Articles

For whom is data literacy empowering? An awareness-action typology

Shiri Mund, New York University, *shiri@nyu.edu*

Yoav Bergner, New York University, *yoav.bergner@nyu.edu*

For whom is data literacy empowering? An awareness-action typology

Abstract

Building on recent empowerment perspectives on data literacy, we examine how students and working adults talk about their understanding of data and report on their own personal-data-related practices. Through a deductive and inductive analysis of interviews with 19 subjects ranging from middle school to middle age, we find that awareness and action with respect to data consumption and production do not necessarily increase in tandem. For example, being more aware of the data that can be used to track them does not make individuals more likely to take action to manage their personal data. While some feel anxiety about the gap between knowledge and action, others resolve the tension by choosing not to care. These findings are synthesized in a typology of personas in the space of data awareness and action. We investigate the relationship between age and educational attainment with location in this awareness-action space and discuss implications for data literacy education.

Keywords: data literacy; lifelong learning; empowerment; social justice; qualitative research.

Introduction

The last two decades have seen an unprecedented growth in the volume and accessibility of digital data. Data are generated by and, subsequently, influence human behaviour in almost all areas of daily life, ranging from politics and policymaking to employment, education, health, entertainment, marketing, shopping, and social interaction. Terms such as 'dataveillance' highlight the fact that our personal lives are continually tracked through our digital traces (Raffaghelli, 2020). The term *surveillance capitalism* (Zuboff, 2015) has emerged to describe the ways in which corporations have monetized the collection of data used to predict and even modify human behaviour. As society contends with the impact of digital data on knowledge, communication, and privacy, it faces a pressing need for students and citizens who are intelligent producers and cautious consumers of this data.

Despite its acknowledged relevance as a key 21st century skill, data literacy has been difficult to define (Koltay, 2015; Wolff et al., 2016). Various disciplines have staked out terms like statistical literacy, information literacy, and critical literacy, each emphasizing different data-related competencies (Matthews, 2016). Data literacy perhaps encompasses more than any one of these disciplines independently. Francois, Monteiro, and Allo (2020) argue that data literacy is about critical and reflective citizenship based on interdisciplinary skills and competences, bringing 'not only statisticians and mathematicians together but also computer scientists, sociologists, lawyers, and philosophers' (p. 202).

In this work, we build on existing definitions in the literature and conceptualize data literacy as *the collection of skills, attitudes, and beliefs needed by individuals in the 21st century to understand the potential and limitations of data, become critical producers and consumers of data in a variety of forms, and thoughtfully engage with data.*

We regard data literacy as a multi-dimensional learning construct including factors related to competency and empowerment (Gebre, 2018; Storksdieck, 2016). Consequently, we describe this larger perspective before focusing, in the present study, on the empowerment factor in the context of personal data. We then describe a qualitative study exploring conceptions of and attitudes towards data in a sample of middle school students, high school students, undergraduates, and adults (*N* = 19). Subjects were chosen from a wide range of ages in order to replicate and extend previous studies and also to draw hypotheses regarding trends across cohorts. As a result of the interview analyses, we arrive at typology in the space of data awareness and action, highlighting personas with different combinations of each variable. The typology offers a novel perspective on a person's learning progression in data literacy, and we discuss implications for data literacy education.

In what sense is data literacy empowering?

Data literacy comprises a set of skills, attitudes, and beliefs and draws from several related disciplines (Bhargava et al., 2015). The construct has been of increasing interest to *information literacy* scholars (e.g., Prado & Marzal, 2013; Koltay, 2015), who emphasize data's role as a source of information. It is also often referenced by *statistical literacy* researchers (Gal, 2002; Schield, 2004, Gould, 2017), who underscore the skills needed to work with data. Data literacy draws from *digital literacy*, the set of cultural competencies and social skills acquired by young people engaged in online communities (Jenkins, 2006) and *media literacy*, the ability to access, analyze, evaluate, create, and act using all forms of communication (Bulger & Davidson, 2018). Some have argued for a clearly defined set of competencies unique to data literacy (Wolff et al., 2016). On the one hand, media and information literacy tend to de-emphasize technical dimensions of data analysis (Livingstone, Van Couvering, & Thumin, 2008; Bhargava et al., 2015). On the other hand, understanding where data come from and how they are collected, reduced, and structured (the critical aspects of data literacy) are topics often outside the scope of statistics instruction. Moreover, statistics education has largely focused on linear models and smaller samples (Ridgway, 2016), but contemporary notions of data literacy commonly prioritize understanding and working with large, messy datasets. Digital data such as web clicks or online searches are often of a different nature in that they are not representatively collected but analyzed in their entirety (Francois, Monteiro, & Allo, 2020).

There is evidence of the long-term benefit of student engagement in data literacy education as it relates to social justice, both to the individual and for society. For example, Lehrer and colleagues (2011) contend that complex data-related notions, such as variability, may 'prove to be of enduring value as students participate in ever widening circles of civic discourses that are governed by measures and models' (p. 735). Additional work in democratic participation underscores this. Merola and Hitt (2015) showed that participants who scored low in numeracy were more likely to support a policy on criminal reform based on the political party making the argument than on the strength of the data, but this was not the case for more quantitatively literate participants. In addition, students who are aware of the subjective influences on data might be better equipped to navigate around bias if they have a stronger sense of their own dispositions on a variety of ethical, civic, and legal issues (Bowler et al., 2017). While the imperative of data literacy development is becoming increasingly accepted, open questions remain as to how to integrate data literacy education in a way that is relevant and developmentally appropriate. As Twidale and colleagues (2013) articulate, there is a need to develop a better sense of levels of data literacy that permit different kinds of engagement with data depending on the context.

