Scholarly Assessment Reports

Carson, J. (2020). Quantification – Affordances and Limits. *Scholarly Assessment Reports*, 2(1): 8. DOI: https://doi.org/10.29024/sar.24

RESEARCH

Quantification – Affordances and Limits

John Carson
University of Michigan, US
jscarson@umich.edu

We live in a world awash in numbers. Tables, graphs, charts, Fitbit readouts, spreadsheets that overflow our screens no matter how large, economic forecasts, climate modeling, weather predictions, journal impact factors, H-indices, and the list could go on and on, still barely scratching the surface. We are measured, surveyed, and subject to constant surveillance, largely through the quantification of a dizzying array of features of ourselves and the world around us. This article draws on work in the history of the quantification and measurement of intelligence and other examples from the history of quantification to suggest that quantification and measurement should be seen not just as technical pursuits, but also as normative ones. Every act of seeing, whether through sight or numbers, is also an act of occlusion, of not-seeing. And every move to make decisions more orderly and rational by translating a question into numerical comparisons is also a move to render irrelevant and often invisible the factors that were not included. The reductions and simplifications quantifications rely on can without question bring great and important clarity, but always at a cost. Among the moral questions for the practitioner is not just whether that cost is justified, but, even more critically, who is being asked to pay it?

Keywords: Quantification; intelligence; measurement; archive; objectivity

Introduction: Quantification as Moral Endeavor

We live in a world awash in numbers. Tables, graphs, charts, Fitbit readouts, spreadsheets that seem never to stop, economic forecasts, climate modeling, weather predictions, journal impact factors, H-indices; the list could go on and on and it would still just scratch the surface. We are measured, surveyed, tracked, and subject to constant surveillance, largely through the quantification of a dizzying array of features of ourselves and the world around us, the ubiquitous 'Big Data' that algorithms then analyze to anticipate our needs, or perhaps bend our wants to their interests. Those working in the fields of scientometrics and informetrics already know this, of course. The power of numbers to illuminate, to reveal hidden connections, to make sense of complex and confusing processes with multiple causes and effects—this is the world these scholars inhabit and try to make sense of in their daily practice. They know that some researchers shape a field much more than others, that certain articles are fundamental while others come and go scarcely leaving an imprint, that some studies produce fundamental advances, while others seem largely to have squandered their resources. At the extremes, the seeing is easy: Nobel Prize winners on one end; those who manage one minor publication before moving along into other careers on the other. But in the middle, it can get very messy; there it is no easy task to determine who or what has had an important impact, and particularly to calculate persuasively the relative significance of one scholar or paper or research project versus another. The case of methodology papers and citation analysis is well known: an article consolidating the methodology for a basic experimental procedure can be cited endlessly in the relevant subfield (Aksnes, Langfeldt & Wouters, 2019; MacRoberts & MacRoberts, 1989). Whether the high citation number means the article is influential or simply convenient is another matter, and probably one less open to resolution through further statistical analysis than via professional judgment. Amassing the data, analyzing it, and employing expertise to draw conclusions about it, this is the practice that data scientists know well. What I would like to do here is to invite practitioners to step back from the challenges of doing quantification in order to engage more fully with the hopes and worries that for many doubtless lurk along the edges of their professional pursuits.

The hopes, it seems to me, are easy enough to articulate: that the researcher has gotten it right. That the quantifications the researcher has chosen, the data sets that have been built, the algorithms that have been employed all lead to conclusions that are not only accurate, but meaningful; that the researcher has seen through the chaos to the order lying buried inside or has constructed an order that is important to see. The worries are also readily specified: that the researcher is seeing the wrong thing, that for all of the sophistication and care employed to render the phenomenon

visible, the final result is not so much reflecting reality as trying to make it. Maybe the data set was too small or even too big; maybe the wrong things were quantified and included; maybe what was most important was simply resistant to quantification?

Does the higher impact factor of the *New England Journal of Medicine* (70.67 in 2018), for example, really mean it is a significantly more important journal than *The Lancet* (59.102 in 2018), especially when the two-year citation per document rating of the *The Lancet* (45.02 in 2018 vs 39.95 in 2018) is higher (Liao, 2019; Scimago, 2020)? And what are the criteria of importance for a journal anyway? The publication of the occasional path-breaking article or the continuous appearance of solid, important pieces, what the historian of science Thomas Kuhn called 'normal science,' that incrementally move a field forward (Kuhn, 1962)? Numbers and analysis may help answer these questions, but they also require professional judgment and trust. Indeed trust, as another historian of science Stephen Shapin has pointed out, is critical, if not always sufficiently acknowledged: trust in colleagues to have done their data collection and analyses correctly, trust in instruments and machines (including computers) to have worked correctly, trust in the coding of the raw material, trust that fatigue or distraction or bias was not at play when aggregating or manipulating or evaluating data, and even trust in the scientists/scholars whose work is being analyzed to follow the appropriate citation practices and thus make their influences visible (Shapin, 1995).

