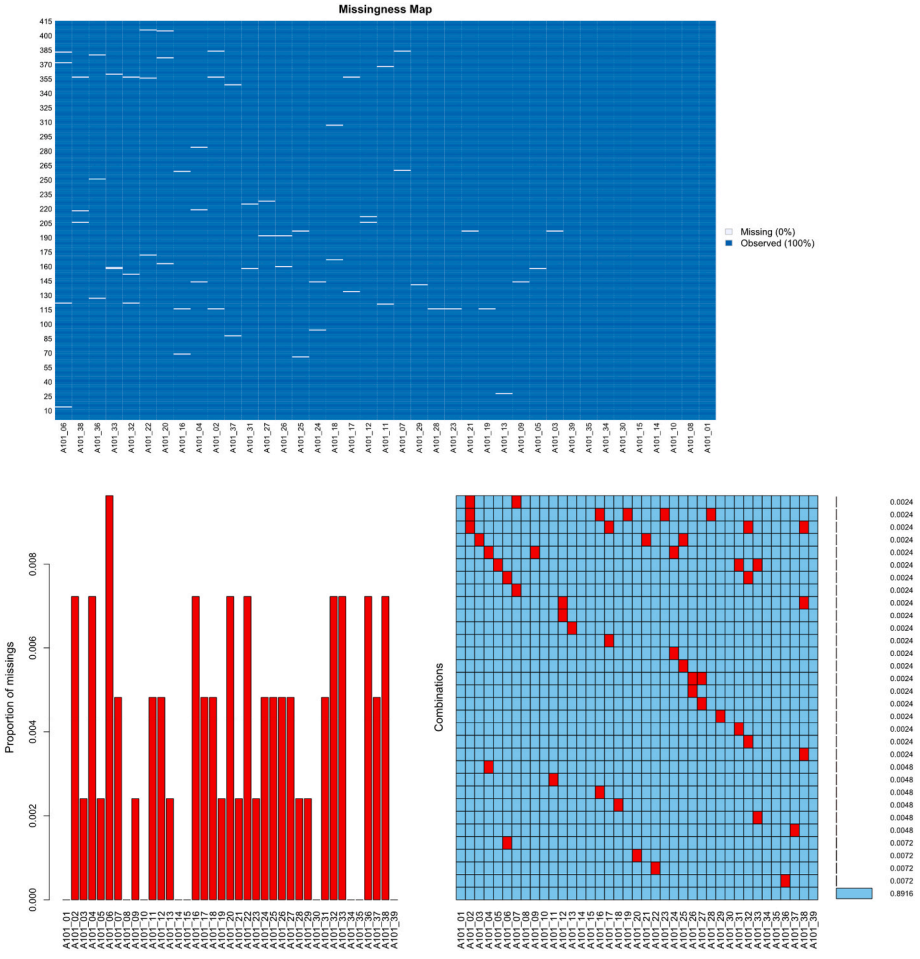


Fig. 2. Distribution and number of missing values in absolute and relative terms across all subjects and variables.

Thus, we could not find any "demonstrable proof [that] indicates that they are truly aberrant and not representative of any observations in the population" (Hair et al., 2019, p. 91), which is why we did not exclude these cases from the data set. In total, each variable missed between 0 and 4 values, which makes up 0–0.96% of all data. In addition, the data was missing completely at random, as demonstrated in Fig. 2. Therefore, no imputation or deletion methods were applied.

Based on Bartlett's test of sphericity, the null-hypothesis that the correlation matrix was an identity matrix could be rejected ($p < .001$). The significant result (i.e., $p < .05$) indicates that there is some redundancy among the variables, which means that they can be reasonably summarized with a smaller number of factors. The overall MSA of the Kaiser-Mayer-Olkin criterion was 0.97, with a range of 0.94–0.98 for each item, which is far above the minimum recommended threshold of 0.5 (Field et al., 2012) or 0.6 (Tabachnik et al., 2019), respectively. A visual inspection of the correlation matrix revealed that a majority of the coefficients were $\geq$0.30, indicating a sufficiently high magnitude of coefficients in the correlation matrix. Based on these measures, we assumed that the correlation matrix was adequate for performing an EFA. (Watkins, 2021; Hair et al., 2019; Tabachnik et al., 2019).

3.2. Number of factors to retain

Horn's parallel analysis, conducted with 20,000 iterations, found two factors to be the optimal solution, regardless whether the reduced or unreduced correlation matrix was used. In contrast, Velicer's minimum average partial reached a minimum of 0.0086 with three factors. A visual inspection of the scree plot supports these results. Depending on subjective preferences, two or three factors could be retained (Fig. 3). Consequently, we analysed models with one, two, three, and four factors for signs of under- or overfactoring, as well as their interpretability and theoretical meaningfulness.

3.3. EFA model evaluation

Following RQ1, the next section evaluates and compares different factor models to identify the most fitting number of factors.

3.3.1. One factor model

The hypothesis that the one-factor model would exhibit signs of underextraction was confirmed. The communalities were rather weak (only two variables had communalities >0.60) and there was no reasonable unifying theme (i.e., meaningful content category/categories) other than that they were evaluating some aspect of AI literacy. Furthermore, 47.8% of the off-diagonal residuals exceeded 0.05 and 15.1% exceeded 0.10. These results were consistent across rotation techniques (i.e., promax and oblimin rotation), therefore strongly indicating the presence of at least one other factor.

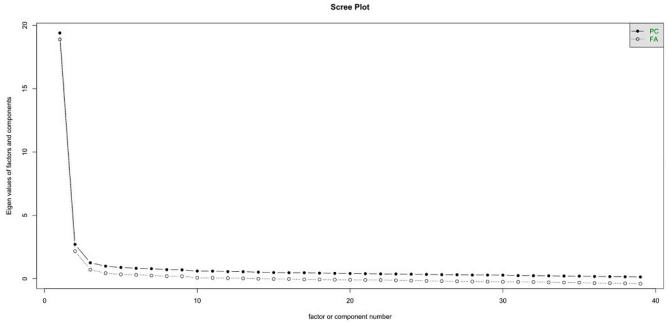


Fig. 3. Screeplot.

3.3.2. Two and three factor models

The difference between the two-factor model and the three-factor model was rather ambiguous, which is consistent with the contrasting results of the parallel analysis and the minimum average partial method, which suggested the extraction of two and three factors, respectively.

Both models had somewhat elevated levels of off-diagonal residuals, with 15.1% of residuals exceeding 0.05 and 3% of residuals exceeding 0.10 in the two-factor model and 11.3% of residuals exceeding 0.05 and 1.08% of residuals exceeding 0.10 in the three-factor model. Although this might indicate underfactoring, it could also be due to the ordinal nature of the data set and the multivariate non-normality. In addition, the RSMR-value of both models (0.04 and 0.03, respectively) lay under the suggested threshold of $\leq$ 0.08.

All models had a sufficient number of pattern coefficients that loaded saliently on each factor (i.e., more than three, Fabrigar & Wegener, 2012; Mulaik, 2009). The only exception is the three-factor oblimin model when applying the conservative salience threshold of $\geq$ 0.32 described above. Here, no variables would load saliently on the third factor. The promax rotation method, on the other hand, comes to a reasonable distribution of salient pattern coefficients on all three factors. The two- and three-factor model both showed marginally acceptable communalities and no Heywood-cases (Harman, 1976). The mean of the communalities was 0.54 (SD = 0.08) for the two-factor model and 0.57 (SD = 0.08) for the three-factor model.

While the one-factor model was only able to explain 48% of the variance, the two-, three- and four-factor models were able to explain 54%, 57%, and 58% of the variance, respectively.

To analyse the internal consistency reliability, we combined every variable that saliently loaded on a factor in a scale and calculated Cronbach's alpha with bootstrapped confidence intervals. The internal consistency of both scales in the two-factor model was excellent, with $\alpha = 0.95$ [CI 0.94, 0.96] for the first scale and $\alpha = 0.94$ [CI 0.93, 0.95] for the second scale. Albeit having slightly lower alpha-values, the internal consistency of the three scales in the three-factor model was also excellent: $\alpha = 0.94$ [CI 0.93, 0.95] for the first scale, $\alpha = 0.93$ [CI 0.91, 0.94] for the second scale, and $\alpha = 0.89$ [CI 0.87, 0.91] for the third scale.

3.3.3. Four factor model

Most of the parameters described above (e.g., RSMR, number of salient pattern coefficients) would also have been acceptable when using the four-factor model. However, fewer variables loaded on each factor, with only three variables loading saliently on the fourth factor, which might be a weak indication of overextraction. Four main reasons speak against the adoption of the four-factor model: First, parallel analysis and minimum average partials have resulted in the recommendation to extract either two or three factors. Second, the increase in explained variance from the three-factor model to the four-factor model is relatively insignificant at less than one percent. Third, the salient loading variables could not be classified into any meaningful content-related categories. And fourth, all other things being equal, the more parsimonious solution is usually the better one (Ferguson, 1954).

3.4. Final model selection and factor names

Since Cattell (1978) and other researchers conclude that the right number of factors is not a question of a correct absolute number, but rather a question "of not missing any factor of more than trivial size" (p. 61), the three-factor model seems to represent a good compromise between parsimony and avoiding the risk of underextraction (see Fig. 4).

As for RQ2, the findings and assessments based on the data coincide well with the content-related examination of the individual factors. With the two-factor solution, a unifying theme could be identified but is rather diffuse and unclear. However, the three-factor solution creates a more plausible classification of the manifest variables to the latent factors in terms of content (see Table 1). Based on the reasons given, the

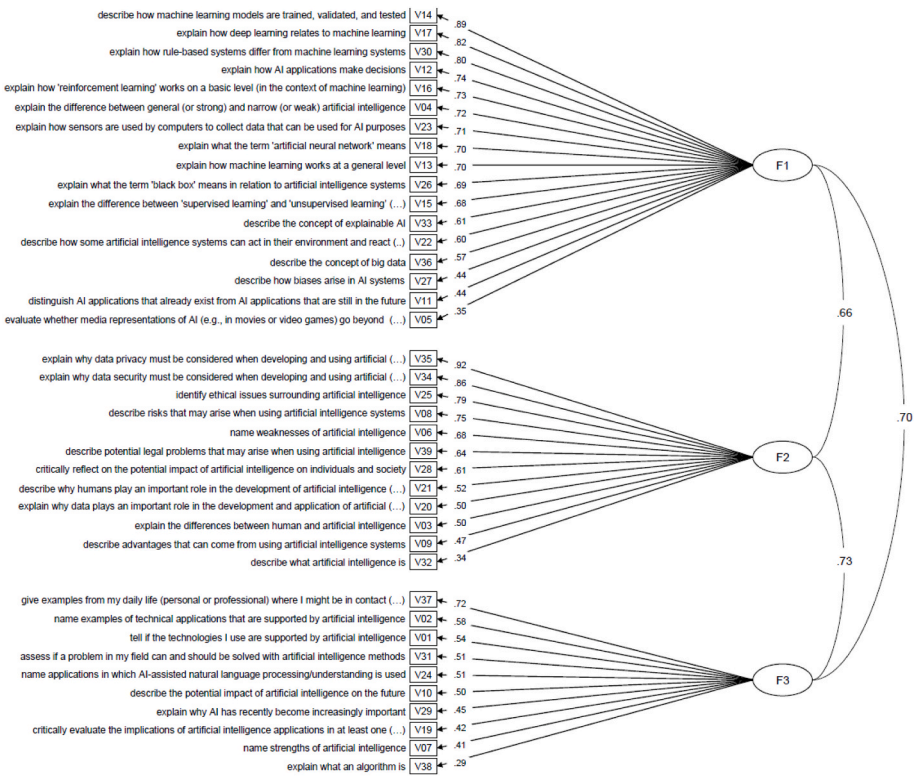


Fig. 4. Path diagram for the 3-factor promax model.

three-factor model was chosen as the best model.

The first factor's highest pattern coefficients were found in variables centred around the understanding of machine learning, e.g. "I can describe how machine learning models are trained, validated, and tested". Other rather technical or theoretical AI competencies such as defining the differences between general and narrow AI or explaining "how sensors are used by computers to collect data that can be used for AI purposes" load saliently on this factor, too. Thus, we propose the first factor's name to be "Technical Understanding". The variables loading saliently on the second factor deal with the recognition of the importance of data privacy and data security in AI, ethical issues related to AI, and risks or weaknesses that may appear when applying AI technologies. Therefore, the second factor is to be called "Critical Appraisal" as it reflects competencies related to the critical evaluation of AI application results. Lastly, the variables with the highest pattern coefficients that load on the third factor are concerned with "examples of technical applications that are supported by artificial intelligence" or assessing "if a problem in [one's] field can and should be solved with artificial intelligence methods". Consequently, the third factor is to be called "Practical Application". Accordingly, the interaction of the three factors could be called the TUCAPA-model of AI literacy.

3.5. Variable elimination

The last section of the results section serves to answer RQ3, which deals with the elimination of items that do not add value and can therefore be excluded. As described above, we excluded variables that loaded saliently on more than one factor and variables with a communality of 2 SD (i.e., 0.08) under the mean communality (0.57). After item elimination, 31 items remained in the final SNAIL-questionnaire. We did not use item parameters (i.e., item difficulty[1] and item discrimination[2]) as exclusion criteria because they were all within the acceptable range (see Table 2).

Overall, five items were eliminated due to diffuse loading patterns; e.g. "I can name strengths of artificial intelligence" (loading saliently on "Critical Appraisal" and "Practical Application"). Furthermore, two variables were deleted because of low communalities; e.g. "I can explain the differences between human and artificial intelligence", and one item that did not load saliently on any factor and had a weak communality; "I can explain what an algorithm is" (see Table 1). We repeated the EFA process with the reduced set of variables and found comparable results. One of the main differences was the decrease in interfactor-correlations, which is somewhat trivial, given that we specifically excluded variables that loaded saliently on more than one factor. The internal consistency of the three scales (i.e., three factors) based on the reduced variable set was excellent. The alpha-values were very similar to the values of the unreduced variable set, with $\alpha = 0.93$ [CI 0.92, 0.94] for the first scale, $\alpha = 0.91$ [CI 0.89, 0.93] for the second scale, and $\alpha = 0.85$ [CI 0.81, 0.88] for the third scale.

For the sake of brevity, all other results and diagrams can be found in the supplementary material (Supplementary Material 1). Consequently, the final SNAIL-questionnaire consists of 31 variables loading on three factors.

4. Discussion

4.1. Relation between TUCAPA and other models

One of the most well-known lists of AI literacy components was certainly published by Long and Magerko (2020), who list 16 competencies that constitute AI literacy. These competencies seem to have only minor relevance for the design of AI literacy assessment questionnaires. This could be due to the large number of 16 competencies, some of which are at the level of latent factors (e.g., competency 11 "Data Literacy") and some at the level of individual manifest variables (e.g., competency 4 "General vs. Narrow [AI]"). Nevertheless, some competencies listed by Long & Magerko (e.g., competency 1 "Recognizing AI") correspond to variables used in SNAIL (e.g., V01 "I can tell if the technologies I use are supported by artificial intelligence.").

Many researchers refer to the literature review by Ng et al. (2021b)

Table 1
List of all variables sorted by factors based on the three-factor TUCAPA-model of AI literacy.
List of all variables sorted by factors based on the three-factor promax model.

