Development of a German Instrument for Self-Rated Data Literacy

An Algorithm-based Approach to Scale Development

Leonie Hagitte

2025-01-25

Contents

Abstract

This study aimed to develop a german, self-rated data literacy questionnaire tailored for non-specialized populations. Drawing on theoretical frameworks and previous operationalizations, the construct of data literacy was defined to include five facets: Comprehension, Evaluation, Integration, Communication, and Statistics. An initial pool of 71 items was refined through cognitive interviews and further narrowed to a final set of 20 items using an algorithm-based selection approach. Data were collected via a cross-sectional online survey (N = 616), and the measurement model was assessed using confirmatory factor analysis (CFA). Model fit indices indicated acceptable fit (CFI = .96, SRMR = .08, RMSEA = .05) of the initial model. A multi-group CFA indicated deterioration of fit (CFI = .92, SRMR = .087, RMSEA = .038), but confirmed scalar invariance across subsamples (random sample split), indicating some stability of the measurement model. The scale exhibited acceptable internal consistency (McDonald's ω = .92) and suggested criterion validity in most aspects, with moderate to strong correlations observed with constructs such as information literacy and need for cognition. In addition to the measurement properties, limitations were noted in factor independence due to multicollinearity among some facets, as well as due to correlated residuals of the scales items. The questionnaire provides the first step towards a novel tool for assessing data literacy in laypeople and warrants further research into the construct's structure, applications and relevance.

Keywords: Data Literacy, Questionnaire Development, Algorithm-Based Item Selection, Genetic Algorithm

Acknowledgements

I want to thank my advisers, Prof. Martin Schultze, Prof. Timo Lorenz, and Prof. Manuel Völkle for their time and patience, and my friends for their resourceful advice.

1 Background

In a world characterized by information overload and rapid technological advancements (Koltay, 2017; Leighton et al., 2021; Roetzel, 2019), the relevance of data literacy for today's society becomes evident. Data literacy serves as a potent tool in navigating the complex data-driven environment (Carmi et al., 2020; Cui et al., 2023; Leighton et al., 2021; Ridsdale et al., 2015) and individuals equipped with strong data literacy skills can discern patterns, critically evaluate information, and make informed decisions (F. Chen et al., 2024; Cui et al., 2023). The exploration of citizens' interaction with media and the cultivation of their agency has traditionally started around concepts such as written literacy and information literacy (Association of College & Research Libraries, 2000; C. Brown & Krumholz, 2002). Information literacy, as defined by the American Library Association in 1989, encompasses the skills required to locate, evaluate, and effectively use information (C. Brown & Krumholz, 2002). This foundational understanding of information literacy has evolved through various frameworks and models, emphasizing its importance in educational settings and lifelong learning (Tomar, 2023). In more recent discussions, data literacy has been approaching relevance among discussed competencies regarding agency in the current society (Carmi et al., 2020; Leighton et al., 2021). Deficiency in data literacy not only exposes individuals to various risks and harms on personal, social, physical, and financial levels but also constrains their capacity to actively engage as informed citizens within an evolving, data-driven society (Carmi et al., 2020; Leighton et al., 2021). Thus, data literacy is a competency that is becoming increasingly important to everyone. Research has acknowledged this in recent years, as more and more has been undertaken in that direction (F. Chen et al., 2024; Cui et al., 2023). This study aims to complement the current research, with a self rating questionnaire for assessing data literacy among citizens.

Data literacy involves the ability to effectively collect, manage, evaluate, and apply data in a critical manner (Ridsdale et al., 2015). According to Wolff et al. (2016), it means being able to ask and answer everyday questions using both small and large datasets while considering ethical aspects. This includes skills such as selecting, cleaning, analyzing, visualizing, criticizing, and interpreting data, as well as communicating insights from data, and using data for various purposes. Frank et al. (2016) distinguishes between cognitive skills, like data collection and analysis, and social skills, which involve trusting data while maintaining skepticism. Calzada-Prado & Marzal (2013) outline five dimensions of data literacy: understanding data, acquiring data, interpreting and evaluating data, managing data, and using data. Understanding data includes knowing their types, roles, and significance, while acquiring data involves evaluating and select-

ing sources. Interpreting and evaluating data encompass understanding different presentation methods and data interpretation. Managing data includes storage, management, and re-use. Using data involves preparation, analysis, communication, and ethical considerations (Calzada-Prado & Marzal, 2013). This small comparison already highlights one prominent attribute of data literacy: it is a heterogeneous concept (F. Chen et al., 2024). Every subject or profession seems to hold their own definition or framework of data literacy (F. Chen et al., 2024; Cui et al., 2023). While heterogeneity is beneficial for assessing specific skills (e.g. in a recruitment test), it limits the generalisability and comparability of data literacy across individuals with different background. It furthermore limits the accuracy of communication about the topic as two people with different background might hold different definitions on data literacy. In the study of Cui et al. (2023) it also becomes apparent, that one group seems to be less focused on in the research on data literacy: citizens or the broader public. Citizens in this case mean people of the general public, that hold no special role or profession, related to data handling or aspects related to data literacy (Wolff et al., 2016). Despite being the largest demographic group, citizens are often overlooked in favor of specific professions such as researchers, librarians, students, or educators (Cui et al., 2023). This trend raises questions about the emphasis on certain aspects of data literacy, many of which tend to align more closely with professional roles than with the needs of laypeople (Schüller, 2020). In their framework Schüller (2020) highlight the different roles, people can hold: Some of the factors or skills regard *data-consumption*, whereas the most are skills *data-producers* would have. This is also reflected when taking a look into related concepts. The definition proposed by Wolff et al. (2016) suggests that data literacy shares some common competencies with statistical and information literacies.

Information literacy, often studied in library sciences, overlaps with data literacy in terms of accessing, critically evaluating, and using data sources (Calzada-Prado & Marzal, 2013; Shields, 2005). Wolff et al. (2016) also emphasize the importance of the data inquiry process, starting from identifying problems, designing studies, acquiring data, conducting analysis, to drawing data-based conclusions. In comparison, Gould (2017) argued that data literacy is essentially the same as statistical literacy but with additional competencies needed due to the increasing importance of data. These added competencies include understanding who collects the data, how and why data is collected, and understanding data privacy and ownership (Gould, 2017).

Additionally, when discussing data literacy, it is essential to understand the distinctions between the terms data and information, and their relationship to one another. The concepts of data and information are foundational in various

fields, yet their precise definitions and relationships are often subject to interpretation (Koltay, 2017; R. Schneider, 2013). While data can be viewed as the raw material (Shannon, 1948) from which information is derived (Bates, 2009), it is the reduction of entropy through organization and interpretation that gives rise to meaningful information (Brillouin, 1953; Jaynes, 1957). Thus, rather than viewing data as synonymous with entropy or information, it is more accurate to consider information as emerging from the structured representation of data.

It is structured representation of data, that also according to the framework of (Schüller, 2020), citizens are primarily concerned with: citizens are holding roles where they mainly consume data or information. Consequently, they may encounter difficulties with tasks or items related to producing facets such as providing or exploiting data. This imbalance could potentially undermine the fairness of tests and questionnaires designed to assess data literacy, particularly, if these assessments prioritize competencies closer to data and statistical literacy over those closer to information literacy.

1.1 Conceptual Integration and Delineation From Other Concepts

Because data literacy is such a heterogeneous construct, it is very prone to construct proliferation. This, for example, becomes evident, when looking into the literature review from Cui et al. (2023). In their study they found over fourteen different definitions of data literacy, incorporating different sets of 75 competencies. Furthermore, very few of those definitions are targeted at citizens, and none of the questionnaires they report target citizens. Thus, the objective of the present study was to formulate a definition of data literacy, based on the current literature, that works for citizens. For the conceptual integration, I used the review of Cui et al. (2023) and assessed what competencies were incorporated in most of the definitions, and would therefore be considered central to the construct. This included counting the same competencies as well as checking what competencies might be named differently, but are considered the same substantively. Furthermore, data literacy encompasses certain characteristics and behaviors from similar constructs. Examples for those convergent constructs are *critical thinking*, *media competency*, *technology competency*, *statistical literacy* as well as *information literacy* (F. Chen et al., 2024; Cui et al., 2023; Leighton et al., 2021). As those constructs share substantive parts, differing in size regarding the respective definition of the constructs, it is to be expected that all of them show moderate to strong positive correlations. Concluding, data literacy can be defined as "the ability to collect, manage, evaluate, and apply data effectively. It involves asking and answering real-world questions from datasets while considering ethi-

cal use. Core skills include selecting, cleaning, analyzing, visualizing, presenting, critiquing, and interpreting data, information and their sources" (Cui et al., 2023; Ridsdale et al., 2015; Wolff et al., 2016).

Most definitions in the literature discern about five factors of general data literacy (Cui et al., 2023), which can be further categorized into three *consumer* and two *producer* facets (Schüller, 2020). Trying to describe the associated behaviors, I labeled these five factors as *Comprehension*, *Evaluation*, *Integration*, *Communication*, and *Statistics*. The *consumer* facets: *Comprehension*, *Evaluation*, and *Integration*, are broadly relevant to nearly everyone in society, including the general public. In contrast, the *producer* facets: *Communication* and *Statistics*, are primarily important for individuals actively working with data. The *consumer* facets are closely aligned with media literacy, emphasizing critical analysis and interpretation of various media formats to support understanding and engagement with media content (cf. Table 1).

1.1.1 Consumer Facets

The factor *Comprehension* encompasses behaviors and skills associated with critical thinking (Payan Carreira et al., 2022; Rear, 2019), such as identifying weaknesses in ones reasoning or actively shaping discourse and public dissemination of information. It involves the ability to comprehend various forms of data presentation, detect inconsistencies, interpret data comprehensively, and identify logical fallacies. Individuals with high scores on this factor demonstrate a strong aptitude for processing and making sense of information across different formats, enabling them to draw accurate conclusions and insights (Carlson et al., 2014; Vahey et al., 2006; Wolff et al., 2016). Both statistical and information literacy involve using data, and information to make informed decisions (Callingham, 2006; Gal, 2002; Webber & Johnston, 2017). Therefore, the *Evaluation* factor involves skills related to critically evaluating information sources and discerning between facts and opinions (Callingham, 2006; Frank et al., 2016; Gould, 2017; Kõuts-Klemm, 2019). It includes elements of information literacy (Association of College & Research Libraries, 2000) by focusing on evaluating the credibility of data sources, considering factors like reputation and biases, similar to assessing the quality of information sources (Evaluation & Integration) (Webber & Johnston, 2017). Individuals scoring high on this factor demonstrate awareness of potential biases or vested interests in information sources, as well as the context in which the information was presented (Callingham, 2006; Kõuts-Klemm, 2019; Lusiyana et al., 2020).

