Towards Benchmarking Data Literacy

Calum Inverarity
Open Data Institute
London, United Kingdom
calum.inverarity@theodi.org

Dr David Tarrant
Open Data Institute
London, United Kingdom
david.tarrant@theodi.org

Emilie Forrest
Open Data Institute
London, United Kingdom
emilie.forrest@theodi.org

Phil Greenwood
Glacis
Bath, United Kingdom
phil.greenwood@glacis.co.uk

ABSTRACT

Data literacy as a term is growing in presence in society. Until recently, most of the educational focus has been around how to equip people with the skills to use data. However the increased impact that data is having on society has demonstrated the need for a different approach, one where people are able to understand and think critically about how data is being collected, used and shared. Going beyond definitions, in this paper we present research on benchmarking data literacy through self assessment based upon the creation of a set of data literacy levels for adults. Although the work highlights the limitations of self assessment, there is clear potential to build on the definitions to create potential IQ-style tests that help boost critical thinking and demonstrate the importance of data literacy education.

CCS CONCEPTS

• **Applied computing** → **Education**.

KEYWORDS

data literacy, competency model, competencies, data skills, self-assessment, benchmarking, assessment

ACM Reference Format:
Calum Inverarity, Dr David Tarrant, Emilie Forrest, and Phil Greenwood. 2022. Towards Benchmarking Data Literacy. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487553.3524695

1 INTRODUCTION

The Open Data Institute (ODI), as an institute, has always promoted the importance of data literacy and skills in society. In this paper we present some of our recent work in the area of data literacy; and our initial explorations into tools and techniques that can help organisations understand – and potentially benchmark – current capability. Section 2 sets the scene for the paper by discussing and considering the various interpretations of data literacy in academia, the public sector and business. In section 3 we discuss why data literacy is so critical, drawing attention to current gaps and provisions to fill these. Section 4 introduces the ODI's interpretation of data literacy, our key focus on comprehension in addition to skills, and how we visualise the balance needed between practical skills and comprehension with the Data Skills Framework. In section 5, we examine current approaches to benchmarking data literacy and present an alternative model and methodology and discuss our findings. We consider how to reduce the subjectivity from benchmarking data literacy in section 6 by exploring how IQ tests and other alternative questioning techniques can be applied to this area. We then conclude in section 7 and discuss next steps in both benchmarking and improving data literacy in society.

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04.
https://doi.org/10.1145/3487553.3524695

2 DEFINING DATA LITERACY

The definition of data literacy has been the subject of considerable academic discussion, particularly within the past decade [12–14, 24, 34]. In parallel, spurred by the increasing importance of data-related capabilities within contemporary work forces, a similarly significant level of discussion has taken place within public administration and business circles as to what it means for a person, business or society to be data literate. Inevitably, there are points of convergence and divergence among these various definitions as a result of the motivations and priorities of these different communities. Consequently, these serve as the roots from which different approaches towards data literacy and skills have grown – with varying levels of emphasis placed upon the entangled branches.

2.1 The data literacy discussion in academia

When discussing data literacy, Frank, Walker, Attand and Tygel [13, pg.5] point out that "as a research topic it has, until recently, been largely confined to the skills that students and researchers need to use data". This, they argue, is something that has necessitated challenge and change as a result of the increasing profile of open data. The implication of which has been an increased attention and emphasis on data literacy beyond certain narrow enclaves, to something that is of importance for broader society. As a result, the idea that data literacy spans beyond a limited set of skills has gained a degree of traction in academia as captured by Wolff et al. [44].

> Data literacy is the ability to ask and answer real-world questions from large and small data sets through an inquiry process, with consideration of ethical use

> of data. It is based on core practical and creative skills, with the ability to extend knowledge of specialist data handling skills according to goals. These include the abilities to select, clean, analyse, visualise, critique and interpret data, as well as to communicate stories from data and to use data as part of a design process. [44, pg.23]

While Wolff et al propose that data literacy is based on skills, this definition extends to include broader contextual considerations, such as the ethical use of data, which serves to ground these skills in context.

There is also an important discussion surrounding the degree of flexibility that should be granted surrounding definitions of data literacy. On the one hand, Frank et al suggest that "although defining data literacy increases rigor and helps clarify its scope it [is] also important to be flexible and not let discussion of definitions inhibit or constrain research. Any attempt to define a term in its formative stage is as much prescriptive as descriptive – it is an implicit recommendation to include certain things and exclude others" [13, pg.6]. Conversely, Khan, Kim and Chang [14, pg.4] propose stricter parameters, suggesting that "confining data literacy to a single discipline or forming multiple literacies under different disciplines with separate terminology only stands to befuddle users", it is evident that agreement on definitions remains a somewhat distant prospect at present. It is therefore important to communicate the parameters clearly, when discussing data literacy. This unresolved discussion also provided pause for thought when analysing our findings, as shall be elaborated on within our results.

In their conference paper in which they discuss their proposed data literacy intervention, Debruyne, Kearns, O'Neill, Colclough, Grehan and O'Sullivan [4, pg.24] note that data literacy is complex for several reasons, including that it "encompasses many activities (both with [active] and [passive] elements such as creating a plot and comprehending it), which draw on many disciplines ranging from maths to art". Their acknowledgement of the *passive*, as well as the *active* elements of data literacy are particularly useful in illuminating the current disconnect between academia and the public service/business communities in both definitions and approaches to measuring and improving data literacy. Within public service/business communities, emphasis appears to focus more heavily upon the *active* elements - such as the skills to gather and process data - while placing lesser emphasis on the more passive, contextual components.

