NOTE: This is a pre-print accepted manuscript, accepted for publication as an entry in the forthcoming Encyclopedia of Social Justice, currently expected to appear in early 2025.

Please cite as: Reigh, E., Arastoopour Irgens, G., McBride, C., Quiterio, A., & Wilkerson, M. H. (In Press). Critical data literacies. To appear in Cortez, A. & Lizarraga, J. (Eds.) Encyclopedia of social justice in education - language, literacy, youth and culture. Bloomsbery Publishing.

Critical Data Literacies

Keywords: critical literacies; data literacy; statistics education, artificial intelligence; machine learning; algorithmic justice

Critical data literacies are practices that involve reading, writing, visualizing, and critiquing data in ways that: (1) recognize the inherently human dimensions of data as socially constructed; and (2) leverage data to identify and address injustices in our social worlds.

Although critical approaches to data are sometimes framed as new in educational research, scholars and activists of color have long used data to expose and dismantle racist policies and social structures. At the World's Fair in Paris in 1900, the African American sociologist Dr. W.E.B DuBois and his collaborators presented photographs and visualizations of demographic data that illustrated the "color line," or the socially constructed spaces that separate racial groups for the purposes of upholding white supremacy (Battle-Baptiste and Rusert 2018). Today, critical race scholars continue this work by using data to fight against injustices such as gentrification, inequities in school funding, and environmental racism.

We define *data* as any systematic observation or measurement that has been recorded and can be used for analysis. While there are many examples of data collection in human history, modern data infrastructures have expanded the ways in which we collect, share, and use data. A quantitative, or numeric, dataset can be as small as a single recorded observation, such as a tally mark carved into a stone plate. In contrast, a dataset can also contain a seemingly endless number of data points, like an algorithmically generated database of a billion global events geotagged with location coordinates. Data can also be qualitative, or non-numeric, including descriptions of social behavior, stories, social media posts, or images. Datasets are often created and/or used to gain insight into patterns of typicality and variability within a system of interest. Increasingly, datasets are being used to train artificial intelligence systems that reproduce or replace human judgements.

Whether quantitative or qualitative, datasets are socially constructed. They are created and transformed through the lens of one's personal and community connections, cultural practices, and the social and political milieu (Lee, Wilkerson and Lanouette 2021). Datasets are therefore shaped both intentionally and unintentionally by social context and can embed, (re)produce, and advance systems of oppression (Noble 2018). Engaging critically with data requires complementing it with social and historical information, particularly when analyzing racialized phenomena (Gillborn, Wilmington, and Demack 2018). Critical data literacy practices

NOTE: This is a pre-print accepted manuscript, accepted for publication as an entry in the forthcoming Encyclopedia of Social Justice, currently expected to appear in early 2025.

Please cite as: Reigh, E., Arastoopour Irgens, G., McBride, C., Quiterio, A., & Wilkerson, M. H. (In Press). Critical data literacies. To appear in Cortez, A. & Lizarraga, J. (Eds.) Encyclopedia of social justice in education - language, literacy, youth and culture. Bloomsbery Publishing.
involve examining datasets' contexts, purposes, and histories of production. Critical data literacy practices also involve using data and data-driven systems to dismantle oppression and work towards justice (D'Ignazio and Klein 2020).

What are critical data literacies?

To conceptualize critical data literacies, we draw from the well-established field of critical literacy, which attends to the ideological functions of text and interrogates relationships between authors and readers (Freire 1970). In this tradition, data and data products are viewed as *texts* that are actively constructed by their authors and re-interpreted by others. The plural literac*ies* emphasizes the interplay of different forms of sensemaking that span multiple practices and semiotic domains (e.g., data visualizations, verbal descriptions, maps). Considering datasets to be texts underscores that they do not neutrally mirror the world; rather, they are integrated into complex activity systems that involve humans and their sociocultural tools.