Data literacy often emphasizes integrating data-driven insights into one's opinions and values, which influence decision-making processes, without focus-

Table 1: The Nomological Network of the Construct and its Factors

Factor	Description	Convergent Constructs	Adjacent Constructs
Data Literacy	The ability to collect, manage, evaluate, and apply data effectively.	Statistical Literacy +++ Information Literacy +++ Media Competency ++ Critical Thinking ++	Need for Cognition + Openness + Conscientiousness +
Comprehension	Individuals with high scores on this factor demonstrate a strong aptitude for processing and making sense of information across different formats, enabling them to draw accurate conclusions and insights.	Statistical Literacy ++ Information Literacy +++ Media Competency ++ Critical Thinking +++	Need for Cognition: + Openness + Conscientiousness +
Evaluation	Represents the ability to evaluate the credibility and reliability of information by considering the reputation of the source, the context in which it is presented, and potential biases or vested interests. Individuals high in this factor demonstrate a heightened awareness of these elements.	Statistical Literacy ++ Information Literacy +++ Media Competency +++ Critical Thinking +	Need for Cognition: ++ Openness + Conscientiousness +
Integration	Reflects the ability to seek a comprehensive understanding of diverse topics by engaging with multiple perspectives and integrating data-driven insights. Individuals high in this factor are open to adapting their opinions based on new evidence, prioritize evidence-based information, and strive to align their values with reliable data.	Statistical Literacy ++ Information Literacy ++ Media Competency ++ Critical Thinking +++	Need for Cognition: ++ Openness ++ Conscientiousness +
Communication	Individuals scoring high in this factor hold the ability to translate data into simple visualizations, present findings confidently, and articulate complex information effectively in written and visual as well as verbal formats.	Statistical Literacy +++ Information Literacy ++ Media Competency +++ Critical Thinking ++	Need for Cognition: + Openness + Conscientiousness +
Statistics	It involves proficiency in organizing and analyzing data using software tools, conducting statistical analyses, and understanding research methodologies.	Statistical Literacy +++ Information Literacy ++ Media Competency ++ Critical Thinking +	Need for Cognition: ++ Openness + Conscientiousness ++

Note. The table illustrates the nomological network of Data Literacy and its factors. Each factor is described in terms of its primary function, along with convergent constructs and adjacent constructs. The constructs are displayed alongside their expected correlation with the data literacy factors (+++ = strong positive, ++ = moderate positive, + = weak positive).

ing on the decision making (Callingham, 2006; Wolff et al., 2016). Thereby it aligns with the goals of statistical and information literacy (cf. Table 1). The *Integration* factor relates to the ability and motivation to integrate data-driven insights into one's worldview and values (Callingham, 2006; Carmi et al., 2020; Wolff et al., 2016). It involves actively seeking comprehensive understanding of various topics, engaging with diverse perspectives, and consciously incorporating data-driven insights. Individuals scoring high in this factor adapt their opinions based on new data, prefer evidence-based information, and ensure their values align with reliable data. They engage with information and perspectives that challenge their existing views, showing a willingness to reassess their opinions and

positions based on new data (Carmi et al., 2020).

1.1.2 Producer Facets

The *Communication* factor revolves around the often referenced skill to effectively communicate and present data through various means, including visual formats, verbal explanations, and written descriptions (Calzada-Prado & Marzal, 2013; Kõuts-Klemm, 2019; Wolff et al., 2016). It requires translating complex data into clear formats, ensuring comprehension by varied audiences. Individuals scoring high in this factor hold the ability to translate data into simple visualizations, present findings confidently, and articulate complex information effectively in written, visual as well as verbal formats (Calzada-Prado & Marzal, 2013; Kõuts-Klemm, 2019; Wolff et al., 2016).

This definition, incorporates statistical literacy by emphasizing data interpretation, analysis, and understanding different types of data representations, such as graphs and tables (Statistics) (Gal, 2002). The *Statistics* factor covers skills related to managing and analyzing data effectively (Gould, 2017; Ridsdale et al., 2015). It involves proficiency in organizing and analyzing data using software tools, conducting statistical analyses, and understanding research methodologies (Gould, 2017; Williams et al., 2014). Individuals scoring high on this factor exhibit competence in conducting interviews or surveys for data collection, performing basic statistical analysis and recognizing trends in graphical representations (Deahl, 2014; Gould, 2017; Williams et al., 2014). In contrast to this presented concept of the *Statistics* factor, statistical literacy often focuses more narrowly on statistical concepts and methods, such as probability, sampling, and hypothesis testing (Gal, 2002; Gould, 2017). The *Statistics* factor encompasses a broader range of skills beyond statistical concepts, such as data visualization, software usage, and understanding data collection methods. While information literacy involves assessing the quality of information sources, this data literacy definition places a particular emphasis on assessing data quality, considering factors like sample size, biases, and data context. This aspect extends beyond traditional information literacy (Association of College & Research Libraries, 2000) and is more specific to data literacy.

The factors *Evaluation* and *Statistics* encompass behaviors and skills related to digital competency, including navigating and critically evaluating online sources and platforms, using information and communication technology (ICT), and utilizing statistical software. While digital competency focuses on the operational aspects of digital tools (Zhao et al., 2021), data literacy emphasizes the critical evaluation and application of data within specific contexts.

1.2 Nomological net

The term data literacy could evoke associations with skills, abilities or knowledge. Ability refers to an individuals potential or aptitude, encompassing innate or developed capacities across domains such as linguistic, mathematical, or motor abilities (Carroll, 1993). Knowledge, by contrast, is the cognitive representation or mental model of information, objects, and relationships. It forms the informational basis for interaction with and interpretation of the environment, encompassing all stored and retrievable data within an individuals mental framework (Norman & Rumelhart, 1975). Skill describes the application of learned and task-specific activities, often categorized into motor, cognitive, or social domains. Unlike abilities, skills imply mastery achieved through practice and are typically assessed based on performance quality (Green, 1998).

Motivational theories have evolved from broad mechanistic views to more nuanced understandings that incorporate social-cognitive aspects of motivation, highlighting the interplay between personality traits and learning outcomes (Dweck & Leggett, 1988; Mischel, 1973; Ross et al., 2005). This reflects a broader trend in psychology where the understanding of learning is increasingly contextualized within the individual learners psychological profile, including their personality characteristics (Elander, 2004; Komarraju et al., 2011; Mischel, 1973). Those constructivist theories advocate for learning as a socially mediated process, where personality traits such as openness to experience and conscientiousness can significantly influence how learners engage with content and interact with peers (Komarraju et al., 2011).

The role of motivation, influenced by personality traits, is particularly relevant in the context of competencies such as data literacy. Data literacy requires not only cognitive skills but also the motivation to engage with data, interpret it, and apply it effectively (Keshavarz, 2021). Research indicates that individuals with high levels of conscientiousness and openness are more likely to pursue learning opportunities related to data literacy, as these traits correlate with intrinsic motivation and a proactive approach to learning (Keshavarz, 2021; Mahmood et al., 2021; Saleh et al., 2018). For example, students who exhibit high conscientiousness tend to set specific learning goals and persist in their efforts to achieve them, which is essential for mastering complex skills like data analysis (Komarraju et al., 2009; Saleh et al., 2018).

Thus, data literacy can be seen as more on the side of proficiencies, because many behaviors incorporated in data literacy go beyond the mere question of whether a person is ‘able to do it’. It includes the question of one’s motivation to do it.

1.2.1 Need for Cognition

The concept of *Need for Cognition* (NFC), can be defined as "a need to structure relevant situations in meaningful, integrated ways. It is a need to understand and make reasonable the experiential world" (Cohen et al., 1955, p. 291). The concept captures individual variations in the engagement and enjoyment of thinking tasks (Bless et al., 1994). As data literacy incorporates several cognitive aspects as well as the motivation to understand data and information one gets presented with, it is expected to correlate positively with NFC. As NFC is more trait-like and data literacy is more a competency and therefore less stable, a small, nonetheless positive correlation is expected (i.e. correlation should be small) (cf. Table 1).

1.2.2 Opennes to new Experiences

Openness to new experiences is one factor of the *Big Five* (Costa & McCrae, 1992) and reflects a broad appreciation for art, emotion, adventure, unconventional ideas, imagination, curiosity, and diverse experiences (John et al., 2008). Individuals high in openness tend to be intellectually curious, receptive to emotions, appreciative of beauty, and eager to explore new possibilities. They are often more creative and emotionally attuned compared to those low in openness. In contrast, individuals low in openness tend to seek fulfillment through perseverance and are characterized as pragmatic and sometimes viewed as dogmatic or closed-minded (John et al., 2008). As already mentioned, data literacy can be thought of as a competency, thus incorporating the individual motivation, leading to a certain behavior. Therefore, openness to new experiences is expected to correlate positively with data literacy. As openness to new experiences is more trait-like and data literacy is not, the correlation is expected to be small (cf. Table 1).

1.2.3 Conscientiousness

Conscientiousness is one factor of the *Big Five* (Costa & McCrae, 1992) and refers to an individuals propensity for self-discipline, dutifulness, and striving for achievement in alignment with external standards or expectations. It encompasses levels of impulse control, regulation, and goal-directed behavior (Costa & McCrae, 1992; John et al., 2008; Toegel & Barsoux, 2012). High conscientiousness is characterized by persistence and focus, often perceived as stubbornness, whereas low conscientiousness is linked to flexibility and spontaneity, potentially manifesting as carelessness and unreliability (Toegel & Barsoux, 2012). Individuals with high conscientiousness tend to prefer planned actions over spontaneous ones (Costa & McCrae, 1992; John et al., 2008).

The cognitive aspects of data literacy as well as the motivation to understand data and information speaks to the conscientiousness of people as well. As being critical and at times detail oriented (e.g. in interpreting results, or spotting inconsistencies in presented information or while examining the credibility of sources) is also integral to data literacy, they are are expected to correlate positively. However, as conscientiousness is expected to be more stable over time, opposing to data literacy, a small, nonetheless positive correlation is expected (i.e. correlation should be small) (cf. Table 1).

1.3 Aim of the Study

The aim was to derive a comprehensive definition of data literacy based on existing approaches and to develop a questionnaire for self-rated data literacy with citizens being the target population. Thus the research question of this study is: "Does the proposed set of items effectively capture the latent factor structure of self-rated data literacy, and can the created scale be considered a reliable and valid measure of this construct?"
H1: The training-data will support the suggested latent factor structure and the proposed measurement model.
H2: The latent factor structure of the initial analysis will be supported by a different sample.
H3: A moderate to high positive correlation with information behavior related self-efficacy (SWE-IV-16) (Behm, 2018) is expected.
H4: A moderate positive correlation with self-perceived competence in using information and communication technology (ICT-SC25) (Schauffel et al., 2021) is expected.
H5: A small positive correlation with need for cognition (NFC-K) (Beißert et al., 2015) is expected.
H6: A small positive correlation with openness to new experiences (BFI-10) (Rammstedt et al., 2014) is expected.
H7: A small positive correlation with conscientiousness (BFI-10) (Rammstedt et al., 2014) is expected.

2 Methods

2.1 Item Creation

Prior to the item creation, a review of the literature was done. Since the goal was to create a self-report questionnaire, the indicators of the latent constructs were decided to be subjective indicators, or *Q-data* (Bühner, 2021). Furthermore

it was set that the target group for the questionnaire are German speaking citizens. There were no clear restrictions regarding age or education, other than the participants being of legal age and that the questionnaire should not be directed at professionals in terms of data literacy, or highly educated people.

The item creation itself was oriented towards the Act-Frequency-Approach (Buss & Craik, 1983), thus the prototype approach. I started with thinking of frequent, relevant behaviors, convictions or believes reflecting the factors of data literacy, hence being prototypical of those latent factors. For the *Comprehension* factor that could be recognizing whether the interpretations of others fit the available data, or recognizing when one is presented with contradictory information or understand the information a graphic contains, when data is presented visually. Regarding the *Evaluation* factor that could be assessing the credibility of information, or consider the reputation of the source. As well as being aware that publishers' own interests can influence the published information or to check information by comparing several sources with each other. The *Integration* factor could be represented by dealing with information that challenges ones views or consciously incorporating data-based findings into ones opinion-forming process. Additionally it could also manifest itself in changing ones mind if new data calls it into question. For the *Communication* factor, that could be represented by feeling confident in presenting data in visual formats in a way that can be understood by different target groups. But also to feel confident in expressing ones point of view in discussions or being able to summarize the most important information from data sets. As for the *Statistics* factor, this could be represented by knowing how to distinguish between causality and correlation or analyzing data sets using simple statistical methods.

