DATA LITERACY COMPETENCIES OF STUDENTS AT TU DORTMUND UNIVERSITY

Henrike Weinert, Alina Künne and Katja Ickstadt
TU Dortmund University, Germany
henrike.weinert@tu-dortmund.de

The Technical University of Dortmund (TU Dortmund) aims to strengthen the data and statistical literacy of all its students. To this end, the Data Competence Network (DaCoNet) programme has been set up to provide a wide range of courses to help students improve their competences in this area. In order to better assess the status of the competences, a questionnaire was developed. It consists of a self-assessment part of the data literacy competences and an associated performance test. We present the developed questionnaire as well as the results of the first survey on data literacy competences of our students using this instrument. We show the influence of background variables such as study programme, study progress or own interest on the results of the self-assessment and the performance test. From the results we derive recommendations for action in statistics and data literacy education.

INTRODUCTION

Data have become an integral part of professional, private and public life. Therefore, the issue of data and statistical literacy is becoming increasingly important in teaching and research. Researchers and teachers agree that data literacy is considered a key competence of the 21st century (Ridsdale et al., 2015; Schüller, 2020). It is therefore important that, in addition to data professionals (e.g. statisticians or data scientists), all other people are given the opportunity to learn data literacy (Knaus, 2020). Data not only influence our daily lives, but also play an important role in finding solutions to both simple and complex problems (Beaulieu & Leonelli, 2022). For this reason, more and more interdisciplinary teaching formats in the field of data literacy are emerging at German universities.

At the Technical University of Dortmund (TU Dortmund), the concept of the Data Competence Network (DaCoNet) has provided the opportunity to take online courses in data literacy. It is based at the TU Dortmund Centre for Data Science and Simulation (DoDaS), an interdisciplinary centre for data science and scientific computing that combines research and teaching in this field. At the TU Dortmund, there are both strong technical-scientific subjects, which traditionally have a strong connection to data, and strong social-humanities subjects, which increasingly need at least basic data skills in a datafied world. The DoDaS offers a perfect basis for training data skills at different levels.

With our Data Literacy programme we are offering new courses and hope to improve existing ones, especially statistics courses for non-statisticians. We hope to achieve cross-fertilisation by linking with existing courses in various disciplines. Initially, courses will be set up at a basic level. Small online units cover statistical basics such as data types, key figures and graphics, and how to recognise applications with errors: What can go wrong? Where are the pitfalls? Software (R or Python), data ethics and critical thinking, research data management (RDM), representativeness, the idea of inferential statistics. There is a strong emphasis on critical reflection and, as an interdisciplinary course, on communication across disciplines.

Since a way of measuring students' competencies in data literacy has been lacking up to now, a questionnaire has been developed as a first step in this direction as part of a Master's thesis.

SURVEY INSTRUMENT

Data literacy encompasses more than one set of competencies, which are presented in Ridsdale et al.'s (2015) matrix. The intention of this matrix is to provide a basis for emerging research in data literacy and to set a standard for the assessment and evaluation of data literacy competencies. Five key skills are identified: *Conceptual framework* or *data foundation* (Bonikowska et al., 2019), *data collection, data management, data evaluation* and *data application*. In order to have a first survey instrument within DaCoNet that measures students' data literacy competences at a low level, a competence framework is used here that is based on this matrix, but does not cover all sub-competences in order to keep the survey time manageable. Aiming at the basics of data literacy, the competency framework used here first covers *data foundation*. This is followed by the basic

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada. ©2023 ISI/IASE http://iase-web.org

competences from the key competence *data evaluation*, with a focus on the sub-competence *data interpretation*, because of the importance of correctly interpreting tables, graphs and charts. Finally, the *critical thinking* and *data ethics* competencies from the *data application* domain are assessed.

At the time of the research, there was no scientifically proven test to measure data literacy. Bonikowska et al. (2019) present a number of test procedures designed to capture an individual's data literacy, including the self-assessment test 'My Databilities 2.0' by the private company 'Data To The People' (2022a & 2022b). This is strongly based on Ridsdale's competency matrix and contains statements for the various sub-competences, ranging from low to high competency in the sub-area, with which one can assess one's own abilities, e.g. "With guidance, I can understand the importance of data" to "I can help others to understand the importance of data". For this work a two-step approach is used for these sub-competences: a self-assessment test and a performance test.