Datasets exist in complex ecosystems where they can be re-used and re-interpreted by others. For example, many datasets that have been collected by researchers, private industries, or state/national organizations have been made publicly available for further analysis and investigation. Such datasets are often deeply personal and are simultaneously entangled within existing power structures. For instance, data collected from "public" social media networks describe a user's personal activities, such as their communications with loved ones or details of personal events. However, these datasets can be also used by companies, governments and other institutions, often without the user's knowledge, to feed algorithms that populate newsfeeds, inform automated chat agents, and predict social behavior.

The field of critical literacies offers analytical tools to deconstruct the features of texts and to explore how they influence others. Examples of these tools include interrogating who collected data and for what purpose, what information or perspectives might be missing from data, and how data would be different if it were collected by someone else. Our personal data may influence what we see online and what others know about us, in ways that may or may not serve our own interests. The ability to make agentive decisions about personal data is central to critical data literacies (Pangrazio and Selwyn 2019).

In his work with marginalized communities, Freire (1970) developed an approach to critical literacy that engages learners in the processes of "reading the world" of their own materials and social lives, including the oppressive power relationships therein, and "reading the word," or engaging with texts. These mutually constitutive processes of critical reflection allow learners to construct new ways to think about and act in the world, both individually and

NOTE: This is a pre-print accepted manuscript, accepted for publication as an entry in the forthcoming Encyclopedia of Social Justice, currently expected to appear in early 2025.

Please cite as: Reigh, E., Arastoopour Irgens, G., McBride, C., Quiterio, A., & Wilkerson, M. H. (In Press). Critical data literacies. To appear in Cortez, A. & Lizarraga, J. (Eds.) Encyclopedia of social justice in education - language, literacy, youth and culture. Bloomsbery Publishing.
collectively. Thus, the development of literacy practices serves the broader purposes of self-examination, transforming one's relationship to the world, and dismantling unjust systems.

Similarly, a critical approach to data involves relating datasets to one's own lived experiences or those of others. For instance, data collection about COVID-19 infection rates at the onset of the 2020 pandemic often did not consider race as a variable. As early anecdotal evidence emerged of disproportionate infection rates along racial lines, affected communities advocated for states to collect racial data related to COVID infections and outcomes. Subsequent analyses confirmed that communities of color were disproportionately impacted, a finding that advocacy groups leveraged to lobby for more resources in these communities.

In another example, Dr. Safiya Umoja Noble discovered that the internet search algorithms that distribute knowledge to billions of people daily can reinforce racial and gender discrimination at large scales. She described an example that took place in 2011, in which she searched for "Black girls" to find activities for her nieces. Instead, the search algorithms displayed stereotyped images and explicit content. Dr. Noble challenges the idea that internet search engines and databases are neutral, arguing that the data used to train the algorithms reflects and reproduces social and historical inequities. This form of algorithmic oppression is perpetuated by people and corporations who are in positions of power.

Importantly, critical data literacy involves not only identifying such inequities, but also addressing them through data. For example, systems can be more fairly built by drawing on different communities' experiences to create inclusive and representative datasets. Desmond Patton worked with youth who were formerly involved in gangs to analyze social media data to understand violence in Chicago. Their analysis method generated new training datasets using a subset of social media posts from African American men in Chicago that were interpreted and coded based on the youths' knowledge of the local context (described in D'Ignazio and Klein 2020). This example shows how erroneous judgements based on limited attention to the experiences of those most affected can be avoided, and how datasets can be developed from the start in ways that challenge oppression and work towards co-liberation.

How do individuals engage in critical data literacies?

Criticality is not simply an add-on at the beginning or end of one's work with data; it should be infused throughout the life cycle of a dataset. We adopt Wild and Pfannkuch's *Problem, Planning, Data, Analysis, and Conclusion* (PPDAC) cycle (cited in Louie 2022), an organizational model that coheres "fragments" of thinking in statistical investigations. While the cycle's original conceptualization encouraged a generally critical approach to data, we extend it to explicitly foreground power across the entirety of the data lifecycle to identify and

NOTE: This is a pre-print accepted manuscript, accepted for publication as an entry in the forthcoming Encyclopedia of Social Justice, currently expected to appear in early 2025.