My point is simple: quantification can never stand alone. The allure of what historians of science Lorraine Daston and Peter Galison have dubbed 'mechanical objectivity,' or the production of data with a minimum of human input, has been strong since the nineteenth century in western science, but has inevitably proven elusive (Daston & Galison, 2010). Machines, instruments, algorithms, and the like do not eliminate human subjectivity, they simply move it to another place: the person who designed or fabricated the machine, the individual operating the instrument, the specialist deciding which algorithm to use and what to include in the data to be analyzed, or the scientist making sense of the results. At the turn of the nineteenth century astronomers discovered that no matter how powerful their telescopes and how careful their observations, there was never exact agreement on the position of a celestial body (Schaffer, 1988). Each marked the body's position in a characteristic way, anticipating more quickly or slowly when the object would exactly cross the hairs of the telescope sight. Eventually this phenomenon came to be known as the 'personal equation,' and Laplace and others developed the mathematics of errors with its famous binomial distribution or bell-shaped curve to try and address it. Even that attempt to adjust for human subjectivity mechanically was only partially successful, however, because it was founded on a model of the human observer—someone who was always a little early or a little late—that does not necessarily correspond to the reality of an individual who is sometimes anxious and sometimes weary, whose attention can be total at one moment and a bit distracted at the next. As Daston (1995) has argued forcefully, objectivity in modern science has gone hand-in-hand with a particular moral economy for the practitioner, an ideal of exactitude and disinterestedness that would be hard for any mere mortal to totally embody. Quantification is thus as much human as it is mechanical, subjective as well as objective.

In the rest of my essay, I want to keep this inextricable mixture of the mechanical and human firmly in mind. Why quantify? The world is messy and complex and full of noise; combining bits of information and analyzing them rigorously is often the only way to detect patterns and truths concealed in the flood of information with which we are inundated. Every act of seeing, however, is also an act of occlusion, of not-seeing. To attend to one part of the visual spectrum one must pay much less attention to other parts; to count one kind of thing, one cannot include other kinds. This is not a profound observation, certainly, but the moral demands it places on practitioners are worthy of consideration. It is not always easy to give weight to what we are not counting or seeing, to give space to what is not in the records as well as to what is staring one in the face. It follows that quantification is not just a technical pursuit, but a moral one; the reductions and simplifications it relies on can bring great and important clarity, but always at a cost.

The moral questions for the practitioner are first, is the cost justified, and second, and even more critically, who will pay it? The powerful or the powerless? The established or those fighting for a place? In the end this essay will be a plea for epistemic modesty, for continually being aware, and making the consumers of the findings aware, that insights are at best but partial, and that essential factors might have eluded the best efforts to take account of them. How to do that practically is not an easy question to answer, but simply ignoring it is not an ethical option.

To provide some substance to this inquiry, I will draw on work I have done on the history of quantifying human intelligence, and in particular levels of intelligence, often expressed as a single number: IQ (Carson, 2007). Intelligence was not always understood as something that could be quantified or, really, as something that existed in degrees at all. Indeed, to this day numerous scientists and lay people alike dispute these very claims. But for many over the course of the nineteenth and first part of the twentieth centuries intelligence did become something that was singular and could be measured. The story of how intelligence was turned into a quantifiable entity and one that was quantified in very particular ways is long and complex, and so I can only tell parts here. Drawing on some of those episodes, supplemented with other examples from the history of quantification, my goal is to illuminate a few key features of quantification, particularly those where its strengths and limitations are most visible. The essay itself will be in three parts. Part one discusses simplifying vision through the reduction and amplification of complex phenomena; part two addresses the consequences, intended and otherwise, of achieving visibility; and part three explores what can be called 'silence in the archive.'

Part One: Reduction and Amplification

In 1839 Philadelphia physician Samuel G. Morton (1839) published *Crania Americana, or a Comparative View of the Skulls of Various Aboriginal Nations of North and South America*. This work and others that followed in its wake helped generate a major controversy within American scientific and literary circles during the middle decades of the nineteenth century. Dispute centered on the issue of the origin of the human races. Were human beings all part of a single species with a single set of ancestors, as was traditionally believed and as Christian Scripture seemed to indicate, or were there multiple creations of distinct species of human beings, the races into which humanity seemed to be divided, as the American school of anthropologists would come to argue? Morton stood firmly in the camp of the polygenicists, as the multiple-species proponents were called. He sought to demonstrate that the human races were independent in origin, character, and development, and chose the human skull as his principal research material.

Over the course of thirty years Morton assembled a collection of some 600 skulls, on which he performed a number of measurements, most notably the calculation of cranial volume. His finding that caused the greatest sensation was that the human races could be arranged in a hierarchy according to mean internal cranial capacity. Truth be told, there was nothing particularly striking or novel about the hierarchy Morton proposed: his quantifications mostly replicated common European biases about the superiority of the 'white' race and the inferiority of all others. What is of greater interest for the purposes of this essay is to examine what he did in order to produce those results (Carson, 1999).

At first, Morton's process was mostly reductive: out of the welter of features available to analyze, he chose a few—most importantly racial or group identity and cranial capacity—for further notice. Then he measured volume, by filling each skull first with shot and later with seed, removing as a byproduct all those peculiarities in shape that failed to register as differences in cranial size. Once reduced to a numerical measure of volume, the individual skulls were eliminated entirely, as the cubic capacities, via basic mathematics, were converted into aggregated means tied solely to racial or group identity. Finally, having before him only his racial averages, Morton could build anew, arraying the quantities into numerical order. He thereby transformed the races or groups into a hierarchy, which, because of a presumed causal link between cranial volume and degree of intelligence, was deemed to represent a scale of overall mental superiority.