Factor 1 (Technical Understanding) I can…	**Factor 2 (Critical Appraisal)** I can…	**Factor 3 (Practical Application)** I can…
describe how machine learning models are trained, validated, and tested. (V14)	explain why data privacy must be considered when developing and using artificial intelligence applications. (V35)	give examples from my daily life (personal or professional) where I might be in contact with artificial intelligence. (V37)
explain how deep learning relates to machine learning (V17)	explain why data security must be considered when developing and using artificial intelligence applications. (V34)	name examples of technical applications that are supported by artificial intelligence. (V02)
explain how rule-based systems differ from machine learning systems. (V30)	identify ethical issues surrounding artificial intelligence. (V25)	tell if the technologies I use are supported by artificial intelligence. (V01)
explain how AI applications make decisions. (V12)	describe risks that may arise when using artificial intelligence systems. (V08)	assess if a problem in my field can and should be solved with artificial intelligence methods. (V31)
explain how 'reinforcement learning' works on a basic level (in the context of machine learning). (V16)	name weaknesses of artificial intelligence. (V06)	name applications in which AI-assisted natural language processing/understanding is used. (V24)
explain the difference between general (or strong) and narrow (or weak) artificial intelligence. (V04)	describe potential legal problems that may arise when using artificial intelligence. (V39)	describe the potential impact of artificial intelligence on the future. (V10) [1]
explain how sensors are used by computers to collect data that can be used for AI purposes. (V23)	critically reflect on the potential impact of artificial intelligence on individuals and society. (V28)	explain why AI has recently become increasingly important. (V29)
explain what the term 'artificial neural network' means. (V18)	describe why humans play an important role in the development of artificial intelligence systems. (V21)	critically evaluate the implications of artificial intelligence applications in at least one subject area. (V19)
explain how machine learning works at a general level. (V13)	explain why data plays an important role in the development and application of artificial intelligence. (V20)	name strengths of artificial intelligence. (V07) [1]
explain what the term 'black box' means in relation to artificial intelligence systems. (V26) [2]	explain the differences between human and artificial intelligence. (V03) [2]	explain what an algorithm is. (V38) [3]
explain the difference between 'supervised learning' and 'unsupervised learning' (in the context of machine learning). (V15)	describe advantages that can come from using artificial intelligence systems. (V09) [1]	
describe the concept of explainable AI. (V33)	describe what artificial intelligence is. (V32)	
describe how some artificial intelligence systems can act in their environment and react to their environment. (V22)		
describe the concept of big data. (V36)		
describe how biases arise in AI systems. (V27) [1]		
distinguish AI applications that already exist from AI applications that are still in the future. (V11) [1]		
evaluate whether media representations of AI (e.g., in movies or video games) go beyond the current capabilities of AI technologies. (V05)		

Note. The variables are sorted by pattern coefficient, with variables loading the highest on each factor appearing at the top of each column. Note that the table shows the model *before* elimination of eight items. Eliminated items have a lighter font. The superscript numbers indicate the reason for elimination, with (1) indicating salient loadings on more than one factor, (2) indicating extraordinarily low communalities, and (3) indicating a combination of (1) and (2).

Note. The variables are sorted by pattern coefficient, with variables loading the highest on each factor appearing at the top of each column. Note that the table shows the model *before* elimination of eight items. Eliminated items have a lighter font. The superscript numbers indicate the reason for elimination, with (1) indicating salient loadings on more than one factor, (2) indicating extraordinarily low communalities, and (3) indicating a combination of (1) and (2).

Table 2
Item parameters sorted by factors based on the three-factor promax model.

F1 – Technical Understanding				
Item	Mean	SD	Item Difficulty	Item Discrimination
V14	1.63	1.51	.27	.76
V17	1.61	1.43	.27	.72
V30	1.88	1.58	.31	.72
V12	2.15	1.58	.36	.70
V16	1.98	1.56	.33	.68
V04	1.73	1.44	.29	.68
V23	1.99	1.61	.33	.69
V18	1.52	1.53	.25	.70
V13	2.48	1.68	.41	.71
V26^{x}	1.69	1.68	.28	.52
V15	2.09	1.61	.35	.66
V33	1.97	1.53	.33	.59
V22	2.26	1.52	.38	.73
V36	2.25	1.72	.37	.64
V27^{x}	2.4	1.75	.40	.67
V11^{x}	2.16	1.56	.36	.67
V05	2.65	1.76	.44	.64
F2 – Critical Appraisal				
V35	3.62	1.57	.60	.70
V34	3.48	1.61	.58	.69
V25	3.62	1.55	.60	.70
V08	3.46	1.5	.58	.74
V06	3.54	1.5	.59	.73
V39	2.97	1.72	.49	.68
V28	3.31	1.6	.55	.70
V21	3.68	1.47	.61	.70
V20	3.42	1.58	.57	.70
V03^{x}	4	1.31	.67	.56
V09^{x}	3.7	1.41	.62	.70
V32	3.94	1.18	.66	.63
F3 – Practical Application				
V37	3.65	1.53	.61	.59
V02	2.84	1.76	.47	.67
V01	2.6	1.55	.43	.60
V31	2.5	1.64	.42	.70
V24	2.22	1.72	.37	.63
V10^{x}	3.46	1.52	.58	.68
V29	3.41	1.49	.57	.65
V19	2.43	1.72	.40	.68
V07^{x}	3.54	1.46	.59	.63
V38^{x}	3.63	1.53	.60	.55

Note. Items are sorted by the magnitude of their pattern coefficients, with items having higher loadings listed first. To calculate item difficulty[1] and item discrimination,[2] the data set was mutated by subtracting 1 from every value in the data set. Consequently, the range of possible values was 0–6 (instead of the aforementioned Likert-scale with values ranging from 1 to 7). Items that were eliminated are indicated by $^{(x)}$.

in developing their models (Carolus et al., 2023, Pinski et al., 2023). The categories identified by Ng et al. seem to fit relatively well with the three factors of the TUCAPA-model, as "know and understand" overlaps with "Technical Understanding" and "use and apply" corresponds to "Practical Application". The last two factors of Ng et al., "evaluate and create" and "ethical issues," could be combined into one factor in our case, "Critical Appraisal".

Karaca et al. (2021) developed MAIRS-MS, a scale designed to assess the so-called AI readiness of medical students. AI readiness is a construct that resembles AI literacy in many ways. In their research project, they also conducted an EFA and found four factors that seem to fit well with the factors described in this paper. Karaca et al.'s "Cognition"-factor is very similar to the "Technical Understanding"-factor, although the underlying items tend to be on a more general level (e.g., "I can define the basic concepts of data science.", p. 5). The "Ability"-factor has some resemblance to the "Practical Application"-factor in our model. And again, the last two factors, "Vision" and "Ethics," could be combined into the "Critical Appraisal" factor of the TUCAPA model.

Future research should investigate whether AI competencies related to "ethics" or "ethical issues" really represent a separate AI literacy factor, or whether this competency is part of a larger construct such as "critical appraisal." In any case, our model contributes to the further development of AI literacy theory, as it differs from other models in terms of its factor count and by following an inductive approach. This inductive approach does not require theoretical considerations in advance, but develops theoretical insights from practical observations.

4.2. Limitations

One of the major limitations of self-assessment questionnaires is that their responses can be influenced by conscious or unconscious biases. For this reason, the current questionnaire should only be used if the results of the survey are not linked to consequences that directly affect the respondents (e.g., grades, job applications). In addition to the development of self-assessment scales, it would therefore be important to develop performance tests that objectively test individuals' AI knowledge and skills, rather than having them subjectively rated by the respondents themselves.

The TUCAPA model is composed of three factors derived from statistical results, as shown earlier. However, other research groups reached a different number of factors in their studies, some of which contained slightly different substantive foci. For example, Wang et al. (2022) and Carolus et al. (2023) found a factor with a focus on "AI ethics" that is not represented as a separate factor in the TUCAPA model. This may be due to several reasons. One possible explanation is that the experts in the Delphi study by Laupichler et al. (2023), in which the items were generated, did not consider ethical aspects of AI and therefore formulated few items on this topic.

In addition, the use of paid and anonymous study participants involves certain risks and might lead to response biases. For example, it could be assumed that the anonymity and incentivization cause the acquired subjects to spend little time and attention on answering the SNAIL-questionnaire. However, we used three different careless responding checks, making it unlikely that participants merely "clicked through" the questionnaire. In addition, several studies have shown that the use of paid online participants does not pose an extraordinary threat to the scientific integrity of research (Buhrmester et al., 2011; Crump et al., 2013). Nevertheless, it may be worth repeating the study with a different sample, as we used a non-probabilistic consecutive sampling technique that could affect the validity of the results described. It is possible that sampling bias has occurred due to the sampling technique. For example, there is a possibility that only people who are already interested in the topic of AI and therefore rate their abilities higher than people who are not interested in AI participated in the study. A new dataset should preferably be representative of the entire population of AI non-experts', or at least differ from the dataset used in this study in terms of participant characteristics, participants' countries of origin, etc. This would also have the advantage of ensuring the reproducibility and reliability of the results reported in this study, as it would enable the execution of a confirmatory factor analysis.

4.3. Future research

Future research projects should test the theoretical validity of the three-factor TUCAPA model through confirmatory factor analysis (CFA). This could simultaneously determine whether there is a separate "AI ethics" factor or whether the aspect of AI ethics is already included in the three factors of the TUCAPA model (e.g., in the Critical Appraisal factor). In addition to the previously mentioned use of the questionnaire in other samples or in specific sub-populations, the use of SNAIL in other cultures would also be important. For this purpose, the questionnaire has to be validly translated into the corresponding languages beforehand. This would help the international applicability of the scale, as the questions are currently only available in English. Moreover, it should be

investigated whether SNAIL can be applied equally well in all subject domains, or whether there are practical differences in AI literacy between different domains. For example, it could be possible that individuals with a high level of technical understanding (e.g., individuals from the field of mathematics or mechanical engineering) would rate the questions of the Technical Understanding factor very positively, while people from fields with less technical affinity (e.g., medicine, psychology) may evaluate the same questions rather negatively. Furthermore, it should be examined whether SNAIL is suitable to investigate the teaching effectiveness of courses that aim to increase the AI literacy of their participants. Since SNAIL is freely available as an open access offering, this would also be interesting for platforms such as "Elements of AI" (University of Helsinki & MinnaLearn, 2018) or "AI Campus" (KI Campus, 2023), which offer open educational resources to improve general AI literacy. Last but not least, the SNAIL-questionnaire should be compared with related constructs such as "attitudes toward AI" (Schepman & Rodway, 2020; Sindermann et al., 2021) or "digital literacy" (Gilster, 1997) to investigate the relationship between each construct. For example, it is possible that more pronounced AI literacy reduces anxiety toward AI (Wang & Wang, 2022), leading to more positive attitudes toward AI.

5. Conclusion

We conducted an exploratory factor analysis to develop the "Scale for the assessment of non-experts' AI literacy" (SNAIL) questionnaire, which is designed to assess AI literacy in non-experts. In doing so, we found that the construct represented by the questionnaire can be divided into three subfactors that influence individuals' response behaviour on AI literacy items: Technical Understanding, Critical Appraisal, and Practical Application. Therefore, the model can be abbreviated as the TUCAPA model of AI literacy. Our study provides initial evidence that the 31 SNAIL items are able to reliably and validly assess the AI competence of nonexperts. However, further research is needed to evaluate whether the results found in our study can be replicated and are representative of the population of nonexperts. Finally, we would like to encourage all researchers in the field of AI literacy to use psychometrically validated questionnaires to assess the AI literacy of individuals and groups as well as to evaluate the learning outcome of course participants.

Funding statement

This work was supported by the Open Access Publication Fund of the University of Bonn.

Ethics approval statement

Data collection for this study took place in February 2023 using the study participant acquisition program Prolific. The subjects received appropriate financial compensation for participating in the study. Participation in the study was voluntary and participants gave their informed consent. The study was approved by the Research Ethics Committee of the University of Bonn (Reference 194/22).

Author contributions

Matthias Carl Laupichler: Conceptualization, Formal Analysis, Writing – Original Draft, Visualization **Alexandra Aster:** Writing – Review & Editing, Data Curation **Nicolas Haverkamp:** Methodology. **Tobias Raupach:** Supervision, Resources.

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.

Data availability

Research data will be published as Supplementary Material (Excel-File)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbr.2023.100338.

References

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3*, 77–85.
Benson, J., & Nasser, F. (1998). On the use of factor analysis as a research tool. *Journal of Vocational Education Research, 23*(1), 13–33.
Bentler, P. M. (2005). *EQS structural equations program manual*. Multivariate software.
Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk. *Perspectives on Psychological Science, 6*(1), 3–5. https://doi.org/10.1177/1745691610393980
Carolus, A., Koch, M., 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*.
Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10
Cattell, R. B. (1978). *Use of factor analysis in behavioral and life sciences*.
Cetindamar, D., Kitto, K., Wu, M., Zhang, Y., Abedin, B., & Knight, S. (2022). Explicating AI literacy of employees at digital workplaces. *IEEE Transactions on Engineering Management*. https://doi.org/10.1109/TEM.2021.3138503
Comrey, A., & Lee, H. (1992). Interpretation and application of factor analytic results. In *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates.
Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), Article e57410. https://doi.org/10.1371/journal.pone.0057410
Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford, UK: Oxford University Press.
Faruqe, F., Watkins, R., & Medsker, L. (2021). *Competency model approach to AI literacy: Research-based Path from initial framework to model. ArXiv Preprint*.
Ferguson, G. A. (1954). The concept of parsimony in factor analysis. *Psychometrika, 19*(4), 281–290. https://doi.org/10.1007/BF02289228
Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE.
Gilster, P. (1997). *Digital literacy*. John Wiley & Sons, Inc.
Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.
Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). University of Chicago Press.
Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447
Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*(1), 31–36. https://doi.org/10.1007/BF02291575
Kandlhofer, M., Hirschmugl-Gaisch, S., & Huber, P. (2016). Artificial intelligence and computer science in education: From kindergarten to university. *2016 IEEE Frontiers in Education Conference (FIE)*, 1–9.
Karaca, O., Çalışkan, S. A., & Demir, K. (2021). Medical artificial intelligence readiness scale for medical students (MAIRS-MS) – development, validity and reliability study. *BMC Medical Education, 21*(1), 112. https://doi.org/10.1186/s12909-021-02546-6
KI Campus. (2023). AI Campus. https://Ki-Campus.Org/.
König, P. D., & Wenzelburger, G. (2020). Opportunity for renewal or disruptive force? How artificial intelligence alters democratic politics. *Government Information Quarterly, 37*(3), Article 101489. https://doi.org/10.1016/j.giq.2020.101489
Laupichler, M. C., Aster, A., & Raupach, T. (2023). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence, 4*, Article 100126. https://doi.org/10.1016/j.caeai.2023.100126
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
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, 1–16*. https://doi.org/10.1145/3313831.3376727
Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4*(2), 73–79. https://doi.org/10.1027/1614-2241.4.2.73
Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Journal of the Society of Bengal, 2*(1), 49–55.
Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519–530.
Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085
Microsoft Corporation. (2018). Microsoft Excel. Retrieved from https://office.microsoft.com/excel.

Mulaik, S. A. (2009). Foundations of factor analysis. *Chapman and Hall/CRC*. https://doi.org/10.1201/b15851

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021a). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology, 58*(1), 504–509. https://doi.org/10.1002/pra2.487

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021b). 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., Leung, J. K. L., Su, M. J., Yim, I. H. Y., Qiao, M. S., & Chu, S. K. W. (2022). *AI literacy in K-16 classrooms*. Springer.

Norman, G. R., & Streiner, D. L. (2014). *Biostatistics: The bare essentials* (4th ed.). People's Medical Publishing.