It is these passive elements – which are often neglected and can escape immediate consideration – that can provide the greatest contribution in improving data literacy more broadly. When taken together, Pinney suggests that these passive elements can contribute towards a data literacy that enables emancipatory participation in society, which can "enable both researchers and populations affected by inequality to ask critical questions around power" [29, pg.233].

2.2 Alternative definitions of data literacy

It is important to note that in spite of the general convergence that has occurred within academic discourse surrounding the definition of data literacy, this is but one sphere in which this nomenclature has progressed towards an arguably firmer, more stable state. Discussion also extends into analysis of what data literacy is for specific purposes, such as how to measure it in public service [2].

The consequence is that there can often end up with a conflation of responses [24]. As an example, within the UK government's 2020 National Data Strategy [5] there is continued reference to boosting data science skills as the solution to data skills *and* literacy gaps in the UK. In the follow up Qualifying the UK Data Skills Gap report from the UK's Department for Digital, Culture, Media and Sport (DCMS) [6], which was developed to elucidate the proposed solutions to the challenges identified in the 2020 National Data Strategy, there was an acknowledgement that "the data revolution has implications not only for experts with advanced analytical skills, but for the entire UK workforce".

While true in itself, this statement stops short in acknowledging the impact that this has for everyday citizens, outside of specific, professional contexts. This line of thinking remains commonplace not only in governmental thinking, but in many areas of business: that improving specific data skills is the panacea to improving data literacy. Later in the same report, it is noted that:

> "A review of curriculum content and mathematical literacy demonstrates the need for rapid change to ensure a 'fully data-literate population'. Despite the UK's well-developed economy, the country is frequently positioned mid-rank in mathematical skills - alarmingly, the numeracy skills of 75% of Britons aged 16-65 may prevent them from comparative price analysis of products and services, as well as household budgeting." [6]

This once again suggests a conflation of skills and literacy as being roughly similar and resolvable, something that would benefit from delineation, clarification and subsequent elaboration.

The consulting firm, KPMG, recognises that data literacy is more than an essential skillset for data scientists – it is about society at large gaining a fundamental understanding of the language and value of data in all of its forms and applications" [17], while similarly succinct definitions from other members of the 'Big Four' (Deloitte, EY, KPMG and PwC) are less evident. There is, however, acknowledgement of the importance of 'digital upskilling' to address data literacy from PwC [21], who place emphasis on technologically-focused solutions such as increasing competency with market-leading software [20]. This is reflective of a sector in which specific challenges and needs are identified and targeted responses are developed to address inefficiencies. Such an approach can, however, have certain drawbacks when the focus on upskilling is not complemented with similar levels of focus on the behaviours necessary for their appropriate utilisation.

3 WHY DATA LITERACY IS SO CRITICAL

The European Data Market study[3] valued the data economy in the EU27 plus the UK to be in excess of 400 billion euros in 2019. The same study references the number of data professionals as 7.6 million in 209, or 3.6% of the total workforce. Yet the European Data Market Monitoring tool "continues to register an imbalance between the demand and the supply of data skills in Europe as the

estimated gap reached approximately 459,000 unfilled positions in the EU27 plus the UK." This equates to 5.7% of the total demand. The study also notes that "the data skills gap is forecast to continue in all the forecast scenarios as demand will continue to outpace supply."

This same message is echoed across many sources. The Winterberry Group [18] identified that 44% of companies surveyed intended to be "extremely data centric" within 18 months, but that only 9% currently felt they were operating at that level. The Data Literacy Index [32] found that 76% of key business decision makers aren't confident in their data literacy skills. In its 2018 report, Gartner [19] cited poor data literacy as the second highest internal roadblock to the success of a Chief Data Officer. At the moment, organisations do not have the skills they need to benefit from their data as they would like to.

One issue in unlocking the full value of data within an organisation is that efforts to increase data literacy have been focused on data practitioners. Focusing on upskilling such employees, such as data scientists, has been shown to lead to lack of understanding by managers and leaders on what can and can't be done with data by the practitioners [36]. This leads to insight (as well as risk) being lost in translation. Thus a focus on a balance of skills is required to help those working with data understand the impact on people and processes, while those who make decisions need to increase their data literacy in relation to the practical opportunities and limitations of working with data.

Pothier and Condon remark: "Data literacy competencies are needed within many (if not most) departments in a business organization. It is not enough to educate data analysts and data scientists only, but to think more broadly about how these competencies can be learned as part of a general business education." [30] (p.6). Further, when discussing the role of human resources, Pothier and Condon[30] note Sinair, Ray and Canwell [35], who suggest that they "must work relentlessly to develop and recruit people who advance digital transformation across their organizations. Yet most have struggled to advance their own digital competencies. This neglect has hindered their ability to leverage data into talent strategies that can help transform their businesses". Whole teams and departments are often left behind when considering how to improve an organisation's data literacy, which is detrimental to the overall aim.

Some of the fault lies with data literacy learning provisions. In her paper looking at those who are attempting to meet the ever-growing demand for data professionals, Jeonghyun Kim [15, pg.169] notes that

> "A number of institutions are responding to the need for data skills in the marketplace by launching new academic programs aimed at boosting the number of qualified data professionals, but the content and focus of such programs varies widely."