Two features of Morton's transformations of his material objects into numbers are particularly relevant for this essay. First, for Morton the transformation of mind into skull, and skull into aggregated volumetric means, was critical. Morton began with the assumption that brain volume was directly linked to mental power. Because of that belief, he systematically eliminated almost all of the physical particularities that each skull manifested—be it thickness of the cranial walls, or pattern of the skull suturing, or size of the nasal cavity—until each embodied solely those characteristics Morton believed significant for his project. In so doing, Morton simultaneously prepared his objects for their translation into quantitative entities and narrowed their range of possible meanings and representations. Once Morton had reduced his skulls to a set of cranial volume averages by race, then arraying them in numerical order became an unproblematic task.

Second, Morton had to remake not just his skulls, but intelligence as well. Intelligence as an activity, as a set of operations of the human mind to be investigated and understood was of little interest to Morton. His goal was not to explain how mind worked or how individuals thought, but to quantify and rank collections of minds. Whatever intelligence may have meant in other contexts, for Morton it had to be contained within a single magnitude, varying from person to person or race to race. If it were allowed to become more complicated, then the possibility of arraying its possessors in one determinate order might have vanished entirely. And so, the connection between mind and skull not only served to make a mental phenomenon more physical and visible, but also to facilitate its transformation into a numerically arrayable quantity.

Shorn of specifics, I suspect that Morton's basic methodology will seem familiar to most engaged in some form of quantitative practice. Morton began with a complex, multivalent phenomenon—in his case, individual irregular skulls—and through a series of reductions and amplifications, converted them into a much simpler quantified result. In that form he and his readers could now see a pattern in the skulls' volumes not readily visible when surveying the crania individually, with all of their idiosyncratic features. Reduced to average volume measurements, the underlying order present in some small way in each skull could be amplified so that the hierarchy of the groups became manifest. All that was required was that the measurements be done accurately, the samples representative of the entire group, and so on.

These caveats are crucially important, of course; most modern researchers would argue that Morton ran afoul of the second, if not the first, and thus that his hierarchy was an artifact rather than real (Gould, 1981; Mitchell, 2018; Lewis, et al, 2011; Fabian 2010; Stanton, 1960). Nonetheless, as science studies scholar Bruno Latour (1987, 1983) has shown, such techniques of reduction and amplification lie at the heart of the practice of much of modern science. To interrogate, manipulate, and understand a phenomenon, it is first necessary to render it small enough to fit on a microscope slide or in a petri dish, or even on a piece of paper. There the feature of interest can be amplified, studied, manipulated, and perhaps eventually understood. IQ scores or H-indices do not capture all of the features of the phenomena from which they are drawn. But if the translations are careful, the quantifications accurate, and the construct itself meaningful, then the simplifications and amplifications they represent allow users to see and understand in ways that would have been difficult if not impossible by other means.

It was a rather long journey from those individual skulls to the ranking of groups by their presumed average level of intelligence. This is true of virtually any empirical result: raw material must be put through a series of

transformations to be converted into usable data that can then be subjected to some analytical procedure. Along the way, decisions are constantly being made as to what to take notice of and what to disregard. Does the thickness of the skull matter? How about its shape, or the number of indentations? Mostly, the process of quantification is one of paring down, of removing elements or possibilities so that as much data can be aggregated as possible (Porter, 1995, 1986). Distinguishing by skull thickness or skull shape would have meant separating the data into multiple sets and thus multiple rankings according to volume. If mental power was solely determined by brain volume, as Morton believed, then ignoring these features made sense. But what if the shape of the brain mattered? Or thicker skulls compacted the same amount of brain into a smaller space? Neither nature nor culture says whether to aggregate or disaggregate a data set; it is professional judgment all the way down. There's nothing wrong with that; without decisions of this type constantly being made the data would remain forever raw, particulars whose individuality would preclude aggregation and thus analysis. If the specifics of every citation mattered, for example, how would citation analysis ever be possible?

What needs to be remembered, though, is that the criteria used to prepare the raw material for analysis may reflect expert experience, disciplinary beliefs, and accumulated practice, but they might not always be the best standards to work by. Or, rather, they might reflect and re-instantiate what is already known and believed about the world and how it operates, rather than leaving space to interrogate the phenomena in new ways. This is the dilemma: one can learn nothing without reducing and amplifying, but in that process, one can also ignore as insignificant features that might, from another perspective, be crucial. Morton was not alone in believing that brain volume was related directly to mental power; the analogy with muscle size and strength was widely asserted. He thus felt justified in ignoring all other features of the skulls, including whether they were derived from men or women.

Later researchers argued that this difference mattered, that skull volume reflected overall size differences as much as it did differences in brain power, and thus promoted providing separate averages by sex (Gould, 1981). Morton's judgment in this regard was not borne out. Should he have worried that he had reduced too much? Undoubtedly. But he was also a leader in the field, his techniques were widely admired, and his results were for many uncontroversial. How does one determine what must be taken account of and what can be safely ignored (Collins, 1992)?

Part Two: Visibility and Its Consequences

As has already been suggested, Morton did not just collect his data and try to render it meaningful, he also published it and sought to bring visibility to his results. Visibility might not strike most practitioners, regardless of their discipline, as a particularly problematic result of their labors. Typically, researchers want their quantifications and analyses to be well regarded, persuasive, broadly adopted, and deemed important. But another episode from the history of the quantification of intelligence illuminates not just the value of achieving visibility, but also some of its costs. It is a reminder that the success of one measure can crowd out the use of others and that this crowding out can have epistemic, disciplinary, and social consequences.