Data Literacy and Empowerment

In a review of existing definitions of data literacy, Gebre (2018) argued that there are two primary research conversations regarding data literacy: a competency- and an empowerment-orientation. The competency-oriented perspective is associated with statistics skills and quantitative fluency and grounded in 'developing technical and conceptual skill of learners so that they become able to deal with mainly quantitative data,' while the empowerment-oriented view focuses on the 'use of data literacy as a means to foster civic engagement and build an equitable and democratic society' (p. 332-333). The present analysis concentrates on the latter perspective, which traces its roots back to Freire’s definition of literacy as not merely as the acquisition of skills but also as a form of emancipation (Tygel & Kirsch, 2015; D’Ignazio, 2017). This approach underscores the importance of data literacy as a consciousness and draws from Critical Data Studies (Dalton, Taylor & Thatcher, 2016), a discipline focused on algorithmic transparency, power dynamics in data, data discrimination, and privacy. The empowerment perspective views data literacy as more than a technical pathway (D’Ignazio, 2017), and research in this area often focuses on *personal data management*, including how personal data is created, where it lives, who owns it, and the awareness and agency needed to control one’s own data, rather than data analysis skills. Educational interventions from the empowerment perspective focus on freedom of choice, awareness, and agency (Raffaghelli, 2020).

Scholars have coined a variety of terms which reference the social and ethical dimensions of data. Pangrazio and Selwyn (2019) introduce the term *personal data literacies* to describe data as socially situated and context dependent and to emphasize the interaction of technical, social, and ethical dimensions. Philip and colleagues (2016) use the play on words 'becoming racially literate about data and data literate about race' to describe both how power operates in the collection, analysis, interpretation, representation, and communication of data and how societal meanings about race are produced, in part, by the collection, storage, conversion, manipulation and representation of data sets. The Data-Pop Alliance white paper introduces the term *data inclusion* and defines data literacy as '*the ability to constructively engage in society through and* *about data*.' Indeed, one can think of data literacy empowerment through a variety of lenses. Two notable examples are open data and quantified-self movements.

The open data movement aims to make data collected by governments and non-governmental organizations accessible to citizens (Ridgway, 2016). Researchers argue that this movement has the potential to re-define notions of democracy, participation, and journalism by bringing values and practices from open-source culture to digital data (Baack, 2015). Open-source data can empower individuals to analyse raw data and make their own interpretations. However, as Baack notes, even though the idea behind the democratization of information is to potentially allow everybody to interpret raw data, most citizens do not have the time or capacity to do so, and the vision of empowerment through open data relies on intermediaries to help the public interpret the data. Another take on data literacy as empowerment may be associated with the 'quantified self' (Lupton, 2016). Quantified-self tools offer individuals the ability to benefit from data collection and analysis in areas ranging from media usage and fitness tracking to emotional wellbeing and personal finance. However, as Lalji (2019) argues, while these applications claim to empower individuals by enabling them to gain personal knowledge through data collection and achieve individual goals, they also obscure how individuals' data can be used in ways that are adverse to their interests.

Ultimately, data literacy empowerment is contingent on individuals having awareness of the data that is around them and collected about them, the knowledge to engage with that data, and the willingness to take action. The present work thus addresses the following research question: *What is the relationship between awareness and action in the context of personal data literacy empowerment?*

Data Awareness and Action

Two variables related to an individual's data literacy and empowerment are their awareness with respect to the existence and use of data in its variety of forms and action taken to engage with those data. The concept of data awareness in the realm of personal data management has been explored in the consumer marketing literature. For example, Graeff and Harmon (2002) looked at consumer awareness and knowledge of data collection practices using loyalty cards. The researchers found that few consumers expressed awareness of the way loyalty programs are used to create data profiles. Moreover, younger consumers were more aware of data collection practices, but they were not any more likely to take action to protect their personal data. The authors suggest that younger consumers might be more accustomed to and familiar with being compensated for their personal information. However, more recent literature shows that young people might not be more aware of data and data practices in all contexts.

In a quantitative study of over 1,000 young adults' attitudes towards data and privacy, Hoofnagle, King, Li, and Turow (2010) found that 42 percent of young Americans answered all five online privacy questions incorrectly. The authors argued that while it might seem like young people are less concerned with maintaining privacy, a gap in privacy knowledge, or awareness, might explain the apparently careless way young behave online. Findings from Stanford's History Education Group in 2016 suggested that young people, often assumed to be proficient in social media, cannot distinguish between an advertisement and a news story and seldom consider the identity, biases, or motivations of a source (Shreiner, 2018).

In the context of the COVID-19 pandemic, Nanni and colleagues (2021) advocate giving individuals more awareness of the kind of data they can collect about themselves to trace virus transmission and limit spread. The authors argued that increased awareness, paired with increased control over how data are shared, can support individuals in taking informed action during a public health crisis.

We set out to use these two variables—awareness and action—to guide and analyze interviews with subjects ranging from middle school to middle age. As we shall report next, we found that awareness and action with respect to data consumption and production do not necessarily increase in tandem. For example, being more aware of the data that can be used to track them does not make individuals more likely to take action to manage their personal data. While some expressed anxiety about the gap between knowledge and action, others resolved the tension by choosing not to care. We hence arrived at a typology of personas in the space of data awareness and action and investigated the relationship between age and educational attainment with location in this awareness-action space. We turn first to our qualitative methods.

Methods

Data collection

We conducted nineteen semi-structured interviews via ZOOM with four cohorts of participants: middle school students, high school students, undergraduate students, and working professionals. The goal of the interviews was to explore how individuals think about data, define data, and interact with it in their day-to-day lives. Each interview lasted between 30 and 60 minutes, with conversations with younger students tending to be shorter. Participants were second and third-degree connections recruited through word of mouth using various contacts in the researchers' networks. We reached out to colleagues and acquaintances to explain the purpose of the study and request suggestions of individuals in their respective networks who might be willing to lend an hour of their time. From this list, we used purposive sampling to ensure diversity in the educational background and data experience of participants.

School-age participants were selected based on their academic year and type of school (public vs private). Undergraduate students were recruited to represent a diversity of majors, serving as a proxy for experience with data. Workforce participants were sampled for variety in age and professional background. Participants were located across the US, including cities such as New York City, Washington DC, New Orleans, Atlanta, and Los Angeles. Table 1 shows the breakdown of participants by cohort and major or profession (where applicable). Given the deliberate variety in age, educational background, and professional expertise, we expected to see variance in familiarity and comfort with data. We expected that an undergraduate student with statistics training would demonstrate higher levels of quantitative data literacy than a middle school student, and similarly for a professional working in a data-focused role compared with other adults. At the same time, we also hypothesized that middle aged participants could be less immersed in the world of digital media and as a result less aware of how their digital footprint is collected and used.