This, in part, can be explained by the continued multiplicity of definitions surrounding data literacy. As a consequence, what we see in the marketplace, within academia and within government is often diverse solutions, many of which place emphasis on the teaching of precise skills for the purpose of equipping professionals with specific tools to undertake particular tasks, as evidenced by earlier discussion of how data literacy is approached in both the public service [2, 5] and business [20]. Gartner[19] discovered in its 2018 report that many data literacy learning provisions are focused (or perceived to be focused) on "just tools training".

This, we believe, only partly addresses the problem.

Furthermore, Kim [15] suggests that many gaps exist within these programmes that prevent professionals from developing the rounded skills that they require to be fully competent – something she notes certain institutions sometimes attempt to resolve through recommending complimentary courses from departments outside of their core focus. This suggestion speaks to the core of the issue, which is developing a critical contextual analysis surrounding specific skills to enable an individual to wield and utilise these with competence and purpose. It is not only about having the ability to be able to undertake a task, but also knowing when to do so, under what circumstances that are appropriate. In other words, to apply critical thinking to data.

Through research conducted by the ODI for the purpose of discovering the data skills required by people in different sectors and those in higher education [41], it was found that the highest level of attainment recorded in responses equated to Level 4 and 5 of the European Qualifications Framework[9]. At these levels people are expected to be able to apply a range of cognitive and practical skills to develop creative solutions to specific (level 4) and abstract (level 4) problems.

Market analysis of data literacy programme offerings available showed that there was a focus on improving compulsory skills (levels 1-3 of EQF) and university-level programmes, which equated to levels 6-7 of the EQF [38]. As such, adults sent on courses focusing on building capacity around data were often encountering a gap within their learning progression (levels 4-5), which served to stymie confidence and the development of competence.

Combined, these analyses paint a familiar picture of an environment in which companies are increasingly looking to optimise their use of data, however they currently lack the means to do this. Comparison of data literacy levels both within and between companies can, however, be a problematic roadblock in beginning the process of improving data literacy within an organisation or, more broadly, within society.

4 DATA LITERACY AT THE ODI

Founded by inventor of the web, Sir Tim Berners-Lee, the ODI sits at the intersection of government, academia and the private sector. At the ODI, we recognise the need to train both the people who are putting data and information out there, as well as those reading it, how to interpret and question it to ensure they understand it and are not being misled or deceived. It is the arts of critical thinking and scepticism that are continuing to prove key in a world where we are bombarded by numbers, statistics and, increasingly, data and machines using that data to make decisions that affect us.

The ODI defines data literacy as:

> "The ability to think critically about data in different contexts and examine the impact of different approaches when collecting, using and sharing data and information." [38]

Figure 1: The ODI Data Skills Framework

We believe increasing data literacy should including helping people to:

- compare and contrast how different people use numbers, graphs and infographics to convey important messages on topics such as climate change, population growth or global pandemics.
- evaluate the impact of bias and limited sampling on important decisions, such as those in the criminal justice system or when hiring decisions are made.
- examine the ways that data is collected and the purposes of this collection, from irregular collection in a census to sensors in trains, cars and busses that keep those vehicles running, to our interactions with digital assistants.

Following research into the broad range of skills required to get the most from data [22, 41], we built the ODI Data Skills Framework (Figure 1). Through this framework we try and communicate the spectrum of areas that require some level of data literacy, while not being too prescriptive about the detail behind each.

Broadly speaking, the skills framework is split into left and right. On the left are the softer skills relating to decision making and delivering benefit while protecting people and data. On the right is the more practical side, relating to understanding how data is used to inform that decision making.

Although not a direct mapping of Debruyne's [4, pg.24] *active* and *passive* model, the skills framework is an attempt to integrate the more human element to data, from that of building communities and managing change to working ethically.

One challenge with the ODI Data Skills Framework is the perception that it is competency model, where each hex can be taken individually and broken down into competencies that can be used to assess data literacy. However, this will often lead to a focus on skills and tools training as opposed to encouraging elements of

problem solving and critical thinking. The temptation to look at each hex as an individual skill that does not rely on the others is also somewhat shortsighted, especially when considering the close connection between aspects like governance and data platforms, analytics and ethics, measuring success and data visualisation.

5 BENCHMARKING DATA LITERACY

Competency models have historically been a popular way to benchmark an employee's personal and professional development. The idea of evaluating someone's ability by competency, rather than traditional academic testing, dates from the 1970s with research to suggest that existing academic testing was not a good predictor of how well someone performed at work[25].

Competency models have since been used across businesses, although defining what a competency actually is, is "one of the most fraught tasks in business research, with little agreement among researchers" [42]. Vazirani[42] goes on to cite Page and Wilson [28] who reviewed 337 definitions and citations and defined competencies as "the skills, abilities, and personal characteristics required by an employee". The personal characteristics can be difficult to measure [23] whereas knowledge, skills and ability may be easier to assess.

Some of the downsides of creating competency frameworks include:

- Competency models are arduous pieces of work that can become outdated very quickly
- There is often an expectation among employees that competency models relate to role, performance and pay.
- In relation to performance and pay, competency models will often relate to measurable skills as opposed to overall knowledge and literacy.
- Competency models often sit in the domain of HR who don't always consult other relevant stakeholders. As such there is a risk that these models may not always be representative.