That way, over 100 potential items have been created, that were then refined in terms of wording, structure but content as well. I avoided inversely worded items, and decided to not ask for specific examples, because those might enhance the difficulty of the respective items, depending on the personal background of the person answering the question. Additionally, I made sure that items were only ever asking for one behavior, conviction or opinion at once. Overall I tried to formulate the items as easy and understandable as possible. Ten cognitive interviews were held to refine those potential items, and to confirm the prototypicality of the behaviors etc. asked in the items. The cognitive interviews comprised the think-aloud technique as well as probing, to get an understanding of how the items are perceived, what comes to mind when reading the items and whether the items really ask for relevant behaviors (Fowler, 1995). Furthermore, that way unclear formulations or difficult wordings could be resolved. The interviews were administered iteratively to refine the items consecutively. In the

interviews it became apparent, that especially the degree to which people can relate to the factors four and five differed heavily. This was expected, since those producer factors are considered to play a lesser role in the every day live of citizens. It also turned out that some words like *data*, *source* or *information* are understood differently regarding the personal background. The ages of the interview participants ranged from 21 to 66 years. Among the ten participants, three were men and seven were women. Their professional backgrounds included diverse occupations: students, an employee at a sports facility, a music teacher, an IT professional, a civil servant, a construction manager, and two nurses (one working in an intensive care unit and the other in an operating unit). Additionally, two of the participants were retired. After the cognitive interviews of the 118 potential items, 71 remained.

2.2 Sample

The participants were recruited through a combination of personal and professional networks, along with outreach on several online social media platforms (e.g. *Instagram*, *LinkedIn*, *Whatsapp*, *Telegram* and via e-mail). Conducted in German, the participation in the study was entirely voluntary, with no external incentives provided. An a-priori power analysis was conducted to determine the necessary sample size for the structural equation modelling. I used the 'semPower' package in R (Moshagen & Bader, 2024) and also took a look into studies with similar goals and methods. The power analysis gave an analytical estimate of $N = 645$, and a simulated estimate $N = 613$, for the respective measurement model. In the literature sample sizes of $N = 500$ up to $N = 1000$ could be found (Algner & Lorenz, 2022; Remmert et al., 2022; J. Schneider et al., 2024). Accordingly, the target sample size for this study was selected to fall within this range, to balance practical feasibility with theoretical recommendations.

Participants had to be of legal age to be included in the study. Furthermore, attention check questions were included (three instructed response items and one seriousness check item at the end) within the survey to assess participants' attentiveness. Participants who failed to correctly answer two out of the four attention check questions were excluded from the analysis.

2.2.1 Demographics

The following characteristics of this study's sample are being made with referral to the respective statistics in the German population of 2022 (Bundesargentur für Arbeit, 2024; Statistisches Bundesamt, 2023a, 2023b, 2024c, 2024b, 2024a). The sample for this study comprised $N = 616$ participants. Within the sample, 48.3%

identified as female (50.65%), 50.2% as male (49.35%), and 1.5% did not identify with binary gender categories. The average age was 40 years (M = 44.6), with an average age of 39 years (M = 38.79; SD = 14.5) amongst women (M = 45.9) and 42 years (M = 41.96; SD = 14.23) amongst men (M = 43.2). The average age of people not identifying with binary gender was 28 years (M = 27.63; SD = 11.0). Regarding education, 31.9% of the participants reported a university degree as their highest educational level, 19.6% reported a degree from a university of applied sciences, 18.0% of the participants reported A-levels and 12.8% of the participants reported a completed apprenticeship as their highest educational level. The remaining participants were distributed across the other categories. Among participants who reported being employed, 64.88% indicated full-time employment. Specifically, 51.75% of women and 77.73% of men reported full-time employment at the time (65.15%; men: 42.40%, women: 22.75%). In contrast, 35.12% of all employed participants reported working part-time. At the time of the survey, 48.25% of women and 22.27% of men reported working part-time (28.22%;men = 6.16%; women = 22.06%). Additionally, 13.52% of all participants indicated that they were not employed at the time (6.63%). The study encompassed every sector within the occupational classification at least once. 25.2% of the participants indicated that they were students at the time of the survey (3.39%).

2.3 Open Science Standards

This project used the reproducibility workflow proposed by Peikert et al. (2021). 'Docker' and 'renv' work together to create a reproducible and portable environment. 'Docker' captures the complete software stack, while 'renv' focuses on managing R package dependencies and providing a clear documentation of the R package environment. This combination ensures that the analysis can be easily reproduced and shared with others in a reliable and transparent manner. The study, including all associated code, has been made openly available on GitHub as part of the commitment to open science. The repository can be accessed at https://github.com/LeonieHagitte/Thesis. The whole text is also hosted on GitHub-pages https://leoniehagitte.github.io/Thesis/. It is to be mentioned, that the 'repro' package from Peikert et al. (2021) has slightly changed in its functionality, namely that it does not ensure any longer, that old versions of the used software get re-installed. Furthermore I used the 'reproducibleRchunks' package from Brandmaier & Peikert (2024). This package enables the verification of computational results in R for reproducibility, ensuring that the same script with the same data produces identical results across different computers or at different times. When knitting the respective document, the user can see the results for the respective chunks, as to whether they are reproducible or not. The study was

preregistered at Zenodo (DOI:10.5281/zenodo.11196495).

2.4 Procedure

A cross-sectional online survey was used to examine a sample from the general population. Participants completed the self-perceived data literacy items alongside demographic questions and additional validation measures. Survey questions of each measurement were randomized for each participant to minimize order effects and response biases. All items, except for the feedback item, were given as a forced choice question, to reduce missing data. Due to the potential for missing responses caused by the length of the questionnaire, such as attention declines or reduced motivation, I addressed this issue by shortening the assessment. To shorten the overall length of the assessment the questions in each factor of the data literacy questionnaire were randomly selected for each participant. That way each participant only answered half of the formulated items, the other half were planned missings. I treated the first 25 participants as a pilot, to check for potential problems in the survey (like length, spelling mistakes that have been overlooked etc.).

2.5 Instruments

2.5.1 Measuring Data Literacy

Regarding data literacy 71 items were created. Each participant should answer 38 items of the 71 which were randomly selected. To answer the items, respondents indicated their agreement on a five-point rating scale (1 = "strongly disagree", 2 = "somewhat disagree", 3 = "neither agree nor disagree", 4 = "somewhat agree", 5 = "strongly agree") with a "don't know" option.

2.5.2 Measuring Information Literacy

The SWE-IV-16 (Behm, 2018) assesses the self-efficacy beliefs of adolescents and adults in their ability to engage in information behavior. This questionnaire measures the process model of information-related problem-solving (Brand-Gruwel et al., 2009). In the present study this construct was used as a proxy for information literacy. It consists of 16 statements addressing self-assessed abilities in searching for and evaluating information, as well as managing information searches effectively. Each statement begins with "When I search for information on a topic or a specific question…" and respondents indicate their agreement on a five-point Likert scale (1 = "strongly disagree", 2 = "somewhat disagree", 3 = "neither agree nor disagree", 4 = "somewhat agree", 5 = "strongly agree"). The

total scale value is computed as the arithmetic mean of the items. In this study, McDonalds ω was .91. For comparative reasons it is mentioned that Cronbachs α was .91 as well.

2.5.3 Measuring Need for Cognition

The NFC-K (Beißert et al., 2015) is a scale used to assess NFC through four items, which represent two facets: *engagement* and *joy*. The NFC-K is measured with a seven-point rating scale, ranging from 1 = "strongly disagree" to 7 = "strongly agree", with a "neither" option in the middle. The German version of the scale was adapted from the original English scale by Cacioppo & Petty (1982) and translated by Bless et al. (1994). To determine an individual's NFC score, a mean value (scale value) is computed from the four raw score points of the responses. The resulting mean values range between 1 and 7. In the current study McDonalds ω was .62 and Cronbachs α was .60.

2.5.4 Measuring Technology Competency

To assess self-perceived competence in using ICT, the five general items of the ICT-SC25 (Schauffel et al., 2021) were used. The ICT-SC25 is a scale consisting of 25 items designed to assess self-perceived competence in using ICT. It is available in both German (ICT-SC25g) and English (ICT-SC25e). The scale measures general and domain-specific ICT competence, including communication, processing and storing, content generation, safe application, and problem-solving skills. Items are measured using a six-point fully-labeled rating scale ranging from 1 = "strongly disagree" to 6 = "strongly agree". Researchers can choose to utilise either the entire scale or individual subscales based on their specific research objectives. Manifest scale scores for the ICT-SC25g/e are calculated separately for each subscale by computing the unweighted mean score of the items within each subscale (Schauffel et al., 2021). In the current study, the McDonalds ω was .93, while Cronbachs α was .93.

2.5.5 Measuring Openness and Conscientiousness

The BFI-10 (Rammstedt et al., 2014) was used to assess personality based on the five-factor model. Only the items on openness and conscientiousness were assessed. The items are answered on a five-point rating scale from 1 = "strongly disagree" to 5 = "strongly agree". To measure the respondent's individual traits on the two personality dimensions, the responses to the two items for each dimension are averaged. First, the negatively worded item need to be recoded (items 1, 3, 4, 5, and 7), then the mean value is calculated for each dimension from both the

recoded and non-recoded items. The values for the five dimensions range from 1 to 5 (see Rammstedt & John (2007) for reference values). In this study for the items on openness, the McDonalds ω was .63; while the Cronbachs α was .63. For the items on conscientiousness, the McDonalds ω was .56; while the Cronbachs α was .56.

3 Analysis

3.1 Data Quality

Careless or inattentive response patterns were assessed using attention-check items. Additionally, the data was visually inspected for outliers through boxplots. The data showed several outliers, that were left in the dataset. The outliers appear consistently across multiple items, which might suggest a systematic bias, where a subset of respondents consistently answers in the lower categories. This could be due to respondents finding the items too difficult or not engaging with the material in a meaningful way. Missing data, due to planned random omissions, was imputed using Full Information Maximum Likelihood (FIML) with 'stuart' in combination with 'lavaan' (Rosseel, 2012; Schultze, 2022). FIML was also applied during model estimation to handle missing values. Aside from the planned random missings in the data literacy questions, there were non-planned missings due to incomplete questionnaire submissions. These non-planned missings were not imputed. For those participants who did not finish the questionnaire, data from their data literacy-related responses were still used in the training sample, as listwise deletion would have resulted in too few data points for proper algorithm functioning.

3.2 Main Analyses

Algorithm based item selection was done via the R package 'stuart'(Schultze, 2022) and the confirmatory factor analysis (CFA) via the R package 'lavaan'(Rosseel, 2012). Inference criteria: Model fit was assessed using established criteria (Hu & Bentler, 1999). Comprising of Chi square significance testing as well as a combination of several fit indices, i.e., Root Mean Square Error of Approximation (RMSEA) < 0.05, Standardized Root Mean Square Residual (SRMR) < 0.07, Comparative Fit Index (CFI) > 0.95 (Hu & Bentler, 1999). Model-specific cutoff values were considered as well, using the 'ezCutoffs' package (Schmalbach et al., 2019).

3.2.1 Rationale for Measurement Model

The design of the measurement model has to be chosen prior to item selection, which includes determining the number of items per factor in the final scale (Schultze & Lorenz, 2024). This decision influences both the parsimonity and the evaluation of the model. The scale was designed to be as parsimonious as possible, while ensuring that a real model fit could be estimated. Statistically, a configuration with three items per factor would result in a just-identified model, leaving no degrees of freedom to test fit. To address this, four items per factor were chosen as the optimal configuration. This choice represents the most parsimonious structure that allows for the assessment of model fit. Furthermore, the selection of four items per factor ensures that each facet can be independently utilized in practical applications while still providing 'local' indicators of fit. This design choice allows researchers to test each facet separately, without requiring the inclusion of additional constructs, thereby maintaining the flexibility and utility of the model across various contexts.