The decision to use a broad sample was deliberate. All groups are involved in and affected by data in everyday life, and the purpose of the study was to provide provisional directions for future inquiry at scale. We recognize that, given our broad approach and small sample size, it is hard to understand different groups in depth through this one study. Because limited previous work offers this sort of anchoring analysis, we set out to record a range of understanding among participants and to inform the development of a framework for further study.

The interview protocol consisted of four sections: 1) Defining Data, which asked questions about how participants define data and think about collecting, using, and producing data; 2) Engaging with Data, which focused on ways in which participants encounter data and engage with data in their everyday lives; 3) Personal Data Management, which explored how individuals think about and manage their own data; and 4) Statistical Thinking, which asked participants to explain common statistical terms. Audio recordings of interviews were transcribed manually with the assistance of ZOOM's natural language processing feature and uploaded for analysis using qualitative data analysis software.

Table 1: Participant breakdown

Cohort	Characteristics	Total
Middle School	3 Private school; 2 Public school	5
High School	2 Private school; 2 Public school	4
Undergraduate	Majors represented: Psychology, Computer Science, Biology, Political Science, Communications	5
Adult	Professions represented: artist, higher education administrator, researcher, lawyer, nurse (3 ages 22-40; 2 ages 41-65)	5

Interpretive procedures

The qualitative analysis was initially driven by a deductive approach but grew to include inductive processes as well. As Armat and colleagues (2018) explain, a deductive or directed approach is used when some views, previous research findings, theories, or conceptual frameworks regarding the phenomenon of interest exist. Because we sought to confirm and possibly expand on previous frameworks, we used a deductive approach to drive the initial analysis. Our semi-structured interview protocol allowed us to address specific questions around conceptions of data and perspectives on privacy and personal data management (including action and awareness. which are the two variables of interest in this paper) while leaving room for flexibility and allowing participants to guide the conversation. The interviews served to probe for different perspectives in efforts to understand how they are adopted by individuals across participant groups. Transcripts were coded using the data literacy perspectives, definitions, and traits identified in the literature. Attitudes of interest were also coded for, such as attitudes towards data privacy and data monetization.

The semi-structured approach also allowed for spontaneity and for insights to emerge organically from the conversation. New codes were created when a new theme or idea emerged that did not directly fit into the predetermined deductive codes, and a thematic analysis process was followed (Braun & Clarke, 2008). This included refining the codes and grouping them into categories. After each new transcript was analyzed, we went back to re-code earlier transcripts. After coding five transcripts representing all four cohorts, we started to group codes into subcategories and categories. The initial set of over 340 codes that emerged from iterative analyses of the data were organized into seven categories in total. The first five consisted of the deductive codes based on predetermined hypotheses and frameworks. The last two code categories included ideas, conceptions, and misconceptions that emerged from the data literacy conversations through inductive analysis.

In the present manuscript, we focus on two codebook subcategories: awareness and action. The complete list of code groups, categories, and subcategories, shown for context, are reproduced in Appendix B.

Analysis

Emergent themes: awareness and action

Two themes emerged as highly relevant to how participants in this study talked about data: *awareness*—of the data all around them, of how data is collected and used, of policies and practices influencing personal data management— and *action* taken to protect personal data or influence one's data footprint. This motivated an analysis to better understand how these two variables are at play in individuals' experiences with data.

The relationship between awareness and action is nuanced. On the one hand, low awareness might confine action. Even if an individual recognizes their limited awareness of data, they may not know how to strengthen that awareness or access data in order to take relevant action. This particular obstacle did not emerge in the conversations in this study, but is plausible and worth considering. Alternatively, it is possible that individuals think they are aware, but their understanding is not accurate. This challenge to the interpretation of awareness was salient across several of the cohorts interviewed, as we will discuss later in this section. Finally, age has a large effect on how awareness is cultivated. Younger students appear to rely on adult figures—parents and teachers. At least this was the case for the participants in this study. For undergraduate students and adults, awareness is cultivated through a mixture of processes including things heard on the news and through word of mouth.

We found it compelling to organize our respondents in a kind of abstract two-dimensional space using these two themes as variables. The process was akin to qualitative persona development in user-centered research studies (Adlin & Pruitt, 2010). The organization of our findings is as follows. We will first describe an emergent typological framework in terms of awareness and action as an abstract space. This presentation may appear to be mostly theoretical at first, with few specific grounding examples from the interviews. After describing how we located subjects in this framework and sought to understand and impose more structure on it, we move on to flesh out the personas themselves with specific evidence from the interviews.

In the context of data and the myriad ways data are used, there are (at least) two ways to think about action. The first form of literacy-as-action concerns whether an individual can obtain their own data or look at data and come to their own conclusions. This type of action may be empowering but also aligns with the competency perspective of data literacy. The second form of action aligns more directly with empowerment and concerns actions relating to controlling one's own data footprint or managing how data are collected and used by third parties. In our data literacy conversations, participants referred to both kinds of action, but the latter form of action was a dominant and recurring theme. The awareness-action typology described below focuses on this second kind of action.

Awareness and action dimensions for a typology

We hypothesized that higher awareness would be associated with higher levels of action, and vice versa. Following this thought, we drew an awareness-action plane on which a two-by-two grid represented a mean split of each variable. Individuals could be located anywhere within the plane, but the four quadrants would help to simplify reasoning about awareness and action. Thus, the top right quadrant corresponded to high awareness and high action, the bottom left to low awareness and low action, the top left to low awareness but high action, and the bottom right quadrant to high awareness but low action. We then rank-ordered each interview transcript according to awareness and action dimensions. The process began as coarse sorting (high, medium, low) with comparative refinements made on a second pass. For example, an individual who expressed acute awareness of the way different forms of data are produced or collected and how third parties might use that data would be ranked high on awareness. A participant who expressed very little interest in managing their data footprint, engaging with sources critically, or using data to make decisions would be ranked low on action. These orderings were used to position participants, as shown in Figure 1. We emphasize that this analysis was intended to be approximate and for purposes of developing the theory. This method was not a precise assessment of participant levels on these two variables. Moreover, the conversations were not expected to comprehensively uncover an individual's data knowledge or practices.