If one wanted to pick a date for the birth of the modern approach to measuring intelligence, 1905 would be hard to top. In that year, the French psychologist Alfred Binet along with his collaborator Théodore Simon published their first articles on a new psychological instrument they had developed, the Binet-Simon Intelligence Scale (Carson, 2007; Binet & Simon, 1905a, 1905b, 1905c, 1905d; Wolf, 1973). The scale was comprised of a series of age-related tasks whose purpose was to reveal whether an examinee's overall intelligence was developing at a rate typical of their peers. After revision in 1908 and 1911 the scale did so by reporting the examinee's result quantitatively as a single number, their 'mental age' (M.A.), which facilitated conceiving of intelligence as a physical and measurable object. With the help of American psychologist Henry Herbert Goddard, the test reached America in 1910, where adoption by psychologists was swift. Soon rival versions of the test proliferated and a vogue for testing swept the discipline. Before World War I, however, the main application for the intelligence test was clinical, to diagnose degrees of mental deficiency (Zenderland, 1998; Samelson, 1979).

In the latter part of the 1910s, two events helped to propel the project of measuring intelligence into national prominence. First, in 1916 Lewis Terman (1917), a young psychologist at Stanford University, completed his own revision of the Binet scale (Minton, 1988; Chapman, 1988). The Stanford-Binet, as it was popularly called, did much to cement the place of intelligence testing within the American intellectual and cultural landscape. Thoroughly overhauled for an American population, the Stanford-Binet popularized a new way of quantifying mental ability first championed by German psychologist William Stern, the Intelligence Quotient or IQ, a ratio of mental age to chronological age (times 100) designed to remain constant over time. Terman also chose questions for the test so that distribution of IQ scores would fit the bell-shaped normal distribution curve. With IQ, Terman fully transformed intelligence into a standardized, quantifiable characteristic applicable to the entire range of human minds, whether child or adult, male or female, white or black. Whatever variation individual intellects of the same IQ might manifest, and even whatever growth they might sustain, were rendered invisible in the process of producing a Stanford-Binet intelligence quotient. Represented as an innate characteristic fundamental to determining an individual's life course, IQ seemed of potentially immense significance and relevant to a variety of social decisions, including an individual's occupational choice, level of appropriate education, and possible need for institutionalization (Terman, 1919, 1922).

Second, the advent of World War I afforded American psychologists the opportunity to demonstrate the relevance of intelligence measures to more mainstream arenas than the clinic or programs for the educationally lagging. Harvard psychologist Robert M. Yerkes assembled a team of mental testers—including Terman and Goddard—to aid the war effort by providing the Army with an efficient way of classifying the millions of new recruits it would need to mobilize for the war. Because the Stanford-Binet was designed to be administered individually, it seemed impractical for the military's needs. And so, Yerkes and his fellow psychologists developed new mental tests that could be given to groups: Army Alpha for literates and Army Beta for English-language illiterates. They then examined over 1.75 million soldiers, with the results used as one means of sorting recruits into various categories of military usefulness—ranging from officer candidates to those deemed unfit for frontline duty—and of justifying those decisions (Carson, 1993).

Intelligence testing gained enormous legitimacy and public exposure from the army testing program. Moreover, one of the key findings derived from analyzing the testing data—that a large percentage of American soldiers scored in the 'feebleminded' range or worse—shocked the nation. The publicity surrounding this finding helped transform an endeavor that had existed mainly on the margins of American culture to one that seemed of critical significance. Leaders in many sectors of American society—education, industry, government—were now confronted with quantitative evidence that seemed to confirm their worst fears about the declining quality of the American population. In response, they turned to intelligence testing as an objective, efficient, and credible means of differentiating students, workers, and applicants for employment or admissions.

After the war, a variety of new multiple-choice tests began to supplant the Stanford-Binet as the most common technology of intelligence assessment and to be used by schools and industry to help rank applicants and sort populations. The nature of intelligence reified in Army Alpha and these new postwar civilian tests, including the SAT for college admissions, perpetuated the Stanford-Binet model: a unitary, inheritable, biological entity that allowed all human beings to be ranked on a single scale as finely graded as the numerical calculation of IQ permitted. In the process, intelligence's status and domain were transformed. Though no one way of understanding intelligence, or instrument for its assessment, completely dominated before World War I, in broad terms after the war intelligence became ever more tightly associated with quantified measurements and precise rankings of entire populations. Intelligence understood as IQ came to be seen as a prime factor in success in virtually all human endeavors, used by many to explain, in almost Darwinian terms, why some individuals were at the top of the social/occupational hierarchy, and others were at the bottom (Carson, 2007; Lemann, 1999).

To summarize: this is the story of how one approach to quantifying intelligence, represented by numbers such as IQ or SAT scores, became the dominant way in which intelligence was understood and operationally used. Its extraordinary visibility after the war, and really because of the war, did not completely preclude other ways of approaching the quantification of intelligence, but it did overshadow most of them. Psychologists continued to debate whether intelligence was unitary, or composed of a few distinguishable elements, or best understood as a large number of semi-independent factors, and largely did so on the basis of statistical analysis of intelligence test data. But the practical value of reducing the measurement of intelligence to a single number so that ranking applicants or candidates was straightforward and seemed objective meant that even those skeptical about the meaning of a single measure often succumbed to its seductions. Thus, in the 1930s–1940s psychologist David Wechsler, dissatisfied with the Stanford-Binet approach to quantifying intelligence, developed his own tests for children and adults, reporting scores for both verbal and nonverbal intelligence. He also, though, provided an IQ score, and it was that score that was used most often in clinical and non-clinical settings (Wechsler, 1939, 1949; Frank, 1983).