Please cite as: Reigh, E., Arastoopour Irgens, G., McBride, C., Quiterio, A., & Wilkerson, M. H. (In Press). Critical data literacies. To appear in Cortez, A. & Lizarraga, J. (Eds.) Encyclopedia of social justice in education - language, literacy, youth and culture. Bloomsbery Publishing.
address oppressive structures. We illustrate our approach using an example in which learners explore local air quality through geotagged monitors (described in Louie 2022).

Defining a *Problem* involves identifying the motivations to explore a given topic using data. For example, learners may seek to investigate air quality in their community because of health issues due to pollution, such as asthma. However, posing questions about inequality or other social issues does not automatically make a data investigation critical. These problems must include analyses of the ways that power structures operate within the phenomenon of interest. For example, learners exploring air quality may notice that polluting industries are concentrated in certain parts of a city. This recognition may lead them to then ask when and why the industries were placed there, which communities are impacted, and how the pollution affects health. Defining a problem critically, then, involves asking: "What is the history of this problem or topic and how does it impact the data?"; "What economic, political, and state/national interests are connected to this problem or topic?"; "Which people, communities, and environments are impacted by this issue, and how should this inform our investigation?"; and, "What aspects of this problem cannot be understood through data alone?"

Planning involves choosing data sources and analytical methods to address the problem. An investigator may collect their own data, or they can select and evaluate existing datasets. Investigators should consider the benefits and drawbacks of different sampling procedures, types of measurements, and data management practices. Then, statistical methods must be chosen to address the questions. Critical approaches to planning involve questioning the purposes and positionality of those who create the datasets, the statistical methods and algorithms that are used, and how different methodologies may privilege certain perspectives while silencing others. For instance, learners exploring air quality may choose to collect data at local playgrounds, schools, and other locations frequented by children, thus missing disproportionate risks to lower-income adults working in certain industries. Critical planning thus involves asking: "Who collected this data and why?"; "What perspectives are captured and missing from our plan?"; and, "How would this data look different if collected by someone else, in another place, or at a different time?" The planning phase also raises important questions about consent and privacy. While collecting geotagged air quality data, learners' location may be recorded. Critical practices related to privacy include asking "Did individuals give informed consent for this data to be collected?"; "What data is being collected about me?"; and "How can I protect my privacy?"

The *Data* phase of statistical inquiry involves generating a dataset – or evaluating existing datasets for their alignment with the research question – and preparing the data for analysis. Whether a dataset is newly created or sourced, it may require additional preparation,

NOTE: This is a pre-print accepted manuscript, accepted for publication as an entry in the forthcoming Encyclopedia of Social Justice, currently expected to appear in early 2025.

Please cite as: Reigh, E., Arastoopour Irgens, G., McBride, C., Quiterio, A., & Wilkerson, M. H. (In Press). Critical data literacies. To appear in Cortez, A. & Lizarraga, J. (Eds.) Encyclopedia of social justice in education - language, literacy, youth and culture. Bloomsbery Publishing.

such as removing incomplete records, reclassifying data, computing new values, and restructuring the dataset to allow comparisons or analyses of trends. Each of these transformations carry implications for what is foregrounded in analyses. For the air quality example, learners may find themselves selecting between datasets that include information on race/ethnicity, or income. Selecting and transforming data is explicitly political; focusing on only income, for example, may obscure the role of racist institutions in the inequitable distribution of environmental stressors. Similarly, learners may need to select between racial categories; broader categories can allow for a more valid statistical analysis, but can also present groups as monolithic and hide important internal differences. In this way, a critical approach to the data phase involves asking "What variability and outliers may be highlighted or obscured by the structure of my data and my choice of analysis methods?" When creating training datasets or leveraging simulation methods for statistical analysis as part of this phase, another critical question is "What are the histories, cultures, and contexts of the training texts?"