Self-Assessment test

A self-assessment test can provide consistent and therefore reliable results (Fitzgerald et al., 2000). This is supplemented by a performance test in order to better assess the students' abilities and to check how well they match the self-assessment.

Based on the My Databilities self-assessment test, four to six sub-items were developed for each sub-competence, each of which was rated on a five-point Likert scale. As part of the test development, the think aloud pre-test (n=21) showed that the Likert scale was more appropriate than the statements from My Databilities 2.0, so these were used for the main survey. Our self-assessment test items are shown in Table 1.

Table 1: Items of self-assessment test (translated from German):
five-point Likert scale from *does not apply at all* to *fully applies*

Data Foundation	
f1	I can understand the importance of data (e.g. why purchase behaviour data is important for product placement).
f2	I contribute to an environment that encourages the use of data (e.g. study, work).
f3	I know the difference between data, information and knowledge.
f4	I can use and recommend appropriate software when dealing with data (e.g. Excel, R).
Data Evaluation	
e1	I can read and understand tables, charts and graphs.
e2	I can produce tables, charts and graphs (e.g. using appropriate software).
e3	I can describe which type of analysis is appropriate (e.g. whether the median or the arithmetic mean is more appropriate).
e4	I can carry out a simple data analysis (e.g. calculate median, arithmetic mean with appropriate software).
e5	I can use the data provided to me to support my decision-making process (e.g. wear a mask depending on the corona incidence).
e6	I can take data-based action (e.g. derive a recommendation for action from the results of a data analysis).
Data Application	
a1	I can recognize legal problems related to data (e.g. whether a data set has been successfully anonymized).
a2	I can identify ethical issues related to data (e.g. whether a dataset contains sensitive data).
a3	I can work with data in a legally correct way (e.g. anonymize data independently).
a4	I can publish data in a legally correct way.
a5	I am prepared for the fact that data analysis requires dealing with ethical questions (Keyword: critical thinking).

An exploratory factor analysis (EFA) was carried out to provide initial indications of the quality of the test. The results confirm that the self-test measures three latent factors. The latent trait *data application* is clustered into one latent factor (factor loadings ranging from .5 to .8). The other two factors are a mixture of *data foundation* and *data evaluation*. One factor combines the items e2, e3, e4 (loadings of .7 and .8) with f4 (.5) and thus all the items related to software. Since dealing with data in principle always requires dealing with software and is less part of the general understanding of data, f4 could also be located in the area of *data evaluation*. The associated factor therefore represents the ability to evaluate data. On the other hand, another factor combines items e1, e5, e6 (loadings .5 to .7), f1, f1 and f3 (.3 to .5) and thus mainly those that deal with understanding data and using data products without working with the data independently. It is therefore difficult to make a precise distinction between *data foundation* and *data evaluation*. The three latent factors have correlations of 0.4 and 0.5 with each other. Cronbach's $\alpha = .66$ for the four *data foundation* items, $\alpha = .82$ for the six *data evaluation* items and $\alpha = .84$ for the five *data application* items, which could not be increased by reducing the number of items.

For further questions, we consider the total score S of the self-assessment, which is the mean of all 15 items and is on a scale from 1 (low assessment of one's own abilities) to 5 (very high assessment).

Performance Test

In the second part of the survey, students' data literacy competences are assessed by means of a performance test. Since this is a basic understanding of data literacy and simple feasibility is to be ensured, single and multiple choice questions are suitable (Bonikowska et al., 2019; Schüller 2020). The test focuses on the basics of the respective sub-competences, i.e. a low level. On the one hand, items are used here that have already been used in courses to test knowledge. These are simple comprehension and knowledge questions that capture the competences to be measured and test the level assumed in this paper, and on the other hand, items from Schüller et al. (2019) are used. Table 2 gives an overview of the items of this performance test. The total score P of the performance test is a weighted normalised mean of the individual items P1 to P8, on a scale from 0 (no points in the test) to 1 (full score).