Pinski, M., & Benlian, A. (2023). AI literacy - towards measuring human competency in artificial intelligence. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 165–174.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine, 112*(1), 22–28. https://doi.org/10.1177/0141076818815510

RStudio Team. (2020). *RStudio*. Boston, MA: Integrated Development for R. RStudio, PBC. http://www.rstudio.com/.

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards artificial intelligence scale. *Computers in Human Behavior Reports, 1*, Article 100014. https://doi.org/10.1016/j.chbr.2020.100014

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. *KI - Kunstliche Intelligenz, 35*(1), 109–118. https://doi.org/10.1007/s13218-020-00689-0

Streiner, D. L. (1998). Factors affecting reliability of interpretations of scree plots. *Psychological Reports, 83*(2), 687–694. https://doi.org/10.2466/pr0.1998.83.2.687

Su, J., & Ng, D. T. K. (2023). Artificial intelligence (AI) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence*. , Article 100124. https://doi.org/10.1016/j.caeai.2023.100124

Tabachnik, B. G., Fidell, L. S., & Ullman, J. B. (2019). In *Using multivariate statistics* (Vol. 7). Pearson.

University of Helsinki, MinnaLearn. (2018). Elements of AI. https://www.elementsofai.com/.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321–327. https://doi.org/10.1007/BF02293557

Verma, J. P. (2019). *Statistics and research methods in psychology with Excel*. Springer Singapore. https://doi.org/10.1007/978-981-13-3429-0

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*. https://doi.org/10.1080/0144929X.2022.2072768

Wang, Y. Y., & Wang, Y. S. (2022). Development and validation of an artificial intelligence anxiety scale: An initial application in predicting motivated learning behavior. *Interactive Learning Environments, 30*(4), 619–634. https://doi.org/10.1080/10494820.2019.1674887

Watkins, M. R. (2021). *A step-by-step guide to exploratory factor analysis with R and RStudio*. Routledge.

Widaman, K. F. (2018). On common factor and principal component representations of data: Implications for theory and for confirmatory replications. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(6), 829–847. https://doi.org/10.1080/10705511.2018.1478730

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity, 2021*, 1–18. https://doi.org/10.1155/2021/8812542

education sciences

MDPI

Article

Evaluating AI Courses: A Valid and Reliable Instrument for Assessing Artificial-Intelligence Learning through Comparative Self-Assessment

Matthias Carl Laupichler [1,*], Alexandra Aster [1], Jan-Ole Perschewski [2] and Johannes Schleiss [2]

[1] Institute of Medical Education, University Hospital Bonn, 53127 Bonn, Germany; alexandra.aster@ukbonn.de
[2] Artificial Intelligence Lab, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany; jan-ole.perschewski@ovgu.de (J.-O.P.); johannes.schleiss@ovgu.de (J.S.)
* Correspondence: matthias.laupichler@ukbonn.de

Abstract: A growing number of courses seek to increase the basic artificial-intelligence skills ("AI literacy") of their participants. At this time, there is no valid and reliable measurement tool that can be used to assess AI-learning gains. However, the existence of such a tool would be important to enable quality assurance and comparability. In this study, a validated AI-literacy-assessment instrument, the "scale for the assessment of non-experts' AI literacy" (SNAIL) was adapted and used to evaluate an undergraduate AI course. We investigated whether the scale can be used to reliably evaluate AI courses and whether mediator variables, such as attitudes toward AI or participation in other AI courses, had an influence on learning gains. In addition to the traditional mean comparisons (i.e., *t*-tests), the comparative self-assessment (CSA) gain was calculated, which allowed for a more meaningful assessment of the increase in AI literacy. We found preliminary evidence that the adapted SNAIL questionnaire enables a valid evaluation of AI-learning gains. In particular, distinctions among different subconstructs and the differentiation constructs, such as attitudes toward AI, seem to be possible with the help of the SNAIL questionnaire.

Keywords: AI literacy; AI-literacy scale; artificial intelligence education; assessment; course evaluation; comparative self-assessment

Citation: Laupichler, M.C.; Aster, A.; Perschewski, J.-O.; Schleiss, J. Evaluating AI Courses: A Valid and Reliable Instrument for Assessing Artificial-Intelligence Learning through Comparative Self-Assessment. *Educ. Sci.* **2023**, *13*, 978. https://doi.org/10.3390/educsci13100978

Academic Editors: Gary K. W. Wong and Ho-Yin Cheung

Received: 28 July 2023
Revised: 20 September 2023
Accepted: 21 September 2023
Published: 26 September 2023

1. Introduction

1.1. AI Literacy

Artificial intelligence (AI) is permeating more and more areas of daily life. While no universal definition of AI exists, most definitions agree that it is "a branch of computer science dealing with the simulation of intelligent behavior in computers" [1] and that AI represents the idea of "computer programs that have some of the qualities of the human mind" [2]. Examples of AI use can be found in a wide variety of fields, including mundane applications such as movie recommendations [3] and video games [4] and applications in highly specialized professions such as medicine [5] and engineering [6,7]. In the course of such developments, more and more people are coming into contact with AI applications, consciously or unconsciously. In order to be able to deal with AI in a meaningful and outcome-oriented way and to be able to assess possible benefits as well as risks, a certain understanding of AI is essential. This basic understanding, which most, or even all, individuals should have, is often referred to as AI literacy.

While different definitions of AI literacy exist, the most commonly cited definition comes from a paper written by Long and Magerko [8]. They described AI literacy 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" [8], (p. 2).

Various educational projects aim to improve individuals' AI literacy. Introductory courses on AI are offered at a wide variety of educational levels, starting as early as kindergarten and elementary school [9,10], continuing through K-12 education [11–13], and ending with higher education and adult education in universities and similar institutions [14,15]. To measure the impact and effectiveness of these educational projects, some of them have been examined in evaluation studies and accompanying research, the results of which have been published.

Many researchers resort to self-created or unvalidated instruments to measure learning success. Other researchers do not measure learning success at all, but limit their findings to the lowest level of the Kirkpatrick model [16] by reporting direct, affective reactions toward a course. However, assessing learning gains with a reliable, objective, and valid instrument is important in uncovering potential problems in the delivery of AI literacy content and in evaluating the quality of AI courses. For this reason, many researchers have called for the development of AI-literacy-assessment instruments of high psychometric quality [15,17,18].

1.2. Assessing AI Literacy

Several relatively well-validated scales already exist that try to capture affective attitudes toward AI [19–21]. However, the measurement of AI literacy that meet psychometric quality standards is still a fairly new field of research. In fact, existing measurement tools are still evolving and an optimal assessment tool has not yet been established. Nevertheless, some promising initial efforts have been made to develop AI-literacy-assessment instruments.

The first AI literacy scale was published by [22] and included four sub-factors of AI literacy: "awareness", "usage", "evaluation", and "ethics". The authors of that study drew on previous research in digital literacy and human–AI interaction to develop their scale. Another study that reported the creation of a set of AI-literacy items was published by Pinski and Benlian [23]. They presented the findings of a preliminary study that distributed the items to 50 individuals. The resulting dataset was then used to draw conclusions about the structure of AI competencies, using structural-equation modeling.

An unpublished manuscript by Carolus et al. [24], describing the development of their scale for AI-literacy assessment, has not yet undergone a peer-review process. Nonetheless, this scale represents an interesting contribution to AI-literacy assessment, as they used a top-down approach and based their item development on the categories introduced in the well-cited review by Ng et al. [25].

In contrast, Laupichler et al. [26] followed a bottom-up approach and used a Delphi study to generate a set of content-valid items that were relevant and representative for the field of AI literacy. Those items were validated in another study, which is currently undergoing a peer-review process. The items generated in the Delphi study were presented to a sample of more than 400 participants and, subsequently, analyzed through an exploratory factor analysis that found three AI-literacy factors: technical understanding, critical appraisal, and practical application [27].

1.3. Using an AI-Literacy-Assessment Instrument to Evaluate Learning Gains

While the aforementioned instruments have been validated using appropriate samples, it has not yet been determined whether they are suitable for assessing learning gains or evaluating the effectiveness of AI courses. However, a valid and reliable evaluation of AI courses is essential for several reasons. First, high-quality assessment tools enable quality-assurance procedures to be implemented. Second, such evaluation tools could be used to identify potential strengths and weaknesses of an AI course in a resource-efficient manner. The information obtained could be used as part of the continuous and iterative improvement of a course. Third, evaluating different courses with the same instrument would allow comparability across individual courses or study programs. Such

comparisons could then be used, for example, for external evaluation of course offerings or for program specialization.

We hypothesize that AI-literacy scales developed to assess the status quo of individuals' AI knowledge can also be used to evaluate the quality of AI courses, with some minor adaptations. It is particularly important that these AI-course-evaluation instruments be validated for this purpose in order to obtain meaningful and comparable results. In this study, the "scale for the assessment of non-experts' AI literacy" (SNAIL), which was developed by Laupichler et al. [27], was used, because it was validated on a sufficiently large sample and the items can be adapted particularly well for course evaluation. However, one of the other scales described above ([22–24]) could just as easily have been used, as the adaptations described below can be applied to all items, regardless of their origin.

The original test instrument was changed only with respect to two parameters, the adoption of the language (optional) and the introduction of self-assessment of differences via a "retrospective assessment" and a "post-assessment". Concerning the first parameter, the original scale by Laupichler et al. [28] was originally validated in English and the course participants were German native speakers. Therefore, the original items were first translated into German (see the Materials and Methods section of this article). This prevented misunderstandings and lowered the cognitive barrier to completing the evaluation questionnaire, which in turn had a positive effect on the response rate.

The second modification of the original scale allows for the measurement of differences in self-assessed AI literacy that may occur by attending the AI course. For this purpose, each item was presented as a retrospective assessment version and a post-assessment version, meaning that the participants had to assess their individual competency on every item, respectively, in a retrospective manner (i.e., looking back to the time before the course) and with respect to their current capability (i.e., after taking part in the course). The retrospective/post method is often more suitable for assessing learning gains than the traditional pre/post test (one assessment before and one after a course, [28]) because it is subject to fewer biases. Especially when assessing skills prior to educational intervention, learners tend to overestimate their competency because they often cannot yet fully grasp the depth of the field, an effect commonly referred to as response-shift bias [29,30].

1.4. Research Questions

The objective of this study was to investigate the applicability of the validated AI literacy scale for the evaluation of AI courses. Specifically, we aimed to assess the scale's effectiveness in measuring changes in learning gains through comparative self-assessment. Additionally, we sought to investigate whether certain items or factors within the scale showed more significant increases in knowledge and skill than others. These distinctions could prove to be valuable in identifying any weaknesses in the evaluated courses, thus facilitating targeted improvements. Therefore, our first research question was:

RQ1: Can the adapted "scale for the assessment of non-experts' AI literacy" be used to reliably and validly assess the learning gains of AI courses?

In addition, we aimed to explore the extent to which course participants' AI literacy was influenced by their attitudes toward AI, and vice versa. If AI literacy and attitudes toward AI are correlated, then it might be advisable to assess attitudes toward AI in future AI-course evaluations. Moreover, if the relationship between the two variables is causal (rather than merely correlative), it might be necessary to take steps to improve participants' attitudes in addition to their learning gains. Thus, our second research question was:

RQ2: Are AI course participants' self-assessed AI literacy and their attitudes toward AI correlated?

Finally, we wanted to investigate the extent to which attending other AI courses prior to the evaluated AI course had an impact on learning success and self-assessment values. Furthermore, AI education does not take place only in formal settings such as courses; students also use other sources, such as educational videos, books, and social media posts,

to learn about AI topics. Therefore, a question on the use of other means of education was added, and its relationship to participants' self-assessments was examined. Accordingly, our third and final research question was:

RQ3: Does AI education outside of an evaluated AI course have an impact on learning gains?

2. Materials and Methods

2.1. AI Course

The course that was evaluated using the SNAIL questionnaire was an interdisciplinary AI course designed to teach AI skills to undergraduate students. Students from different study programs were allowed to register for the course, which meant that both students who studied computer science or related subjects and students who had relatively little contact with programming and computer science content in their previous studies took part in the course. Although it could be argued that computer science students are experts in the field of AI, it was still reasonable to use the "scale for the assessment of non-experts' AI literacy". Although these students could be expected to have a high technical literacy, they had little or no education that was focused on fostering their AI literacy. Furthermore, the sample did not only consist of computer science students, but also students from other disciplines, so that comparisons between individuals with low and high technology literacy were possible.

The course had a rather technical focus—teaching how artificial neural networks work. It consisted of a lecture, instructor-led exercises, and self-study content and was structured in an application-oriented way. The course scope of all activities amounted to approximately 150 h. The learning outcomes of the course included the application of methods of data analysis with neural networks for solving classification, regression and other statistical problems; the evaluation and application of neural learning techniques for the analysis of complex systems; and the capability of developing neural networks. Thus, the course corresponded mostly to the technical-understanding and practical-application dimensions of the SNAIL questionnaire.

2.2. Translation of the "Scale for the Assessment of Non-Experts' AI Literacy" (SNAIL)

To ensure a valid and systematic translation of the SNAIL items, we followed the international recommendations for translating psychological measurement instruments wherever possible [31–33]. Two bilingual speakers whose native language was German independently translated the SNAIL items from English into German. Subsequently, these two translators compared the items and analyzed the differences in the translations in order to reach a common consensus. Thereafter, two additional bilingual speakers (one of whom was a native English speaker) independently translated the items from German back into English. Subsequently, all translators, as well as two methodological experts who were experienced in developing questionnaires, met to identify problems and differences in the scale translation. This expert panel was able to produce a final SNAIL version in German (see Supplementary Material S1).

2.3. Evaluation Procedure

We presented each of the 31 SNAIL items and asked the participants about their current self-assessment (after attending the course) and their retrospective self-assessment (at the beginning of the course). Participants were then presented with the five items of the "Attitudes Toward Artificial Intelligence" scale by Sindermann et al. [22]. This attitude scale was used because it is one of the shortest and, thus, one of the most resource-efficient instruments designed to capture attitudes; it had already been validated in the native language of the course participants. In addition, some socio-demographic questions about the participants' ages, fields of study, etc. were collected. Finally, the instrument asked to what extent the participants had already educated themselves on the topic of AI prior to the course, in other courses or with other methods.

2.4. Data Analysis

All analyses were performed using Microsoft Excel or IBM SPSS Statistics (version 27). To determine the AI-literacy-learning gains, the mean of the retrospective items was compared to the mean of the post-items, using *t*-tests. The one-tailed *t*-test was used, as it was assumed that students' AI literacy could only improve by attending the course. This was done both at the item level and at the factor level. Since *t*-tests easily become statistically significant, especially in the area of teaching effectiveness and learning-gain evaluation, even for practically irrelevant increases, an additional analysis method was used. The so-called student comparative self-assessment [28,34] is a more valid tool to assess the actual increase in competence or knowledge, as it accounts for the initial level of participants' AI literacy. The calculation of the comparative self-assessment (CSA) gain is described in Raupach et al.'s 2011 article, "Towards outcome-based programme evaluation: using student comparative self-assessments to determine teaching effectiveness" [34], and CSA gain values can range from −100% to +100% (although in reality, negative values are rare). We used Spearman's rank correlation for correlations between metric and ordinal variables (such as AI-literacy-learning gains and the amount of AI education outside of the course) and Pearson correlation for correlations between metric variables. The reliability of the scale was evaluated by assessing the internal consistency (Cronbach's alpha) of the three factors.

