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.

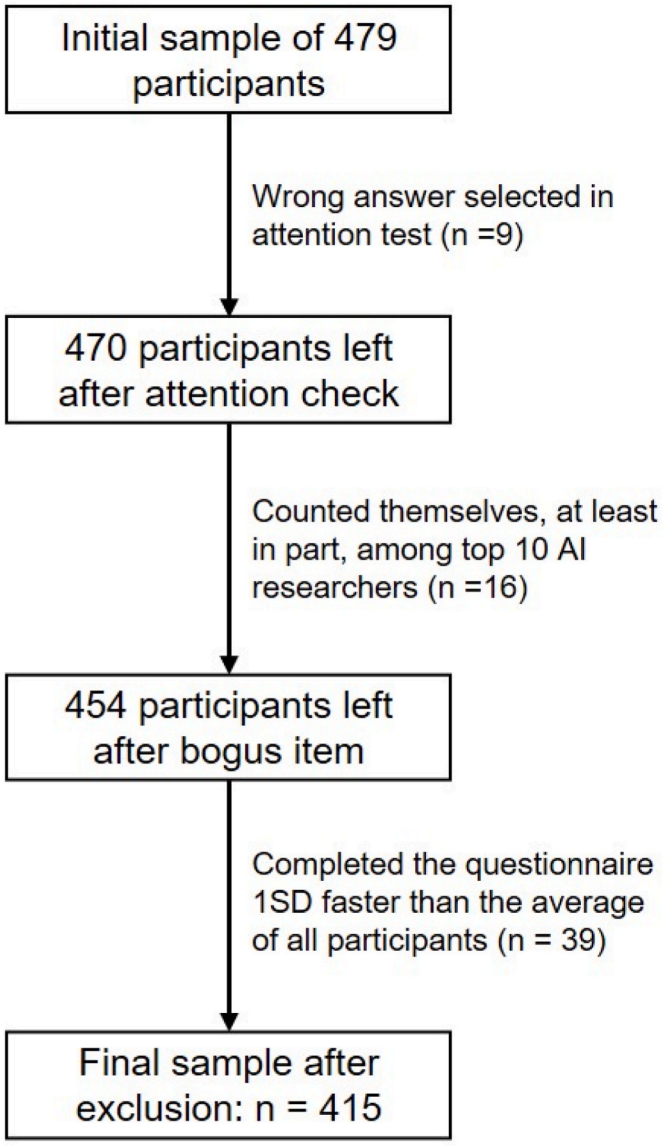

Fig. 1. Number of participants excluded from data analysis based on three exclusion criteria.

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.

