Biol Theory (2015) 10:237–245
DOI 10.1007/s13752-015-0217-z

THEMATIC ISSUE ARTICLE: QUALITY & QUANTITY

A Plea for "Shmeasurement" in the Social Sciences

Olivier Morin[1]

Received: 21 September 2014 / Accepted: 22 June 2015 / Published online: 24 July 2015

Abstract Suspicion of "physics envy" surrounds the standard statistical toolbox used in the empirical sciences, from biology to psychology. Mainstream methods in these fields, various lines of criticism point out, often fall short of the basic requirements of measurement. Quantitative scales are applied to variables that can hardly be treated as measurable magnitudes, like preferences or happiness; hypotheses are tested by comparing data with conventional significance thresholds that hardly mention effect sizes. This article discusses what I call (with tongue in cheek) "shmeasurement." To "shmeasure" is to fail to apply quantitative tools to quantitative questions. We "shmeasure" when we try to measure what cannot be measured, or, conversely, when we ask binary questions of continuous measurements. Following the critics of standard statistical tools, it is argued that our statistical toolbox is indeed less concerned with the measurement of magnitudes than we take it to be. This article adds, however, that measurement is not all there is to scientific activity. Most techniques of proof do not resemble measurement as much as voting—a practice that makes frequent use of numbers, figures, or measurements, yet is not chiefly concerned with assessing quantities. Measurement is only one among three functions of the scientific toolbox, the other two being collating observations and deciding which hypotheses to relinquish. I thus make a plea for "shmeasurement": the mismeasure of things starts to make more sense once we take into account the nonquantitative side of scientific practice.

✉ Olivier Morin
olivier@cognitionandculture.net

[1] The KLI Institute, Klosterneuburg, Austria

Keywords Measurement · Physics envy · Quantity · Significance testing · Tallies

Introduction

The scientific revolution excepted, it is hard to think of a time when more promises were made on behalf of measurement tools. The standardized statistical toolbox of experimental science seems poised to take over the parts of psychology, anthropology, or economics that still elude its grasp. Even history and literature are urged to go quantitative. The trend is not just an academic one: data analysis is entering our homes and lives, via our smartphones and computers. We are invited to count our calorie intake, our popularity on Twitter, our productivity. *Measure yourself* is the new *gnothi seauton*. This trend generates celebrations and worries in equal measure. Are we seeing the dawn of a new Gradgrindian age?

A recurrent charge against quantitative science is "physics envy." In a nutshell, standard measurement tools are suspected to imitate the appearances of quantitative science, without the substance. This point of view is backed by strong arguments. As we shall see, empirical scientists tend to confound scales of measurement, assuming that the variable they study possesses quantitative properties, like additivity, without sufficient proof. They rely on significance tests whose validity is widely debated. Is quantitative science, as currently practiced, measurement or "shmeasurement"? This article argues that it is, in fact, neither. I agree with the critics that many tools in the standard statistical toolbox are not quantitative in the strictest sense of that word. They often fail to measure anything at all. On the other hand, it would be unfair to see them only in this light. If they do not perform well as

measurement tools, that is because they have other functions.

The argument is based on a distinction between three kinds of practices: measurements, collations, and tallies. *Measurements* aim at assessing quantities. *Tallies* aim at making decisions, excluding certain possibilities and promoting others. Tallies often make use of measurements, but only as a means to the end of putting various views on trial. *Collations* are ways of bringing a collection of observations or judgments together under a common format.

According to Plutarch,[1] the Spartan senate was elected by acclamation. The people were assembled, and each candidate was produced before the crowd, one by one. Meanwhile, a group of blind referees, seated where they could not see the candidates, had the task of rating the loudness of popular acclamations. The most acclaimed candidates went to form part of the Council of Elders. (See Girard 2010 for an analysis.) This example shows all three processes at work. Acclamation serves to collect individual judgments on candidates, gather them, and express them in a single commensurate format (the loudness of an acclamation). Acclamations *collate* individual opinions about each candidate. As for the blind referees, they were rating the intensity of acclamations, a physical magnitude. Although they did not express their assessment with numbers (if they did, Plutarch does not say), the fact that they are assessing a magnitude is enough to say that they were engaging in *measurement*. Together with the assembled people, whose cheering they were rating, the referees produced a ranking of candidates to the senate. Thanks to that, the voting procedure produced winners and losers: it decided between possible outcomes. The whole voting procedure was a *tally*. That tally made use of a means of measurement (the blind referee ratings). It also used collations of opinions (the acclamations).

This article contends that measurements on the one hand, collations and tallies on the other, have fundamentally different demands. Standard scientific method, however, is built around measurement, and our statistic toolbox is, ostensibly at least, based on the assumption that researchers are trying to assess magnitudes. Seen in this light, much of the current practice in the quantitative sciences is deficient, as a growing number of critics points out. More precisely, many statistical instruments concern themselves with empirical observations that do not have much to do with magnitudes, yet usually appear or pretend to deal with quantitative matters. Critiques of mainstream science in the tradition of Stephen Jay Gould see this as "physics envy." I will argue, instead, that a variety of problems raised by the critics can be seen in a more constructive light, once we see that much of what passes as quantitative work really has to do with collating observations and organizing tallies, rather than measuring magnitudes.