One of the items of the technical-understanding factor had to be excluded due to a technical error ("I can describe the concept of explainable AI"), which resulted in a total of 30 items being used to assess AI literacy (13 instead of 14 items in the technical-understanding factor).

3. Results

3.1. Participants

Because there was no formal enrollment for the course, it was not possible to determine how many people officially attended the course. However, an average of about 40 people attended the lectures and exercises. In total, 25 students (62.5% of all attendees) took part in the study. Study participants were, on average, 22.9 years old (SD = 2.3) and in their sixth semester (M = 6.0, SD = 2.9). More men (n = 16, 64%) than women (n = 9, 36%) participated in the course. As mentioned above, the course was attended by participants from different study programs. Six participants (24%) came from computer science, six (24%) came from a program called "Philosophy Neuroscience Cognition", four (16%) studied statistics, three (12%) studied medical engineering, and two (8%) studied electromobility, computer visualistics, or did not specify their main study program, respectively.

The mean time it took participants to complete the study was 8:01 min (SD = 1:11 min). Participants responded to almost every question and missing values were relatively rare, with an average of 0.2 missing values per respondent (max = 3).

3.2. Learning Gains and Reliability

For all 30 AI literacy items used, the mean values of all participants' retrospective assessments were compared to the mean values of all participants' post-assessments, using independent *t*-tests. To test the null hypothesis that the variances were equal, a Levene test was calculated prior to each *t*-test. The homoscedasticity assumption was only violated for one comparison ("I can explain how sensors are used by computers to collect data that can be used for AI purposes"). In this individual case, a Welch test was performed. Based on a significance level of $\alpha = 0.05$, a significant improvement in performance was found for a majority of the items. Only three items failed to show a statistically significant improvement in the corresponding AI competency: "I can explain the difference between general (or strong) and narrow (or weak) artificial intelligence"; "I can explain why data privacy must be considered when developing and using artificial intelligence applications"; and "I can identify ethical issues surrounding artificial intelligence". The effect size, expressed by Cohen's d, painted a similar picture. Cohen's d was below 0.5 for only five items, indicating

a small effect. All other items had at least a medium effect (d > 0.5), with 14 items showing a strong effect (d > 0.8).

When the analysis was performed at the factor level, similar results were found. There was a significant difference between the retrospective assessments and the post-assessments for all three factors, $t(48) = 4.38$, $p < 0.001$, $d = 1.25$ for the technical-understanding (TU) factor, $t(48) = 3.47$, $p < 0.001$, $d = 1.21$ for the critical-appraisal (CA) factor, and $t(48) = 3.30$, $p < 0.001$, $d = 0.93$ for the practical-application (PA) factor, respectively (see Figure 1).

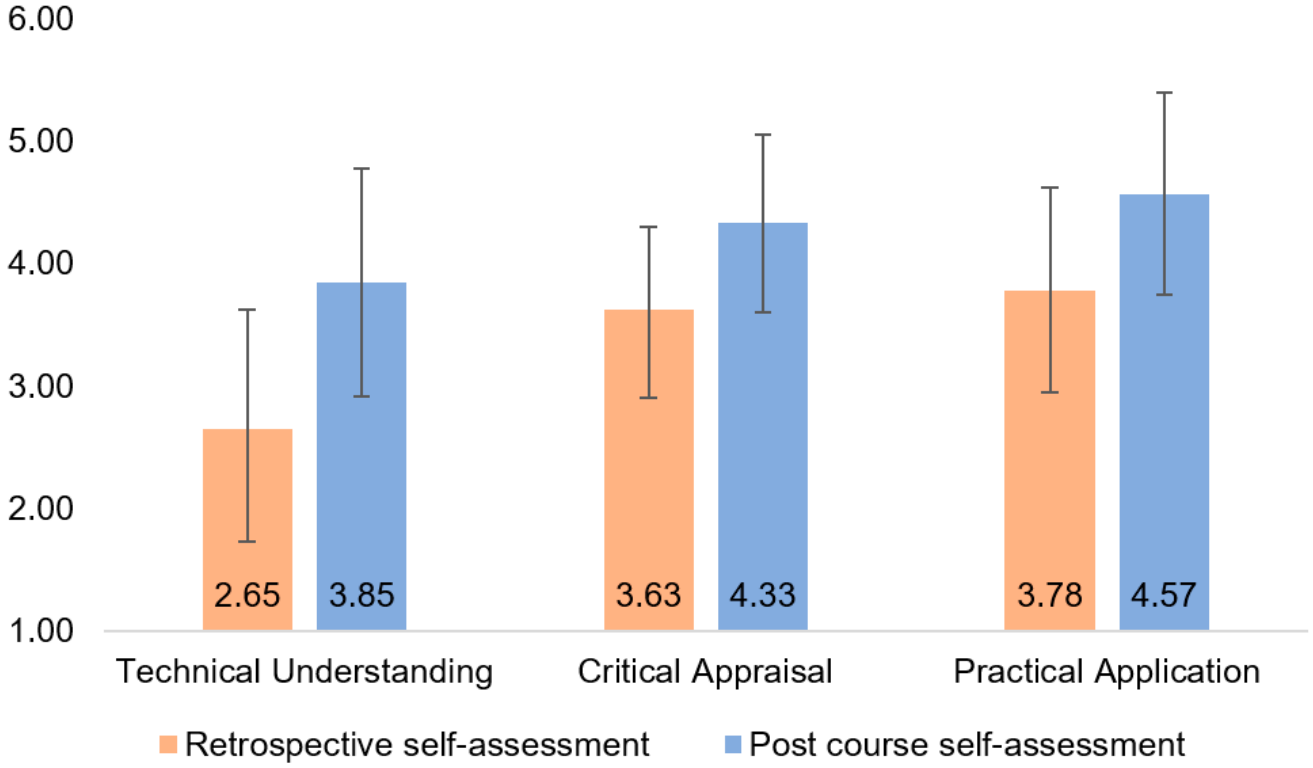


Figure 1. Mean values of retrospective assessment and post-self-assessment of all three factors.

As described above, the more informative change parameter CSA gain is reported in the following section. It may be due to the retrospective/post assessments being less susceptible to bias than the traditional pre/post assessments that all CSA gain values were positive, as negative values would imply a loss of AI literacy over the course. However, the actual height of CSA gain varied greatly from item to item (see Figure 2). Some items showed rather small improvements, in the range of 15 to 30%, which could have been due to several reasons. First, students may have already assessed themselves as relatively confident in their relevant competence before attending the course (i.e., the retrospective values were already fairly high). For example, items such as "I can identify ethical issues surrounding artificial intelligence" were rated relatively highly even before attending the course, with an average retrospective-assessment rating of M = 4.20 (SD = 1.02) and a post-assessment rating of M = 4.48 (SD = 0.94) on a six-point Likert scale, leading to a CSA gain of 15.6%. Second, the AI course may have failed to teach the corresponding aspects that are represented by the item. An example of this is the item, "I can describe potential legal problems that may arise when using artificial intelligence", which had an average retrospective-assessment rating of M = 2.88 (SD = 1.14) and a post-assessment rating of M = 3.44 (SD = 1.39). For other items, however, acceptable to good CSA gain was found in the range of 40%, up to more than 50%. A positive example was the item, "I can explain what the term 'artificial neural network' means", for which a CSA gain of 56.8% was found. This was somewhat unsurprising, because one of the main aspects of the course was the teaching of competencies related to artificial neural networks.

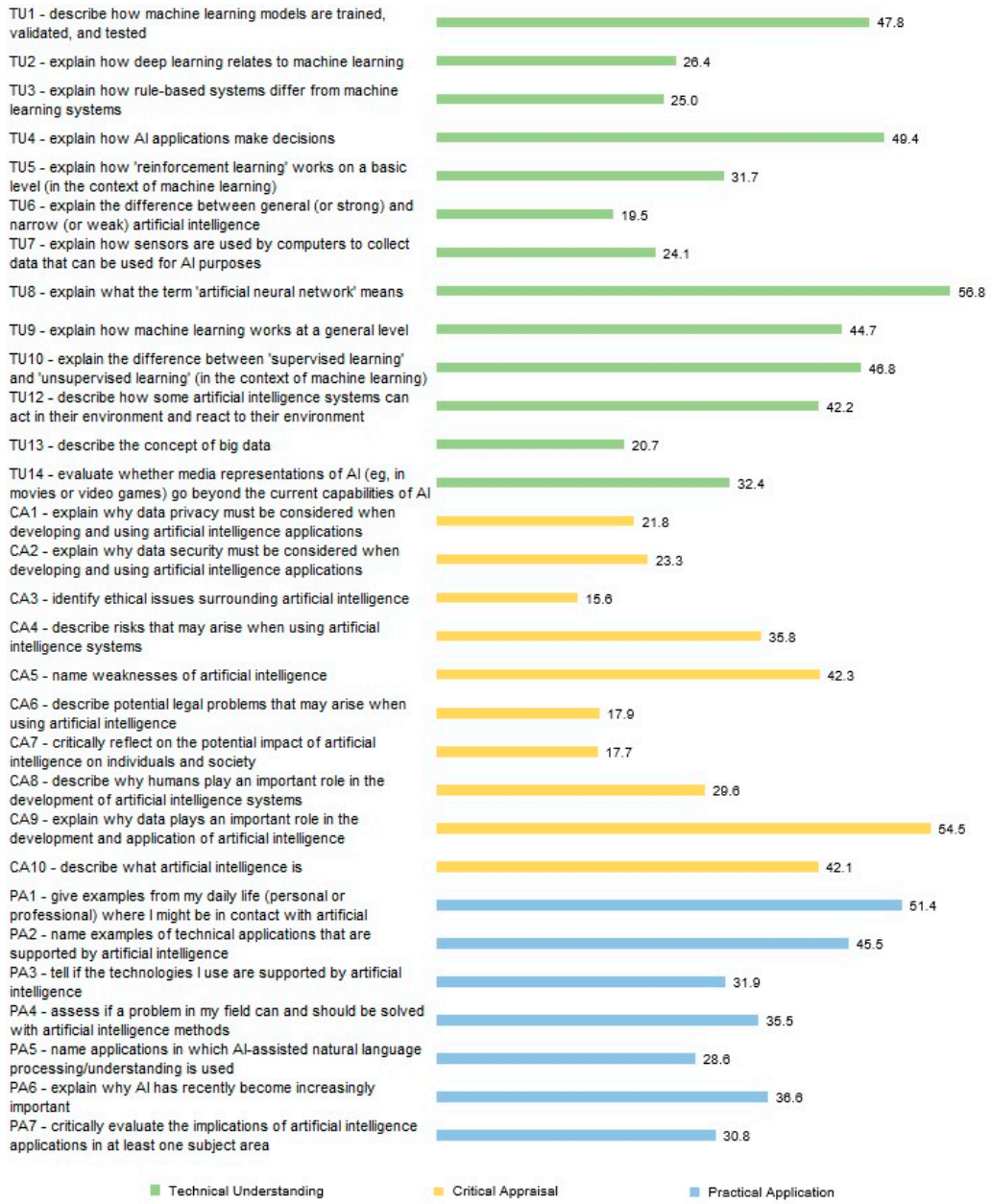


Figure 2. Mean CSA gain values, in percentages, for all three factors. Note: CSA gain can reach values from −100% to +100%. Item TU11 was excluded, due to a technical error.

By calculating the CSA gain average over all items of the respective factor, the differences of the CSA gain between the individual items were removed. The CSA gain values were 36.9% for the TU factor, 30.1% for the CA factor, and 37.2% for the PA factor.

The reliability of the individual subscales (i.e., factors) of the SNAIL questionnaire can be rated as good (>0.80) to excellent (>0.90). Cronbach's α for the retrospective assessments was 0.90, 0.92, and 0.85 for the three factors TU, CA, and PA, respectively. The internal consistency of the scales at post-assessment was slightly lower at 0.83 (TU factor), 0.90 (CA factor), and 0.88 (PA factor), but still in the good-to-excellent range.

3.3. Relationship between AI Literacy and Attitudes toward AI

The Pearson product–moment correlations between the SNAIL factors (TU, CA, and PA) and the two factors of the "Attitudes Toward Artificial Intelligence" scale, namely "fear" and "acceptance", did not reach statistical significance. In addition, the correlations

between the two "Attitudes Toward Artificial Intelligence" scale factors and the mean scores of the retrospective assessment and the post-assessment of the three SNAIL factors were not significant.

3.4. Relationship between AI Education Prior to the Course and Learning Gains

Participants were asked whether and to what extent they had attended other AI courses prior to attending the evaluated course (variable name "other courses"). They were also asked whether and to what extent they had used other means of AI education (such as instructional videos and books.; variable name "other AI education").

Although the correlations between "other courses" and CSA gain were negative for all three factors, this correlation did not reach statistical significance (significance level of $\alpha = 0.5$). Interestingly, the effect was reversed for the variable "other AI education", as all correlations were positive. However, these correlations did not reach statistical significance.

Thereafter, we examined in detail how attending other courses or using other AI educational opportunities affected the absolute retrospective self-assessment and the post-self-assessment. Attending other courses was strongly positively correlated with the assessment scores on the TU factor, with Spearman's $\rho = 0.556$, $p < 0.01$ and Spearman's $\rho = 0.402$, $p = 0.046$ for the retrospective assessment and post-assessment scores, respectively. However, the correlations between "other courses" and the CA or PA factor did not reach statistical significance (see Table 1). Using other means of AI education was also strongly correlated with assessment scores on the TU factor, with Spearman's $\rho = 0.557$, $p < 0.01$ (retrospective assessment) and Spearman's $\rho = 0.684$, $p < 0.001$ (post-assessment). In this case, however, significant positive correlations were also present for the other two factors in the post-assessment (Spearman's $\rho = 0.492$, $p = 0.013$ for the CA factor and Spearman's $\rho = 0.524$, $p < 0.01$ for the PA factor), but not in the retrospective assessment.

Table 1. Correlations between CSA gain, retrospective/post assessment for each factor, and the usage of other courses or other AI education.

Variable	1	2	3	4	5	6	7	8	9	10	11
1. CSA TU	—										
2. CSA CA	0.382 0.059	—									
3. CSA PA	0.556 ** 0.004	0.573 ** 0.003	—								
4. TU—retrospective	−0.106 0.614	0.004 0.985	−0.107 0.611	—							
5. TU—post	0.416 * 0.039	0.385 0.058	0.321 0.118	0.678 ** <0.001	—						
6. CA—retrospective	−0.254 0.220	−0.234 0.260	−0.169 0.419	0.357 0.080	0.271 0.191	—					
7. CA—post	0.233 0.263	0.471 * 0.018	0.415 * 0.039	0.146 0.486	0.505 * 0.010	0.588 ** 0.002	—				
8. PA—retrospective	−0.347 0.089	−0.023 0.914	−0.096 0.648	0.239 0.250	0.159 0.448	0.523 ** 0.007	0.266 0.198	—			
9. PA—post	0.206 0.323	0.519 ** 0.008	0.593 ** 0.002	0.073 0.730	0.447 * 0.025	0.302 0.142	0.729 ** <0.001	0.589 ** 0.002	—		
10. Other courses	−0.111 0.597	−0.348 0.088	−0.244 0.239	0.556 ** 0.004	0.402 * 0.046	0.364 0.074	0.066 0.755	0.043 0.838	−0.147 0.482	—	
11. Other AI education	0.376 0.064	0.309 0.133	0.292 0.157	0.557 ** 0.004	0.684 ** <0.001	0.271 0.190	0.492 * 0.013	0.332 0.105	0.524 ** 0.007	0.171 0.413	—

* $p < 0.05$. ** $p < 0.01$. Note. TU: technical-understanding factor, CA: critical-appraisal factor, PA: practical-application factor. Retrospective: mean retrospective assessment; post: mean current assessment (after taking the course).