3.2.2 Meta-Heuristics

Algorithm-based item selection was used to choose the most relevant items, reducing the initial item pool. In classical approaches, items are evaluated within the overall item pool and are then often selected based on their individual properties (e.g. difficulty, discrimination, item-scale-correlations). Compared to classical approaches, algorithms are more objective and efficient in finding a good solution with regard to certain criteria (Leite et al., 2008; Olaru et al., 2015). Furthermore, some empirical studies suggest that the use of algorithms leads to similar or better results in scale construction than traditional approaches (Olaru & Danner, 2021; Sandy et al., 2014; Schroeders et al., 2016). The automated approach takes the opposite perspective to the classical approach, the one of meta heuristics, by repeatedly estimating CFAs for a multitude of possible item-combinations (Schultze, 2017). Thus, a pool of items with some constraints and the goal to find the one combination that best fits the suggested purpose (e.g. equation 1) of the final scale is estimated (Schultze, 2017). Hence, the selection of items and construction of a questionnaire can be viewed as a combinatorial problem, like the knapsack problem ("Choose a set of objects, each having a specific weight and monetary value, so that the value is maximized and the total weight does not exceed a predetermined limit") (Kerber et al., 2022; Schroeders et al., 2016, p. 4; Schultze, 2017). For this study, a set of 20 items from a set of 71 items was selected, to form a questionnaire. The data literacy self rating scale was optimized for model fit criteria (RMSEA, SRMR & CFI). Those criteria were defined in the

objective function (cf. Equations 1 & 2) in 'stuart'(Schultze, 2017, 2022). Equation 1 is a conceptual expression indicating that Φ is a function of the fit indices RMSEA, SRMR, and CFI. It leaves the specific functional forms of F undefined. Equation 2 specifies the exact transformations for these fit indices, where each $F()$ in the general formulation corresponds to a specific transformation in the computational form. This formulation reflects an objective where models with smaller RMSEA and SRMR values and higher CFI values are preferred in the optimization process.

$$\Phi = F(\text{RMSEA}) + F(\text{CFI}) + F(\text{SRMR}) \quad (1)$$

$$\Phi = (1 - \text{RMSEA}) + (1 - \text{SRMR}) + (1 + \text{CFI}) \quad (2)$$

This objective function used deviated from the one that was planned in the preregistration, because the planned function did not result in acceptable model fits. Because of that, the demands on the algorithm were reduced solely in favor of model fit. I used the genetic algorithm of 'stuart' for the item selection. Genetic algorithms are based on Darwinian evolution principles – selection, crossover, mutation, and survival of the fittest (Holland, 1992; Schroeders et al., 2016). With the genetic algorithm, the initial set of 71 items was reduced based on the evolutionary process of selection, but opposing to evolution with a goal: A near-optimal 'solution'. The survival of an item is determined by its quality (i.e. "fitness") (Galán et al., 2013).

3.3 Validation

The provided sample was split into two subsets (i.e. training data and test data) using the 'holdout' function in 'stuart'. The specified item-selection procedure is applied to the training data. The training data underwent k-fold cross validation (k=3), using the 'kfold' function in 'stuart'. The number of folds (k) was determined based on the total sample size, ensuring that each fold was sufficiently large to support a CFA with adequate statistical power. To this end, a power analysis was conducted using the 'semPower' package in R (Moshagen & Bader, 2024) to calculate the required sample size per fold, which was determined to be n = 128. This calculation ensured that dividing the training sample (n = 373) into three folds would be the maximum feasible configuration. Those k-folded selections were then again iterated three times, to enhance the stability of the solution even more. The other sample, the test data, is used for evaluation of the final model's performance, as well as the latent correlations with the convergent measures. Validation with the test data is conducted (using an multi-group confirmatory factor analysis (MGCFA) in 'lavaan') to assess the invariance of the

measurement models between the training and testing datasets. Invariance levels are assessed using the criteria of F. F. Chen (2007). Invariance is necessary to claim that the scale validation has worked.

3.3.1 Assessing the Nomological Net

Criterion validity was examined through correlation analyses, using Kendall's rank correlation coefficient (Kendall, 1938), between the data literacy factors and the latent factors from the other administered questionnaires. Bartlett scores (Bartholomew et al., 2009) were used for the data literacy factors to account for substantial variation in factor loadings across items. For the factors from the other questionnaires, the respective scoring procedures outlined in their manuals were applied to ensure accurate computation of scores for the correlation analysis.

3.4 Exploratory Analyses

Because of the complexity of the model, as well as the several objectives of the initially planned objective function, the genetic algorithm had problems converging into a stable solution, let alone a solution with sufficient model fit. Therefore, some objectives were dropped (the composite reliability and the varying item intercepts per factor), in favor of model fit. Different solutions were systematically explored, dependent on the estimator (Maximum likelihood with robust standard errors (MLR) or robust Weighted Least Squares (WLSMV) and the respective data structure (data treated as metric or ordinal data), as well as on the objective function, to lead to the best solution (cf. code in the supplementary Material ("Analysis.Rmd"; starting at line 2340)). Furthermore, the data was checked for multicollinearity because of convergence issues.

3.4.1 Analysis of Confounding Variables

To check for possible external influences of the categorical, demographic variables (age; gender; highest educational level; the pursued degree, if studying & occupation), a dummy-coded multiple linear regression was calculated for every selected item. The reference categories for the categorical demographic variables (e.g., "Gender," "Education") are chosen automatically by R. For this analysis the 'lm()' function was used.By default, this function uses the first level of a factor variable as the reference category unless otherwise specified.

4 Results

The descriptive analysis of the selected items showed several results to be highlighted. The item *F1F11* ("Ich kenne unterschiedliche Arten von Grafiken") had a high average score (M = 4.48, SD = 0.75), indicating that respondents rated their knowledge of different types of graphs very positively. Similarly, *F3F9* ("Ich kann neue Informationen in meinen Wissensstand integrieren") also demonstrated a strong ceiling effect with a mean of 4.32 (SD = 0.64), suggesting that integrating new information is a skill respondents deemed well-developed. Furthermore, several items exhibited high variability in responses. The item *F4F1* ("Ich kann Ergebnisse in Streudiagrammen darstellen") had the lowest mean (M = 2.88) but also the highest standard deviation (SD = 1.45), indicating substantial differences in respondents' perceived abilities to use scatterplots, or in their understanding of the item. Other items with notable variability included *F4F3* ("Ich kann Daten in Grafiken so präsentieren, dass sie für verschiedene Zielgruppen verständlich sind," M = 3.45, SD = 1.23), *F5F3* ("Ich bin in der Lage, einfache Datenbanken zu verwalten," M = 3.44, SD = 1.31), and *F5F4* ("Ich kann Datensätze mit einfachen statistischen Methoden analysieren," M = 3.28, SD = 1.26), reflecting diverse self-assessments of these specific skills. Additionally, item *F2F2* ("Wenn Ich die Glaubwürdigkeit von Informationen beurteile, berücksichtige ich den Ruf der Quelle"), had a relatively high average (M = 4.04) as well as variability (SD = 1.04), suggesting differing abilities to evaluate source credibility. Similarly, *F2F6* ("Ich kenne Merkmale zur Bewertung von Datenqualität," M = 3.19, SD = 1.13) and *F5F18* ("Wenn ich Daten analysiere, achte ich auf ein systematisches Vorgehen," M = 3.74, SD = 1.13) showed considerable variation, highlighting mixed levels of competency in assessing data quality and conducting systematic analyses. Table 1 presents the translation of the final set of items along with their means and standard deviations. The original German versions of the items are available in the Appendix (cf. Table A1).

4.1 Model Fit and Measurement Invariance

Of all explored conditions and respective models, the following model is to be highlighted, as it showed the best overall fit. With an objective function only optimizing for model fit criteria (RMSEA, SRMR, CFI), the algorithm selected 20 of the 71 original items representing the five factors Comprehension, Evaluation, Integration, Communication and Statistics with four items each (Figure 3). The solution exhibits good model fit (according to Hu & Bentler (1999)) with data treated as metric and MLR as estimator: Satorra-Bentler-X^2 (df = 414) = 582.582, p < 0.001, CFI = .96, SRMR = .08, RMSEA = .05, 90%-CIRMSEA [.036; .054]. This

Table 2: The selected item pool

Tag	Item	Mean	*SD*
F1F6	When evaluating graphics, I am able to recognize contradictions.	3.74	.83
F1F8	I recognize whether the interpretations of others fit the available data.	3.66	.97
F1F11	I know different kinds of graphics.	4.48	.75
F1F14	I can recognize the central thesis of a scientific text.	3.92	.88
F2F2	When I assess the credibility of information, I consider the reputation of the source.	4.04	1.04
F2F6	I know features for evaluating data quality.	3.19	1.13
F2F15	I check the qualifications of authors before relying on the information.	3.18	.94
F2F20	I can tell when data is of poor quality.	3.25	.97
F3F2	I deal with information that challenges my views.	3.57	.80
F3F4	I can find information in databases quickly and easily.	3.38	.92
F3F6	I prefer data-based information when forming an opinion.	3.75	1.09
F3F9	I can integrate new information into my knowledge base.	4.32	.64
F4F1	I can present results in scatter plots.	2.88	1.45
F4F3	I can present data in graphics in such a way that they are understandable for different target groups.	3.45	1.23
F4F6	I can choose the most suitable form of presentation.	3.54	1.01
F4F8	I can use programs to create graphics to present results.	3.73	1.17
F5F3	I am able to administer simple databases.	3.44	1.31
F5F4	I can analyze data sets using simple statistical methods.	3.28	1.26
F5F8	When I am confronted with extensive data sets, I can gain insights from them.	3.47	1.08
F5F18	When I analyze data, I pay attention to a systematic approach.	3.74	1.13

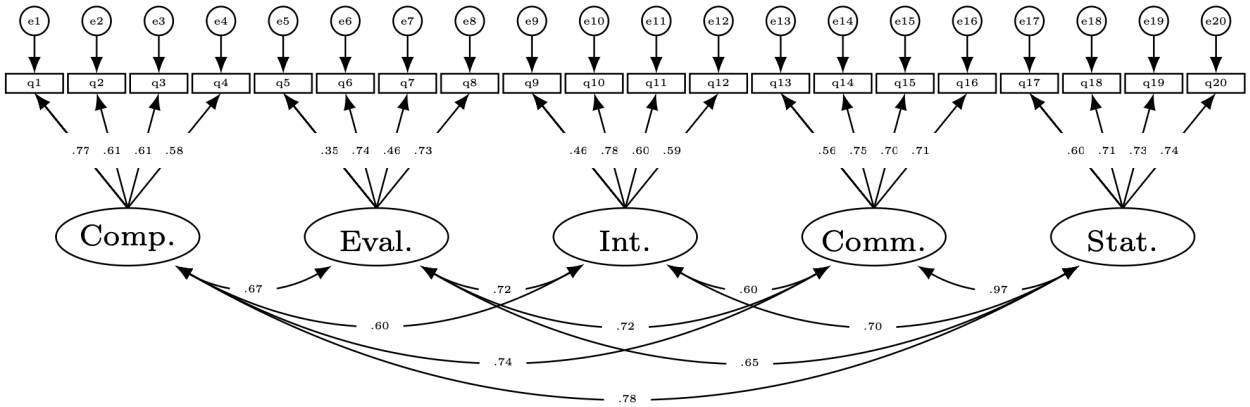
Note. Translated items sorted by latent factor, selected by the algorithm. The tag refers to the name the item had in the programming. The First part of the tag is always indicative of the factor (F1 = Comprehension, F2 = Evaluation, F3 = Integration, F4 = Communication and F5 = Statistics). For every item the mean with the respective standard deviation (*SD*) is displayed.

supports the first hypothesis (H1).

For the configural model of the training sample in the MGCFA, standardized loadings of the factor Comprehension ranged from .58 to .77. For the factor Evaluation loadings ranged from .35 to .74. For the factor Integration loadings ranged from .46 to .78. For the factor Communication loadings ranged from .56 to .75. For the factor Statistics loadings ranged from .60 to .74. The complete model is also displayed in Figure 3.

