ALL DATA IS BIASED

HEATHER KRAUSE

In this chapter, I am going to share some stories with you that show how the worst equity problem we are dealing with in data at the moment is that we are making prejudiced choices but don't understand how. Most of us are reading this because we know that math, science, and data can improve the world. One of the reasons many people like the idea of data in the mission-driven sector is that we believe data offers an objective, bias-free way to make decisions. I have good news and bad news for you. The bad news is that this is a data myth. At every single step of a data project, we are making choices. Choices about whose lived experience to center; choices about whose worldviews get prioritized; choices about who gets reflected in the work. The good news is that, once we move past this myth, we can get to some valuable, grounded work on using data for racial equity.

Here is my favorite story about making choices in the way we use data. What is the average number of students in these classrooms? There are three students in classroom A, six students in classroom B, and nine students in classroom C.

If you said the average classroom size is 6, you are right. If you said the average classroom size is 7, you are right. These answers use the same math; they just embed a different perspective. Let's look at the math from the teacher's perspective.

FIGURE 2.2.1 What is the average number of students across these three classrooms?

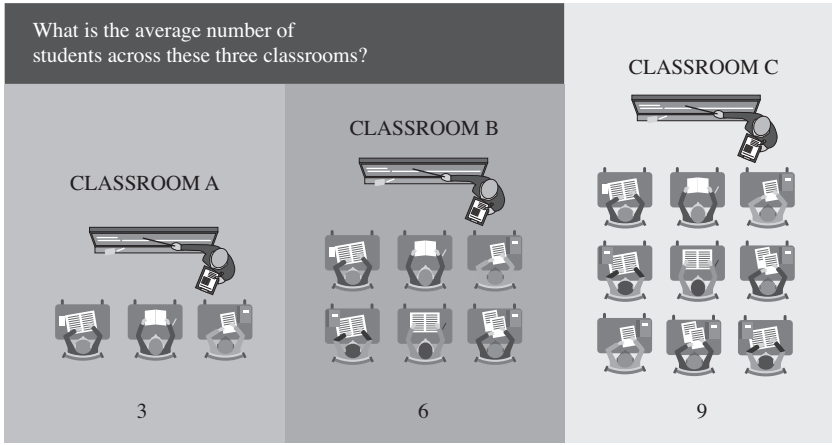
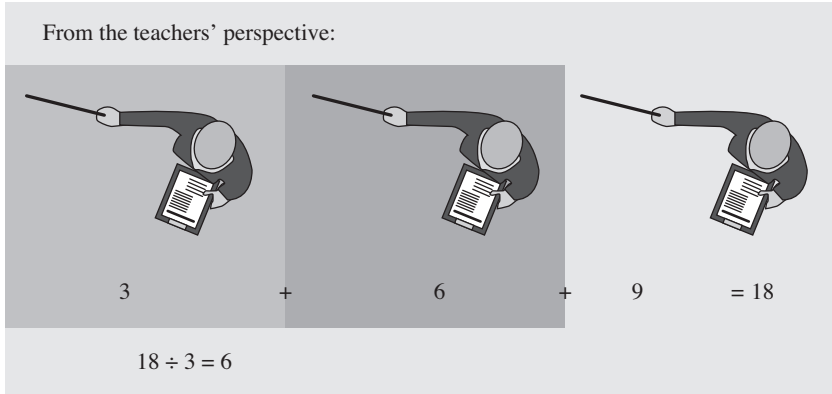


FIGURE 2.2.2 From the teacher's perspective

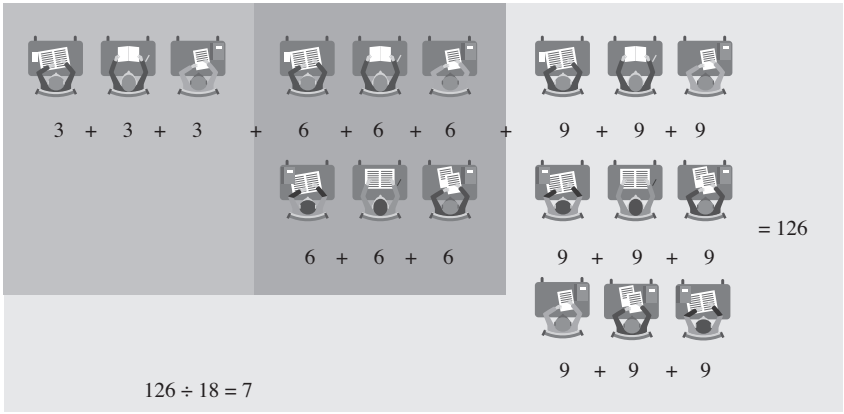


The first teacher takes a look around and sees three students, the next sees six students, and the last sees nine students. 3 + 6 + 9 is 18. Divided by the three teachers is six. The average students per classroom from the teachers' perspective is six students.