The difficulty in escaping the powerful utility of intelligence as IQ—the extraordinary visibility and importance it achieved in various domains of American society—highlights some of the consequences good and bad that successful quantification can entail. As is well known, a debate rages to this day over what intelligence tests actually measure, if anything at all, as well as whether they are so biased in terms of race, class, gender, and/or socio-economic status as to be meaningless. Regardless of where one comes down on these issues, the social power of this form of quantification is undeniable. Intelligence rendered as something like IQ helped spawn an entire subfield of psychology, was taken up by a wide range of institutions, and continues to influence the lives of millions. In the process, other ways of assessing intelligence have been pushed to the margins, even when there are as good statistical arguments in favor of, say, eight major components to intelligence as there are for it being something unitary (Carson, 2015). Moreover, while no one denies that an individual's mental abilities develop as they progress through childhood, such changes are rendered invisible when intelligence is quantified as IQ as opposed to, say, its Binet-Simon form, mental age.

A successful form of quantification can thus crowd out other possibilities, and as it is adopted by and integrated into more and more areas of practice, becomes ever harder to dislodge, however compelling or imperfect one might find it (Fraser, 1995). These downstream effects seem especially true of quantifications that simplify complex phenomena so that clear-cut comparisons can be made. They can do important work, but they can also become easy ways for institutions to make decisions that might better be based on more holistic assessments. If the H-index were to become a stand-in for scholarly productivity, for example, might departments or disciplines move toward basing tenure or promotion decisions on meeting a particular cutoff score, rather than assessing the work the individual actually produced (Chapman, 2019; Moher, et al, 2018)? No researcher can fully anticipate, much less control, all of the

possible consequences of their scholarly endeavors. Nonetheless, given the power routinely accorded to quantitative assessments and to the algorithms that make sense of data, it is imperative that practitioners consider what becomes less visible as a given form of quantification becomes more dominant, as well as the consequences adoption of that quantification might have.

Part Three: Silence in the Archive

Data scientists probably spend little time in archives and may not even think of their work as having much to do with an archive. Historians are another matter. Historians have long seen the archive as central to their scholarly legitimacy. Where once historians spoke mostly of going to the archive—be it the National Archives in Washington DC, the Vatican archives in Rome, or the Proquivemex corporate archives in Mexico City—now they talk more about constructing their archive. They mean by this that one needs to pull together diverse materials relevant to the research question at hand, often from a range of places. From this perspective, the archive is not some pristine other, but rather something built for a particular purpose based on human judgement, with all of the attendant possibilities for idiosyncrasy, error, and bias. It is, fundamentally, a kind of data set. Indeed, whatever the name, an essential part of scholarly practice in many fields is to assemble the empirical material that will then be analyzed to answer some question or determine the plausibility of some hypothesis. And in both cases, it is not nature that decides what is in or out, but the practitioner, who must not only choose but almost inevitably clean up their data before it can be used.

Historians in recent years have begun to worry about the archive and what it means for the knowledge that can be produced. A number of different concerns have been articulated, most circling around the issues of who and what gets into the archive and who and what does not, as well as the perspective represented by the archive as an institution (Blouin & Rosenberg, 2011; Farge, 2015). Archives, historian Michel-Rolph Trouillot (1995), among others, has pointed out, are not neutral; rather, they have a politics, and that politics typically reflects the interests of the established and the powerful, since those are the groups who have typically established the archive and whose records are most likely to be preserved. *The Papers of Thomas Jefferson* (1950–2019), third President of the United States, for example, now runs to forty-four printed volumes and are still not finished, while the extant material on Sally Hemmings, the enslaved woman with whom Jefferson had four children, would barely fill a single file folder (Gordon-Reed, 2009). National archives in almost every country are overflowing with official documents and often allow careful reconstruction of the actions of political leaders, but may say little directly about the concerns and lives of average citizens. Police and court records are an important source for many historians, but represent by and large the perspective of those in power, their view of what crime is, when it occurs, and who perpetrates it. Bankers swindling millions with deceptive packaging of subprime mortgages, for example, will not appear in the police archive, because few were charged, while a young person jumping a subway turnstile could easily end up as part of a data set on crime.

Closer to home, the many volumes of *Science* or *Nature* are filled with articles that represent what the editors and peer reviewers—the elite in the relevant fields—deemed not only persuasive, but important enough to be published in these flagship journals. Analyzed retrospectively, these volumes would tell a story about the development of modern science, but one that necessarily reflects the judgements of the journals' gatekeepers. Is that the only tale important to tell?

No archive, and indeed no data set, is neutral; they present points of view, and inevitably leave out much more than they could conceivably include. The silence that thus permeates data repositories thus can have profound effects on what is understood to be significant, included in the quantifications and calculations being performed on the material, and used to understand the world. If something or someone is not present from the start, how can their contribution be properly weighted? And if the biases of power and privilege help to shape the archive, can they fail to shape the knowledge that is being produced? Intelligence testing pioneer Henry Goddard's archive contains the case history of a twelve-year-old boy recommended for psychological examination because of repeated run-ins with the law (Goddard, n.d.). Goddard administered the Binet test, but was surprised to discover that the boy scored in the range of normal intelligence. Unable to explain the boy's criminal behavior on the basis of low intelligence, Goddard looked for other signs of pathology. But what if the boy had not done so well on the test? Further investigation would have been deemed unnecessary and the possibility that there were other explanations for the boy's actions unexplored. The archive would have remained silent.