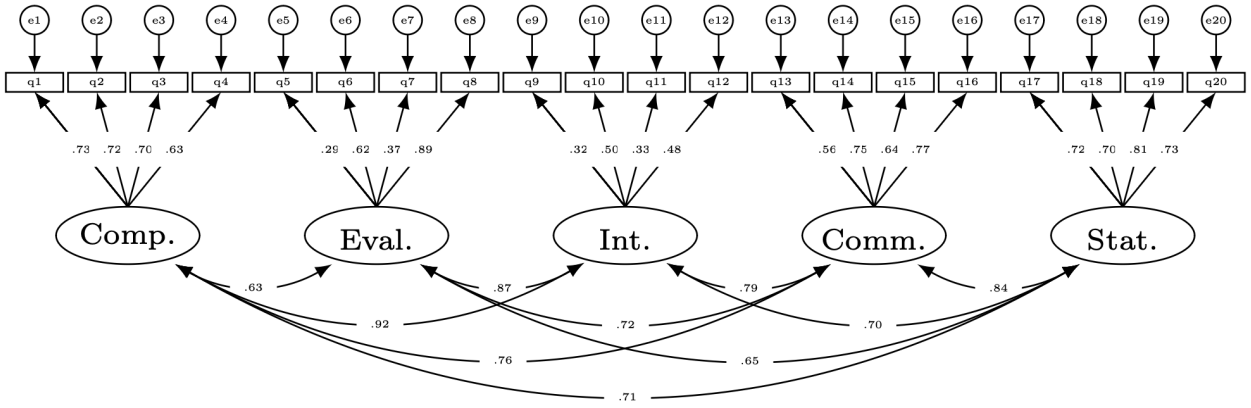
For the configural model of the test sample in the MGCFA, standardized loadings of the factor Comprehension ranged from .63 to .73. For the factor Evaluation loadings ranged from .29 to .89. For the factor Integration loadings ranged from .32 to .77. For the factor Communication loadings ranged from .56 to .77. For the factor Statistics loadings ranged from .70 to .81. The complete model is also displayed in Figure 4. Cross-validation of the MGCFA with the test-data indicated that the assumption of scalar invariance holds across the two subsamples:

Figure 1: Measurement Model of the MGCFA Model in the training sample.



Note. This figure displays the configural measurement model of the data literacy items within the training sample, showing standardized factor loadings and latent covariances. The abbreviations q1 up to q20 represent the items, the abbreviations e1 up to e20 their respective residuals. The abbreviations, Comprehension (Comp.), Evaluation (Eval.), Integration (Int.), Communication (Comm.), and Statistics (Stat.) indicate which is the respective latent factor.

Figure 2: Measurement Model of the MGCFA Model in the test sample.



Note. This figure displays the configural measurement model of the data literacy items within the test sample, showing standardized factor loadings and latent covariances. The abbreviations q1 up to q20 represent the items, the abbreviations e1 up to e20 their respective residuals. The abbreviations, Comprehension (Comp.), Evaluation (Eval.), Integration (Int.), Communication (Comm.), and Statistics (Stat.) indicate which is the respective latent factor.

$X^2(df = 389) = 659.746$, $p < 0.001$, CFI = .92, SRMR = .087, RMSEA = .038; ΔCFI < .00 , ΔSRMR < .00, ΔRMSEA = .001 (cf. Table 2). This indicates support for the second hypothesis (H2).

Table 3: Fit Indices for Model and Results of MGCFA testing

Model	Invariance level	CFI	RMSEA	SRMR	Δ CFI	Δ RMSEA	Δ SRMR
Model 1	Configural	.92	.041	.086	-	-	-
	Metric	.92	.039	.087	.002	.002	.001
	Scalar	.92	.038	.087	-	.001	-
	Residual	.91	.041	.092	**.015**	.003	.005

Note. The table shows the essential fit indices of the model created with 'stuart', alongside their change in terms of Invariance testing, per invariance level. All Indices shown were estimated with ML as estimator.

The model of the final set of items shows McDonald's ω_{total} = .92. The composite values McDonald's ω_{total} of the factors are Comprehension = .76, Evaluation = .65, Integration = .67, Communication = .81 and Statistics = .68. While overall fit indices (e.g., RMSEA, CFI) suggest a good fit, an analysis of correlated residual was done to reveal possible specification errors. The correlated residuals from the training sample as well as from the testing sample, can be seen in the appendix (cf. Table A2 & Table A3). Both samples show several residuals to be correlated.

Table 4: Latent Factor correlations training sample

	Comprehension	Evaluation	Integration	Communication	Statistics
Comprehension	1.000				
Evaluation	0.672	1.000			
Integration	0.596	0.715	1.000		
Communication	0.735	0.720	0.604	1.000	
Statistics	0.782	0.649	0.700	**0.971**	1.000

Note. The table shows the correlations between the latent factors for the training sample.

4.2 Multicollinearity

To furthermore investigate the model, the covariances of the latent factors were examined. The factor correlations for the training sample are displayed in Table 2. The correlation of .97 between Communication and Statistics is to be highlighted. The correlations for the test sample are displayed in Table 3. In the test sample the correlation of .84 between Communication and Statistics is to be highlighted, alongside the correlation of .92 between Comprehension and Integration, as well as the correlation of .87 between Evaluation and Integration. Especially the correlations, that exceed .90 suggest multicollinearity in the data.

Table 5: Latent Factor correlations test sample

	Comprehension	Evaluation	Integration	Communication	Statistics
Comprehension	1.000				
Evaluation	0.627	1.000			
Integration	**0.918**	**0.870**	1.000		
Communication	0.760	0.432	0.799	1.000	
Statistics	0.707	0.361	0.694	**0.837**	1.000

Note. The table shows the correlations between the latent factors for the test sample.

4.3 Criterion Validity

For the correlations, Kendall's rank correlation coefficient (τ) was estimated, because the data was not normally distributed. The factors (Comprehension, Evaluation, Integration, Communication & Statistics) of the data literacy scale correlated moderately to highly with the SWE-IV-16 (τ =.47, τ =.46, τ =.40, τ =.48, τ =.39 ; all p < .01). The factors of data literacy showed small to moderate correlations with the NFC-K (τ =.22, τ =.24, τ =.30, τ =.37, τ =.42; all p < .01). The factors of the data literacy scale showed moderate correlations with the general items of the ICT-SC25 (τ =.31, τ =.17, τ =.23, τ =.43, τ =.32; all p < .01). The factors Comprehension, Evaluation and Integration correlated slightly negative with the openness of the BFI-10 (τ = -.15, p < .05; τ = -.18, p < .01; τ = -.26, p < .01). Openness did not correlate statistically significant with the other factors. The factors Comprehension, Evaluation, Communication and Statistics correlated slightly up to moderate with conscientiousness of the BFI-10 (τ = .13, p < .05; τ = .25, p < .01; τ = .16, p < .05; τ = .17, p < .01). The latent correlations with respective confidence intervals are also displayed in table 6. Those results indicate support for the hypotheses 3, 4 and 7 (H3, H4 & H7), while not supporting the hypotheses 5 and 6 (H5 & H6).

4.4 Control Variables

The analysis of control variables (gender, educational level, aspired degree, and occupation classification) revealed that most items showed no statistically significant effects. Singularity issues, particularly in degree and occupation categories, further limited the interpretability of some results.

For 'Education 5' (A-levels), a statistically significant negative effect was found on F2F15 (b = -3.00, p = 0.03), indicating an average score 3.0 points lower than individuals within the reference group. Conversely, A-levels had a positive effect on F4F3 (b = 3.91, p = 0.04). Education also influenced F4F8, where A-levels (b = 2.82, p = 0.02), 'Education 7' (degree from a university of applied sciences) (b = 7.46, p = 0.00), and 'Education 8' (university degree) (b = 7.04, p = 0.00) also

Table 6: Means with standard deviations, and correlations with confidence intervals

Var	*M*	1	2	3	4	5	6	7	8	9
1. Con.	3.68									
	[0.76]									
2. Open	2.18	-0.09								
	[0.87]	[-0.21, 0.04]								
3. Nfc	4.89	0.16*	-0.06							
	[1.07]	[0.04, 0.28]	[-0.19, 0.06]							
4. Ict	4.03	0.03	-0.04	0.18**						
	[0.72]	[-0.10, 0.16]	[-0.16, 0.09]	[0.06, 0.30]						
5. Swe	3.66	0.38**	-0.24**	0.30**	0.34**					
	[0.53]	[0.26, 0.48]	[-0.36, -0.12]	[0.19, 0.41]	[0.23, 0.45]					
6. Comp.	-0.03	0.13*	-0.15*	0.22**	0.31**	0.47**				
	[0.85]	[0.01, 0.26]	[-0.28, -0.03]	[0.10, 0.34]	[0.19, 0.43]	[0.37, 0.56]				
7. Eval.	-0.00	0.25**	-0.18**	0.24**	0.17**	0.46**	0.32**			
	[0.54]	[0.12, 0.36]	[-0.30, -0.05]	[0.12, 0.36]	[0.04, 0.29]	[0.35, 0.55]	[0.20, 0.43]			
8. Int.	-0.01	0.05	-0.26**	0.30**	0.23**	0.40**	0.28**	0.26**		
	[0.54]	[-0.07, 0.18]	[-0.37, -0.14]	[0.18, 0.41]	[0.11, 0.35]	[0.28, 0.50]	[0.16, 0.40]	[0.13, 0.37]		
9. Comm.	0.01	0.16*	-0.10	0.37**	0.43**	0.48**	0.52**	0.27**	0.34**	
	[0.92]	[0.04, 0.28]	[-0.23, 0.02]	[0.26, 0.48]	[0.33, 0.53]	[0.37, 0.57]	[0.42, 0.61]	[0.15, 0.38]	[0.22, 0.45]	
10. Stat.	0.03	0.17**	-0.04	0.42**	0.32**	0.39**	0.40**	0.18**	.30**	0.61**
	[1.00]	[0.04, 0.29]	[-0.17, 0.09]	[0.30, 0.52]	[0.20, 0.43]	[0.27, 0.49]	[0.29, 0.51]	[0.05, 0.30]	[0.18, 0.41]	[0.53, 0.69]

Note. The table displays values of Kendalls tau. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets beneath the M show the respective standard deviation. The other values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

showed statistically significant positive effects.
'Degree 2' (Masters) showed a statistically significant negative effect on F4F8 (b = -3.20, p = 0.05), while 'Degree 3' (state examination) had a positive effect (b = 2.60, p = 0.03). On F5F8, pursuing a state examination showed a positive effect (b = 3.00, p = 0.05), while pursuing a Master and 'Degree 4' (doctorate) encountered singularity issues and could not be estimated.
Finally, for 'Age 22', a statistically significant negative effect was found on F5F8 (b = -3.00, $p < 0.05$), indicating individuals aged 22 scored, on average, 3.00 points lower than the reference group.

5 Discussion

This study meant to examine whether 20 items out of the initial item pool, would reflect the suggested measurement model within the current sample. Furthermore, the aim was to find a solution utilizing the genetic algorithm as well as a 3-fold crossvalidation in 'stuart', that yields good model fit as well as reliability, and that shows measurement invariance with a random split-sample. Additionally, correlations with other scales and constructs were to be examined to locate the construct within the nomological net. The model demonstrating the best overall fit was selected, based on a set of 20 items representing five factors: Comprehension, Evaluation, Integration, Communication, and Statistics, each with four items and respective measurement errors. The model also included inter-factor correlations between all latent factors. Construct validity was evaluated through CFA, using MLR as estimator, as well as correlation analyses with related constructs.