Table 2: Items of Performance test (translated from German, always the possibility to tick "don't know")

Data Foundation - Introduction to data	
P1	Decide whether the example is data, information or knowledge (four examples)
P2	Assign the correct scale level (nominal scale, ordinal scale, ratio scale) to the following characteristics (two characteristics)
Data Evaluation - Data Interpretation	
P3	Interpretation of a given bar chart (multiple choice for five statements)
P4	Distribution form (six histograms): decide for which form the arithmetic mean gives a misleading value (multiple choice for six statements)
P5	Assignment of which information about data is needed to be able to make certain statements. Item from Schüller et al. (2019), p. 51, example 6.3.1
Data Application - Data Ethics	
P6	Decide whether the example is one for anonymised data (multiple choice for four statements)
Data Application - Critical thinking	
P7	Decide whether statements on the report on studies are true or false (five statements)
P8	Decide whether statements on ethical difficulties in statistical hypothesis testing are true or false (five statements)

The survey also collected background variables, including field of study, level and semester of study, and interest in data, the latter measured on a five-point scale from low to very high. The questionnaire was tested in the main survey by means of an online survey. The survey was clicked on 799 times, resulting in a non-representative, non-random sample in which n = 496 students provided complete information (62%). The survey took a median of 11 minutes (mean 14 minutes) to complete, meeting the aim of creating a short survey that would also be completed voluntarily. Students from all faculties are represented, although in very different numbers. For this reason, the summary of the results makes only a rough distinction between traditionally more data-oriented subjects, in this case STEM, and less data-oriented subjects in the humanities and social sciences. For the progression of studies, a distinction is made between beginners (1-2 semesters) and advanced students among Bachelor's students, and Master's students are considered separately.

RESULTS

From the individual items we calculated total scores for the self-test *S* and the performance test *P*, as well as scores related to the sub-competences Data Foundation (*SF* and *PF*), Data Application (*SA* and *PA*) and Data Evaluation (*SE* and *PE*). The results are shown in Table 3, also subdivided according to the classification described above.

Table 3: Results form self-assesment *(S)* and performance test *(P)*

	whole sample (n=496)		more data-oriented (n=297)		less data-oriented (n=199)		BA beginners (n=127)		BA advanced (n=239)		MA (n=107)	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
SF	3.79	0.72	3.94	0.68	3.58	0.72	3.62	0.67	3.80	0.70	3.89	0.79
SE	4.02	0.70	4.14	0.63	3.85	0.77	3.77	0.69	4.04	0.68	4.17	0.68
SA	3.06	0.88	3.09	0.83	3.01	0.94	2.90	0.78	3.04	0.84	3.14	1.00
S	3.62	0.63	3.72	0.57	3.48	0.68	3.43	0.56	3.63	0.61	3.73	0.67
PF	0.58	0.27	0.59	0.27	0.56	0.26	0.50	0.26	0.57	0.26	0.67	0.26
PE	0.62	0.18	0.65	0.16	0.57	0.19	0.59	0.17	0.61	0.18	0.65	0.17
PA	0.66	0.17	0.66	0.16	0.66	0.19	0.63	0.17	0.66	0.17	0.70	0.18
P	0.62	0.14	0.64	0.13	0.60	0.15	0.58	0.14	0.62	0.14	0.67	0.14
interest	3.59	0.92	3.81	0.85	3.26	0.92	3.54	0.9	3.55	0.82	3.64	1.08

On the basis of the data, it was possible to investigate questions H1 to H4 below, although due to our (non-representative) sample and the rough division into groups, we can only evaluate them descriptively.

- *(H1) Is there an influence of the self-assessment on the performance test?*

For our sample, the linear model shows a positive correlation between the results of the self-assessment (S) and those of the achievement test *(P): $P = .55{\cdot}S + 5.011$* (r^2=.17). This confirms that students are quite good at self-assessing their abilities and that such a test is suitable for assessing students' abilities before a course starts and for adapting the content of the course accordingly.

- *(H2) How does progress in studies affect the self-assessment and achievement test?*

On average, progress in studies has a positive effect on both the self-assessment of the data competences and the performance test. For all sub-competences, the lowest mean value was found among Bachelor beginners (1-2 semesters) than among Bachelor students in higher semesters and again the highest mean among Master's students, cf. Table 3 (d=.34 between BA inexperienced and experienced in the self-test, d=.26 between BA inexperienced and experienced in the performance test; d=.54 between BA and MA in the performance test).