5.1 Current approaches to benchmarking data literacy

A self-assessment questionnaire is one way to assess data literacy. Based on the competencies identified by Ridsdale et al (2015) [34], the Databilities® self-assessment tool consists of a series of multiple choice questions where participants have to select the statement that best describes them, as depicted in the example in Figure 2.

All questions are aligned to the same six levels as presented in (Table 1).

While organisations applying the Databilities® self-assessment tool have the opportunity to add questions, it is otherwise intended as a one-size-fits-all approach.

The advantages of a self-assessment approach to assessing organisational data literacy are that it is easy to standardise across an organisation and relatively easy to implement given that everyone should use the tool in the same way.

There are, of course, limitations to self assessment, both against the literacy levels above and the skills framework. The primary limitations relate to the ability of individuals to accurately undertake self assessment of their data skills and literacy. Discussion of this has been previously considered by Bonikowska, Sanmartin and Frenette [2], who note in relation to the assessment of data literacy that subjective tools such as self-assessment surveys "may produce a distorted picture of the actual skill distribution", highlighting how the work of Ehrlinger et al [8] has previously shown that "low performers substantially overestimate their performance on intellectual tasks".

The questions themselves are also open to interpretation. Consider question 4f in Figure 2: "use visual methods and tools to understand and explore data." Not stating specific methods and tools is beneficial as it doesn't limit the scope of the question, however someone with a lower level of data literacy might rank themselves highly based on their ability to use basic spreadsheet functioning, whereas someone else might rank themselves lower using other tools. Simply put, respondents don't know what they don't know. It is also harder to focus on the critical thinking aspect with these questions, as they focus on doing and/or teaching tasks.

4f. Which of these statements best describe you?

(a) With guidance, I can use visual methods and tools to understand and explore data provided to me.
(b) I can use visual methods and tools to understand and explore data provided to me.
(c) I can use visual methods and tools to understand and explore a range of data sources.
(d) I can assist others to use visual methods and tools to explore a range of data sources.
(e) I can teach and assist others to use visual methods and tools to explore a range of data sources.
(f) None of these describe me.

Figure 2: Question 4f from the Databilities® self-assessment tool

Level	Criteria
Level 1	At this level of competency, an individual **can complete simple tasks with instruction**..
Level 2	At this level of competency, an individual **can complete simple tasks on their own, with guidance where needed**.
Level 3	At this level of competency, an individual **can complete well defined tasks on their own**.
Level 4	At this level of competency, an individual **can complete complex problems and tasks on their own**.
Level 5	At this level of competency, an individual **can assist others to complete simple tasks and problems**.
Level 6	At this level of competency, an individual **can teach and assist others to complete complex problems and tasks**.

Table 1: Databilities® levels

5.2 An alternative model to benchmarking data literacy

The ODI has begun developing a method to benchmark levels of adult data literacy. In addition to drawing on the skills definition from the European Qualifications Framework – where skills are described as the combination the cognitive (use of logical, intuitive and creative thinking) and practical (involving manual dexterity and the use of methods, materials, tools and instruments) – we looked at the Organisation for Economic Co-operation and Development's (OECD) definition of literacy. In their Programme for the International Assessment of Adult Competencies (PIAAC), which also encompasses numeracy and problem solving [11], the OECD considers literacy as "the ability to identify, understand, interpret, create, communicate and compute, using printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve their goals, to develop their knowledge and potential, and to participate fully in their community and wider society"[10].

Drawing upon the PIAAC and its emphasis on literacy for the purpose of facilitating participation in society, rather than solely economic prosperity [11], the Data Literacy Framework proposed by the ODI aims to help individuals assess their own, or their teams, data literacy level. Table 1 presents these levels and their corresponding criteria.

As previously discussed, there are limitations to self-assessment relating to individuals' ability to assess their own skills. To further investigate this, we carried out a brief survey within our own network to establish the perceived level of data literacy.

5.3 Survey methodology

For the purposes of trying to understand current perceptions around data literacy, we created a short survey to gather responses to self-assessed questions regarding respondents' levels of data literacy.

A convenience sample was used for the survey, in which we approached contacts within the ODI network for responses. This involved sending the survey directly to a targeted mailing list consisting of contacts who had previously engaged with the ODI through activities such as training and had consented to being contacted with regards to future ODI data literacy efforts. To supplement this targeted approach, a description of the survey and link for interested respondents was included within the weekly ODI newsletter. This contributed to a diversification of the responses from respondents of more diverse backgrounds, however this remains a convenience sample given that the respondents had all expressed either an interest in the ODI's work. The respondents were asked two questions about their job and the primary responsibilities this includes [1] before then being presented with the ODI Data Skills Framework (Figure 1) and asked to identify which side they most closely align with (left hand side, middle, right hand side) within their role. Finally, respondents were presented with criteria statements for each of the 6 levels of the data literacy benchmarking and asked to select all of the statements that included tasks that the respondent would be comfortable undertaking in a professional capacity.