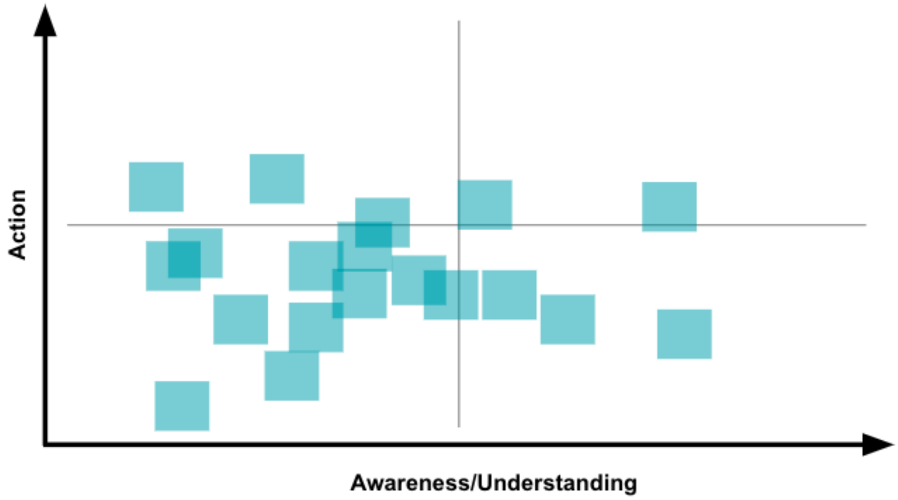

Figure 1. Participants placed in an awareness-action grid based on their interview responses. The working hypothesis was that the upper left and lower right quadrants would be empty.

During discussions between the researchers engaged in this process, it became clear that we would have expected the 'off-diagonal' quadrants in Figure 1 to be empty. It seemed counterintuitive for individuals to have high awareness and take low action or to take more action than their level of awareness would imply. However, after ordering the research participants relative to one another on these two scales, we found that all quadrants were populated. There were a small number of interviewees with levels of awareness higher than the action they were willing to take. There were also some with higher levels of action than their level of awareness would justify. We explore these two surprising findings shortly.

To make more sense of the relationship between action and awareness, and to begin to flesh out the emergent typology, we superimposed two kinds of guiding structure onto the plane. First, distinct zones representing combinations of levels of awareness and action were marked as 'personas'. These personas, symbolizing clusters of participants, will be detailed in the next section. Second, as seen in Figure 2, the three diagonal lines were drawn for thinking about data literacy and its potential progression. All three lines begin at the 'zero' of awareness and action but rise (with awareness) at different slopes.

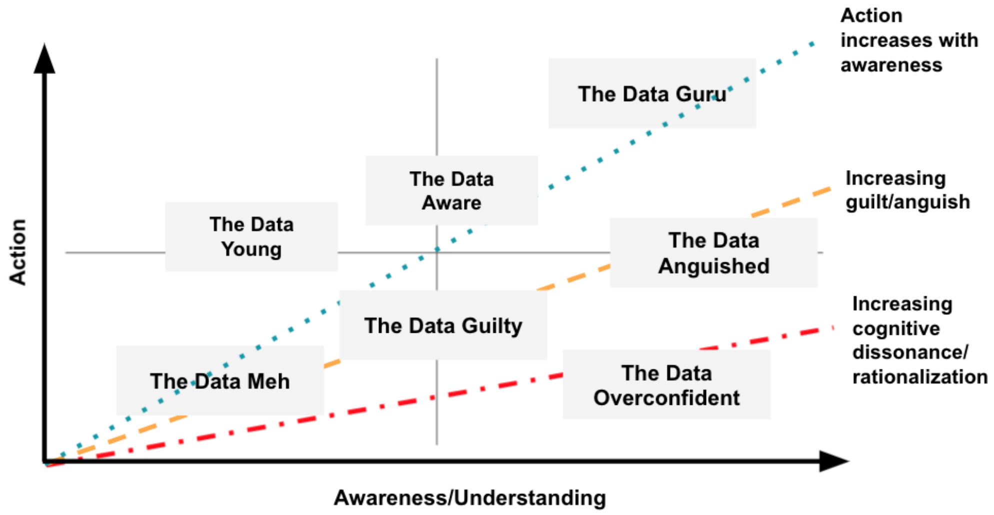

Figure 2. Structure and personas in the awareness-action plane

The dotted line in Figure 2 indicates a trajectory where action increases in tandem with awareness (on some normalized scale). Participants who fall along that 'one-to-one' line reported taking action commensurate with their level of awareness. The dashed line has a shallower slope. Participants who fall along the dashed line tend to take less action relative to their level of awareness. The widening gap between action and commensurate action (the dotted line) as awareness increases may be thought of as indicating an increase in cognitive dissonance. Participants with higher awareness experience higher levels of dissonance, which manifests in descriptions of guilt or even anguish. The third line represents participants furthest from the one-to-one line. Interestingly, these participants no longer expressed high levels of guilt. Some participants felt perfectly comfortable with their level of action and awareness, even if the relationship between the two was lopsided. It is possible that participants who fell along this line based on their data practices were more likely to trivialize or rationalize their behaviors to justify the gap between their awareness and action.

The data literacy personas

In this section, we detail the seven different 'personas' identified from the data. In the field of user experience (UX) design, personas are fictitious but specific, concrete representations of target users (Adlin & Pruitt, 2010). Each persona represents an aggregate of participants who share common behavioral characteristics. These personas help designers understand users' needs, experiences, and behaviors and are used to create products tailored to those specific users. A similar approach has been used in the data privacy literature. For example, Dupree and colleagues (2016) clustered users based on their attitudes and behaviors towards security practices to create 'Privacy Personas.' The authors identified five user clusters that emerge from end-user behaviors and argued that these clusters can support the design of new computer security technologies.

The personas described here emerged as clusters of participants with respect to the awareness-action typology. It is important to acknowledge that, as with any generalization, these are imperfect ways of capturing the nuances in the knowledge, attitudes, and beliefs of any participant in this study or individual in the broader population. As with the Dupree et al's privacy personas, these personas are intended only to serve as a guide for designing and evaluating potential data literacy interventions.