Consider for a moment the case of Rosalind Franklin (Sayre, 1975; Gibbons, 2012). As is now well known, Franklin's work in x-ray crystallography was crucial to the discovery of the structure of DNA. James Watson and Francis Crick were shown a DNA x-ray image that Franklin had made, and used the insight they gained from Franklin's work to propose a new model for the DNA molecule, the now famous double helix (Watson, 1980). Watson and Crick (1953) won Noble Prizes for their work; Franklin's critical contribution, however, was largely invisible. If one looks at the 1953 paper in *Nature* announcing the proposed structure for DNA, there are no citations to Franklin's work, and only one cryptic reference to her near the end of the paper: 'We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers.'

Citation analysis of the 1953 article would reveal the importance of Linus Pauling, Erwin Chargaff, and Maurice Wilkins, among others. But Franklin would remain invisible by such a metric, and indeed it was not until many years after the publication of the DNA article that the significance of her work began to be understood. A woman in a man's

world; an experimentalist rather than more theoretically inclined, as were Watson and Crick—the politics of knowledge made it easy for Franklin to be pushed to the margins and her work consigned to an afterthought, a silence in the archive even for someone who contributed so importantly to one of the fundamental scientific discoveries of the twentieth century.

In Franklin's case, Watson (1980) himself inadvertently hinted at Franklin's importance in 1968 when he published his own account of the discovery of the structure of DNA, *The Double Helix*, and in it made a few, mostly disparaging remarks, about Franklin. Other actors eventually provided a few more details (Franklin herself had died tragically early in 1958 at the age of 37) and historians have since largely redressed the silence regarding Franklin and restored her to the historical record. Silence can be filled in, but it takes work, as well as the realization that something important has been missed.

Silence filling can come in a variety of forms and quantification can serve a variety of purposes. At the turn of the century, African American journalist Ida B. Wells-Barnett (1892, 1895) sought to bring to light a national travesty, the lynching of African American men and women (Bay, 2009). Lynchings were a recurring open secret in post-Reconstruction America. African Americans were a thoroughly marginalized group, and wrongs against them were accorded scant attention in the mainstream white press. At real personal risk, Wells-Barnett dedicated her efforts to ending the silence surrounding lynching and used quantification as her weapon of choice. Drawing data from newspaper accounts (particularly the *Chicago Tribune*), informants, and the like, she created an archive of incidents of lynching and published extraordinary tables detailing the number of lynchings by year and by state. Wells-Barnett's efforts did not alone finally incite a public outcry against lynching, but her determination to remake the public archive and her use of quantifications, with all the implications of objectivity that they brought, were crucial. If quantification can reify the silence in the archive, it can also be used to challenge that silence.

Conclusion: Toward an Epistemics of Modesty

Silence, visibility, and reduction/amplification—this essay has focused on these three aspects of the practices surrounding quantification in order to highlight both the potentials and perils of seeing with numbers, as well as the inextricably human and moral elements of this pursuit. The power of quantitative assessments to clarify and to reveal patterns difficult to detect otherwise is undeniable. At the same time, the immense reductions that quantification requires, the removal of so many specifics and details so as to render the material numerical, carry with them the danger that they might obliterate what is most important. Moreover, because quantitative analysis can appear to be not only technically daunting but also unproblematically objective, finding ways to contest quantifications can be challenging, even for experts. IQ in one form or another continues to fill a variety of social roles, even though its critics are numerous, precisely because it melds the utility of a single quantified entity with the presumed objectivity of how that number was produced.

Quantifying thus brings with it a set of moral demands. Certainly, the work must be done scrupulously and claims for the meaning of a finding not exceed its evidentiary base. The practitioner, however, must also look critically at how they decided what would be included in their data—their archive—and what was removed from the material in the process of transforming it into something usable. Furthermore, they should consider how their form of quantification might render other possible approaches, and thus other ways of seeing, less viable. This is a tall order. Doing the work is often difficult enough; how does one also critically appraise one's own reductions, look laterally, worry about silences, and present results in a way that opens up possibilities rather than closes them off?

There are no easy answers. By way of a conclusion, though, one possible starting point might be to adopt a stance that can be called 'epistemic modesty,' something that science studies scholar Sheila Jasanoff (2003) argues can be achieved through 'technologies of humility.' Epistemic modesty is meant to suggest that scholars in all fields should be aware of the limits of their work, its partialness, and the particular standpoint from which they are coming. They should then try to find ways in the act of presenting findings to also suggest that what they have arrived at is no more than a part of the story. Science studies scholar Donna Haraway (1991) has argued that we must all be aware of the partialness and situatedness of the perspectives we bring to our work and that there is no mechanism that can allow us to inhabit what is sometimes called the 'god's eye perspective.' Our strength comes rather from linking together multiple fragmentary views, keeping in mind that some vantage points, such as those from positions of marginality, might have particularly rich insights to offer.