[1]Respondents were asked for this additional layer of information in order to contextualise the activities and responsibilities involved in their roles - given the diverse interpretations and applications that exist of roles such as 'data analyst'

Level	Criteria
Below Level 1	Able to recall a single piece of specific information as presented in a graph or chart. Not required to understand the structure or meaning of the data.
Level 1	Able to understand the meaning of information or data presented to you. Able to explain what a simple graph means.
Level 2	Able to consider where the information has come from and how this impacts the message being presented. Able to paraphrase and make low level inferences from data and how it is presented. Able to interpret data to state a new fact or existing one in a new way.
Level 3	Able to question the validity of claims, such as a misleading graph, and can spot fake news. Able to interrogate data and information from a variety of sources. Able to understand wider context and subtleties, and some limitations or bias in how data is collected, used and shared.
Level 4	Able to evaluate the methodology behind how the data was collected and interpreted and what impact that has on the conclusions drawn. Able to bring together data from different sources or collected with different methodologies to draw new conclusions and make informed decisions about current or future direction.
Level 5	Able to synthesize or create original ideas based upon a thorough evaluation of a broad range of data/information sources combined with specialist knowledge. Able to recognise your own biases and limitations, as well as those already present in the data/information resulting from how it was originally collected or interpreted. Also able to recognise what biases or limitations that might result in, including deeper societal biases that might not have been corrected for, and take appropriate steps to mitigate for that bias.

Table 2: ODI Levels of Adult Literacy

The objectives of the survey were to:

- establish how people identify with the skills framework in their role
- validate assumptions related to self-assessment

Furthermore, it is acknowledged by the authors that respondents will likely have varying interpretations of the skills included within the Data Skills Framework, or may similarly under/over estimate the relative importance of these skills within their professional roles. This serves to evidence another of the limitations of using a self-assessed survey in isolation from other benchmarking techniques.

Efforts were made to proactively mitigate the potential influencing of respondents answers. For example, reference to 'levels' of

data literacy were omitted. Instead, respondents were asked to "tick all statements that describe tasks that you would be comfortable undertaking in a professional capacity", in the question that assessed respondents' data literacy. Similarly, words such as 'ability' and 'capability', which the authors considered emotive and likely to encourage respondents to desire to be considered more capable or able than they are were avoided. [2]

Cognisant of these limitations, the authors are therefore aware that the results of the survey will be imperfect, however they have utility in an illustrative capacity.

5.4 Results

In total, 93 fully completed responses to the survey were received. Based on the responsibilities and competencies required within their roles, 17 respondents (18%) identified predominantly with the left-hand-side of the Data Skills Framework, 56 (60%) with the middle of the framework (including a limited mix of both left and right) and 20 (22%) respondents with the right hand side of the framework.

Overall, respondents assessed themselves as holding high levels of data literacy. In total, 46 of the 93 (49%) respondents rated themselves as having the highest level of data literacy assessed. This ratio was generally replicated based on the respondents' identification with the ODI Data Skills Framework, with 9 of the 17 (53%) respondents who identified as sitting primarily on the left hand side, 27 of the 56 (48%) identifying with the middle and 10 of the 20 (50%) respondents who identified with the right hand side all assessing themselves at the highest level of data literacy.

5.5 Discussion

An immediate takeaway based on the results is that respondents appear to have a tendency to be generous in their assessment of their data literacy. This result was somewhat anticipated, based on the discussion raised earlier in the paper, including from Bonikowska, Sanmartin and Frenette [2].

This observation further evidences that self-assessment of data literacy is a sub-optimal approach for the purpose of benchmarking, particularly when considered in isolation of other means to determine an individuals data literacy. Presently, other means of measuring data literacy also rely heavily upon self-assessment, such as Data To The People's Databilities[40].

An extension of this analysis, which would require further evidence and scrutiny, is that individuals' self-assessment of their data literacy may be influenced by their domain-specific understanding; ie the context in which they are familiar. Given the cumulative nature of the data literacy levels, this consideration becomes particularly acute towards the upper end of the scale at level 5, in which the criteria lends itself to more precise articulation of an individual's data literacy for a specific purpose.

6 REDUCING SUBJECTIVITY IN BENCHMARKING

To assess data literacy more effectively, we consider it necessary to try to minimise the tendency that respondents inherently have to apply their own contextual understanding to a given problem. Through encouraging objectivity, this can have the emancipatory effect of enabling individuals to apply their specialist expertise within unfamiliar contexts.

This is not an unfamiliar approach in education, especially as related to EQF levels 4-5 where individuals are expected to apply their knowledge to solve specific or abstract problems [9]. It is also very similar to the approach of problem based learning (PBL); a teaching method in which complex real-world problems are used as the vehicle to promote the development of critical thinking skills, problem-solving abilities, and even communication skills [7].

Combining concepts from PBL with the need to reduce subjectivity to benchmark literacy means potentially trying to create a series of simple to understand, but challenging to solve questions through which to assess literacy.

Such techniques are not new. Intellectual Quotient (IQ) tests have been in use for over a century and have become widely regarded as a convenient go-to measure of reasoning and problem solving abilities. These tests benefit from providing a snapshot of current academic abilities based on problems in which all the information required to solve them is presented and there is only one correct answer [27]. As a result, they are not intended as a measure of a person's practical intelligence [37], but rather a person's developed skills and should therefore provide a more objective picture of their problem-solving abilities. Such means of measurement lend themselves to the assessment of data literacy in this regard, in that they minimise the utility of the content and context that a person could bring into the answer and are designed to minimise the influence of a person's native intelligence.