- *(H3) Do students from less data-oriented subjects tend to have poorer self-assessment and performance than students from more data-oriented subjects?*

Students from subjects close to data rate their competences higher than those from subjects far from data and also tend to have poorer test results, although here the differences are not as clear-cut as one might expect, cf. Table 3. (d=.39 for the self-test and d=.31 for the performance test). This shows that STEM subjects have a higher overlap with data literacy content than humanities/social sciences subjects. A more fine-grained investigation would be desirable here, as the division into data-related and data-distant is very rough and it was also not recorded to what extent the students had already had introductory courses in statistics or other subjects that might be important for data literacy. Unsurprisingly, statistics students, for example, had high scores in the self-assessment (n=46, mean= 4.04, sd=0.56) and the performance test (n=46, mean= .70, sd=.11), but so did students in the Faculty of Arts and Humanities, which includes Journalism (n=35, mean =3.54, sd=.59 in the self-assessment;

n=35,mean=.68, sd=.12 in the performance test). The reason for this could be that many of these students have taken the course "Statistics for Journalists", which is a pilot within the framework of DaCoNet to modernise introductory statistics courses towards more data literacy and is being adapted for this purpose (but the data do not allow to investigate this).

- *(H4) Does a higher interest in data lead to a better performance test?*

Interest *(I)* in data, measured on a scale from 1 (low interest) to 5 (very high interest), has a positive effect on performance test scores *P*: $P=.97\cdot I + 7.46$ (r^2=.13). A more extensive investigation would be desirable here, as interest may also have an effect on individual key competences, e.g. a high interest in ethics may have a particular effect on *data application* and less influence on the other competences.

The results show on the one hand that hardly any students have no data literacy at all, but also that hardly any students performed extremely well in all areas (e.g. no one achieved the highest score in the achievement test).

CONCLUSION

A questionnaire was designed to assess students' data literacy skills and tested with a pre-test and an initial voluntary online survey. Students first completed a self-assessment test, and then basic data literacy was measured more objectively through a performance test, keeping the overall length of the questionnaire as short as possible.

The results show that the students in the sample are able to assess themselves well in terms of their data literacy and at the same time achieve an acceptable performance on average in the achievement test, i.e. about half of the points. As the performance test only measures a basic knowledge of data literacy, these results show that there is a need for teaching in the area of data literacy. The results suggest that the factors investigated - self-assessment, field of study (data-related or not), previous study experience and interest in data - influence performance in the achievement test. In particular, student experience seems to have a strong influence on data literacy. The classification into near and far from data seems to have less influence than expected.

It should be noted that the sample is not representative, that the new questionnaire can only capture a fraction of data literacy, and that the classification into near and far data literacy is very rough. Further research is needed to confirm and at best generalise the results, but a basis for the use of a questionnaire is now available. This can be used especially in courses on data literacy or statistical literacy to better assess the field of participants and to tailor the teaching accordingly. As the questionnaire is deliberately kept short, it can also be used for voluntary surveys outside of courses, as it usually takes no more than 11 minutes to complete.

In summary, data literacy is important in many areas and the results of this work contribute to a small but growing area of scientific research.

ACKNOWLEDGEMENTS

This work was supported by the Stifterverband and the state of North Rhine-Westphalia through the *data literacy education.nrw* funding line.

REFERENCES

Beaulieu, A. & Leonelli, S. (2022). *Data and Society: A Critical Introduction.* SAGE Publications Sage CA: Los Angeles, CA.

Bonikowska, A., Sanmartin, C. & Frenette, M. (2019). *Data Literacy: What It Is and How to Measure It in the Public Service.* Analytical Studies: Methods and Referneces (22).

Fitzgerald, J. T., Gruppen, L. D. & White, C. B. (2000). *The Influence of Task Formats on the Accuracy of Medical Students' Self-Assessments.* Academic Medicine, 75(7), 737–741.

Knaus, T. (2020). *Technology Criticism and Data Literacy: The Case for an Augmented Understanding of Media Literacy.* Journal of Media Literacy Education, 12(3), 6–16. https://doi.org/10.23860/JMLE-2020-12-3-2

Ridsdale, C., Rothwell, J., Smit, M., Bliemel, M. Irvine, D., Kelley, D., Matwin, S., Wuetherick, B. & Ali-Hassan, H. (2015). *Strategies and Best Practices for Data Literacy Education*. Knowledge Synthesis Report, DOI:10.13140/RG.2.1.1922.5044.

Schüller, K., Busch, P. & Hindinger, C. (2019). *Future Skills: Ein Framework für Data Literacy – Kompetenzrahmen und Forschungsbericht.* Arbeitspapier Nr. 47. Berlin: Hochschulforum Digitalisierung. DOI: 10.5281/zenodo.3349865

Schüller, K. (2020). *Future Skills: A Framework for Data Literacy*. Working Paper No. 53. Berlin: Hochschulforum Digitalisierung.