The *Analysis* phase involves describing regularity, variability, and trends in the data and connecting the data to real-world phenomena. Variables in a dataset can be compared to understand the relationships between them, or modeled using simulations or predictive algorithms. Moreover, particular data points may be isolated to investigate their features in relation to others. While distinguishing correlation from causation is a common concern in data analysis practice, critical analyses further distinguish between levels of analysis—such as between personal attributes and actions and collective/institutional actions or policies, and between snapshots of time and longer social histories. For example, learners investigating air quality may find that communities of color have higher asthma rates than white communities. Rather than assume this is the result of only personal behaviors or risk factors, learners should consider additional variables, such as pollution burden on these communities stemming from racist regional planning practices and inequitable access to preventative medical care. Assuming a critical perspective in analysis involves asking "How might histories of this phenomenon add nuance to this data?"; "What might be the structural reason for this pattern?"; and "What aspects of history, social context, and power can be examined by cases, trends, or patterns?"

Finally, the *Conclusion* phase of the PPDAC cycle involves choosing which results to share with others and how to communicate those results. Conclusions may be supported with other types of information, like narrative vignettes or social analysis. In the air quality example, learners could integrate historical and political analysis to draw conclusions about why communities of color have been disproportionately impacted. Or, they could highlight a particular community to elaborate on the unique features of the social, political, and geographic

NOTE: This is a pre-print accepted manuscript, accepted for publication as an entry in the forthcoming Encyclopedia of Social Justice, currently expected to appear in early 2025.

Please cite as: Reigh, E., Arastoopour Irgens, G., McBride, C., Quiterio, A., & Wilkerson, M. H. (In Press). Critical data literacies. To appear in Cortez, A. & Lizarraga, J. (Eds.) Encyclopedia of social justice in education - language, literacy, youth and culture. Bloomsbery Publishing.
context. Going further, they might compare areas with similar geography but different levels of risk, asking critical questions about the systems that protected one community but not another. These questions might include "Is sufficient social and historical context given to understand this phenomenon?"; "Who is impacted by the conclusions of this analysis or the outputs of these data-driven algorithms?"; and "Are the positionality and interests of the authors disclosed, and how might they influence the analysis?"

While our examples and questions above are not exhaustive, they demonstrate how critical practices can enhance the full data life cycle. Leveraging the PPDAC cycle and presenting specific questions to be addressed at each phase highlights many concrete opportunities to engage learners in critical practices when working with data.

How do we educate toward critical data literacies?

In recent years, numerous initiatives have investigated curriculum that provides opportunities for learners to engage in critical practices (Arastoopour Irgens et al. 2022; Lee, Wilkerson, and Lanouette 2020; Louie 2022, Reigh et al. 2022). Researchers from the Writing Data Stories project designed a middle-school data science workshop for learners to explore environmental racism (Reigh et al. 2022)–the disproportionate impact of pollution and other environmental stressors on communities of color. In the workshop, learners generate a collective definition of environmental racism by looking at the similarities and differences in a range of case studies from around the world. Then, they engage with a local case study of West Oakland, where a thriving African American community was uprooted in the 1950s through eminent domain to construct overland public transportation and highways. Learners study the history of this area through journalistic accounts and the stories of local activists, and put these sources in conversation with recent data on environmental and social factors, such as air pollution and asthma rates. As a culminating project, learners generate a research question about an environmental injustice of their own choosing. They conduct interviews and other forms of journalistic research, and use what they learn to explore quantitative datasets using the CODAP data analysis tool. Finally, learners create a multimedia "data story" that integrates the learners' own motivations and experiences with different types of data to make an argument about an environmental injustice, explore its causes, and propose possible solutions.

Arastoopour Irgens and colleagues (2022) co-designed a set of Critical Machine Learning (CML) activities for upper elementary and middle school aged children. Learners begin by creating physical representations of algorithms and reflecting on how algorithms are created by humans. Pivoting to digital technologies, learners then explore machine learning through image searches on Google, reflecting on representation and bias in searches such as "basketball

NOTE: This is a pre-print accepted manuscript, accepted for publication as an entry in the forthcoming Encyclopedia of Social Justice, currently expected to appear in early 2025.