And this one is from the students' perspective.

In classroom A, each student counts three students in their classroom, including themself. In classroom B, each student sees six, including themself. And in classroom C, it is nine. Adding the total up and dividing by the

FIGURE 2.2.3 From the students' perspective



number of perspectives is eighteen students. We get an average classroom size of seven.

We do the math in exactly the same way. Both processes are valid; they just center a different lived experience. The way most of us are taught to think about math can make this example, by turns, confusing, enraging, or mind-blowing. You might need to actually get out some dolls and test it.

It is important to note that we are not using a different kind of math. Both means are calculated in the same way: the sum of the units divided by the number of units. We just had to make a choice about which unit to use, where to put the locus of power, whose experience to prioritize.

Many of you (myself included for most of my life) will have felt they did not make a choice, that this is just how math works. That is the most insidious myth with all data and research. The dominant perspective is so deeply ingrained in much of data and models that it seems like the only perspective—or no perspective at all.

Let's look at another example where the math is completely correct but there is a choice to be made. This graph is looking at outcomes in an income improvement project. In this graph, we clearly see that the people in the project had a huge average increase in their income. Project success!

But the people in this project are from three different zip codes. If we are interested in the equity between these groups, we might want to measure them individually to see how they contributed to the average.

Uh oh . . . though we have increased the average income significantly, it has not been the same for everyone. We can see from these different

FIGURE 2.2.4 Measuring a project's impact on average monthly income

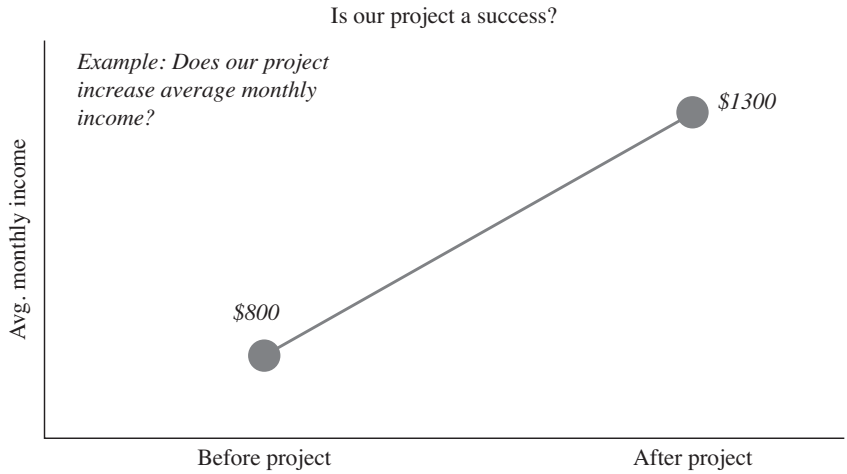


FIGURE 2.2.5 Measuring a project's impact on average monthly income by zip code

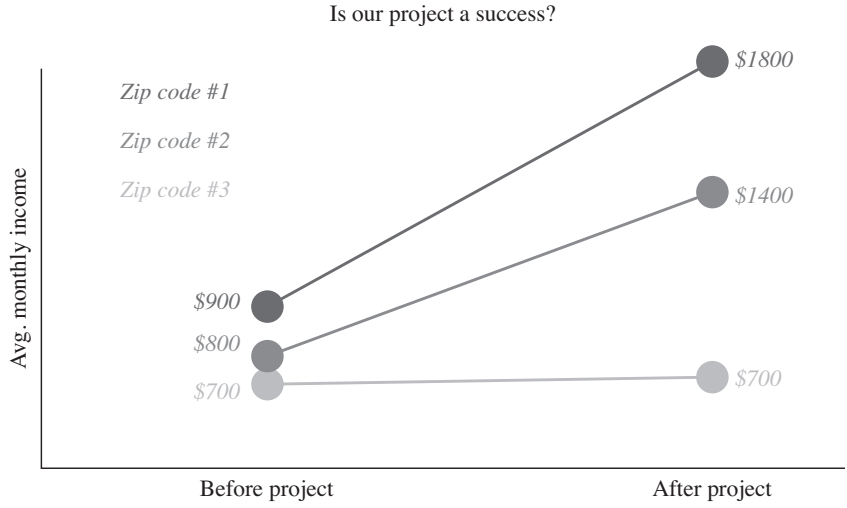
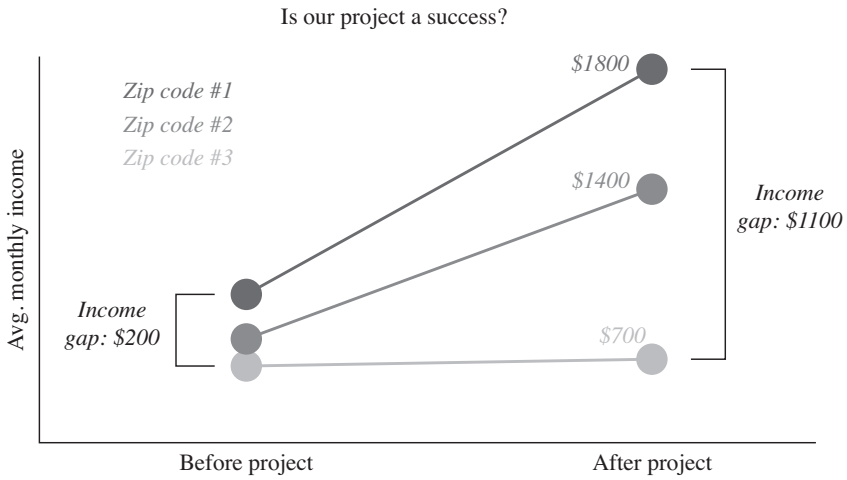