5.1 Model fit Initial Item Selection

Although the RMSEA was a bit low, the initial model exhibited an acceptable to good fit, with indices suggesting good model fit (RMSEA = 0.05, SRMR = 0.08, CFI = 0.96), according to Hu & Bentler (1999). However, recent advancements in scientific methods, such as the "ezCutoffs" package (Schmalbach et al., 2019), propose simulated cutoffs that adapt to specific model characteristics, potentially offering a more rigorous and context-sensitive approach to assessing model fit. According to this, the following cutoffs are suggested: scaled CFI: .986; scaled RMSEA: .022 and SRMR: .041. Respectively, the model exhibits non-acceptable fit. It is argued, that fixed cutoffs are derived from specific confirmatory factor analysis models and may not perform well under varying conditions, such as sample size, model complexity, and response types (Goretzko et al., 2023; McNeish

& Wolf, 2023). This woult suggests that while traditional cutoffs provide a useful starting point, they may not adequately capture the nuances of different modeling scenarios. The factor loadings ranged from 0.50 to 0.81 across the five factors (cf. Figure 1), and the models McDonald's ω_{total} was 0.92, indicating reliable internal consistency. Regarding the first hypothesis (H1) it can be concluded, that the algorithm was able to find a solution, exhibiting good model fit according to Hu & Bentler (1999), in the training sample, for the suggested measurement model. However, according to the cutoffs from the "ezCutoffs" package (Schmalbach et al., 2019), the H1 would not be supported.

5.2 Measurement Invariance and Model Fit

For further validation, the model was tested for measurement invariance against the test-sample. The results suggests that invariance levels up to scalar invariance hold (cf. Table 3). This conclusion is drawn based on the fit indices remaining within the acceptable range proposed by F. F. Chen (2007). However, given that fit indices are regarded more as rough guidelines than as precise or universally applicable cut-off values, an additional perspective was considered. Specifically, the differences between the (robust) confirmatory fit indices were examined to ensure they remained below 0.01, as recommended by Cheung & Rensvold (2002). In practical terms, when scalar invariance holds, the factor loadings, intercepts, and measurement scales can be considered equivalent across groups. Thus, the data indicates that all factor loadings are consistent across groups (Cheung & Rensvold, 2002). This suggests that the data literacy items and their underlying factors are associated with the same strengths in both samples. Additionally, scalar invariance was established, implying that beyond the factorial structure and factor loadings, the item intercepts are also invariant across groups. This allows for meaningful comparisons of latent means between groups rather than observed differences being due to measurement bias or differences in how the construct is understood or measured across groups (Cheung & Rensvold, 2002; Riordan & Vandenberg, 1994). The lack of residual invariance suggests that the residuals (unexplained variance in the indicators) are not equivalent across groups. This implies that there are group-specific differences in how much of the variance in the observed variables remains unexplained by the latent factors. Thus, this lack of residual invariance complicates the interpretation of differences between groups (Cheung & Rensvold, 2002). While the factors themselves may be measured similarly (because scalar invariance holds), the amount of unexplained variability in the responses differs across groups, indicating potential unmodeled differences in how the groups respond to certain items. As a result, any observed differences in the latent factors could be influenced by differing error variances

across groups, making it challenging to draw definitive conclusions about true group differences. Thus, while latent means can be compared, the comparisons may be confounded by measurement error that varies across groups.
Regarding the second hypothesis (H2) it can be concluded that the latent factor structure of the initial analysis was supported by the test-sample. It is also to be noted that the model fit on the configural level (RMSEA = 0.041, SRMR = 0.09, CFI = 0.92), according to Hu & Bentler (1999) and the 'ezCutoffs' package (Schmalbach et al., 2019), is worse than the fit of the training sample, except for the RMSEA (cf. Table 3). The changing of the fit of the test sample is also reflected in the measurement model (cf. Figure 2). There does not appear to be a consistent direction of change, as some factor loadings on the items increase while others decrease. The same applies to the latent correlations between the factors, which could be interpreted as an indication of the solution's instability.

Across the versions of this model, the CFI decreases the most and the RMSEA suffers the least. As the CFI is comparing the tested models against a base-line model, it could be that the strong decrease in this index reflects the divergence of the model respective to the differing sample and thus indicates overfitting. Furthermore, some items showed high correlations (cf. Appendix Table A4), as well as high residual correlations (cf. Appendix Tables A2 & A3) (shared variance not explained by the latent factor). Those correlated residuals, could indicate potential specification errors, both in the training and test sample, therefore the model fit could deteriorate and reduce the CFI. Furthermore, the latent correlations among factors were overall strong, especially between Communication and Statistics (0.91 in the training sample and 0.84 in the test sample) (cf. Tables 4 & 5). This could indicate redundancy among the factors, that needs to be addressed in terms of items and factor specification. The SRMR showed mediocre fit, but degrades in the MGCFA and progressively so with increasing the invariance levels. This could in part be due to the fact, that the SRMR makes no correction for parsimony, thus improves the more paths a model contains. As invariance is increased and paths are constrained to equality, it could make sense for the SRMR to increase. The RMSEA in turn accounts for some complexity of the model, and also performs better in this case as the SRMR.

5.3 Construct Validity

The moderate to high correlations between the newly created data literacy scale and the SWE-IV-16, and the general items of the ICT-SC25 (cf. Table 6) suggest that the scale measures a concept that is related but distinct. This supports the expectations outlined in H3 and H4. The results consistent with H3 and H4, which concern the alignment of the measured latent factors with competency-based

constructs, further support the interpretation that the scale captures perceived competency rather than a stable personal character trait. Those correlations could indicate convergent validity (Gregory, 2004).
The strong correlation between the data literacy scale and NFC-K are not in line with H5 (cf. Table 6). This might reflect conceptual proximity between the two measures, and could indicate that the data literacy questionnaire measures a construct more aligned with trait characteristics than planned. It is plausible that individuals who enjoy understanding concepts in detail (as measured by NFC-K) also display behaviors associated with critically examining data and information. This finding underscores the relevance of motivation as a contributing factor to the observed behaviors, suggesting conceptual overlapping elements between data literacy and NFC.
The data literacy factors also showed small to moderate correlations with personality traits, including a negative correlation with openness and a positive correlation with conscientiousness. The positive association with conscientiousness aligns with H7, indicating that individuals who are organized and diligent may also exhibit behaviors related to data literacy. This correlation could be an indicator for discriminant validity (Cronbach & Meehl, 1955; Hubley & Zumbo, 1996).
However, the negative correlation with openness contradicts expectations (H6). Given the conceptualization of data literacy as a proficiency that incorporates motivation to engage in related behaviors, a slight positive correlation with openness was anticipated, as openness typically reflects curiosity and a willingness to explore new ideas, or a persons motivation to show certain behaviors. The opposite was found. This unexpected negative relationship could suggest that the items of the data literacy scale emphasize ability or skill over proficiency, thereby capturing behaviors less influenced by intrinsic motivation.
Another plausible explanation for the findings would be a mismatch between the items and the study sample. Motivation can be linked to expertise (Earley et al., 1990; Paletz et al., 2013), and participants may perceive the assessed behaviors as routine rather than requiring deliberate effort or motivation. For these individuals, data literacy behaviors might represent habitual actions rather than aspirational or exploratory tendencies.

Paletz et al. (2013) highlight that individuals engaged in routine behaviors can identify problems and adapt their processes as they gain experience. This suggests that, as expertise develops, individuals may rely more on established routines rather than on motivation to initiate behaviors. Similarly, Jong & Fodor (2017) suggest that with increased familiarity, task-related behaviors may become less influenced by external motivational factors and instead governed by ingrained, automated routines. The development of expertise also frequently

leads to cognitive simplification of complex tasks. As individuals gain familiarity, they tend to schematize their behaviors into automatic processes, reducing the reliance on deliberate motivational triggers (Earley et al., 1990). This cognitive simplification allows individuals to focus less on initiating behaviors and more on refining and adapting their routines as needed.
Taken together, these perspectives suggest that the role of motivation in data literacy behaviors may diminish with increasing expertise, as participants shift toward automated and habitual patterns of engagement.

5.4 Control Variables

The control variable analysis revealed statistically significant effects of educational level, degree pursuit, and age on specific items. These findings could help to explain how individual characteristics influence self-reported behaviors and competencies and offer insights for educational interventions. The negative effect of A-levels on F2F15 ("I check the qualifications of authors before relying on the information.") suggests that individuals with A-levels as their highest educational attainment are less inclined to evaluate an author's qualifications before trusting their information. This may reflect insufficient emphasis on critical evaluation skills at the A-level stage, which are typically developed more extensively in tertiary education. This observation underscores a potential gap in curricula, which could be addressed by integrating critical appraisal training earlier in the educational trajectory. Notably, some German schools have already reformed their curricula to emphasize media competency (Media Education Lab, 2020; Richter & Scheiter, 2023), potentially serving as a model for broader implementation.

For F3F2 ("I deal with information that challenges my views."), the analysis was inconclusive. Singularities in the data, likely caused by collinearity or insufficient variability in the predictor variables, limit interpretability. This finding highlights a methodological limitation, suggesting that the item F3F2 may lack robustness for statistical modeling in its current form. Future studies should consider refining or reformulating this item to enhance its validity.

A-levels exhibited a positive effect on F4F3 ("I can present data in graphics in such a way that they are understandable for different target groups."), indicating that A-level graduates report confidence in presenting data graphically to diverse audiences. This may reflect the inclusion of basic communication and data visualization skills in the A-level curriculum, underscoring the value of these foundational competencies at this stage of education.

The results for F4F8 ("I can use programs to create graphics to present results.") demonstrate a progressive relationship between educational attainment and proficiency with data visualization software. While A-levels are associated

with a modest increase (+2.82 points), higher educational levels such as a degree of a university of applied sciences, and a university degree, show substantially larger effects (+7.46 and +7.04 points, respectively). These results likely reflect the greater exposure to specialized tools and training in higher education. Intriguingly, currently pursuing a Master's degree was associated with a negative effect on F4F8 (-3.20 points), potentially indicating transitional challenges or curricular variability in graduate programs. In contrast, currently pursuing a state examination showed a positive effect (+2.60 points), suggesting that this pathway may place greater emphasis on developing these skills.

For F5F8 ("When I am confronted with extensive data sets, I can gain insights from them."), a statistically significant negative effect was observed for 'Age22' (-3.00 points), indicating lower self-reported confidence in extracting insights from large datasets among this age group. This result may reflect developmental factors, such as limited experience or confidence, or contextual influences unique to this stage. However, as no other items showed age-related effects, this finding should be interpreted cautiously. Additionally, while no significant educational variables were identified for F5F8, having a state examination as highest education, demonstrated a significant positive effect (+3.00 points). However, the absence of other degree categories makes a comparison impossible, warranting further investigation.

5.5 Limitations

The results of this study should be interpreted with several limitations in mind. The sample deviates from the general population in multiple demographic variables, potentially compromising its representativeness and generalizability. Occupational distribution among participants shows clustering in fields such as "Gesundheit, Soziales, Lehre und Erziehung", "Buchhaltung, Recht und Verwaltung", "Kaufmännische Dienstleistungen, Vertrieb, Tourismus" and especially "Naturwissenschaft, Geografie und Informatik". This indicates a selection bias, likely due to recruitment methods (who is reached) and implicitly favoring individuals more interested in data literacy.

The item pool for the questionnaire was specifically trained on this non-representative sample, which will likely affect its validity. Because it could result in the creation and measurement of a latent construct that is specific or unique even to this particular sample. Ideally, this approach would have been used with a sample representative of the general public or citizens. That way the results would be more valid and could more likely be generalized. A more representative sample might enhance the indicated effects, making them become clearer.

The visual analysis of item boxplots revealed several outliers in the data.

These outliers were retained in the sample for further investigation, as their clustering in the higher categories (4 and 5) suggests that the items may have been too easy for most respondents, resulting in high levels of agreement. This concentration of responses at the upper end of the scale indicates potential ceiling effects, where the items fail to adequately discriminate across a range of respondent abilities. Conversely, the outliers in the lower categories (0, 1, or 2) are relatively rare and unlikely to significantly influence the overall conclusions. However, their presence suggests that a small subset of participants either found the items unclear or disagreed with them, potentially pointing to systematic factors. The measure was designed with a general population in mind, which may have limited its ability to differentiate at higher levels of item difficulty, particularly among more data-literate participants. Addressing this limitation in future studies would improve the instrument's sensitivity to varying levels of data literacy. Also, as data literacy is a heterogeneous construct, this complicates global instrument development and understanding across all participants. Ideally the questionnaire would incorporate a broader content to better reflect the constructs full scope, thereby increasing content validity. However, a broader tool should be designed to maintain sufficient power to detect indicated factor-specific effects. Expanding the questionnaire with additional items could address this need, although it would deviate from the principle of parsimony.