The Data Meh

The Data Meh persona is characterized by low levels of both awareness and action. While this individual likely has high internal *alignment* (their level of action is proportional to their level of awareness) and therefore low cognitive dissonance, one could argue that their level of data literacy is relatively low. In the data literacy conversations, participants who aligned with the 'Data Meh' persona expressed sentiments such as: *'Kind of kind of creepy in a way, but it's just like the world that I'm in...and I don't think about it a lot. Until someone brings it up and then I'm like yeah, I can't do anything about it'* (RA, High School). Another participant emblematic of this persona noted: *'Actually I have no clue what they do with it. I just know, like based on what I searched'* (EA, High School).

The Data Young

Characterized by relatively low levels of awareness/understanding but average levels of action are the Data Young. One participant described taking action to confirm the trustworthiness of data in articles, but her specific approach of looking for quotes might be described as naive: *'Because when I read an article everything that they're saying it can be true or not… Because one article can say one thing and then one article can say another. So, when it comes to that I get really confused. And I'm like, well, I'm going to the quotations, the quotes and things like that, because those are real people.'* (LF, Middle School). Another participant shared some of the strategies she uses when assessing the reliability of data, but the actions she was taking involved misunderstandings: *'It helps when you're looking at the website and you go up to the search bar. You could see a little lock on the top and the first bit of the search bar. If it unlocks then it's unsecure but if it's locked it's secure. And if it's secure, then you're free to use it. And if it's not, you should probably go to a different website'* (KM, Middle School). Interestingly, these participants took more deliberate action than their degree of awareness/understanding would imply or justify. Most participants associated with this persona were in fact middle school students. We shall return to this point in the following section.

The Data Aware and the Data Guru

At approximately the center of the awareness-action grid is the Data Aware persona. With average awareness/understanding and action commensurate with that understanding, the Data Aware are internally consistent. If we imagine the individuals in this cluster having arrived at this point by following the 1:1 line, as their knowledge and understanding increase so might the level of action they take. For example, one participant, an undergraduate computer science major, described the following practices: *'When someone says like this website offers cookies or whatever, I usually try to say no. And I try to not share a lot of information online. And so, I'm not very present on social media. That's a huge one. Like, I think a lot of people over share on social media. Where else do I try to be careful? I like to turn my location services off when I don't need it. Yeah, I try. I try to do my best'* (JSc, Undergraduate).

The Data Guru represents an individual who takes substantial action commensurate with their high understanding and awareness of data. While there were no participants who clustered around this category, we include this persona for the sake of completeness.

The Data Guilty and the Data Anguished

Participants clustered around the Data Guilty region had average levels of awareness and understanding but below-average engagement with data management. They described feeling guilty when talking about this discrepancy. For example: '*I know that I should be better… every time I read about it, a pang of guilt of, I should be better at maintaining my own privacy. But I really just don't know if I've ever made an active change… in that regard, which is not great. But I guess that's the honest answer.*' (AM, Adult). Looking at the awareness-action plane, these feelings of guilt may be related to a certain level of 'underperformance.'

A more extreme (and less common) version of the Data Guilty persona is the Data Anguished. If awareness level increases without commensurate increase in action, the dissonance gap widens. One adult researcher described spiraling thinking: *'That's the sort of thing that's like sort of the gear starts turning is thinking about this passive data collection as a form of I never gave actual consent because I had no idea what you were going to be doing to me or doing with my data...But if we actually care about the ethics of this we have to acknowledge how much of this data was delivered for the purpose giving us something like some sort of benefit like Facebook. Which now I'm like I don't know if that was, I can't figure out if it was worth it for me to have a Facebook'* (NH, Adult).

The Data Overconfident

The final data literacy persona is one we call the Data Overconfident (although another candidate name for this persona might be Data Detached). Individuals positioned in the bottom right of the chart have a level of awareness and understanding that substantially exceed their level of action. Rather than despair, however, these individuals are at ease. One participant expressed his awareness of targeted advertising alongside his lack of desire to take action: *'So I* *think there probably are like more malicious ways of collecting data that I'm not aware of, but the one that's sort of commonly talked about is this targeted advertising. To be honest, I don't really care'* (JS, Undergraduate Psychology). Only one of our interviewees closely matched this persona. However, we consider this position to merit further investigation. One the one hand, it is consistent with known approaches to reducing cognitive dissonance, such as trivialization and rationalization (McGrath, 2017). On the other hand, this persona provides a counter to the idea that 'knowledge is power' when it comes to data literacy. Choosing to trivialize or rationalize one's non-action in the face of consequential awareness is also consistent with learned helplessness (Hiroto & Seligman, 1975).

The personas across age groups

Figure 3 shows a version of the data literacy awareness-action typology with study participants color-coded by age group: middle school students are shown in blue, high school students in green, undergraduate students in red, and adults in yellow. Some suggestive patterns emerge in this view that suggest a relationship between age, background, and levels of action and awareness

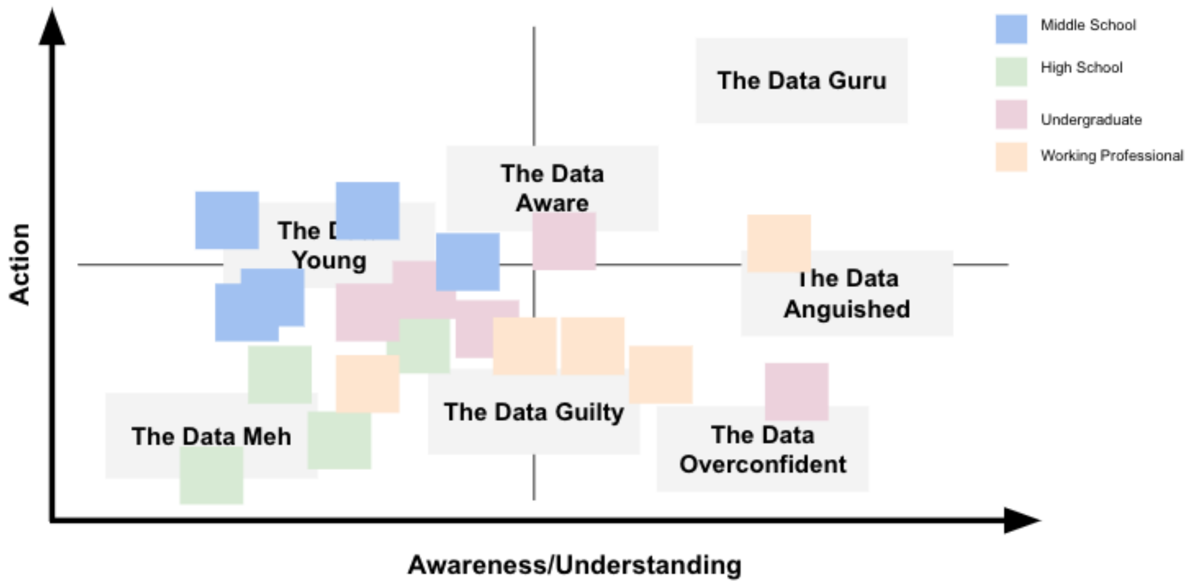

Figure 3. Interview subjects arranged, color-coded by age group, in the awareness-action plane.