FIGURE 2.2.6 Did we increase or decrease the income gap?



zip code lines that the project has been a success for some people and not for others.

If we define success by "Did we increase or decrease the income gap?" then the project has failed badly. It is not that one of these is right or wrong; you may care only about the overall average, but the math is correct in both.

But it is not just a math problem. Put yourself in the shoes of someone from zip code #3. You have participated in a project, watched your neighbors' incomes rise, and heard the organization running the project using data to proclaim it a huge success. How do you feel? Probably, you are either ashamed because there must be something wrong with you or you are furious because you can see that the project did not work for you and the researchers did not include your perspective in the way they used their data.

In this case, we had a choice to make: What data should we use to measure success? All the work downstream of that choice, from the analysis to the design of the graphics, is affected by the equity implications of the initial choice.

We want to think of quantitative research as a situation in which we make one important choice, which research question to look at. And then we follow that research question through a trail of building an objective research design, collecting objective data, doing an objective analysis, and, then, hopefully, creating an objective data visualization.

Not so, though. This is a myth.

Instead, research, even, and sometimes especially, quantitative research, involves dozens and even hundreds of choices at every single step in the research process.

And there is no way to avoid making these choices. Our only option is to continue to hide behind a false narrative about "objective quantitative research" or "value-free evidence," or to figure out how to make choices in our research that better reflect the equity we want to embed.

Note that I did not say, "Now you have to learn how to make the right choice." There is no "right choice." There is no objective and bulletproof equitable data project. My clients come to me searching for that like it is the holy grail or the fountain of youth. Data projects can be intentional and transparent in their choices, but they cannot be objective or choice-free. Equity is a process, not a binary state, between equitable projects or inequitable projects.

Let's look at another time I was trying to use data for equity. I was working with a school district struggling with the way they used data about student outcomes and race. One of the issues that was particularly tense was the reporting on expulsion data. The data was being used to show that more Black and Latinx boys were being expelled from school than white boys. More often than not, this data was analyzed and displayed in a way that emphasized "the equity gap" between Black/Latinx boys and white boys.

The district wanted to improve both situations—the way they were using data about racial equity and the experiences these young men were having in school.

The district launched a project aimed at reducing the rate of expulsion of specific groups of young men. At the outset of the project, they established the research question as: "Has our initiative reduced the rate of Black and Latinx boys being expelled relative to white boys?" Unsurprisingly, this initiative, and its research, was not welcomed by the community.

When framing a research question, there are two key choices we make. The first is where we place the onus of change. The second is how we define success. In this case, the researchers had placed the onus to change on Black and Latinx boys and defined success as the rate of white boys. Neither of these choices was in alignment with the stated equity goals of the project. Essentially, this original research question can be boiled down to: "How good is our project at getting Black and Latinx boys to be like white boys."

To start making choices to align research with racial equity objectives, we needed to put the onus of change somewhere other than on the marginalized people, and define success differently. After conversation with the community and deeper reflections on the actual equity goals of the project, the research question was changed. The questions became: "Has our initiative disrupted the processes in our district that are most strongly related with us pushing out Black and Latinx boys?" and "Has our initiative improved the school characteristics that are most strongly related with creating environments that encourage Black and Latinx boys to fulfill preexisting desires to be in school?" These questions and the way they frame research were welcomed by the community. Data started to go from a weapon of disaggregation and separation to a tool that could be used to reach a common goal.