Please cite as: Reigh, E., Arastoopour Irgens, G., McBride, C., Quiterio, A., & Wilkerson, M. H. (In Press). Critical data literacies. To appear in Cortez, A. & Lizarraga, J. (Eds.) Encyclopedia of social justice in education - language, literacy, youth and culture. Bloomsbery Publishing.
player" or "computer scientist." Next, learners test out a tool that uses machine learning algorithms to guess a user's drawing. Children learn the concepts of training data, input, and output and also reflect on how input data can be biased and exclude certain populations. Next, learners make their own image classification system using a webcam and Google's Teachable Machine. Learners are unknowingly provided with a set of images that are biased, and they then explore the source of bias in the training set and reflect on the consequences of its use. Last, learners engage in a guided discussion with adult facilitators about discrimination in facial recognition systems and systemic racism. After participating in this set of reflective activities, learners are invited to design and program their own machine learning robots for social good. They create a fantastical visual story about a "superhero" robot that can help people, and reflect on how it can be designed in responsible ways to mitigate bias. Children create a prototype using Scratch block-based programming, Google Teachable Machine, and a micro:bit powered robot. As a group, they decide how they would like to showcase their robots to the broader public.

Conclusion

Among increasing calls for "data literacy education" at the K-12 level, focusing on *critical* data literacies draws from a scholarly tradition in literacy studies and critical pedagogy to deepen how teachers engage learners with data. This approach offers concrete tools for learners to leverage data in ways that identify and address injustice, at a time when data and data-driven systems are accelerating the reproduction of social inequity.

Author's Note

We rely on literature reviews and secondary sources that represent the state-of-the-art in data literacy education. We hope this decision will give readers exposure to current conversations as well as access to foundational works cited by those review pieces.

References

Arastoopour Irgens, Golnaz, Hazel Vega, Ibrahim Adisa, and Cinamon Bailey. "Characterizing children's conceptual knowledge and computational practices in a critical machine learning educational program." *International Journal of Child-Computer Interaction* 34, no. 4 (2022).

Battle-Baptiste, Witney, and Britt Rusert, *W. E. B. Du Bois's Data Portraits: Visualizing Black America*. Hudson, NY: Princeton Architectural Press, 2018.

Noble, Safiya Umoja. *Algorithms of oppression*. New York: New York University Press, 2018.

NOTE: This is a pre-print accepted manuscript, accepted for publication as an entry in the forthcoming Encyclopedia of Social Justice, currently expected to appear in early 2025.

Please cite as: Reigh, E., Arastoopour Irgens, G., McBride, C., Quiterio, A., & Wilkerson, M. H. (In Press). Critical data literacies. To appear in Cortez, A. & Lizarraga, J. (Eds.) Encyclopedia of social justice in education - language, literacy, youth and culture. Bloomsbery Publishing.

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. MIT press, 2020.

Freire, Paulo. *Pedagogy of the Oppressed*. New York: Continuum, 1970.

Gillborn, David, Paul Warmington, and Sean Demack. "QuantCrit: education, policy,'Big Data' and principles for a critical race theory of statistics." *Race Ethnicity and Education* 21, no. 2 (2018): 158-179.

Lee, Victor R., Michelle Hoda Wilkerson, and Kathryn Lanouette. "A call for a humanistic stance toward K–12 data science education." *Educational Researcher* 50, no. 9 (2021): 664-672.

Louie, Josephine. "Critical data literacy: Creating a more just world with data." Commissioned paper for the National Academy of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6-12. Washington, DC, 2022.

Pangrazio, Luci, and Neil Selwyn. "'Personal data literacies': A critical literacies approach to enhancing understandings of personal digital data." *New Media & Society* 21, no. 2 (2019): 419-437.

Reigh, Emily, Meg Escudé, Michael Bakal, Edward Rivero, Xinyu Wei, Collette Roberto, Damaris Hernández, Amber Yada, Kris Gutiérrez, and Michelle Hoda Wilkerson. "Mapping Racespace: Data Stories as a Tool for Environmental and Spatial Justice." *Occasional Paper Series*, no. 48 (2022): 79-95.

Authors

Emily REIGH, School of Education, University of California, Berkeley, USA
Golnaz ARASTOOPOUR IRGENS, College of Education, Clemson University, USA
Cherise McBRIDE, School of Education, University of California, Berkeley, USA
Ashley QUITERIO, School of Education and Social Policy, Northwestern University, USA
Michelle Hoda WILKERSON, University of California, Berkeley, USA