Jasanoff (2003, 238–242) contends that a more humble and open form of knowledge production can occur if we follow four precepts: (1) take care to make the framing of problems to be explored expansive and inclusive; (2) pay particular attention to the most vulnerable; (3) insure that our work is distributed to a wide network of those who might be affected by it so that they have a chance to respond; and (4) consider whatever we come up with to be provisional and open to revision as we learn more. None of these techniques can ensure that mistakes are not made, but they should at least remind us to think hard about the silences our archive will contain and to think carefully about who might bear the costs of whatever we have rendered invisible. Rather than just think with our archive, we should also try to read against the grain and push on it to reveal the traces, such as the contributions of Rosalind Franklin, that reading with the grain might miss.

Acknowledgements

I would like to thank Cinzia Daraio, Henk F. Moed, and Giancarlo Ruocco for the invitation to speak at the XVII International Conference on Scientometrics and Informetrics in Rome in 2019 and for their warm hospitality; this article is a direct response to their invitation and prompting to think big about quantification.

Competing Interests

The author has no competing interests to declare.

References

Aksnes, D. W., Langfeldt, L., & **Wouters, P.** (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open, 9*(1). DOI: https://doi.org/10.1177/2158244019829575

Bay, M. (2009). To tell the truth freely: The life of Ida B. Wells. New York: Hill & Wang.

Binet, A., & **Simon, T.** (1905a). Méthodes nouvelles pour diagnostiquer l'idiotie, l'imbécillité et la débilité mentale. In Atti del V congresso internazionale di psicologia. Rome: Foranzi.

Binet, A., & **Simon, T.** (1905b). Sur la nécessité d'établir un diagnostic scientifique des états inférieurs de l'intelligence. *L'Année psychologique, 11*, 163–190. DOI: https://doi.org/10.3406/psy.1904.3674

Binet, A., & **Simon, T.** (1905c). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année psychologique, 11*, 191–244. DOI: https://doi.org/10.3406/psy.1904.3675

Binet, A., & **Simon, T.** (1905d). Application des méthodes nouvelles au diagnostic du niveau intellectuel chez des enfants normaux et anormaux d'hospice et d'école primaire. *L'Année psychologique, 11*, 245–336. DOI: https://doi.org/10.3406/psy.1904.3676

Blouin, F. X., & **Rosenberg, W. G.** (2011). Processing the past: Contesting authority in history and the archives. New York: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199740543.001.0001

Carson, J. (1993). Army Alpha, army brass, and the search for army intelligence. *Isis, 84*, 278–309. DOI: https://doi.org/10.1086/356463

Carson, J. (1999). Minding matter/mattering mind: Knowledge and the subject in nineteenth-century psychology. *Studies in History and Philosophy of the Biological & Biomedical Sciences, 30*, 345–76. DOI: https://doi.org/10.1016/S1369-8486(99)00016-3

Carson, J. (2007). The measure of merit: Talents, intelligence, and inequality in the French and American republics, 1750–1940. Princeton: Princeton University Press. DOI: https://doi.org/10.1515/9780691187679

Carson, J. (2015). Intelligence: History of a concept. In *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed., (vol. 12, pp. 309–312). Oxford: Elsevier. DOI: https://doi.org/10.1016/B978-0-08-097086-8.03094-4

Chapman, C. A., et al. (2019). Games academics play and their consequences: How authorship, *h*-index and journal impact factors are shaping the future of academia. *Proceedings of the Royal Society, B. 286,* 2019–2047. DOI: https://doi.org/10.1098/rspb.2019.2047

Chapman, P. D. (1988). Schools as sorters: Lewis M. Terman, applied psychology, and the intelligence testing movement, 1890–1930. New York: New York University Press.

Collins, H. M. (1992). Changing order: Replication and induction in scientific practice. 2nd. ed. Chicago: Chicago University Press.

Daston, L. (1995). The moral economy of science. *Osiris, 10*, 2–24. DOI: https://doi.org/10.1086/368740

Daston, L., & **Galison, P.** (2010). Objectivity. New York: Zone Books.

Fabian, A. (2010). The skull collectors: Race, science, and America's unburied dead, 79–120. Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226233499.001.0001

Farge, A. (2015). The allure of the archives. New Haven: Yale University Press.

Frank, G. (1983). The Wechsler enterprise: An assessment of the development, structure, and use of the Wechsler tests of intelligence. Oxford: Pergamon.

Fraser, S., ed. (1995). The bell curve wars: Race, intelligence, and the future of America. New York: Basic Books.

Gibbons, M. G. (2012). Reassessing discovery: Rosalind Franklin, scientific visualization, and the structure of DNA. *Philosophy of Science, 79*, 63–80. DOI: https://doi.org/10.1086/663241

Goddard, H. H. (ND). No title. *File: Case History, box M614, Goddard Papers, Archives of the History of American Psychology*. Akron, OH: University of Akron, p. 2.

Gordon-Reed, A. (2009). The Hemingses of Monticello. New York: W. W. Norton.

Gould, S. J. (1981). The mismeasure of man. New York: W. W. Norton.

Haraway, D. (1991). Situated knowledges: The science question in feminism and the privilege of partial perspective. In *Simians, cyborgs, and women: The reinvention of nature* (pp. 183–201). New York: Routledge.

Jasanoff, S. (2003). Technologies of humility: Citizen participation in governing science. *Minerva, 41*, 223–44. DOI: https://doi.org/10.1023/A:1025557512320

Jefferson, T. (1950–2019). The papers of Thomas Jefferson. Vols. 1–44. Princeton: Princeton University Press. DOI: https://doi.org/10.1515/9780691194400

Kuhn, T. (1962). The structure of scientific revolutions, 1–42. Chicago: University of Chicago Press.