These issues regarding the misfit of items and sample, as well as the heterogeneity of the construct further result in limitations regarding the reliability of the instrument. Although the McDonald's ω_{total} would indicate good reliability, the question of reliability stretches beyond a single measure for internal consistency. The reliability of a measure must be evaluated in relation to its target audience, as its validity and generalizability depend on how well the sample reflects the intended population. Therefore, it is crucial to determine which sample the measure is based on and the population it aims to represent.

In discussing the characteristics and demographics of the sample, the randomization of the data literacy items among respondents also introduced certain limitations. The primary limitation lies in the unequal demographic distribution across items. For example, while the overall sample has an average age of 40 years, this demographic balance may not hold for each individual item. This issue extends to other demographic variables as well, resulting in the diversity of the sample not being consistently reflected in the responses to individual items. Consequently, the selected items would require further and coherent testing of the item set as a whole on a diverse sample to evaluate how demographic variability influences responses. The limited demographic balance in the items also results in limited control for confounding variables in those variables. It furthermore

limits the informative value of comparisons regarding the demographic variables among the items.

Additionally, the training and testing data sets differed in size, which could influence measurement invariance testing (F. F. Chen, 2007). While the sample sizes were appropriate, they were at the lower threshold of the prior power analysis (Hu & Bentler, 1999; e.g., Kass & Tinsley, 1979), suggesting that larger samples might have been better. Also the 3-folding was on the lower threshold of power with the used sample size and possibly, a larger number of folds could have helped to increase the stability of the solution. Lastly, it should be noted that algorithm-based item selection is a heuristic approach, rather than deterministic, and may not always yield the optimal solution (Blum & Roli, 2003; Schultze, 2017). In this specific study, the results of the algorithm based selection with the genetic algorithm in 'stuart' appeared to be unstable across different runs. I tried to account for that via the k-folding and multiple iterations of the selection, but when the process of the data imputation was changed, from 'mice' to FIML, different items were selected, indicating unstable selections. This could potentially be accounted for, by using the bruteforce implementation in 'stuart'. But it could also be a hint at model misspecification, because of which the algorithm finds multiple local minima.

The models yielded rather bad fit measures, when using the initially planned objective function (with the aim of optimising for model fit, reliability, and variability in the difficulty of items. Therefore, it was changed in favor of model fit criteria exclusively. Thus, the final objective function that was used, did not incorporate terms to optimize for reliability McDonald's ω nor variability in the difficulty of items.

Furthermore WLSMV could not be used, although being the appropriate estimator for rating scales, because of the ceiling effects in the items and the scarcely used lower answer categories. I tried to account for that by collapsing the respective answer categories, but this seemingly resulted in an introduction of better fit to the model. Because of that, I decided to use MLR as estimator and were therefore able to use FIML as imputation method, implemented in 'lavaan'. Ideally future studies, with a sample without ceiling effects, would use WLSMV as estimator, since modified weighted least square estimators for ordered-categorical indicators (MWLS_C) provide accurate estimates of the model parameters given a stable weight matrix (Wirth & Edwards, 2007) and furthermore, WLSMV is a robust estimator which does not assume normally distributed variables and provides the best option for modelling categorical or ordered data (T. A. Brown, 2006).

5.6 Future directions

Future research in this domain should focus on expanding the qualitative development of items to ensure comprehensive coverage of the various dimensions of data literacy. A more extensive qualitative item creation process, informed by interviews with a diverse group of individuals, would contribute to a more nuanced understanding of the construct, particularly across different occupations as well as educational levels. Such qualitative insights could help identify representative behaviors and cognitive processes, forming a more solid foundation for item development. Additionally, these insights would contribute to the establishment of a more robust theoretical model for self-perceived data literacy. Expert interviews would be particularly valuable, integrating multiple perspectives on the construct and providing a more grounded theoretical base.

Future studies should consider assessing whether items in data literacy scales are sensitive to the distinction between routine and aspirational behaviors. Additionally, capturing the progression from motivated to routine actions could provide deeper insights into how expertise shapes self-perceived data literacy. These findings invite further refinement of the scale to better balance the measurement of skills, proficiencies, and the motivational aspects of data literacy. Future research should also consider whether the scale captures routine versus intentional behaviors, as this distinction may vary across samples and contexts.

This study clearly highlights that both the measurement model and the items of the scale require revision. Given that data literacy is a heterogeneous and multi-faceted construct, future research should rigorously explore the dimensionality of the scale to ensure its accuracy and utility. A thorough investigation of this dimensionality is essential for the validity of any future attempts at scale validation. A careful and thorough development of items is necessary to address the multi-faceted nature of the construct. This is particularly important for adapting the scale to measure individual skill levels accurately, which could be achieved through methods such as Item Response Theory (IRT). An IRT-based framework could significantly enhance the scale's precision by allowing for variability in item difficulty, thereby improving the adaptability and accuracy of the measurement. The introduction of adaptive testing, based on IRT principles, could lead to a more efficient scale, requiring fewer items while maintaining or even improving measurement precision.

5.7 Conclusion

The present study represents a first step towards addressing the gap in tools for measuring self-rated data literacy among the general public. The final instrument

demonstrated acceptable psychometric properties that could be approved upon, including strong internal consistency and promising criterion validity. While the results validate the scale's potential utility, challenges remain, particularly concerning multicollinearity among producer facets and limitations in the broader generalizability due to sample-specific characteristics. These issues necessitate further refinement of the scale and the respective measurement model. Replication studies would be necessary to establish its reliability and invariance across diverse populations. Future research should also explore the practical implications of enhancing data literacy in educational and societal frameworks, addressing the critical role of such competencies in an increasingly data-driven world. In conclusion, this thesis provides a foundational contribution to the emerging field of data literacy research. It lays a solid groundwork for future studies, emphasizing the need for ongoing refinement and broader validation efforts to fully realize the scale's potential in fostering data-literate societies.

References

Algner, M., & Lorenz, T. (2022). You're prettier when you smile: Construction and validation of a questionnaire to assess microaggressions against women in the workplace. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.809862

Association of College & Research Libraries. (2000). *Information literacy competency standards for higher education.* Brochure; American Library Association.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, thomson and bartlett. *British Journal of Mathematical and Statistical Psychology*, *62*, 569–582. https://doi.org/10.1348/000711008x365676

Bates, M. J. (2009). *An introduction to metatheories, theories, and models.* https://api.semanticscholar.org/CorpusID:62396882

Behm, T. (2018). *SWE-IV-16: Skala zur erfassung der informationsverhaltensbezogenen selbstwirksamkeitserwartung [verfahrensdokumentation, fragebogen deutsche und englische version (SES-IB-16)]* [Open Test Archive]. Leibniz-Institut für Psychologie (ZPID). https://doi.org/10.23668/psycharchives.4598

Beißert, H., Köhler, M., Rempel, M., & Beierlein, C. (2015). Deutschsprachige kurzskala zur messung des konstrukts need for cognition NFC-k. *Zusammenstellung Sozialwissenschaftlicher Items Und Skalen (ZIS)*. https://doi.org/10.6102/zis230

Bless, H., Wänke, M., Bohner, G., Fellhauer, R., & Schwarz, N. (1994). Need for cognition: Eine skala zur erfassung von engagement und freude bei denkaufgaben [presentation and validation of a german version of the need for cog-

nition scale]. *Zeitschrift Für Sozialpsychologie*, *25*, 147–154.

Blum, C., & Roli, A. (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surv.*, *35*(3), 268–308. https://doi.org/10.1145/937503.937505

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education*, *53*(4), 1207–1217. https://doi.org/https://doi.org/10.1016/j.compedu.2009.06.004

Brandmaier, A. M., & Peikert, A. (2024). Automated reproducibility testing in r markdown. *Preprint*.

Brillouin, L. (1953). Negentropy principle of information. *Journal of Applied Physics*, *24*(9), 1152–1163.

Brown, C., & Krumholz, L. R. (2002). Integrating information literacy into the science curriculum. *College & Research Libraries*, *63*, 111–123. https://doi.org/10.5860/crl.63.2.111

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Bühner, M. (2021). *Einführung in die test- und fragebogenkonstruktion* (4., korrigierte und erweiterte Auflage, p. 752). Hogrefe Verlag.

Bundesargentur für Arbeit. (2024). *Arbeitslosenzahl in deutschland im jahresdurchschnitt von 2005 bis 2024*. https://de.statista.com/statistik/daten/studie/1223/umfrage/arbeitslosenzahl-in-deutschland-jahresdurchschnittswerte/#:~:text=Im%20Monat%20Juni%202024%20waren,um%20rund%20178.800%20Personen%20höher; Statista.

Buss, D., & Craik, K. (1983). The act frequency approach to personality. *Psychological Review*, *90*, 105–126. https://doi.org/10.1037/0033-295X.90.2.105

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Callingham, R. (2006). Assessing statistical literacy: A question of interpretation? *International Conference on Teaching Statistics (ICOTS7)*.

Calzada-Prado, F. J., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, *63*(2), 123–134. https://doi.org/10.1515/libri-2013-0010

Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2014). Determining data information literacy needs. In J. Carlson & M. Johnston (Eds.), *Data information literacy: Librarians, data, and the education of a new generation of researchers*. Purdue University Press.

Carmi, E., Yates, S. J., Lockley, E., & Pawluczuk, A. (2020). Data citizenship: Re-

thinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1481

Carroll, J. B. (1993). *Human abilities: Their nature and measurement*. Cambridge University Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F., Cui, Y., Lutsyk-King, A., Gao, Y., Liu, X., Cutumisu, M., & Leighton, J. P. (2024). Validating a novel digital performance-based assessment of data literacy: Psychometric and eye-tracking analyses. *Education and Information Technologies*, *29*(8), 9417–9444. https://doi.org/10.1007/s10639-023-12177-7

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cohen, A. R., Stotland, E., & Wolfe, D. M. (1955). An experimental investigation of need for cognition. *The Journal of Abnormal and Social Psychology*, *51*(2), 291–294. https://doi.org/10.1037/h0042761

Costa, P. T., & McCrae, R. R. (1992). The five-factor model of personality and its relevance to personality disorders. *Journal of Personality Disorders*, *6*(4), 343–359. https://doi.org/10.1521/pedi.1992.6.4.343

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Cui, Y., Chen, F., Lutsyk, A., Leighton, J., & Cutumisu, M. (2023). Data literacy assessments: A systematic literature review. *Assessment in Education: Principles, Policy & Practice*, *30*, 1–21. https://doi.org/10.1080/0969594X.2023.2182737

Deahl, E. (2014). *Better the data you know: Developing youth data literacy in schools and informal learning environments* [M.S. Thesis]. Massachusetts Institute of Technology.

Dweck, C., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256–273. https://doi.org/10.1037/0033-295X.95.2.256

Earley, P. C., Lee, C., & Hanson, L. A. (1990). Joint moderating effects of job experience and task component complexity: Relations among goal setting, task strategies, and performance. *Journal of Organizational Behavior*, *11*(1), 3–15. https://doi.org/10.1002/job.4030110104

Elander, J. (2004). Student assessment from a psychological perspective. *Psychology Learning & Teaching*, *3*(2), 114–121. https://doi.org/10.2304/plat.2003.3.2.114

Fowler, Jr., F. J. (1995). *Improving survey questions: Design and evaluation.* Sage Publications, Inc.

Frank, M., Walker, J., Attard, J., & Tygel, A. (2016). Data literacy: What is it and how can we make it happen? editorial. *The Journal of Community Informatics*, *12*(3), 4–8.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review / Revue Internationale de Statistique*, *70*(1), 1–25. https://doi.org/10.2307/1403713

Galán, S. F., Mengshoel, O. J., & Pinter, R. (2013). A novel mating approach for genetic algorithms. *Evolutionary Computation*, *21*(2), 197–229. https://doi.org/10.1162/EVCO_a_00067

Goretzko, D., Siemund, K., & Sterner, P. (2023). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement*, *84*(1), 123–144. https://doi.org/10.1177/00131644231163813

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Green, F. (1998). *The value of skills* (Department of Economics Discussion Paper No. 9819). University of Kent, Department of Economics.