One risk with multiple choice questions, such as those used in IQ tests, is that they often simply test someones ability to either recall what they have seen before or their ability to eliminate the obvious wrong answers [45], which is why they are not effective for measuring practical intelligence. However many argue that when carefully constructed they can be used effectively, especially when the reasoning problem has to consider multiple factors or variables [26].

Question 1 (shown in Figure 3) is a reasonable example of a question where you could change the input data (shown in brackets) to vary the question and avoid the element of recall as well as increasing some of the complexity of reasoning required.

Q: You have been given data relating to [people's salaries] and asked to calculate the average?
How would you do this?

(a) Find the total of all the [salaries] and divide by the number of [people] there are
(b) Sort the [salaries] in ascending order and pick the one in the middle of the sorted list
(c) Sort the [salaries] in ascending order and find the most common one
(d) All of the above are equally valid

Figure 3: Multiple choice question with variance on input

[2] See Annex A for full list of survey questions

While the example shown in Figure 3 has a single correct response, answering this question requires understanding that salary data is never normally distributed. The mean is pulled high by the high earners and the modal value represents something towards minimum wage. If the context was changed, eg favourite colour, or average exam grade, then the answer would be different.

As another example, take the graph shown in Figure 4 from the University of Washington's course on the art of scepticism in the world of data [43].

Q: The graph below shows 11 points representing the relationship between crop yield and salt concentration in the soil. There is space for another 10 points to be added, choosing from A, B, C or D which is most likely to represent the value of Y (the yield) when X (salt concentration) reaches to 20?

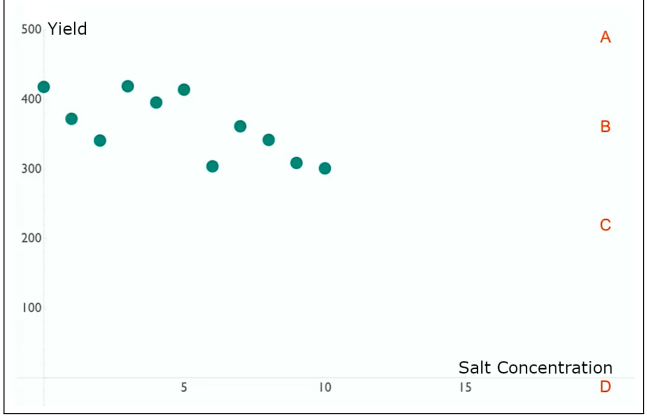

Figure 4: Multiple choice question where variance comes from the labels on the axis

In this example (B), (C) and (D) are all trend lines that fit the 10 data points shown, to pick the correct one, you need to apply knowledge of the domain (in this case agriculture). So like in the first example, changing the domain would be a good way to assess the ability for someone do demonstrate their practical ability to apply their knowledge.

Sometimes however, there may not even be a right answer, such as choosing measures for success, where a multitude of ways of representing and reporting on data may exist. Figure 5, shows one such example.

This question makes people think sceptically about summary stats and how data is used to inform decision making. Here, the ideal reaction of the learner is a recognition of the challenge presented by the learning experience, rather than through a test of knowledge. It is anticipated that they can relate to this when they come across similar scenarios in real life. In this way, this type of question is more suited as a starting point in the experiential learning cycle [16].

We are not alone in trying to think this way about data literacy as the work of Questionmark and Cambridge Assessment demonstrates[33]. There is however still some way to go to evaluate the effectiveness of such techniques to both benchmark and drive adoption of approaches with increased focus on literacy over skills.

Two recruitment agencies sift applications and put candidates forward for interviews. The success rate of candidates is used as a performance measure. A summary dashboard indicates that Recruits R Us are performing slightly better. (Table 1)

Drilling into the various applicant categories (Table 2) shows that Acme Recruiting actually performs better in all categories.

What does this result tell us?

Agency	**Success Rate**
Acme Recruiting (Acme)	43%
Recruits R us (RRS)	44%

Table 1: Summary statistics

Category	**Agency**	**Interviews**	**Offers**	**Success**
Leavers	Acme	35	14	40%
	RRS	16	6	38%
Graduates	Acme	39	17	44%
	RRS	19	8	42%
Experienced	Acme	3	2	67%
	RRS	10	6	60%

Table 2: Applicants by category

Figure 5: Demonstrating risks of summary statistics

7 CONCLUSIONS AND NEXT STEPS

The definition of data literacy has been the subject of considerable discussion, both within academia, the public and the private sector. While it has been recognised that data skills will be important for the entire workforce, until recently, most of the focus in data has been around how to equip people with the skills to use data. However the increased impact that data is having on society has demonstrated the need for a different approach, one where people are able to understand and think critically about how data is being collected, used and shared.

Going beyond a definition, defining levels of literacy has helped us shape the conversation beyond high-level definitions. However, challenges remain in using this as a benchmarking tool. Further work remains to investigate if cognitive abilities, as related to data literacy, can be measured, or if the levels should simply be used as a reference for those looking to build effective data literacy learning experiences[1, 31, 39, 43], including novel concepts such as data escape rooms.

REFERENCES

[1] Southampton Data Science Academy. 2020. *Fundamentals of Data Science (Non-Technical).* Retrieved February 3, 2022 from https://southamptondata.science/courses/fundamentals-of-data-science-non-technical.htm

[2] Aneta Bonikowska, Claudia Sanmartin, and Marc Frenette. 2019. *Data Literacy: What it is and how to Measure it in the Public Service.* Statistics Canada, Analytical Studies Branch.