Middle school students clustered around the Data Young persona; they expressed relatively lower levels of awareness, but tended to report taking more action than their level of awareness would suggest. At this age, students are hearing about the importance of data literacy and taking action based on what they are told to do or what they believe adults are doing. Despite limited knowledge, these young people are nevertheless optimistically engaged.

High school students clustered around the Data Meh persona. They portrayed limited understanding of data and a low desire to take action. While some did express an interest in learning more about data, others were at ease with their current (consistently low) levels of understanding and engagement. It is perhaps a cliché regarding high school students that they are inclined to exude confidence and disinterest more than they are interested in modeling responsible data strategies. College students and adults were more widely distributed across the typology, especially as awareness levels ranged higher. While the patterns were less clear, undergraduate students were closer to the Data Young and Data Aware—indicating relatively congruent levels of awareness and action—while adults were more likely to cluster around the Data Guilty, meaning that their awareness tended to be higher than the level of action taken. The two participants with the highest demonstrated levels of awareness expressed feelings of anguish (adult) or detachment/overconfidence (undergraduate).

Discussion

For whom is data literacy empowering?

> *I can give you perfect knowledge and it won't change your behavior one iota. People choose not to change their behavior because the culture and the imperatives of the organization make it too difficult to act upon the knowledge. Knowledge is not the power. Power is power. The ability to act on knowledge is power. Most people in most organizations do not have the ability to act on the knowledge they possess.* (Attributed to Michael Schrage; Gurteen, 2003)

The empowerment-oriented view of data literacy, focusing on data literacy to address issues of equity, citizenship, and community engagement, has gained prominence in the literature, but few participants in our study conceptualized data literacy from an empowerment perspective. In fact, only one (adult) participant made any sort of reference to the relationship between data literacy and empowerment, suggesting that data can be used to capture disparities and injustices in society. Strengthening the empowerment perspective in data literacy can help replace guilt and detachment with a sense of agency. But this will require some shifts in perspective. The data literacy awareness-action typology and personas may be useful in guiding the designs of data literacy education pathways.

Awareness and action should be seen as two distinct outcomes of educational experience, and levers may be operated differently to help bring these into alignment for different groups of individuals. We have seen that awareness can (and does sometimes) increase without corresponding increases in action. If data literacy education is to be directed towards goals of social justice, then it is not enough to move the needle on awareness. The quote at the top of this section reminds us that *the ability to act on knowledge* is where empowerment resides.

Interestingly, along this view, middle school students in our sample appeared the most empowered. Granted, they had limited understanding. But they were prepared, and eager to act. Middle school students may benefit from interventions that elevate awareness and support reflecting on actions they are excitedly taking. Older students and adults in our sample were more likely to present as anguished or detached. High school students, who may tend towards the 'data meh', may nevertheless be receptive to experiences that are highly aligned with their passions and interests. In other words, these experiences should focus on the will to act. For adults, it is particularly important to recognize that gaps between awareness and action can lead to cognitive dissonance, guilt, or detachment. To anticipate this, high school and college educators might want to think about success stories where organized individuals have been able to push back against privacy-compromising practices by the government or private interests. Although it tends to be treated rather dryly, the 2018 European legislation on General Data Protection Regulation (GDPR; https://gdpr.eu/) may be amenable to more motivating examples.

Recognizing a more expansive definition of data

Participants in this study tended to begin conversations about data by alluding to statistics and scientific inquiry, but, over time, they revealed broader conceptions of data (e.g., resources for solving problems and understanding human experience) than prior literature has suggested. These findings are significant for several reasons. First, they show that individuals—students in particular—do tend to recognize the broader role that data plays in their lives. However, these definitions are usually not top of mind compared to what they learn in math or science class. Educators can play an important role in helping students become more aware of the impact that data has in their lives outside of school, its influence on what they do and the decisions they make, and the ways it can be a source of empowerment.

One way that education researchers have revealed diverse sources and types of data is by having students collect and analyze their own data. For example, in 'Quantified Recess', Lee and Drake (2013) prompted pairs of fifth grade students to develop summary and comparative measures using their own step counts from Fitbit fitness trackers. Taylor (2017) designed the 'counter-mapping the neighborhood' activity in which high school students used geospatial technologies on their mobile phones to gather data and imagine how their neighborhoods could change in the future. Bergner and colleagues (2020) examined the potential for quantitative literacy development when competitive high-school step dancers had access to motion and audio capture data from their own routines. These are just a few examples of projects that help students see themselves as both producers and consumers of data. When engaging with their own student-generated data, increased competence is accompanied by increased ability to act. These formative experiences can thus have an important role in cultivating a sense of empowerment.

Limitations and future directions

Our purposive sample included participants from high- and low-income backgrounds who identify as Black, Hispanic, and Asian (as well as immigrants to the US), but most of the sample was White (79%). Research with a larger sample of minority participants is needed to further explore topics around data and trust in institutions. Findings in this study, for example, point to a high degree of 'blind trust' in third parties collecting personal data. A majority-minority sample might uncover more nuance regarding the relationship between individual experience, data literacy, and trust in parties collecting and making use of personal data. One might hypothesize that groups historically disadvantaged by mainstream institutions may be less trusting of government and private third parties collecting and using their data. Design-based research paired with data literacy interventions for different age groups would also be a valuable next step to building on and testing some of these initial findings.