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications.* Pearson Education India.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence.* The MIT Press. https://doi.org/10.7551/mitpress/1090.001.0001

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology*, *123*(3), 207–215.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630.

John, O., Naumann, L., & Soto, C. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of Personality: Theory and Research, 3rd Edn.* (pp. 114–158).

Jong, J. P. d., & Fodor, O. C. (2017). Attuning to individual work routines and team performance. *Team Performance Management*, *23*(7/8), 385–406. https://doi.org/10.1108/tpm-01-2017-0001

Kass, R. A., & Tinsley, H. E. A. (1979). Factor analysis. *Journal of Leisure Research*, *11*(2), 120–138. https://doi.org/10.1080/00222216.1979.11969385

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93. http://www.jstor.org/stable/2332226

Kerber, A., Schultze, M., Müller, S., Rühling, R. M., Wright, A. G. C., Spitzer, C., Krueger, R. F., Knaevelsrud, C., & Zimmermann, J. (2022). Development of a short and ICD-11 compatible measure for DSM-5 maladaptive personality traits using ant colony optimization algorithms. *Assessment*, *29*(3), 467–487. https://doi.org/10.1177/1073191120971848

Keshavarz, H. (2021). *Personality factors and knowledge sharing behavior in information services: The mediating role of information literacy competencies*. https://doi.org/10.1108/VJIKMS-05-2020-0095

Koltay, T. (2017). Information overload in a data-intensive world. In A. J. Schuster (Ed.), *Understanding information: From the big bang to big data* (pp. 197–217). Springer International Publishing. https://doi.org/10.1007/978-3-319-59090-5_10

Komarraju, M., Karau, S. J., & Schmeck, R. R. (2009). Role of the big five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences*, *19*(1), 47–52. https://doi.org/10.1016/j.lindif.2008.07.001

Komarraju, M., Karau, S. J., Schmeck, R. R., & Advic, A. (2011). The big five personality traits, learning styles, and academic achievement. *Personality and Individual Differences*, *51*(4), 472–477. https://doi.org/10.1016/j.paid.2011.04.019

Kõuts-Klemm, R. (2019). Data literacy among journalists: A skills-assessment based approach. *Central European Journal of Communication*, *12*(24), 299–315. https://doi.org/10.19195/1899-5101.12.3(24).2

Leighton, J. P., Cui, Y., & Cutumisu, M. (2021). Key information processes for thinking critically in data-rich environments. *Frontiers in Education*, *6*. https://doi.org/10.3389/feduc.2021.561847

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*, 411–431. https://doi.org/10.1080/00273170802285743

Lusiyana, A., Festiyed, F., & Yulkifli, Y. (2020). Measuring the physics students' data literacy skill in the era of industry 4.0 by using mirecal learning model. *International Journal of Scientific and Technology Research*, *9*(1), 1203–1205. https://api.semanticscholar.org/CorpusID:216552579

Mahmood, M., Frolova, Y., & Gupta, B. (2021). The HEXACO, academic motivation and learning approaches: Evidence from a central asian country. *Education + Training*, *63*(6), 920–938. https://doi.org/10.1108/ET-11-2019-0257

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory

factor analysis models. *Psychological Methods*, *28*(1), 61–88. https://doi.org/10.1037/met0000425

Media Education Lab. (2020). *Medialogues on propaganda: Final report.* Media Education Lab, University of Rhode Island; Media Education; Educational Technology Lab, University of Wuerzburg; Media Literacy Now. https://mediaeducationlab.com/sites/default/files/FINAL%20REPORT%2C%20Medialogues%20on%20Propaganda.pdf

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review*, *80*(4), 252–283. https://doi.org/10.1037/H0035002

Moshagen, M., & Bader, M. (2024). semPower: General power analysis for structural equation models. *Behavior Research Methods*, *56*, 2901–2922. https://doi.org/10.3758/s13428-023-02254-7

Norman, D. A., & Rumelhart, D. E. (1975). Memory and knowledge. In D. A. Norman, D. E. Rumelhart, & the LNR Research Group (Eds.), *Explorations in cognition.* Freeman.

Olaru, G., & Danner, D. (2021). Developing cross-cultural short scales using ant colony optimization. *Assessment*, *28*(1), 199–210. https://doi.org/10.1177/1073191120918026

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale big five assessments. *Journal of Research in Personality*, *59*, 56–68. https://doi.org/10.1016/j.jrp.2015.09.001

Paletz, S. B. F., Kim, K. H., Schunn, C. D., Tollinger, I., & Vera, A. (2013). Reuse and recycle: The development of adaptive expertise, routine expertise, and novelty in a large research team. *Applied Cognitive Psychology*, *27*(4), 415–428. https://doi.org/10.1002/acp.2928

Payan Carreira, R., Sacau-Fontenla, A., Rebelo, H., Sebastião, L., & Pnevmatikos, D. (2022). Development and validation of a critical thinking assessment-scale short form. *Education Sciences*, *12*, 938. https://doi.org/10.3390/educsci12120938

Peikert, A., Van Lissa, C. J., & Brandmaier, A. M. (2021). *Reproducible research in r: A tutorial on how to do the same thing more than once.* https://doi.org/10.31234/osf.io/fwxs4

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2014). Big five inventory (BFI-10). *Zusammenstellung Sozialwissenschaftlicher Items*

Und Skalen (ZIS). https://doi.org/10.6102/zis76

Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking. *Assessment & Evaluation in Higher Education*, *44*, 664–675.

Remmert, N., Schmidt, K. M. B., Mussel, P., Hagel, M. L., & Eid, M. (2022). The berlin misophonia questionnaire revised (BMQ-r): Development and validation of a symptom-oriented diagnostical instrument for the measurement of misophonia. *PLOS ONE*, *17*, 1–27. https://doi.org/10.1371/journal.pone.0269428

Richter, D., & Scheiter, K. (2023). *Kompetenzverbund lernen.digital*. Bertelsmann Stiftung. https://schule21.blog/2023/10/11/kompetenzverbund-lernendigital/

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D. R., Matwin, S., & Wuetherick, B. (2015). *Strategies and best practices for data literacy education knowledge synthesis report*. https://doi.org/10.13140/RG.2.1.1922.5044

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, *20*(3), 643–671. https://doi.org/10.1177/014920639402000307

Roetzel, P. G. (2019). Information overload in the information age: A review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research*, *12*(2), 479–522. https://doi.org/10.1007/s40685-018-0069-z

Ross, M. E., Blackburn, M., & Forbes, S. (2005). Reliability generalization of the patterns of adaptive learning survey goal orientation scales. *Educational and Psychological Measurement*, *65*(3), 451–464. https://doi.org/10.1177/0013164404272496

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

Saleh, S., Ashari, Z. M., & Kosnin, A. M. (2018). Personality traits and intrinsic motivation on academic performance. *International Journal of Engineering & Technology*, *7*(4.28), 317–322. https://doi.org/10.14419/ijet.v7i4.28.22607

Sandy, C. J., Gosling, S. D., & Koelkebeck, T. (2014). Psychometric comparison of automated versus rational methods of scale abbreviation: An illustration using a brief measure of values. *Journal of Individual Differences*, *35*, 221–235. https://doi.org/10.1027/1614-0001/a000144

Schauffel, N., Schmidt, I., Peiffer, H., & Ellwart, T. (2021). ICT self-concept scale (ICT-SC25). *Zusammenstellung Sozialwissenschaftlicher Items Und Skalen (ZIS)*. https://doi.org/10.6102/zis308_exz

Schmalbach, B., Irmer, J. P., & Schultze, M. (2019). *Fit measure cutoffs in SEM*.

https://doi.org/10.1080/10705519909540118

Schneider, J., Striebing, C., Hochfeld, K., & Lorenz, T. (2024). Establishing circularity: Development and validation of the circular work value scale (CWVS). *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1296282

Schneider, R. (2013). Research data literacy. *Worldwide Commonalities and Challenges in Information Literacy Research and Practice*, *397*, 134–140. https://doi.org/10.1007/978-3-319-03919-0_16

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLOS ONE*, *11*(11), 1–19. https://doi.org/10.1371/journal.pone.0167110

Schüller, K. (2020). *Future skills: A framework for data literacy* (Working Paper No. 53). Hochschulforum Digitalisierung. https://doi.org/10.5281/zenodo.3946067

Schultze, M. (2017). *Constructing subtests using ant colony optimization* [Doctoral dissertation]. Freie Universität Berlin.

Schultze, M. (2022). *Stuart: Subtests using algorithmic rummaging techniques*.

Schultze, M., & Lorenz, T. (2024). *I choo-choo-choose you: A tutorial on automated item selection in scale construction*. https://doi.org/10.31234/osf.io/pkm3q

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shields, M. (2005). Information literacy, statistical literacy, data literacy. *IASSIST Quarterly*, *28*(2–3), 6. https://doi.org/10.29173/iq790

Statistisches Bundesamt. (2023a). *Durchschnittsalter der bevölkerung in deutschland nach geschlecht von 2011 bis 2022*. https://de.statista.com/statistik/daten/studie/1084446/umfrage/durchschnittsalter-der-bevoelkerung-in-deutschland-nach-geschlecht/; Statista.

Statistisches Bundesamt. (2023b). *Leichter rückgang: Vollzeitbeschäftigte arbeiteten 2022 durchschnittlich 40,0 wochenstunden*. [Pressemitteilung Nr. N047 vom 28. August 2023]. https://www.destatis.de/DE/Presse/Pressemitteilungen/2023/08/PD23_N047_13.html

Statistisches Bundesamt. (2024a). *Anzahl der studierenden an hochschulen in deutschland in den wintersemestern von 2002/2003 bis 2023/2024*. https://de.statista.com/statistik/daten/studie/221/umfrage/anzahl-der-studenten-an-deutschen-hochschulen/; Statista.

Statistisches Bundesamt. (2024b). *Bevölkerung nach dem gebietsstand und durchschnittsalter 1990 bis 2023*. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Tabellen/bevoelkerungsstand-gebietsstand-werte.html.

Statistisches Bundesamt. (2024c). *Bevölkerung nach nationalität und*

geschlecht 1970 bis 2023 in deutschland. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Tabellen/deutsche-nichtdeutsche-bevoelkerung-nach-geschlecht-deutschland.

Toegel, G., & Barsoux, J.-L. (2012). How to become a better leader. *MIT Sloan Management Review*, *53*, 51–60.

Tomar, M. (2023). Assessing information literacy programs in academic libraries: A comprehensive review. *International Journal of Information Studies*, *15*, 108–118. https://doi.org/10.6025/ijis/2023/15/4/108-118

Vahey, P., Yarnall, L., Patton, C., Zalles, D., & Swan, K. (2006). Mathematizing middle school: Results from a cross-disciplinary study of data literacy. *Annual Meeting of the American Educational Research Association*.

Webber, S. A., & Johnston, B. (2017). Information literacy: Conceptions, context and the formation of a discipline. *Journal of Information Literacy*, *11*. https://doi.org/10.11645/11.1.2205

Williams, S., Deahl, E., Rubel, L., & Lim, V. (2014). City digits: Local lotto: Developing youth data literacy by investigating the lottery. *Journal of Digital Media Literacy*.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Wolff, A., Gooch, D., Montaner, J. J. C., 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. https://doi.org/10.15353/joci.v12i3.3275

Zhao, Y., Pinto Llorente, A. M., & Sánchez Gómez, M. C. (2021). Digital competence in higher education research: A systematic literature review. *Computers & Education*, *168*, 104212. https://doi.org/10.1016/j.compedu.2021.104212