[3] Gabriella Cattaneo, Giorgio Micheletti, Mike Glennon, Carla La Croce, and Chrysoula Mitta. 2020. *The European Data Market Monitoring Tool: Key facts & figures, first polocy conclusions, data landscape and quantified stories.* Retrieved January 26, 2022 from https://datalandscape.eu/sites/default/files/report/D2.9_EDM_Final_study_report_16.06.2020_IDC_pdf.pdf

[4] Christophe Debruyne, Anne Kearns, Ciaran O'Neill, Mary Colclough, Laura Grehan, and Declan O'Sullivan. 2021. DALIDA: Data Literacy Discussion Workshops for Adults. In *13th ACM Web Science Conference 2021*. 23–25.

[5] Media Department for Digital, Culture and Sport. 2020. *National Data Strategy*. Retrieved January 26, 2022 from https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy#data-1-2

[6] Media Department for Digital, Culture and Sport. 2021. *Quantifying the UK Data Skills Gap - Full report*. Retrieved January 26, 2022 from https://www.gov.uk/government/publications/quantifying-the-uk-data-skills-gap/quantifying-the-uk-data-skills-gap-full-report

[7] Barbara J Duch, Susan E Groh, and Deborah E Allen. 2001. *The power of problem-based learning: a practical" how to" for teaching undergraduate courses in any discipline*. Stylus Publishing, LLC.

[8] Joyce Ehrlinger, Kerri Johnson, Matthew Banner, David Dunning, and Justin Kruger. 2008. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes* 105, 1 (2008), 98–121.

[9] Europass. 2022. *The European Qualifications Framework*. Retrieved January 27, 2022 from https://europa.eu/europass/en/european-qualifications-framework-eqf

[10] Organisation for Co-operation and Development. n.d.. *Adult Literacy*. Retrieved January 31, 2022 from https://www.oecd.org/education/innovation-education/adultliteracy.htm

[11] Organisation for Co-operation and Development. n.d.. *Survey of Adult Skills (PIAAC)*. Retrieved January 31, 2022 from https://www.oecd.org/skills/piaac/

[12] Mark Frank and Johanna Walker. 2016. Some key challenges for data literacy. *The Journal of Community Informatics* 12, 3 (2016).

[13] Mark Frank, Johanna Walker, Judie Attard, and Alan Tygel. 2016. Data Literacy-What is it and how can we make it happen? *The Journal of Community Informatics* 12, 3 (2016).

[14] Hammad R Khan, Jeonghyun Kim, and Hsia-Ching Chang. 2018. Toward an Understanding of Data Literacy. *iConference 2018 Proceedings* (2018).

[15] Jeonghyun Kim. 2016. Who is teaching data: meeting the demand for data professionals. *Journal of Education for Library and Information Science* 57, 2 (2016), 161–173.

[16] David Kolb. 1984. *Experiential Learning: Experience As The Source Of Learning And Development*. Vol. 1.

[17] KPMG. 2021. *Data Literacy - Meeting the need for a data-literate society*. Retrieved January 27, 2022 from https://home.kpmg/xx/en/home/insights/2021/11/data-literacy.html

[18] Winterberry Group LLC. 2018. *The Data-Centric Organization 2018*. Retrieved January 26, 2022 from https://www.iab.com/wp-content/uploads/2018/02/DMA-IAB-Winterberry-Group-The-Data-Centric-Org-2018-February-2018.pdf

[19] Valerie Logan and Alan D. Duncan. 2018. *Getting Started With Data Literacy and Information as a Second Language: A Gartner Trend Insight Report*. Retrieved January 26, 2022 from https://emtemp.gcom.cloud/ngw/globalassets/en/doc/documents/3892877-getting-started-with-data-literacy-and-information-as-a-second-language.pdf

[20] PricewaterhouseCoopers Auditing Ltd. 2021. *Data Analytics Academy - PwC*. Retrieved January 27, 2022 from https://www.pwc.com/hu/pwc_digital_learning_solutions/digital_skills/Data_Analytics_Online_Academy.html

[21] PricewaterhouseCoopers Auditing Ltd. 2022. *Data Analytics Academy - PwC*. Retrieved January 27, 2022 from https://www.pwc.com/hu/pwc_digital_learning_solutions/digital_badges/Data_Analytics_Academy.html

[22] Leonard Mack and David Tarrant. 2016. *European Data Science Academy: D1.4 Study Evaluation Report 2*. Retrieved January 27, 2022 from https://edsa-project.eu/edsa-data/uploads/2015/02/EDSA-2016-P-D14-FINAL-withouthPrivateAppendix.pdf

[23] Anne F Marrelli, Janis Tondora, and Michael A Hoge. 2005. Strategies for developing competency models. *Administration and Policy in Mental Health and Mental Health Services Research* 32, 5-6 (2005), 533–561.

[24] Paul Matthews. 2016. Data literacy conceptions, community capabilities. *The Journal of Community Informatics* 12, 3 (2016).

[25] David C McClelland. 1973. Testing for competence rather than for" intelligence.". *American psychologist* 28, 1 (1973), 1.

[26] Susan Morrison and Kathleen Walsh Free. 2001. Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education* 40, 1 (2001), 17–24.