4. Discussion

4.1. Contextualizing of Results

This study investigated the suitability of Laupichler et al.'s AI-literacy scale, SNAIL [27], for evaluating AI courses. First, evidence was found that suggested that a simple adaptation of the original SNAIL questionnaire allows its use in the context of course evaluations. The

adapted version of SNAIL seems to be able to differentiate between learning objectives and to identify strengths as well as weaknesses of AI courses, providing a balance in evaluating AI-literacy courses. The results indicate that the adapted version of SNAIL is valid and corresponds to the actual AI competencies of course participants. This was supported by several lines of evidence derived from answers to the three research questions.

First, the average learning gain, represented by CSA gain, was particularly pronounced for technical items such as "I can describe how machine learning models are trained, validated, and tested." This was to be expected, as the course focused mainly on the technological methods of AI. On the other hand, items that covered content that did not occur in the course had low CSA gain values. Accordingly, the critical-appraisal factor had a lower overall learning gain because it included some items that dealt with the ethical, legal, and social aspects of AI, which were not covered in the evaluated course. Thus, our study provided initial evidence for an affirmative answer to RQ1. The adapted SNAIL questionnaire can be used to assess the learning gains of AI courses in a valid and reliable way.

Second, RQ2 asked whether AI literacy and attitudes toward AI were correlated, as this might be expected but would not be helpful for a criterion-valid assessment of learning gain. However, the learning gain scores of the three SNAIL factors correlated only very weakly with the two factors of the "Attitudes Toward Artificial Intelligence" scale. This could be an indication of discriminant validity because, in theory, AI literacy and attitudes toward AI are assumed to be two different constructs, representing cognitive/skill and affective aspects, respectively [9,16,20–22].

Third, RQ3 was asked to determine the extent to which participation in other AI courses (in addition to the AI course evaluated here) influenced learning gains. The use of other educational opportunities correlated significantly with retrospective and current self-assessment scores (especially on the TU factor), but not with learning gains. Accordingly, people who had already taken part in many AI courses (especially with a focus on technical understanding) tended to rate their AI literacy higher than did people who had comparatively little AI education. At the same time, however, the amount of actual learning (CSA gain) was unaffected by attending other AI courses. This made sense, because AI education before the course should already have had a positive influence on the retrospective assessment of one's own AI literacy, which in turn should have led to lower learning gains, due to a ceiling effect.

Furthermore, the reliability of the subscales of the adapted SNAIL also seemed to be satisfactory, as illustrated by the good-to-excellent internal consistency. Cronbach's α was high enough for the retrospective items, as well as for the post-assessment items, to justify the use of the adapted scale. In fact, the internal consistency of the scale was so satisfactory that one could consider removing some items to improve test efficiency. This would reduce the length of the questionnaire, which could increase participation rates, especially in the context of course evaluation.

The retrospective/post assessment seemed to yield valid and reliable results. However, it should once again be emphasized that the use of comparative self-assessment gains is particularly suitable for identifying between-subject differences, as well as differences between individual items [28,34]. If future research projects seek to apply adapted AI-literacy scales, it might, therefore, be advisable to calculate the comparative self-assessment gain rather than the traditional mean comparison via *t*-tests. If *t*-tests are nevertheless (additionally) conducted, the effect size, expressed, for example, by Cohen's d, should be included in any case. In this way, the strength of the learning effect can be estimated, at least in relative terms.

While the primary objective of this study was the validation of the SNAIL measurement instrument, a brief examination of the course itself and potential areas for improvement in course content was warranted. As previously noted, the course in question was not a general AI-literacy course, but focused on the technical aspects of AI. Consequently, most items that were subsumed under the technical-understanding factor yielded favorable

results. However, the findings also raised the question as to why course participants did not feel confident, for instance, in explaining the relationship between deep learning and machine learning (item TU2). Therefore, these aspects may merit heightened attention in future course iterations. Given this study's concentration on technical AI methods, our primary aim was not to enhance the learning outcomes regarding ethical items in the future. Nevertheless, course instructors may contemplate providing students with additional resources if they wish to further their knowledge in these areas or provide a broader AI-literacy course.

4.2. Limitations

As with any research, this study had some limitations. Even though students from different disciplines and backgrounds participated in the course, this study examined only one course (i.e., a single sample). In addition, some of the participants came from computer science backgrounds and, thus, they were both technically inclined and likely to have been familiar with some of the terminology. The original (un-adapted) SNAIL, however, was aimed at non-experts, i.e., individuals who had received little formal AI education. Furthermore, the selection of the SNAIL questionnaire was, although not completely without reason, relatively arbitrary. For reasons of evaluation efficiency, it was not possible to examine how respondents would have responded to other adapted questionnaires (e.g., [23–25]).

4.3. Future Research Directions

Future research projects should test the adapted SNAIL questionnaire in additional contexts and for larger courses. For example, the adapted scale should be used to evaluate courses in which the focus is not on technological AI aspects but on ethical, legal, and social features. The evaluation of AI teaching in specific disciplines would also be of interest, as promoted by Schleiss et al. [35]. For instance, whether the learning effects for medical students and engineering students differ for certain items could be investigated. In addition, it would be important to investigate whether complete novices would have produced a similar response scheme as that of the sample in this study. Furthermore, future evaluation studies should compare different AI-literacy-assessment tools to identify similarities and differences. Moreover, the relationship between AI literacy and attitudes toward AI should be further investigated in larger, preferably experimental, research projects to explore the causal direction of possible correlations. Finally, whether the number of items could be reduced without decreasing the internal consistency of the questionnaire should be examined, as this could increase testing efficiency. In addition to internal consistency, test–retest reliability could be investigated. This would be possible, for example, by carrying out a retention test some weeks or months after an AI course is completed.

5. Conclusions

This study presented preliminary evidence suggesting that even small adaptations of existing AI-literacy scales enables their use as AI-course-evaluation instruments. The combination of retrospective self-assessments on one's own competencies before starting a course and self-assessment after attending a course seems to lead to valid results, while simultaneously ensuring test efficiency. Overall, this study contributes to the development of valid, reliable, and efficient AI-course-evaluation instruments that allow a systematic assessment and improvement of AI education.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/educsci13100978/s1, Supplementary Material S1: Adaptation of the original questionnaire used in the study. Supplementary Material S2: Data Set.

Author Contributions: Conceptualization, M.C.L. and J.S.; methodology, M.C.L. and A.A.; formal analysis, M.C.L.; investigation, J.-O.P. and J.S.; resources, J.-O.P.; writing—original draft preparation, M.C.L.; writing—review and editing, J.S. and A.A.; visualization, M.C.L. and J.S.; project administration, J.-O.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The de-identified dataset on which the analyses in this study are based is available in Supplementary Material S2.

Acknowledgments: We want to express our appreciation to our supervisors Tobias Raupach and Sebastian Stober for their trust and support in implementing this project. Moreover, we also want to thank the reviewers for their feedback and constructive comments in the review process.

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

References

1. Merriam-Webster. Artificial Intelligence. Merriam-Webster: Springfield, MA, USA, 2023. Available online: https://www.merriam-webster.com/dictionary/artificial%20intelligence (accessed on 14 September 2023).
2. Cambridge Dictionary. Artificial Intelligence. Cambridge University Press: Cambridge, UK, 2023. Available online: https://dictionary.cambridge.org/dictionary/english/artificial-intelligence (accessed on 14 September 2023).
3. Bennett, J.; Lanning, S. The Netflix Prize. In Proceedings of the KDD Cup and Workshop, San Jose, CA, USA, 12 August 2007; Volume 2007, p. 35.
4. Skinner, G.; Walmsley, T. Artificial intelligence and deep learning in video games—A brief review. In Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 23–25 February 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 404–408. [CrossRef]
5. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef]
6. Li, B.H.; Hou, B.C.; Yu, W.T.; Lu, X.B.; Yang, C.W. Applications of artificial intelligence in intelligent manufacturing: A review. *Front. Inf. Technol. Electron. Eng.* **2017**, *18*, 86–96. [CrossRef]
7. Schleiss, J.; Bieber, M.; Manukjan, A.; Kellner, L.; Stober, S. An interdisciplinary competence profile for AI in engineering. In *Towards a New Future in Engineering Education, New Scenarios That European Alliances of Tech Universities Open Up*; Universitat Politècnica de Catalunya: Barcelona, Spain, 2022; pp. 1601–1609. [CrossRef]
8. Long, D.; Magerko, B. What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–16. [CrossRef]
9. Kandlhofer, M.; Steinbauer, G.; Hirschmugl-Gaisch, S.; Huber, P. Artificial intelligence and computer science in education: From kindergarten to university. In Proceedings of the 2016 IEEE Frontiers in Education Conference (FIE), Erie, PA, USA, 12–15 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–9. [CrossRef]
10. Su, J.; Yang, W. Artificial intelligence in early childhood education: A scoping review. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100049. [CrossRef]
11. Eguchi, A.; Okada, H.; Muto, Y. Contextualizing AI education for K-12 students to enhance their learning of AI literacy through culturally responsive approaches. *KI Künstl. Intell.* **2021**, *35*, 153–161. [CrossRef] [PubMed]
12. Casal-Otero, L.; Catala, A.; Fernández-Morante, C.; Taboada, M.; Cebreiro, B.; Barro, S. AI literacy in K-12: A systematic literature review. *Int. J. STEM Educ.* **2023**, *10*, 29. [CrossRef]
13. Ng, D.T.K.; Leung, J.K.L.; Su, M.J.; Yim, I.H.Y.; Qiao, M.S.; Chu, S.K.W. *AI Literacy in K-16 Classrooms*; Springer International Publishing: Berlin/Heidelberg, Germany, 2023. [CrossRef]
14. Southworth, J.; Migliaccio, K.; Glover, J.; Reed, D.; McCarty, C.; Brendemuhl, J.; Thomas, A. Developing a model for AI Across the curriculum: Transforming the higher education landscape via innovation in AI literacy. *Comput. Educ. Artif. Intell.* **2023**, *4*, 100127. [CrossRef]
15. Laupichler, M.C.; Aster, A.; Schirch, J.; Raupach, T. Artificial intelligence literacy in higher and adult education: A scoping literature review. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100101. [CrossRef]
16. Kirkpatrick, D.; Kirkpatrick, J. *Evaluating Training Programs: The Four Levels*; Berrett-Koehler Publishers: Oakland, CA, USA, 2006.
17. Ng, D.T.K.; Leung, J.K.L.; Chu, K.W.S.; Qiao, M.S. AI literacy: Definition, teaching, evaluation and ethical issues. *Proc. Assoc. Inf. Sci. Technol.* **2021**, *58*, 504–509. [CrossRef]
18. Weber, P. Unrealistic Optimism Regarding Artificial Intelligence Opportunities in Human Resource Management. *Int. J. Knowl. Manag.* **2023**, *19*, 1–19. [CrossRef]
19. Schepman, A.; Rodway, P. Initial validation of the general attitudes towards Artificial Intelligence Scale. *Comput. Hum. Behav. Rep.* **2020**, *1*, 100014. [CrossRef] [PubMed]

20. Schepman, A.; Rodway, P. The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with personality, corporate distrust, and general trust. *Int. J. Hum. Comput. Interact.* **2022**, *39*, 2724–2741. [CrossRef]
21. Sindermann, C.; Sha, P.; Zhou, M.; Wernicke, J.; Schmitt, H.S.; Li, M.; Sariyska, R.; Stavrou, M.; Becker, B.; Montag, C. Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. *KI Künstl. Intell.* **2021**, *35*, 109–118. [CrossRef]
22. Wang, B.; Rau PL, P.; Yuan, T. Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behav. Inf. Technol.* **2022**, *42*, 1324–1337. [CrossRef]
23. Pinski, M.; Benlian, A. AI Literacy-Towards Measuring Human Competency in Artificial Intelligence. In Proceedings of the 56th Hawaii International Conference on System Sciences, Maui, HI, USA, 3–6 January 2023.
24. Carolus, A.; Koch, M.; Straka, S.; Latoschik, M.E.; Wienrich, C. 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. *arXiv* **2023**, arXiv:2302.09319. [CrossRef]
25. Ng, D.T.K.; Leung, J.K.L.; Chu, S.K.W.; Qiao, M.S. Conceptualizing AI literacy: An exploratory review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100041. [CrossRef]
26. Laupichler, M.C.; Aster, A.; Raupach, T. Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Comput. Educ. Artif. Intell.* **2023**, *4*, 100126. [CrossRef]
27. Laupichler, M.C.; Aster, A.; Raupach, T. *Development of the "Scale for the Assessment of Non-Experts' AI Literacy"—An Exploratory Factor Analysis*; Institute of Medical Education, University Hospital Bonn: Bonn, Germany, 2023.
28. Raupach, T.; Münscher, C.; Beißbarth, T.; Burckhardt, G.; Pukrop, T. Towards outcome-based programme evaluation: Using student comparative self-assessments to determine teaching effectiveness. *Med. Teach.* **2011**, *33*, e446–e453. [CrossRef]
29. Howard, G.S. Response-shift bias: A problem in evaluating interventions with pre/post self-reports. *Eval. Rev.* **1980**, *4*, 93–106. [CrossRef]
30. Sibthorp, J.; Paisley, K.; Gookin, J.; Ward, P. Addressing response-shift bias: Retrospective pretests in recreation research and evaluation. *J. Leis. Res.* **2007**, *39*, 295–315. [CrossRef]
31. Tsang, S.; Royse, C.F.; Terkawi, A.S. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi J. Anaesth.* **2017**, *11*, 80–89. [CrossRef] [PubMed]
32. Harkness, J.; Pennell, B.E.; Schoua-Glusberg, A. Survey questionnaire translation and assessment. In *Methods for Testing and Evaluating Survey Questionnaires*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2004; pp. 453–473. [CrossRef]
33. Chang, A.M.; Chau, J.P.; Holroyd, E. Translation of questionnaires and issues of equivalence. *J. Adv. Nurs.* **2001**, *29*, 316–322. [CrossRef] [PubMed]
34. Schiekirka, S.; Reinhardt, D.; Beibarth, T.; Anders, S.; Pukrop, T.; Raupach, T. Estimating learning outcomes from pre-and posttest student self-assessments: A longitudinal study. *Acad. Med.* **2013**, *88*, 369–375. [CrossRef] [PubMed]
35. Schleiss, J.; Laupichler, M.C.; Raupach, T.; Stober, S. AI Course Design Planning Framework: Developing Domain-Specific AI Education Courses. *Educ. Sci.* **2023**, *13*, 954. [CrossRef]

BMC Medical Education

RESEARCH Open Access

Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments

Matthias Carl Laupichler[1*], Alexandra Aster[1], Marcel Meyerheim[2], Tobias Raupach[1] and Marvin Mergen[2]

Abstract

Background Artificial intelligence (AI) is becoming increasingly important in healthcare. It is therefore crucial that today's medical students have certain basic AI skills that enable them to use AI applications successfully. These basic skills are often referred to as "AI literacy". Previous research projects that aimed to investigate medical students' AI literacy and attitudes towards AI have not used reliable and validated assessment instruments.