Conclusion

This study set out to explore the relationship between data, data literacy, and empowerment from the perspective of students and adults at various stages of life. In synthesizing prior frameworks from media, information, and statistical literacy communities and combining those perspectives with findings from personal interviews, we found that individuals rarely use empowerment language, although they do often see data as a valuable resource. Adult respondents recognize, to varying degrees, that extracting value from data requires skill and that controlling who extracts value from one's personal data traces requires vigilance. However, action does not increase in tandem with awareness. Some respondents, especially younger ones, may be feel exuberant about the value of data to themselves and to society. Knowledgeable adults may feel guilty, anxious, or willfully over-optimistic about their actions not taken. The work undertaken in this study was part of a broader effort to understand data literacy skills and dispositions, including awareness as well as usefulness and self-efficacy. The development of a quantitative instrument for assessing these factors on a larger scale is the subject of ongoing research. Larger-sample research will help inform on the relationship between knowledge, beliefs, and action with respect to personal data practices.

References

Adlin, T., & Pruitt, J. (2010). *The essential persona lifecycle: Your guide to building and using personas*. Morgan Kaufmann.

Armat, M., Assarroudi, A., Rad, M, Sharifi, H, & Heydari, A. (2018). Inductive and Deductive: Ambiguous Labels in Qualitative Content Analysis. *The Qualitative Report*, 23(1), 219-221.

Baack, S. (2015). Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation, and journalism. *Big Data & Society*. https://doi.org/10.1177/2053951715594634

Bergner, Y., Mund, S., Chen, O., & Payne, W. (2020). Leveraging interest-driven embodied practices to build quantitative literacies: A case study using motion and audio capture from dance. *Educational Technology Research and Development*, 1-24.

Bhargava, R., Deahl, E., Letouzé, E., Noonan, A., Sangokoya, D., & Shoup, N. (2015). Beyond data literacy: Reinventing community engagement and empowerment in the age of data. Data-Pop Alliance White Paper Series. Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Lad and Overseas Development Institute) and Internews.

Bowler, L., Acker, A., Jeng, W., & Chi, Y. (2017). 'It lives all around us': Aspects of data literacy in teen's lives. *Proceedings of the Association for Information Science and Technology*, 54(1), 27-35.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77-101.

Bulger, M., & Davison, P. (2018). The promises, challenges, and futures of media literacy. *Journal of Media Literacy Education*, *10*(1), 1-21.

Dalton, C. M., Taylor, L., & Thatcher, J. (2016). Critical data studies: A dialog on data and space. *Big Data & Society*, *3*(1), 2053951716648346.

D'Ignazio, C. (2017). Creative data literacy. *Information Design Journal*, 23(1), 6-18.

Dupree, J. L., Devries, R., Berry, D. M., & Lank, E. (2016). Privacy personas: Clustering users via attitudes and behaviors toward security practices. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5228-5239).

Francois, K., Monteiro, C., & Allo, P. (2020). BIG-DATA LITERACY AS A NEW VOCATION FOR STATISTICAL LITERACY. *Statistics Education Research Journal*, 19(1).

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International statistical review*, 70(1), 1-25.

Gebre, E. H. (2018). Young Adults' Understanding and Use of Data: Insights for Fostering Secondary School Students' Data Literacy. *Canadian Journal of Science, Mathematics and Technology Education*, 18(4), 330-341.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, 16(1), 22-25.

Hiroto, D. S., & Seligman, M. E. (1975). Generality of learned helplessness in man. *Journal of personality and social psychology,* 31(2), 311.

Hoofnagle, C. J., King, J., Li, S., & Turow, J. (2010). How different are young adults from older adults when it comes to information privacy attitudes and policies?. Available at SSRN 1589864.

Jenkins, H. (2006). *Fans, bloggers, and gamers: Media consumers in a digital age.* New York University Press.

Koltay, T. (2015). Data literacy: In search of a name and identity. *Journal of Documentation, 71*(2), 401-415.

Lee, V. R., & Drake, J. (2013, June). Quantified recess: Design of an activity for elementary students involving analyses of their own movement data. *In Proceedings of the 12th international conference on interaction design and children* (pp. 273-276)

Lehrer, R., Kim, M. J., & Jones, R. S. (2011). Developing conceptions of statistics by designing measures of distribution. *ZDM*, 43(5), 723-736.

Livingstone, S, Van Couvering, E & Thumim, N. (2008). Converging traditions of research on media and information literacies: Disciplinary, critical, and methodological issues. *Handbook of Research on New Literacies*. New York, USA: Routledge, 2008, pp. 103-132.

Matthews, P. (2016). Data literacy conceptions, community capabilities. *The Journal of Community Informatics*, 12(3).

McGrath, A. (2017). Dealing with dissonance: A review of cognitive dissonance reduction. *Social and Personality Psychology Compass*, 11(12), e12362.

Mérola, V., & Hitt, M. P. (2015). Numeracy and the persuasive effect of policy information and party cues. *Public Opinion* Quarterly, 80(2), 554-562.

Pangrazio, L., & Selwyn, N. (2019). 'Personal data literacies': A critical literacies approach to enhancing understandings of personal digital data. *New Media & Society*, 21(2), 419-437.

Philip, T. M., Olivares-Pasillas, M. C., & Rocha, J. (2016). Becoming racially literate about data and data-literate about race: Data visualizations in the classroom as a site of racial-ideological micro-contestations. *Cognition and Instruction*, 34(4), 361-388.

Prado, J. C., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri*, 63(2), 123-134.

Resnick, I., Kastens, K. A., & Shipley, T. F. (2018). How students reason about visualizations from large professionally collected data sets: A study of students approaching the threshold of data proficiency. *Journal of Geoscience Education*, 66(1), 55-76.

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review,* 84(3), 528-549.

Schield, M. (2004). Information literacy, statistical literacy and data literacy. In *Iassist Quarterly* (IQ).

Selwyn, N., & Pangrazio, L. (2018). Doing data differently? Developing personal data tactics and strategies amongst young mobile media users. *Big Data & Society*, 5(1), 2053951718765021.

Storksdieck, M. (2016). Critical information literacy as core skill for lifelong STEM learning in the 21st century: reflections on the desirability and feasibility for widespread science media education. *Cultural studies of science education*, *11*(1), 167-182.

Twidale, M.B., Blake, C. & Gant, J. (2013). Towards a data literate citizenry. *iConference 2013 Proceedings* (pp. 247-257).