[27] Ulric Neisser, Gwyneth Boodoo, Thomas J Bouchard Jr, A Wade Boykin, Nathan Brody, Stephen J Ceci, Diane F Halpern, John C Loehlin, Robert Perloff, Robert J Sternberg, et al. 1996. Intelligence: knowns and unknowns. *American psychologist* 51, 2 (1996), 77.

[28] C. Page, M.G. Wilson, D. Kolb, and New Zealand. Ministry of Commerce. 1994. *Management Competencies in New Zealand: On the Inside, Looking In?* Ministry of Commerce. https://books.google.co.uk/books?id=xGkjAgAACAAJ

[29] Lulu Pinney. 2020. 14. Is literacy what we need in an unequal data society? In *Data Visualization in Society*. Amsterdam University Press, 223–238.

[30] Wendy Girven Pothier and Patricia B Condon. 2020. Towards data literacy competencies: Business students, workforce needs, and the role of the librarian. *Journal of Business & Finance Librarianship* 25, 3-4 (2020), 123–146.

[31] Barton Poulson. 2019. *Data Fluency: Exploring and Describing Data: Gather greater insight and make better decisions with your data*. Retrieved February 3, 2022 from https://www.linkedin.com/learning/data-fluency-exploring-and-describing-data/gather-greater-insight-and-make-better-decisions-with-your-data?autoAdvance=true&autoSkip=false&autoplay=true&resume=true

[32] Accenture Qlik and the Data Literacy Project. 2020. *The Human Impact of Data Literacy - A leader's guide to democratizing data, boosting productivity and empowering the workforce*. Retrieved January 26, 2022 from https://www.accenture.com/_acnmedia/PDF-115/Accenture-Human-Impact-Data-Literacy-Latest.pdf

[33] question mark. 2022. *Questionmark Data Literacy by Cambridge Assessment*. Retrieved February 3, 2022 from https://www.questionmark.com/platform-services/questionmark-data-literacy-by-cambridge-assessments/

[34] Chantel Ridsdale, James Rothwell, Mike Smit, Michael Bliemel, Dean Irvine, Dan Kelley, Stan Matwin, Brad Wuetherick, and Hossam Ali-Hassan. 2015. Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report. (01 2015). https://doi.org/10.13140/RG.2.1.1922.5044

[35] E Sinar, R Ray, and A Canwell. 2018. HR leaders need stronger data skills. *Harvard Business Review* (2018), 2–5.

[36] Data Society. 2018. *The plight of the Frustrated Data Scientist*. Retrieved January 26, 2022 from https://medium.com/@datasocietyco/the-plight-of-the-frustrated-data-scientist-320428bde6e0

[37] Robert J Sternberg, George B Forsythe, Jennifer Hedlund, Joseph A Horvath, Richard K Wagner, Wendy M Williams, Scott A Snook, Elena Grigorenko, et al. 2000. *Practical intelligence in everyday life*. Cambridge University Press.

[38] David Tarrant. 2021. *Data literacy: what is it and how do we address it at the ODI?* Retrieved January 27, 2022 from https://theodi.org/article/data-literacy-what-is-it-and-how-do-we-address-it-at-odi/

[39] Dave Tarrant. 2022. *Applying Machine Learning and AI Techniques to Data*. Retrieved February 3, 2022 from https://theodi.org/event_series/machine-learning-ai-and-ethics/

[40] Data to the People. 2020. *myDatabilities®*. Retrieved January 26, 2022 from https://www.mydatabilities.com/

[41] Emily Vacher and David Tarrant. 2017. Open Data VET course for private and public sector employees. https://erasmus-plus.ec.europa.eu/projects/eplus-project-details#product6

[42] Nitin Vazirani. 2010. Review paper: Competencies and competency model–A brief overview of its development and application. *SIES Journal of management* 7, 1 (2010), 121–131.

[43] Jevin West and Carl Bergstrom. 2020. *Calling Bullshit: The Art of Skepticism in a Data-Driven World*. American Association for the Advancement of Science.

[44] Annika Wolff, Daniel Gooch, Jose J Cavero Montaner, Umar Rashid, and Gerd Kortuem. 2016. Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics* 12, 3 (2016).

[45] Nikki L Bibler Zaidi, Karri L Grob, Seetha M Monrad, Joshua B Kurtz, Andrew Tai, Asra Z Ahmed, Larry D Gruppen, and Sally A Santen. 2018. Pushing critical thinking skills with multiple-choice questions: does Bloom's taxonomy work? *Academic Medicine* 93, 6 (2018), 856–859.

Appendix

A SURVEY QUESTIONS

For the purpose of this paper a short survey was circulated for the purpose of understanding current perceptions around data literacy. Survey respondents were asked the following questions:

(1) What is your job title?
(2) Please provide a brief description of your role in relation to data (eg to me data is...) [note: we are only going to use this to build a better picture of roles that may have a broad remit, ie Data Scientist]
(3) Based on the skills framework[3], in your role, which side of the framework do you most closely associate with?
(4) Based on the options below [4], please tick all statements that describe tasks that you would be comfortable undertaking in a professional capacity.

[3] Prior to being asked this question, respondents were provided with the ODI Data Skills Framework - Figure 1

[4] Respondents were provided with criteria for levels <1-5, as denoted in Table 1