Methods We used two validated self-assessment scales to measure AI literacy (31 Likert-type items) and attitudes towards AI (5 Likert-type items) at two German medical schools. The scales were distributed to the medical students through an online questionnaire. The final sample consisted of a total of 377 medical students. We conducted a confirmatory factor analysis and calculated the internal consistency of the scales to check whether the scales were sufficiently reliable to be used in our sample. In addition, we calculated t-tests to determine group differences and Pearson's and Kendall's correlation coefficients to examine associations between individual variables.

Results The model fit and internal consistency of the scales were satisfactory. Within the concept of AI literacy, we found that medical students at both medical schools rated their technical understanding of AI significantly lower (M_{MS1} = 2.85 and M_{MS2} = 2.50) than their ability to critically appraise (M_{MS1} = 4.99 and M_{MS2} = 4.83) or practically use AI (M_{MS1} = 4.52 and M_{MS2} = 4.32), which reveals a discrepancy of skills. In addition, female medical students rated their overall AI literacy significantly lower than male medical students, $t(217.96) = -3.65$, $p < .001$. Students in both samples seemed to be more accepting of AI than fearful of the technology, $t(745.42) = 11.72$, $p < .001$. Furthermore, we discovered a strong positive correlation between AI literacy and positive attitudes towards AI and a weak negative correlation between AI literacy and negative attitudes. Finally, we found that prior AI education and interest in AI is positively correlated with medical students' AI literacy.

Conclusions Courses to increase the AI literacy of medical students should focus more on technical aspects. There also appears to be a correlation between AI literacy and attitudes towards AI, which should be considered when planning AI courses.

Keywords Artificial intelligence, AI literacy, Attitudes towards AI, Confirmatory factor analysis, Medical students, Questionnaire

*Correspondence:
Matthias Carl Laupichler
matthias.laupichler@ukbonn.de

[1]Institute of Medical Education, University Hospital Bonn, Venusberg Campus 1, 53127 Bonn, Germany
[2]Department of Pediatric Oncology and Hematology, Faculty of Medicine, Saarland University, Homburg, Germany

Background

The rise of artificial intelligence in medicine

The potential benefits of using artificial intelligence (AI) for the healthcare sector have been discussed for decades [1–3]. However, while in the past the focus was predominantly on theoretical considerations and ambitious future scenarios, AI and its most important subfield, machine learning, have now become an integral part of healthcare [4]. In addition to clinical practice, AI applications have reached medical schools and are being used by students, educators and administrators alike to improve teaching and learning [5–6].

At the same time, a "consensus on what and how to teach AI" [7, p1] in the medical curriculum appears to be lacking, and although there are individual elective courses attempting to foster AI competencies [8–9], the majority of medical students still receive very little AI education [10–11]. However, learning basic AI skills will be critical for all future physicians to fulfill their roles as professionals, communicators, collaborators, leaders, healthcare advocates, and scholars, as all of these roles will be increasingly permeated by AI [12].

Medical student's "AI literacy" and related constructs

In recent years, basic AI skills have often been referred to as AI literacy [13]. AI literacy can be 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" [13, p2]. Thus, AI literacy for medical professionals is less about the ability to develop AI programs or to conduct clinical research with AI, but rather about the ability to interact with AI and use AI applications in the day-to-day provision of healthcare services.

Despite the large number of studies investigating the attitudes and feelings of medical students towards AI (i.e., the affective component of AI interaction [14–16]),, research projects have rarely focused on AI knowledge (i.e., conceptual understanding of AI) or even AI skills (i.e., ability to identify, use, and scrutinize AI applications reasonably). Mousavi Baigi et al. [17] found that all 38 studies they included in their literature review reported some kind of investigation on healthcare students' "attitudes towards AI" (ATAI), while only 26 of the included studies stated that they had asked participants about their AI knowledge. However, a closer look at the studies showed that most of them assessed AI knowledge superficially and focused more on familiarity with AI. Furthermore, only six of the included studies looked at the AI skills of medical students. However, since the concept of AI literacy not only encompasses AI knowledge, but also includes practical AI competencies (such as the ability to recognize the use of AI applications in technical systems), this empirical foundation is not sufficient to make reliable statements about the AI literacy of medical students.

Karaca et al. [18] were among the few who took a systematic approach to studying a closely related but not identical concept to AI literacy. They developed the so-called MAIRS-MS questionnaire instrument specifically designed to assess the "AI readiness" of medical students. AI readiness can be interpreted as a link between attitudes towards AI and knowledge and skills for dealing with AI. Aboalshamat et al. [19] used the MAIRS-MS instrument and found that medical students in a Saudi Arabian sample rated their AI readiness rather poorly with an average score of 2.5 on a Likert scale of 1 (negative) to 5 (positive). Due to the influence of socio-cultural differences and the country-specific characteristics of the medical curricula on the data, these results can only be transferred to other countries to a limited extent.

While the assessment of medical students' AI readiness is an important endeavor, only few studies are currently dealing with competence-focused AI literacy. Evaluating these competences, however, could provide a sufficient baseline to identify knowledge gaps and, if necessary, to revise the medical curricula by developing and implementing appropriate AI courses.

The importance of validated assessment instruments

A major disadvantage of the few available studies on the AI literacy of medical students is the attempt to assess AI literacy with self-developed and non-validated questionnaires. Thus, accuracy and reliability of their measures have not been established. In this study, we therefore used the "Scale for the assessment of non-experts' AI literacy" (SNAIL), which was validated in several peer-reviewed studies. In a pilot study, the scale's items were generated, refined, and subsequently evaluated for their relevance through a Delphi expert survey. As a result, a set of content-valid items covering the entire breadth of AI literacy was available to researchers and practitioners alike [20]. Subsequently, the itemset was presented to a large sample of non-experts who assessed their individual AI literacy. Based on this dataset, an exploratory factor analysis was conducted, which firstly identified the three subscales "Technical Understanding" (TU), "Critical Appraisal" (CA), and "Practical Application" (PA), and secondly excluded some redundant items [21]. In another study, it was demonstrated that the final SNAIL questionnaire is also suitable for assessing AI literacy among university students who have just completed an AI course [22].

Even though medical students' ATAI has been assessed in multiple instances (as described above), very few studies have attempted to investigate the correlative (let alone causal) relationship between medical students' AI literacy and ATAI. Furthermore, to our knowledge, the

studies that have recorded both constructs did not use validated and standardized measurement instruments to investigate ATAI. In this study, the ATAI construct was therefore assessed using the "Attitudes towards Artificial Intelligence" scale [23], which has been validated in several languages. This scale was also developed in a systematic way, using principal component analysis and multiple samples. In addition, the reliability of the ATAI scale was evaluated and found to be acceptable. A major advantage of the scale is its efficiency, since the instrument comprises only 5 items that load on two factors ("fear" and "acceptance" of AI) in total.

Research objective

With this study we wanted to answer five research questions (RQs). RQ1 deals with medical students' assessment of their individual AI literacy. In particular, we aimed to assess the AI literacy sub-constructs described above (TU, CA, PA), as the identification of literacy gaps is paramount for the development of appropriate medical education programs.

RQ1: How do medical students rate their individual AI literacy overall and for the factors "Technical Understanding", "Critical Appraisal", and "Practical Application"?

Regarding RQ2, we wanted to investigate the extent to which the assessment of one's own AI literacy is associated with factors such as gender, age or semester. It is conceivable, for example, that older medical students would rate their AI skills lower than younger students, as younger students might consider themselves to be more technically adept. On the contrary, older medical students might generally consider themselves to be more competent across various competence areas, as they have already acquired extensive knowledge and skills during their academic training.

RQ2: Are there statistically significant differences in AI literacy self-assessment between (a) older and younger, (b) male or female and (c) less and more advanced students?

Furthermore, the medical students' ATAI is covered by RQ3. It is important to know whether medical students have a positive or negative attitude towards AI, as this can have a decisive influence on the acceptance of teaching programs designed to foster AI literacy.

RQ3: How do medical students rate their individual attitudes towards AI?

RQ4 follows from the ideas presented in RQ3, as it is possible that the two constructs AI literacy and ATAI are related. In addition to efforts to increase AI literacy, interventions might be required to change attitudes towards AI.

RQ4: Are the two constructs AI literacy and attitudes towards AI and their respective sub-constructs significantly correlated?

The last RQ deals with previous education and interest in AI, since both aspects might increase AI literacy. We asked if the medical students had attended courses on AI in the past or if they had already educated themselves on the topic independently. In addition, interest in the subject area of AI was surveyed.

RQ5: Is there a correlative relationship between AI education or interest in AI and the AI literacy of medical students?

Methods

Questionnaires

We used the "Scale for the assessment of non-experts' AI literacy" (SNAIL) by Laupichler et al. [20] to assess the AI literacy of medical students. The SNAIL instrument assesses AI literacy on three latent factors: Technical Understanding (14 items focusing on basic machine learning methods, the difference between narrow and strong AI, the interplay between computer sensors and AI, etc.), Critical Appraisal (10 items focusing on data privacy and data security, ethical issues, risks and weaknesses, etc.), and Practical Application (7 items focusing on AI in daily life, examples of technical applications supported by AI, etc.). Each item represents a statement on one specific AI literacy aspect (e.g., "I can give examples from my daily life (personal or professional) where I might be in contact with artificial intelligence."), which is rated on a 7-point Likert scale from 1 ("strongly disagree") to 7 ("strongly agree"). Furthermore, we integrated the "Attitudes towards Artificial Intelligence" scale (ATAI scale) by Sindermann et al. [23]. The ATAI scale assesses participants' "acceptance" of AI with three items and the "fear" of AI with two items. Although an eleven-point Likert scale was used in the original study, we decided to use a 7-point scale (as in SNAIL) to ensure that the items were presented as uniformly as possible. Since the sample described here consisted of German medical students, the validated German questionnaire version was used for both SNAIL [22] and ATAI [23]. All SNAIL and ATAI items were presented in random order.

We included an attention control item ("mark box 3 here.") and a bogus item for identifying nonsensical

responses ("I consider myself among the top 10 AI researchers in the world."), which were randomly presented. Additionally, we used 4-point Likert scales to gather information on whether the students had previously taken AI courses or had educated themselves about AI through other sources. The values ranged from 1 ("I have never attended a course on AI." and "I haven't used other ways to learn about AI yet.") to 4 ("I have already attended AI courses with a workload of more than 120 hours." and "I have informed myself very extensively about AI in other ways."). In addition, we used a 7-point Likert scale to assess students' interest in the field of AI, with lower values indicating less interest in AI. Finally, we inquired about the participants' age, gender, and the total number of semesters they were enrolled in their study program.

Procedure

The study was conducted at two German medical schools (MS1 and MS2) between October and December 2023 after receiving positive ethical approval from the local ethics committees (file number 151/23-EP at medical school 1 and 244/21 at medical school 2). Invitations to participate in the study were distributed via university-exclusive social media groups and online education platforms, mailing lists, and advertisements in lectures. Medical students who were at least 18 years old were eligible for the study and could access the online questionnaire after giving their informed consent to participate. The questionnaire was accessible via a QR code on their mobile device and participants received no financial incentive to take part in the study. The average time it took respondents to complete the questionnaire was 05:52 min (SD = 02:27 min).

Data analysis

The data were analyzed using RStudio (Posit Software, Version 2023). The visual presentation of the results was carried out using Microsoft Excel (Microsoft, Version 2016). Significance level was set at $\alpha = 0.05$ for all statistical tests.

Independent two-sample t-tests were carried out to evaluate differences between groups (e.g., differences in AI literacy between MS1 and MS2). To check the requirements of t-tests, the data were examined for outliers, Shapiro-Wilk tests were carried out to check for normal distribution and Levene tests were run to check for variance homogeneity. In case of variance heterogeneity, Welch's t-test was used. To check for differences considering age and semester distribution between MS1 and MS2, the Mann-Whitney-Wilcoxon-Test was used. Fisher's test served to examine if there was a difference in the gender ratio.

Pearson's correlation was calculated to determine the correlative relationship between continuous variables and Kendall's τ coefficient was computed for ordinal variables. In addition, the factor structure of the two validated instruments (SNAIL and ATAI) was analyzed using a confirmatory factor analysis (CFA). We checked the prerequisites for conducting a confirmatory factor analysis, including univariate and multivariate skewness and kurtosis (using Mardia's test for the multivariate analyses), the number and distribution of missing values, and whether the data differed significantly between the two medical schools, which would necessitate separate CFAs for each subsample. Due to the ordinal scaled variables and multivariate non-normality, we used polychoric correlation matrices to perform the CFA. We calculated the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR) as measures of model fit. As part of this analysis, the internal consistency, represented as the reliability measure Cronbach's alpha, was also calculated for the overall scales as well as for the corresponding subscales.

Results

Participant characteristics

Of 444 completed questionnaires, 28 (6%) participants had to be excluded since they omitted more than 3 (10%) of the SNAIL items. In addition, 8 (2%) participants were excluded because they indicated that they did not study medicine. Furthermore, 24 (5%) participants were excluded since they did not answer or answered incorrectly to the attention control item. Finally, 7 (2%) participants had to be excluded because they agreed, at least in part, to the bogus item (i.e., counting themselves among the "Top 10 AI researchers"). Accordingly, the final sample consisted of a total of 377 (85%) subjects, of which 142 (38% of the final sample) came from MS1 and 235 (62% of the final sample) from MS2.

The participants were on average 22.5 years old ($Mdn = 22$, $Min = 18$, $Max = 36$, $SD = 3.2$) and on average in their 5th semester ($M = 4.7$, $Mdn = 5$, $Min = 1$, $Max = 13$, $SD = 2.6$). Of the participants, 259 (69%) identified as female, 114 (30%) as male and one person as diverse. A Mann-Whitney-Wilcoxon test showed that the two medical schools differed significantly from each other in terms of the age of the participants, $U = 13658.00$, $Z = -2.63$, $p < .01$. The participants in MS1 were on average 0.9 years younger than the participants in MS2. There was no significant difference regarding participants' semesters between the two medical schools, and according to a Fisher's test, the gender distribution was similar.

Most participants stated that they had received little or no AI training. Of all participants, 342 (91%) stated

that they had never attended an AI course. Only 28 (7%) had attended a course of up to 30 h and 6 (2%) people had attended a course of more than 30 h. In addition, a total of 338 (90%) of the participants stated that they never (n=177; 47%) or only irregularly (n=161; 43%) educated themselves on AI using other sources (such as videos, books, etc.). Only 32 (8%) respondents stated that they regularly educated themselves on AI with the help of other sources, and only 5 (1%) participants stated that they had already educated themselves in great detail on AI.

SNAIL and ATAI model fit

The univariate skewness and kurtosis values for the SNAIL were −1.06 to 1.50 and −1.08 to 1.73, which is in the acceptable range of -2.0 and +2.0 for skewness and −7.0 and +7.0 for kurtosis, respectively [24]. The univariate skewness and kurtosis for the ATAI scale was also acceptable, with skewness values between −0.45 and 0.56 and kurtosis values between −0.68 and 0.77. Mardia's test for multivariate skewness and kurtosis were both significant (p<.001), which is why multivariate non-normality had to be assumed. Due to the non-normality and the fact that the values were ordinal (because of the 7-point Likert scale), we used a polychoric correlation matrix instead of the usual Pearson correlation matrix [25]. The polychoric correlation matrix is robust against a violation of the normal distribution assumption. Since participants with a high number of missing answers were excluded before analyzing the data (see Sect. 3.1), the final data set only had an average of 1.1 missing values per variable (0.3%), which is why no data imputation was necessary.