Latour, B. (1983). Give me a laboratory and I will raise the world. In K. D. Knorr-Cetina & M. Mulkay (eds.), *Science observed: Perspectives on the social study of science* (pp. 141–70). London: Sage.

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Milton Keynes: Open University Press.

Lemann, N. (1999). The big test: The secret history of the American meritocracy. New York: Farrar, Straus and Giroux.

Lewis J. E., DeGusta D., Meyer M. R., Monge J. M., Mann A. E., & **Holloway R. L.** (2011). The mismeasure of science: Stephen Jay Gould versus Samuel George Morton on skulls and bias. *PLoS Biology, 9*, e1001071. DOI: https://doi.org/10.1371/journal.pbio.1001071

Liao, Y.-M. (2019). Top 500 journals in 2019 JCR Journal Impact Factor Released in 2019. https://www.researchgate.net/publication/333972052_Top_500_Journals_in_2019_JCR_Journal_Impact_Factor_Released_in_2019, accessed July 18, 2020

MacRoberts, M. H., & **MacRoberts, B. R.** (1989). Problems of citation analysis: A critical review. *Journal of the American Society of Information Science, 40*, 342–48. DOI: https://doi.org/10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U

Minton, H. L. (1988). Lewis M. Terman: Pioneer in psychological testing. New York: New York University Press.

Mitchell, P. W. (2018). The fault in his seeds: Lost notes to the case of bias in Samuel George Morton's cranial race science. *PLoS Biology, 16*, e2007008. DOI: https://doi.org/10.1371/journal.pbio.2007008

Moher, D., Naudet, F., Cristea, I. A., Miedema, F., Ioannidis, J., & **Goodman, S. N.** (2018). Assessing scientists for hiring, promotion, and tenure. *PLoS Biology, 16*(3), e2004089. DOI: https://doi.org/10.1371/journal.pbio.2004089

Morton, S. G. (1839). Crania Americana, or a comparative view of the skulls of various aboriginal nations of North and South America, to which is prefixed an essay on the varieties of the human species. Philadelphia: Pennington. DOI: https://doi.org/10.5962/bhl.title.51431

Porter, T. M. (1986). The rise of statistical thinking, 1820–1900. Princeton: Princeton University Press.

Porter, T. M. (1995). Trust in numbers: The pursuit of objectivity in science and public life. Princeton: Princeton University Press. DOI: https://doi.org/10.1515/9781400821617

Samelson, F. (1979). Putting psychology on the map: Ideology and intelligence testing. In A. R. Buss (ed.), *Psychology in social context*. New York: Irvington Publishers.

Sayre, A. (1975). Rosalind Franklin and dna. New York: W.W. Norton.

Schaffer, S. (1988). Astronomers mark time: Discipline and the personal equation. *Science in Context, 2,* 115–45. Scimago Journal and Country Rank https://www.scimagojr.com/journalrank.php?area=2700&year=2018&type=j, accessed July 18, 2020. DOI: https://doi.org/10.1017/S026988970000051X

Shapin, S. (1995). A social history of truth: Civility and science in seventeenth-century England, 3–41. Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226148847.001.0001

Stanton, W. (1960). The leopard's spots: Scientific attitudes toward race in America, 1815–59. Chicago: University of Chicago Press.

Terman, L. M., et al. (1917). The Stanford revision and extension of the Binet-Simon scale for measuring intelligence. Baltimore: Warwick & York. DOI: https://doi.org/10.1037/13873-000

Terman, L. M. (1919). The intelligence of school children. Boston: Houghton Mifflin.

Terman, L. M. (1922). The psychological determinist; or democracy and the I.Q. *Journal of Educational Research, 6*, 57–62.

Trouillot, M.-R. (1995). Silencing the past: Power and the production of history. Boston: Beacon Press.

Watson, J. D. (1980). The double helix: A personal account of the discovery of the structure of DNA. New York: W. W. Norton.

Watson, J. D., & **F. H. C. Crick.** (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 4356, April 25, 737–738. DOI: https://doi.org/10.1038/171737a0

Wechsler, D. (1939). The measurement of adult intelligence. Baltimore: Williams & Witkins. DOI: https://doi.org/10.1037/10020-000

Wechsler, D. (1949). Wechsler intelligence scale for children. New York: The Psychological Corporation.

Wells-Barnett, I. B. (1892). Southern horrors: Lynch law in all its phases. New York: New York Age Print.

Wells-Barnett, I. B. (1895). A red record: Tabulated statistics and alleged causes of lynching in the United States. Chicago: NP.

Wolf, T. H. (1973). Alfred Binet. Chicago: University of Chicago Press.

Zenderland, L. (1998). Measuring minds: Henry Herbert Goddard and the origins of American intelligence testing. Cambridge: Cambridge University Press.

How to cite this article: Carson, J. (2020). Quantification – Affordances and Limits. *Scholarly Assessment Reports*, 2(1): 8. DOI: https://doi.org/10.29024/sar.24

Submitted: 22 July 2020 **Accepted:** 23 July 2020 **Published:** 07 August 2020

Scholarly Assessment Reports is a peer-reviewed open access journal published by Levy Library Press. **OPEN ACCESS**