Tygel, A. F., & Kirsch, R. (2015). Contributions of Paulo Freire for a critical data literacy. In *Web Science 2015 Workshop on Data Literacy*.

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).

Appendix A

INTERVIEW PROTOCOL

Thank you for taking the time to speak today! It is widely recognized that data literacy is becoming an essential skill for the 21st century. As education researchers, we are trying to learn more about how people think about data and how data literacy develops. The goal of the conversation is for you to share your opinions and understand how you think, not to test what you know. The interview will be recorded so I can refer back to it but all data will be anonymous. Do you have any questions before we start?

Part 1: Defining Data

1. What do you think of when you hear the word data?
a. What are some examples of that data?
b. Are there other definitions of data you can think of?
c. Where do data come from?
2. Who produces data?
3. What do you think of when you hear the term 'collecting data'?
4. What do you think of when you hear the term 'using data'?
5. Who uses data?
6. What is data used for?

Part 2: Engaging with Data

1. In what ways do you use data in your every-day life?
2. What tools do you use to make sense of that data?
3. In what ways are data helpful to you?
4. In what ways do data pose a challenge?
5. Do you see data as a good thing or a bad thing or some combination of the two?
6. When are you more likely to trust data or interpretations made about the data? When are you less so?
7. Describe what you do when you are reading an article or listening to a news story and you come across a graph or a statistic
a. What kinds of questions might you ask?
b. What kinds of sources are you most likely to trust?
c. Which kinds of sources might you be skeptical of?

Part 3: Personal Data Management

1. Who collects your data?
2. Who uses your data?
3. How do you feel about websites collecting and/or using your data? How do your friends feel?

a. Do you know what data they collect?
b. Do you know what they do with this data?
c. Do you feel that consent should be required?
d. Who should have access to that data?
4. Often websites or browsers will collect your data and use it in ways that improve convenience for you, such as storing your password for easy log-in or browsing history.
a. What do you think about this tradeoff?
b. What would you be willing to give up for convenience? For privacy?
5. Who do you think should have access to government data?
a. Have you ever accessed government data?
b. How might government data be useful to the general public?
6. Do you feel the need to protect your data?
a. If so, what steps do you take to do so?
b. Do you feel you are doing enough to protect your data?
c. Do you feel the government or companies are doing enough to protect your data?
7. Do you feel you know as much as you want to know about how your personal data are used?
a. Would you want to know more about how your data are used?
b. Would you want to act differently?
8. If you had the opportunity to set policy on this, what policies would you put in place? Where would you draw the line?
9. How has your perspective on all this changed over time?

Part 4: Statistical Thinking

1. Draw a picture of what a correlation looks like on a graph
2. Does the term *correlation does not equal causation* mean anything to you?
3. Are you familiar with the term statistically significant? What does it mean?
4. A 2019 report states that the average American household has 1.93 children. Can you explain this number?

Appendix B

CODEBOOK AND EXAMPLES

Code Category	Description of Category	Example Subcategory	Excerpt Example
Definitions of Data	Predetermined definitions of data derived from literature review	Data as digital bandwidth	*'Well, on a phone like this, the phone has data on it. I know I can do things on it with the data.'*
		Data as facts to back up a claim	*'I think of evidence based solutions and stuff. And like using real numbers to justify the way we think.'*
Data Literacy Perspective s	Predetermined data literacy perspectives adapted from Gebre (2018)'s data literacy framework	Competency perspective	*'It's always been much more like: make a spreadsheet and fill in whatever you want.'*
		Empowerment perspective	*'But I am, I wouldn't be able to make the best decisions for me without data. So I acknowledge that that's important for my personal empowerment.'*
Personal Data Practices	Individual uses of data and practices around personal data management	Relying on blissful ignorance (practices around trust)	*'It's a combination of fatigue and also it's like ignorance. Ignorance being bliss or just kind of saying, if I don't need to know it's fine.'*
		Taking steps to manage data privacy (protecting personal data)	*'I used a browser installed through Chrome because I didn't want that data on Facebook and I wanted to be taken out of their servers.'*
Attitudes Towards Data	Attitudes and beliefs about the usefulness of data and opinions on concepts surrounding data, including privacy,	Data as ultimately a good thing (overall evaluation of data)	*'While I don't deny that data can be used in a bad way or for an evil purpose, overall I think having more data is always better than having less data; having more data leads to greater transparency.'*

	convenience, and ownership	Collection of digital traces as creepy/invasive (stance on collection of digital traces)	*'I noticed that the advertisement started to be directed exactly to what I was looking at. And I was like, oh my goodness YouTube knows what I looked at yesterday. And that was a little scary at first, you know, that's a little invasive.'*
Individual Traits	Personal traits that influence attitudes about data and personal data practices, such as trust and self-efficacy	High trust in institutions (trust)	*'If it's on like New York Times or some article like that or like something government approved. Then I'll probably trust it, because there's like a lot of reasons why I should trust it and not that many why I shouldn't.'*
		Low data literacy self-efficacy (self-efficacy)	*'Just because I don't really go into data that much. I kind of know the basics of data. But that's just kind of how I look at data and what it means. Um, yeah, I just think I don't know data.'*
Errors and Misconceptions	Misconceptions around statistical concepts or data practices and constraints	Incomplete understanding of statistical significance (statistics concepts)	*'It's notable data that should be taken into consideration. I'm not sure exactly how the intricacies work because I'm not super well versed in it, but I am at least a bit familiar with that.'*
		Confusing producing and collecting data (data collection, analysis, and interpretation)	*'Depending on the job I'd say. People that are in math probably produce data and have data. Scientists have a lot of data. Anyone trying to figure out a percentage of something, which one higher, lower, better, worst producing data'*
Conceptions of Data	*Individual definitions of what is data and the different types*	*Data as part of everything we do* (nature of data)	*'But data is just essentially something that's created in everyday life. Data is created by*

	and concepts that comprise data		*living and I feel like I've also sort of centered on the human side of things, but data comes from anything that is alive in our world.'*
		Data as a way to answer a question/make a decision (uses of data)	*'Well, it's helpful for a lot of choices you make. Just because you obviously want to pick the better one. And if you have enough data, you can figure out which team to choose or anything, really.'*