A t-test was performed for the SNAIL overall score, the TU, CA, and PA subscores, as well as the ATAI subscores (fear and acceptance) to check whether the data sets of the two medical schools differed significantly from each other. As the group size was much larger than n=30, it could be assumed that the normal distribution assumption was not violated following the central limit theorem. A Levene test for variance homogeneity was performed for all SNAIL and ATAI scores. Since the Levene test was significant (p<.05) for the TU factor of the SNAIL instrument and the fear factor of the ATAI instrument, Welch's t-test was used. Welch's t-test showed that the overall SNAIL score, $t(277.15)=2.32$, p=.02, the TU subscore, $t(260.14)=2.60$, p<.01, and the fear subscore, $t(331.36) = -2.06$, p=.04, differed statistically significantly between the two medical schools (see Fig. 1). It was therefore decided that a separate CFA had to be carried out for the data sets of the two medical schools.

We found an equally acceptable to good model fit of the three factor model proposed by [20] for both medical schools. For MS1, the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were both 0.994, the Root Mean Square Error of Approximation (RMSEA) was 0.059 and the Standardized Root Mean Square Residual (SRMR) was 0.071. Accordingly, the three-factor solution fitted slightly better than a one-factor solution (i.e., a single latent factor "AI literacy"), as the latter had the following values: CFI=0.988, TLI=0.987, RMSEA=0.084, SRMR=0.083. The CFA of the MS2 data set led to comparable results. The 3-factor structure seemed to fit better with CFI=0.994, TLI=0.994, RMSEA=0.059, SRMR=0.071 than the 1-factor structure with CFI=0.959, TLI=0.956, RMSEA=0.130, SRMR=0.112. However, as expected, there is a high interfactor correlation of 0.81 between TU and CA, 0.90 between TU and PA and 0.93 between CA and PA.

Regarding ATAI, the two-factor solution proposed by Sindermann et al. [23] appears to have an excellent model fit. The following fit indices were found for MS1: CFI=1.000, TLI=1.012, RMSEA<0.001, SRMR=0.027. Excellent values were also found for MS2: CFI=1.000, TLI=1.016, RMSEA<0.001, SRMR=0.008. We found a negative interfactor correlation between "fear" and "acceptance" of −0.83.

The internal consistency of the SNAIL subscales, expressed by the reliability measure Cronbach's α, was

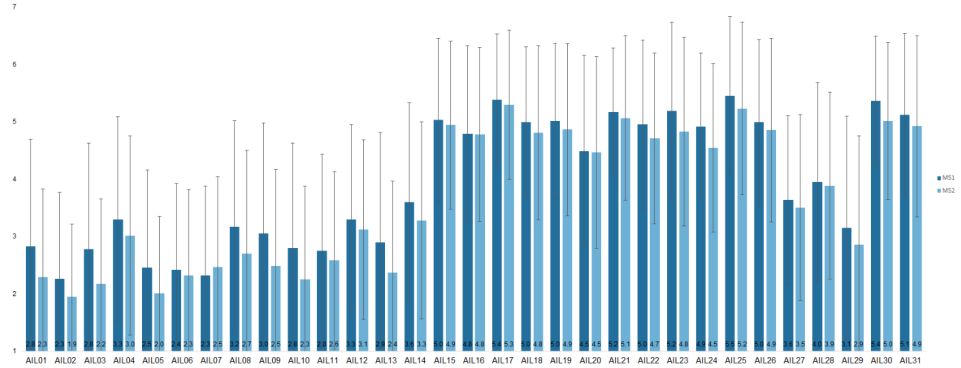


Fig. 1 Mean score for each SNAIL item for both medical schools. *Note* Number of participants in MS1 = 142, number of participants in MS2 = 235, total N = 377

Table 1 Mean, standard deviation, skew, and kurtosis for the TU, CA, PA, and overall SNAIL score for both medical schools

		TU score	CA score	PA score	SNAIL score (all items)
MS1	*M*	2.85	4.99	4.52	3.92
	SD	1.33	1.00	1.07	1.08
	Skew	0.59	-0.67	-0.23	0.14
	Kurtosis	-0.49	0.85	-0.34	-0.33
MS2	*M*	2.50	4.83	4.32	3.66
	SD	1.33	1.07	1.11	0.99
	Skew	1.00	-0.55	-0.09	0.32
	Kurtosis	0.82	0.60	-0.18	0.27

Note Number of participants in MS1 = 142, number of participants in MS2 = 235, total *N* = 377. MS = medical school, TU = Technical Understanding factor, CA = Critical Appraisal factor, PA = Practical Application factor, SNAIL = Scale for the assessment of non-experts' AI literacy

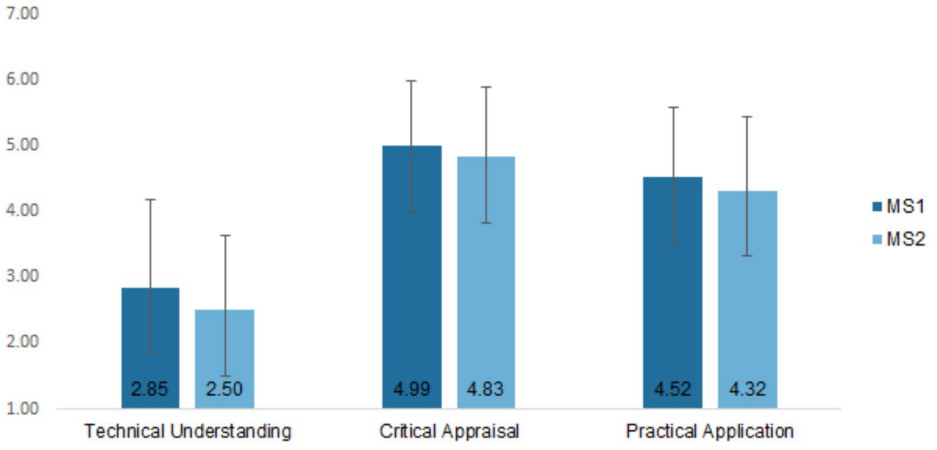


Fig. 2 Mean score for each SNAIL factor for both medical schools. *Note* Number of participants in MS1 = 142, number of participants in MS2 = 235, total *N* = 377. MS = medical school

good to excellent in both samples (MS1 and MS2). In the MS1 sample, the subscales had the following internal consistencies: TU, α = 0.94 [CI 0.93, 0.96]; CA, α = 0.89 [CI 0.86, 0.92], and PA, α = 0.83 [CI 0.78, 0.87]. In the MS2 sample, a Cronbach's α of α = 0.93 [CI 0.91, 0.94] was found for the TU subscale, α = 0.89 [CI 0.87, 0.91] for the CA subscale, and α = 0.81 [CI 0.77, 0.85] for the PA subscale. However, the internal consistency of the ATAI subscales was rather low, with α = 0.53 [CI 0.35, 0.67] for the "acceptance" subscale and α = 0.61 [CI 0.48, 0.71] for the "fear" subscale in the MS1 sample and α = 0.60 [CI 0.48, 0.69] for the "acceptance" subscale and α = 0.64 [CI 0.56, 0.72] for the "fear" subscale in the MS2 sample.

Medical students' AI literacy (RQ1)

To determine how medical students rated their overall AI literacy, the average score of each participant was calculated for each factor as well as for the overall SNAIL scale (see Table 1). The mean TU score was 2.26 points lower than the mean CA score, $t(734.68) = -27.26$, $p < .001$, and 1.77 points lower than the mean PA score, $t(744) = -20.86$, $p < .001$. The mean CA score was 0.49 points higher than the mean PA score, $t(750.08) = 6.28$, $p < .001$. Thus, the differences between the mean values of the subscales are all statistically significant. The results of the individual analyses of the two medical schools were very similar to the overall analysis (see Fig. 2), which is why they are not reported in more detail. In the further course of this paper, the results of the individual medical schools are only given if the values differ significantly between the schools.

Differences in medical students' AI literacy due to moderator variables (RQ2)

There was no statistically significant association between the age and the average SNAIL score of participants. This applies both to the overall sample, $r = .07$, $p = .16$, as well as to the MS1 and MS2 sample, $r = .05$, $p = .59$ and $r = .12$, $p = .07$, respectively. In the overall sample, women rated their AI literacy on average 0.413 points lower than men, $t(217.96) = -3.65$, $p < .001$. There were no differences within the separate samples of the two medical schools in this respect (i.e., in both medical schools, male participants rated themselves as more AI literate). The association between the general SNAIL score and medical students' current semester was statistically significant for the overall sample, $\tau_c = 0.08$, $p < .05$. However, there was a notable difference between the two medical schools: In

Table 2 Mean, standard deviation, skew, and kurtosis for the "acceptance" and "fear" score for both medical schools

		acceptance score	fear score
MS1	Mean	4.32	3.27
	Standard deviation	0.87	0.92
	Skew	-0.48	0.07
	Kurtosis	0.09	0.05
MS2	Mean	4.19	3.49
	Standard deviation	0.96	1.07
	Skew	-0.16	0.15
	Kurtosis	0.28	-0.01

Note Number of participants in MS1 = 142, number of participants in MS2 = 235, total $N = 377$. MS = medical school

MS1, the association between SNAIL score and semester was not statistically significant, $\tau_c = 0.04$, $p = .52$, while it was significant in MS2, $\tau_c = 0.13$, $p < .01$.

Medical students' attitudes towards artificial intelligence (RQ3)

The participants rated their "acceptance" of AI 0.83 points higher than their "fear" of AI, $t(745.42) = 11.72, < .001$. The calculations for the MS1 and MS2 subsets led to very similar results (see Table 2).

Relationship between medical students' AI literacy and attitudes towards AI (RQ4)

The SNAIL total score and the TU, CA and PA factor scores were all significantly correlated (all correlations $r = .64$ to $r = .92$, $p < .001$; see Table 3). This result indicated that the 31 items of the SNAIL questionnaire measure a common main construct, namely AI literacy.

In addition, the "acceptance" subscale of the ATAI questionnaire was also significantly positively correlated with the subscales of the SNAIL questionnaire and with the total SNAIL score. The correlation between the ATAI subscale "fear" and the SNAIL scales, on the other hand, was lower and negative. "fear" correlated strongly negatively with the TU score and weakly (but still significantly) negatively with the SNAIL total score and the PA score. However, the correlation between "fear" and the CA score was not significant. Lastly, the "fear" factor of the ATAI scale correlated strongly negatively with the "acceptance" factor.

Effect of AI education and interest on medical students' AI literacy (RQ5)

Medical students who had attended at least one shorter AI course of up to 30 h rated their AI literacy on average 1.47 points higher than medical students' who stated that they had never attended an AI course, $t(42.492) = 9.90$, $p < .001$. The association between the two variables "Time spent attending AI courses" (ordinally scaled) and the SNAIL total score was significant, $\tau_c = 0.31$, $p < .001$. In addition, students who at least irregularly used other ways to educate themselves about AI rated their AI literacy on average 0.92 points higher than students who never did so, $t(373) = 9.70$, $p < .001$. As expected, the association between the two variables "Regularity with which students train themselves on AI" (ordinally scaled) and the SNAIL total score was significant, $\tau_c = 0.43$, $p < .001$. Finally, medical students' interest in AI also appeared to be a good predictor of their AI literacy (although the causal direction of this association is not clear). Students who rated their interest in AI as rather high (5 to 7 on a 7-point Likert scale) rated their AI literacy on average 0.94 points higher than students who were less interested in AI (1 to 3 on a 7-point Likert scale), $t(373) = 8.68$, $p < .001$. The association between "Interest in AI" and the SNAIL total score was significant, $\tau_c = 0.37$, $p < .001$ (see Fig. 3).

Discussion

In this study, we assessed AI literacy and attitudes towards AI among medical students at two German medical schools using validated assessment instruments. Remarkably, medical students rated their ability to critically appraise AI and to use AI in practice as relatively high, while they rated their technical understanding of AI as rather low. In addition, although both positive and negative attitudes towards AI were evident, positive attitudes (acceptance of AI) seemed to outweigh negative attitudes (fear of AI). While the correlation between medical students' AI literacy and acceptance of AI was clearly positive, the link between AI literacy and negative attitudes appears to be more complex.

Table 3 Correlation matrix for correlations between SNAIL and ATAI scores according to Kendall's τ coefficients

	M	SD	1	2	3	4	5	6
1. SNAIL score	3.76	1.03						
2. TU score	2.63	1.22	0.92***					
3. CA score	4.89	1.05	0.90***	0.64***				
4. PA score	4.40	1.10	0.87***	0.73***	0.83***			
5. acceptance score	4.24	0.93	0.29***	0.29***	0.20***	0.28***		
6. fear score	3.41	1.02	−0.12*	−0.15**	−0.03	−0.11*	−0.45***	

*$p < .05$ **$p < .01$ ***$p < .001$

Note All correlations shown in the table are based on the total sample ($N = 377$)

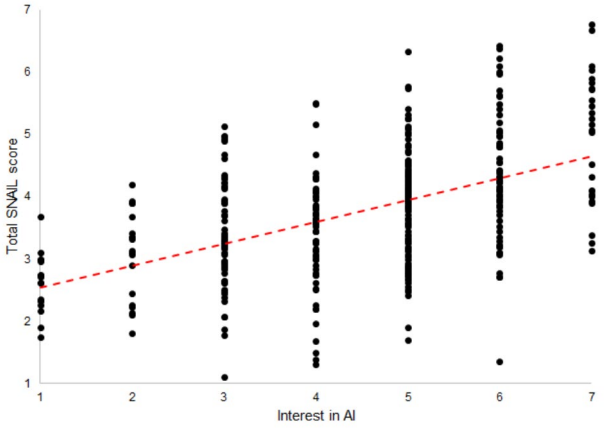


Fig. 3 Scatterplot of Kendall's rank correlation between the total SNAIL score and medical students' interest in AI. *Note* The associations shown in the figure are based on the total sample ($N = 377$)

Interpretation and implications of the results

By using the CFA, we were able to show that the SNAIL questionnaire instrument was suitable for assessing the three latent AI literacy factors TU, CA and PA. This is evident from the good model fit of the three-factor model, but also by the excellent Cronbach's α values for the three subscales. While the model fit was even better for the ATAI measuring instrument, Cronbach's α of that scale was rather low, although this does not necessarily question the usefulness of the ATAI scale [26]. The low alpha values of the ATAI scale are somewhat unsurprising, considering that scales with a very small number of items also tend to have low internal consistency [27]. While the small number of items ensured good questionnaire efficiency, we could not conclusively clarify whether the five ATAI items were able to reliably assess medical students' ATAI in our sample. Finally, we wonder whether the model fit of the ATAI model is not artificially increased, as the two subscales "acceptance" and "fear" measure practically opposite constructs. In future studies, it might therefore be advisable to recode one of the two subscales and conduct a CFA again to determine whether the two-factor structure still results in a good model fit.

RQ1 addressed the level of AI literacy and the AI literacy subconstructs TU, CA and PA of medical students. While the values of all three subscales differ statistically significantly from each other, the difference between TU and the other two factors is particularly interesting. Considering that the midpoint of a 7-point Likert scale is 4, it is surprising that the participants rated their CA and PA skills higher but their TU skills lower than the midpoint. This difference is particularly interesting because it could be assumed that a certain level of technical understanding is crucial for the practical use of AI applications. One possible explanation for the lower self-assessment score of the TU scale could be that aspects such as AI ethics, data security in connection with AI, or the recent AI hype are discussed in popular media, while technical aspects of AI, such as the function of machine learning or the difference between strong and weak AI are rather neglected.