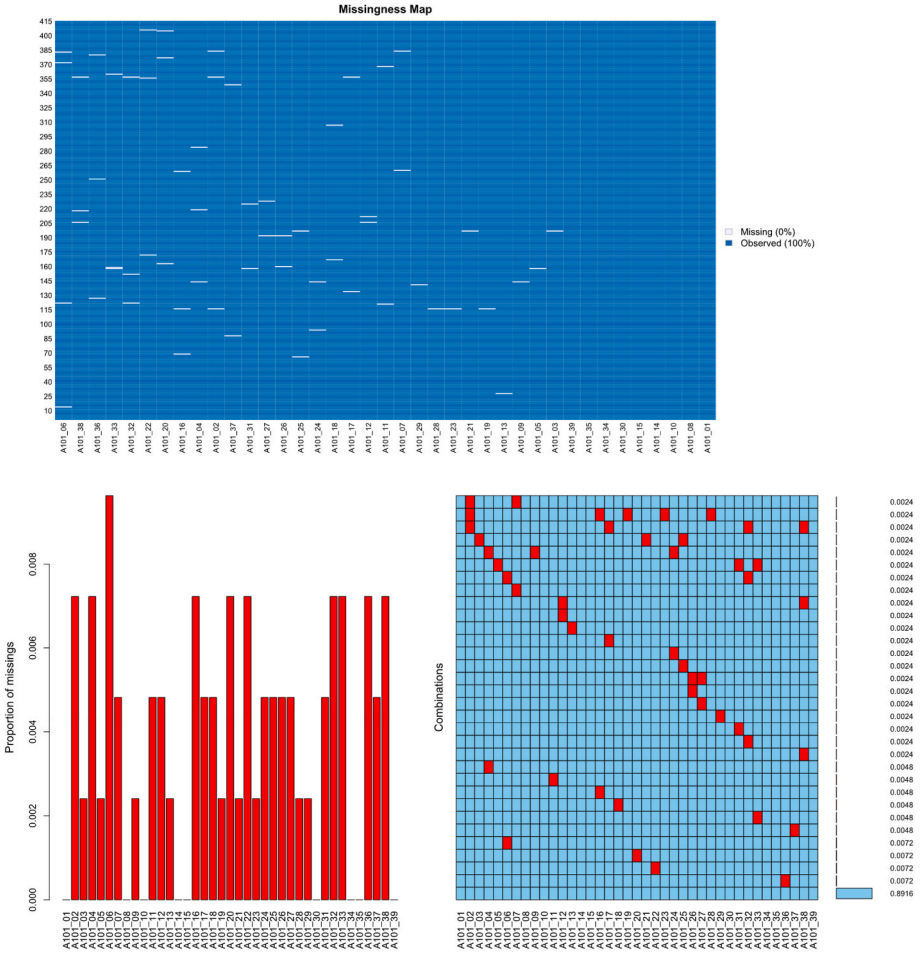

Fig. 2. Distribution and number of missing values in absolute and relative terms across all subjects and variables.

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.

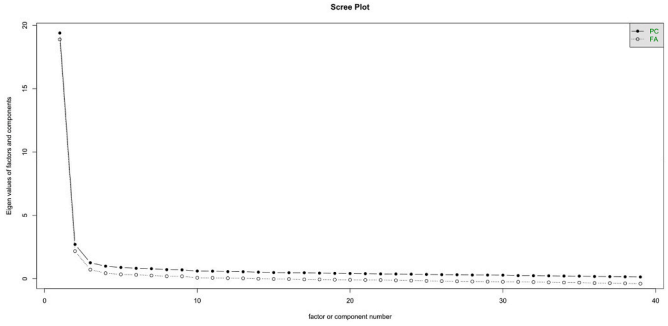

Fig. 3. Screeplot.

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

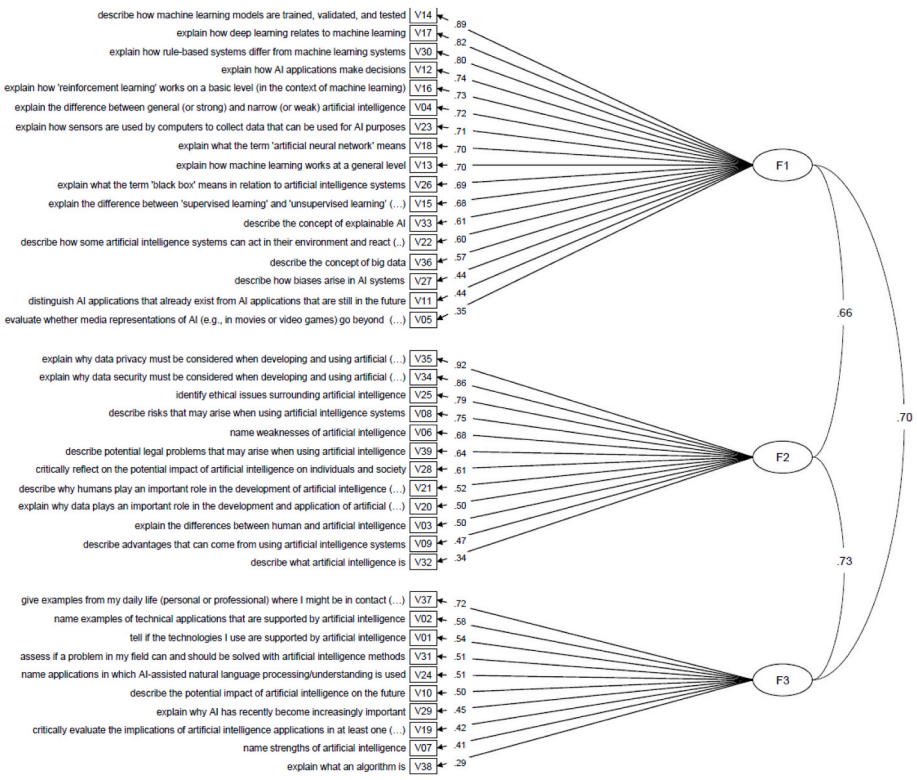

Fig. 4. Path diagram for the 3-factor promax model.

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