Even the smallest choices in the data process can have huge impacts. This example illustrates one of most important equity issues in research: there is a lot of power in getting to make these data choices. When we realize that data, evidence, and research are not completely objective processes, we discover they are a series of choices about whose lived experiences and worldviews we are going to center in the design, question, methodology, analysis, visualization, and more.

To equalize the power of these choices, you need to start by at least informing people in meaningful and useful ways that you are making them and explain your reasoning. This provides us all with the choice to agree or disagree, and is the gateway to getting better feedback and more nuanced perspectives.

Doing this involves vulnerability from usually privileged people and letting go of the power of the "black box" in your data process. This is a practical issue. If your data decisions are made under a veil of mathematical objectivity, "the data doesn't lie" kind of stuff, no one can even tell what your data actually means.

The truth is that even if you do not want to embed more equity in your data, it is about to be demanded of you. Research is losing its sheen of automatic objectivity. When you say this is how it is and our numbers do not lie, not everyone believes you. The kid in the crowded classroom does not agree. The project participant in the blue zip code does not think your project was a success. The Black and Latinx families do not want to participate in inequitable research. When our work does not match the lived reality of the very people the data comes from, people do not buy it, and they are right not to.

Let's talk a little bit more about feeling not seen in the data.

For example, if we are showing survey results about levels of satisfaction with our network of food banks and we have a large amount of data from white clients, a medium amount of data from Black clients, and a small amount of data from American Indian clients, we often don't even show the results from the American Indian respondents, because there are too few and, instead, we say those findings are "not statistically significant." We think we have to do this, because that is what we have been taught to do, but it is a choice. It is a choice with harmful equity consequences. It stops people from being counted, and in a data-based world, that is like saying they don't matter.

There is no math-based reason in this case that supports saying something with a small sample size is insignificant. It is a statement that is both technically and humanly incorrect. This is another data myth. It is a norm so entrenched that it feels like a rule. What we should be doing is talking about levels of uncertainty.

When we say "not statistically significant" in this case, what we mean is that we have a high level of uncertainty about this result. See how much less comfortable it is to say that? "We are uncertain" puts the responsibility where it should be, on us, the data analysts. It leads to the next natural questions of: "Why are you so uncertain about this group?" and "Could you have used a different way to be more equally certain about all the groups?"

"Not statistically significant," in this example, is a shield we hide behind instead of being transparent about our process and the meaning of our results. There are a thousand shields like it in data science. And people are figuring that out.

Data literacy and an understanding of the power structures involved in data is exploding. That is a great thing. The bar is being raised, and we need to rise up to it.

So, we have blown apart the myth that data is objective, that we can get to a "right" answer. We can get only to answers that reflect our inputs and the process we use. Our perspectives are the main shaping force behind those things. We see that our data is selective, our models are malleable, our results can be validly interpreted in more than one way, and almost all of our "data rules" are arbitrary and often unfair.

Should we abandon data and quantitative research? No. Can it still be used for good? Yes.

If we are willing to admit that we are making choices, then we can uncover them, improve them, and communicate them effectively. Then we

really can use data for good. We can estimate and quantify and understand things from an equity lens.

If we value equity in our policies, practices, and systems, it is essential that the next generation of practitioners be supported with tools and training that equips them to succeed in embedding equity in their work with data.

Here are five steps to get started:

1. Include in all trainings the essential task of recognizing that we are making subjective, human choices in our data work.
2. Develop research frameworks that identify as many choice points in the research and evidence creation process as possible. Each of these choice points is the place in which a practitioner can increase the equity and center the voice they intend to represent in their work.
3. Build into emerging research best practices expectations that these choices will be made transparent—both to research subjects and evidence consumers.
4. Teach the importance of statistical methods in aligning the perspectives of the community, the learning agenda, and the world views.
5. Learn to communicate in an accessible and transparent manner about the world views, lived experiences, and quantitative choices that have been used to build the evidence.

We need to talk about what choices we are making in data and why. This is the only way.

Sometimes, it feels like people are losing trust in "science," but, actually, they are losing trust in scientists. They are losing trust in the scientists who won't even hold themselves to the standard of a high school science project: being honest about what they do know and what they don't know, and showing their work about how they are making choices.

Many researchers, practitioners, and analysts are trying their hardest to be good and just, to add valuable, truthful information to what we know about the world. But we cannot hide from criticism behind the idea that all data is objective or that numbers do not lie. Our data reflects the way we see the world, but that is a good thing.

It means that, instead of unsuccessfully trying to pretend that there is no worldview in our data projects, we can acknowledge that there are many choice points at which we embed world views and perspectives in our data projects, and then we can make these choices with purpose.