While the age of the medical students did not appear to have any effect on their AI literacy, gender in particular had an important influence on the self-assessment of AI literacy. This is in line with a wealth of evidence suggesting that women rate themselves more negatively than men in self-assessments [28]. This effect appears to be even more pronounced for technical or scientific subjects, and negative self-assessment may even be associated with objectively lower performance [29]. Nevertheless, it is advisable to use objective AI literacy tests in addition to pure self-assessment scales in order to avoid response biases as far as possible. Furthermore, the semester also seemed to have had an influence on the self-assessment of participants' AI literacy. The correlative relationship between the SNAIL overall score and the participants' semester was particularly pronounced in MS2. However, a closer look reveals that in the MS2 sample, 120 participants (51% of the MS2 sample) were in semester 3 and 67 participants (29% of the MS2 sample) were in semester 7. Since 80% of the MS2 sample therefore stems from one of these two semesters, the association between semester and SNAIL score could be attributed to a sample effect.

The analyses conducted regarding RQ3 showed that medical students' AI literacy is significantly positively correlated with their acceptance of AI, and significantly negatively correlated with their fear of AI. Thus, either AI literate medical students are more likely to accept (and less likely to fear) AI applications than AI illiterate students, or medical students who accept AI are more likely to be AI literate than students who do not accept AI. This finding complements the literature review published by Mousavi Baigi et al. [17], which found that 76% of studies reported positive attitudes towards AI among healthcare students. However, the scale midpoint of 4 should be emphasized again at this point. The medical students only "accept" AI with an average of 4.32 (MS1) and 4.12 (MS2) points and "fear" AI with 3.27 (MS1) and 3.49 (MS2) points. Although we found a statistically significant difference, it is obvious that both the negative and positive attitudes towards AI are relatively close to the midpoint. This may indicate that medical students have nuanced attitudes towards AI.

The investigation of the correlation between AI literacy and ATAI (RQ4) yielded interesting results. In the past, it has been shown for various constructs such as financial literacy [30] or scientific literacy [31] that there is a positive correlation between knowledge about a topic

and positive attitudes towards it. A comparable effect was found in our study for the relationship between AI literacy and ATAI. Medical students who had a higher AI literacy were more likely to have a positive attitude towards AI (and vice versa). However, it should be mentioned again that the causality cannot be evaluated in this cross-sectional study. It is possible that medical students with a positive attitude are more willing to inform themselves about AI, resulting in a higher AI literacy. Nevertheless, it is also possible that students who are well versed in AI are better able to assess the real benefits and risks of AI, which leads to a more critical perception of exaggeratedly negative portrayals of AI.

The results regarding RQ5 indicate that courses and programs to increase AI literacy do indeed appear to have a positive effect on the AI literacy of medical students. This is an important finding as it illustrates that even relatively short AI courses (up to 30 h) are associated with higher AI literacy scores. This is particularly important in the very tightly scheduled medical curriculum, as medical AI education might be perceived as an additional burden by medical students and medical educators alike. Finally, our results indicate that the further development of curricula should arouse medical students' interest in AI. As depicted in Fig. 3, interest in AI seems to have a strong influence on the AI literacy of medical students.

Limitations

We have identified three main limitations: Firstly, this study was designed as a cross-sectional study which serves well to provide an initial picture of the AI literacy and ATAI of medical students. However, the correlative relationships presented here cannot provide any information about the causality of the effects. Secondly, the data was collected from two different medical schools in order to prevent sampling effects from influencing the validity of the results. Nevertheless, it is not possible to draw conclusions from the results of the two medical schools to all medical schools in Germany or even internationally, as various location factors can have an influence on AI literacy and ATAI, e.g. the current status of AI education in the medical curricula. Thirdly, all the instruments used were self-assessment questionnaires. It is conceivable that medical students' self-assessment was subject to response biases that shifted the response behavior in one direction or the other. A bias that is particularly significant in this context is social desirability, which "refers to the tendency of research subjects to choose responses they believe are more socially desirable or acceptable rather than choosing responses that are reflective of their true thoughts or feelings" [32] (Grimm, 2010, p.1). Given that AI is a hyped topic due to recent developments such as the release of OpenAI's ChatGPT, medical students may feel that they have at least somewhat engaged with the topic, which could potentially positively bias their response tendency. Another potential bias is the so-called acquiescence bias, which "describes the general tendency of a person to provide affirmative answers" [33]. This bias might be particularly problematic in the case of the SNAIL, as this scale has only "positive" items (i.e., higher self-assessment ratings equal higher AI literacy). However, at least the latter bias is mitigated by the fact that the SNAIL items are worded neutrally (i.e., not suggestively), which should mitigate the acquiescence tendency to some extent.

We also presented the SNAIL and ATAI items in random order and used a 7-point Likert scale for all items, as opposed to the 11-point Likert scale used by Sindermann et al. [23]. However, we believe that these adjustments to the original scales do not limit the ability of the scales to capture AI literacy and ATAI.

Future research directions

Future studies should firstly attempt to overcome the limitations of this study and secondly continue research on AI literacy and ATAI of medical students to contribute to their better acquisition of such crucial skills.

In order to determine the causal relationships between AI literacy and ATAI or other variables (such as interest in AI), experiments should be conducted that manipulate the ATAI of medical students while establishing a control group. Longitudinal studies or randomized controlled trials would also be suitable for investigating the direction of these effects. In addition, the study should be conducted at other locations and in other countries in order to verify the generalizability of the results considering different medical curricula. Objective testing of medical students' AI literacy [34] would also be desirable for future research projects, as objective performance measurements using knowledge or skill tests are subject to significantly less response bias. Last but not least, the development of AI education programs for medical students should be further supported and their effectiveness measured using validated scales. In this way, courses could be continuously improved to ensure that all medical students have a chance to reach a certain level of AI literacy which is required given the technological advancements. The difference between voluntary elective courses on AI and AI education as part of medical schools' compulsory curricula would also be an important research endeavor. We call for the implementation of AI education for *all* medical students and believe that in the future all medical students should have a certain level of AI literacy in order to continue to fulfill their various professional roles in an effective and safe manner. However, this theory should be empirically tested.

Conclusion

To our knowledge, we were the first to use validated questionnaire instruments to assess the AI literacy and ATAI of medical students. We found that medical students' technical understanding of AI in particular was still relatively low compared to their confidence in critically evaluating and practically using AI applications. This study sheds crucial light on the AI literacy landscape among medical students, emphasizing the necessity for tailored programs. These initiatives should accentuate the technical facets of AI while accommodating students' attitudes towards AI.

Abbreviations

AI	Artificial Intelligence
ATAI	Attitudes towards AI
CA	Critical Appraisal
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CI	Confidence Interval
MAIRS-MS	Medical Artificial Intelligence Readiness Scale for Medical Students
MS	Medical School
PA	Practical Application
QR	Quick-response
RMSEA	Root Mean Square Error of Approximation
RQ	Research Question
SNAIL	Scale for the assessment of non-experts' AI literacy
SRMR	Standardized Root Mean Square Residual
TLI	Tucker-Lewis Index
TU	Technical Understanding

Acknowledgements

The authors express their gratitude to everyone who contributed to the execution of the research project, with special appreciation for the medical educators who encouraged participation in our study.

Author contributions

M.C.L. analyzed the data and wrote the first draft of the manuscript. A.A., M.Mey. and M.Mer. co-wrote the manuscript. M.Mer. was significantly involved in the planning, organization and execution of the study. A.A. and M.Mey. assisted with the data analysis. T.R. provided feedback on the manuscript's content and assisted with the linguistic refinement of the manuscript. All authors read and approved the final manuscript.

Funding

M.C.L., A.A. and T.R. received no financial funding to conduct this study. M.Mey. and M.Mer. were funded by the German Federal Ministry of Education and Research (research grant: 16DHBKI080).
Open Access funding enabled and organized by Projekt DEAL.

Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

The study was approved by the Research Ethics Committee of the University of Bonn (Reference 194/22) and of Saarland University (Reference 244/21). Medical students who were at least 18 years old were eligible for the study and could access the online questionnaire after giving their informed consent to participate.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 15 January 2024 / Accepted: 8 April 2024
Published online: 10 April 2024

References

1. Schwartz WB, Patil RS, Szolovits P. Artificial Intelligence in Medicine. N Engl J Med. 1987;316(11):685–8. https://doi.org/10.1056/NEJM198703123161109.
2. Ramesh AN, Kambhampati C, Monson JRT, Drew PJ. Artificial intelligence in medicine. Ann R Coll Surg Engl. 2004;86(5):334–8. https://doi.org/10.1308/147870804290.
3. Hamet P, Tremblay J. Artificial intelligence in medicine. Metab Clin Exp. 2017;69:36–40. https://doi.org/10.1016/j.metabol.2017.01.011.
4. Haug CJ, Drazen JM. (2023). Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. *New England Journal of Medicine*, *388*(13), 1201–1208. https://doi.org/10.1056/nejmra2302038.
5. Chan KS, Zary N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: integrative review. JMIR Med Educ. 2019;5(1):e13930. https://doi.org/10.2196/13930.
6. Mergen M, Junga A, Risse B, Valkov D, Graf N, Marschall B, medical.training. consortium. Immersive training of clinical decision making with AI driven virtual patients - a new VR platform called medical. GMS J Med Educ. 2023;40(2). https://doi.org/10.3205/zma001600.
7. Lee J, Wu AS, Li D, Kulasegaram K, Mahan. Artificial Intelligence in Undergraduate Medical Education: a scoping review. Acad Med. 2021;96(11):62–70. https://doi.org/10.1097/ACM.0000000000004291.
8. Laupichler MC, Hadizadeh DR, Wintergerst MWM, von der Emde L, Paech D, Dick EA, Raupach T. Effect of a flipped classroom course to foster medical students' AI literacy with a focus on medical imaging: a single group pre- and post-test study. BMC Med Educ. 2022;22(1). https://doi.org/10.1186/s12909-022-03866-x.
9. Hu R, Fan KY, Pandey P, Hu Z, Yau O, Teng M, Wang P, Li T, Ashraf M, Singla R. Insights from teaching artificial intelligence to medical students in Canada. Commun Med. 2022;2(1). https://doi.org/10.1038/s43856-022-00125-4.
10. Frommeyer TC, Fursmidt RM, Gilbert MM, Bett ES. (2022). The Desire of Medical Students to Integrate Artificial Intelligence Into Medical Education: An Opinion Article. *Frontiers in Digital Health*, *4*. https://doi.org/10.3389/fdgth.2022.831123.
11. Sit C, Srinivasan R, Amlani A, Muthuswamy K, Azam A, Monzon L, Poon DS. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. Insights into Imaging. 2020;11(1). https://doi.org/10.1186/s13244-019-0830-7.
12. Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. Lancet Digit Health. 2020;2(3):111–2. https://doi.org/10.1016/S2589-7500(20)30023-6.
13. Long D, Magerko B. (2020). What is AI Literacy? Competencies and Design Considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727.
14. dos Pinto D, Giese D, Brodehl S, Chon SH, Staab W, Kleinert R, Maintz D, Baeßler B. Medical students' attitude towards artificial intelligence: a multicentre survey. Eur Radiol. 2019;29(4):1640–6. https://doi.org/10.1007/s00330-018-5601-1.
15. Stewart J, Lu J, Gahungu N, Goudie A, Fegan PG, Bennamoun M, Sprivulis P, Dwivedi G. Western Australian medical students' attitudes towards artificial intelligence in healthcare. PLoS ONE. 2023;18(8):e0290642. https://doi.org/10.1371/journal.pone.0290642.
16. Kimmerle J, Timm J, Festl-Wietek T, Cress U, Herrmann-Werner A. Medical students' attitudes toward AI in Medicine and their expectations for Medical Education. J Med Educ Curric Dev. 2023;10. https://doi.org/10.1177/23821205231219346.
17. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: a systematic review. Health Sci Rep. 2023;6(3). https://doi.org/10.1002/hsr2.1138.
18. Karaca O, Çalışkan SA, Demir K. Medical artificial intelligence readiness scale for medical students (MAIRS-MS)– development, validity and reliability study. BMC Med Educ. 2021;21(1). https://doi.org/10.1186/s12909-021-02546-6.
19. Aboalshamat K, Alhuzali R, Alalyani A, Alsharif S, Qadhi H, Almatrafi R, Ammash D, Alotaibi S. Medical and Dental professionals readiness for

Artificial Intelligence for Saudi Arabia Vision 2030. Int J Pharm Res Allied Sci. 2022;11(4):52–9. https://doi.org/10.51847/nu8y6y6q1m.
20. Laupichler MC, Aster A, Raupach T. (2023). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence*, *4*. https://doi.org/10.1016/j.caeai.2023.100126.
21. Laupichler MC, 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*. https://doi.org/10.1016/j.chbr.2023.100338.
22. Laupichler MC, Aster A, Perschewski JO, Schleiss J. Evaluating AI courses: a Valid and Reliable Instrument for assessing Artificial-Intelligence Learning through Comparative Self-Assessment. Educ Sci. 2023;13(10). https://doi.org/10.3390/educsci13100978.
23. Sindermann C, Sha P, Zhou M, Wernicke J, Schmitt HS, Li M, Sariyska R, Stavrou M, Becker B, Montag C. Assessing the attitude towards Artificial Intelligence: introduction of a short measure in German, Chinese, and English Language. KI - Kuenstliche Intelligenz. 2021;35(1):109–18. https://doi.org/10.1007/s13218-020-00689-0.
24. Curran PJ, West SG, Finch JF. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychol Methods. 1996;1(1):16–29. https://doi.org/10.1037/1082-989X.1.1.16.
25. Wang WC, Cunningham EG. Comparison of alternative estimation methods in confirmatory factor analyses of the General Health Questionnaire. Psychol Rep. 2005;97(1):3–10.
26. Taber KS. The Use of Cronbach's alpha when developing and Reporting Research Instruments in Science Education. Res Sci Educ. 2018;48(6):1273–96. https://doi.org/10.1007/s11165-016-9602-2.
27. Kopalle PK, Lehmann DR. Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. Organ Behav Hum Decis Process. 1997;70(3):189–97. https://doi.org/10.1006/obhd.1997.2702.
28. Torres-Guijarro S, Bengoechea M. Gender differential in self-assessment: a fact neglected in higher education peer and self-assessment techniques. High Educ Res Dev. 2017;36(5):1072–84. https://doi.org/10.1080/07294360.2016.1264372.
29. Igbo JN, Onu VC, Obiyo NO. Impact of gender stereotype on secondary school students' self-concept and academic achievement. SAGE Open. 2015;5(1). https://doi.org/10.1177/2158244015573934.
30. Dewi V, Febrian E, Effendi N, Anwar M. Financial literacy among the millennial generation: relationships between knowledge, skills, attitude, and behavior. Australasian Acc Bus Finance J. 2020;14(4):24–37. https://doi.org/10.14453/aabfj.v14i4.3.
31. Evans G, Durant J. The relationship between knowledge and attitudes in the public understanding of science in Britain. Public Underst Sci. 1995;4(1):57–74. https://doi.org/10.1088/0963-6625/4/1/004.
32. Grimm P. Social desirability bias. Wiley international encyclopedia of marketing; 2010.
33. Hinz A, Michalski D, Schwarz R, Herzberg PY. (2007). The acquiescence effect in responding to a questionnaire. *Psychosocial Medicine, 4*. PMID: 19742288.
34. Hornberger M, Bewersdorff A, Nerdel C. What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. Computers Education: Artif Intell. 2023. https://doi.org/10.1016/j.caeai.2023.100165. *5*.

Publisher's Note