I will start (in the next section) by presenting the classical conception of magnitudes and measurements. It lays down the condition for the appropriate use of ratio scales, as opposed to interval scales: a truly quantitative variable needs to have an additive structure and a nonarbitrary zero point. In the two parts of the second main section, we shall see how mainstream quantitative methods fail to be properly quantitative. They do so in several ways. One is by assuming that a variable has a quantitative structure in cases where that assumption is not proven (or even patently wrong). Another is by performing "yes/no tests" that turn quantitative measurements into binary answers to qualitative questions. The most popular explanation for the misuses of quantitative tools will then be presented, and countered (in the "What "Physics Envy" Does Not Explain" section). This explanation takes the shape of a series of charges against scientism, positivism, and other hopeless attempts by the softer sciences to emulate the harder ones. In the following section, I argue for a different view. Many statistical tools used by scientists fail to meet the standards of good measurement because their actual function has nothing to do with the assessment of magnitudes. While they appear superficially like measurement tools, they really work as collations or tallies. Some features that help collations and tallies fulfill their functions also make them poor forms of measurement. When collating observations, it sometimes makes sense to prefer a rough indicator to a fine-grained one. When setting up a tally, blind obedience to entrenched traditions can be a good thing (see "The Inverted Values of Collations and Tallies" section). True, the scientific toolbox is less quantitative than it thinks; but it should be assessed by other standards, as a collection of recipes for solving problems that may or may not involve the assessment of magnitudes (discussed in the "Conclusion" section).

A word, before proceeding, on what this article is not about: it is not centrally concerned with models or mathematical formalisms. It will only deal with the empirical side of quantitative methods: measuring magnitudes, testing hypotheses, collating observations. These things can, of course, be related to modeling, when a model's predictions are tested on real data (simulations are another matter). Models, however, also have a theoretical life of their own. They are sometimes used just for the purpose of clarifying a hypothesis, or as "toy models" that are not meant to be tested. Empirical measurements, collations, and tallies can exist without any mathematical formalism, at least some of the time, as the Spartan example shows. This seemed

[1] I took this example from O. D. Duncan's 1984 book *Social Measurement*, on which more below. The interpretation (in terms of collations, measurements, etc.) is mine.

reason enough to dissociate the two topics. The conclusion will come back to this issue.

Measurement: The Standard View

To describe what is called here the classic view of measurement, I rely a great deal on Joel Michell's *Measurement in Psychology*[2] (2005). That book takes psychology to task for neglecting the classical conception of quantity that has informed the sciences since Euclid. A measure as classically defined is the ratio of a magnitude over a unit. Two meters, for instance, is a length of something, related to an arbitrary unit (the meter). A measure thus assesses a magnitude, or a quantity (I will use these two words interchangeably). What, then, are magnitudes? Their main defining features are an additive structure, and the existence of a unique and nonarbitrary zero point. The additive structure of our calendar year is what allows us to say, for instance, that 1950 happened 100 years later than 1850. The existence of a single and nonarbitrary zero point is what allows us to say that 200 K is twice as much as 100 K. The same cannot be said of 100 versus 200 °C: the zero point on the Celsius scale is an arbitrary convention, that is to say, its absolute value cannot be interpreted quantitatively. On such a scale, "temperature zero" is not equivalent to "zero temperature." Thus, a true magnitude is something that we should be able to plot on a ratio scale, as opposed to an interval scale, in Stanley Stevens' standard partition of scales (1946).

Stevens' notion of measurement encompassed all four kinds of scale, not just ratio scales. A nominal scale (where numbers are treated as simple labels with no quantitative import) was a kind of measurement. So were ordinal (where data points are merely ranked), and interval scales (where relative magnitudes can be compared, but no single nonarbitrary zero point is defined, as in the Celsius scale). Most critics of Stevens' work have resisted this view as too broad (Michell concentrates and summarizes most grievances; see also Duncan 1984, Chap. 4). A nominal scale (like a library catalogue or a list of bar codes for products in a supermarket) is simply a collection of names. As such, it does not deal with a magnitude of any kind, and thus, does not intuitively qualify as measurement.

The question whether to include ordinal or interval scale in a definition of measurements is less easily solved, and, insofar as definitions are a matter of conventions, we should not expect any clear-cut resolution. It seems clear, though, that the ratio scale is, so to speak, the most quantitative of all. It is the only one that allows us to perform the four basic operations of arithmetic. The interval scale, insofar as it allows us to compare relative magnitudes, is, in turn, more quantitative than the interval scale (which allows no such thing). The nominal scale, as argued above, is hardly a quantitative scale at all. Stevens' theory of levels of measurement was meant to clarify and regulate the making of quantitative assumptions, but, as Michell observes, its underlying philosophy encouraged a much more permissive use of scales (e.g., Michell 2011, p. 249).

"Measurement, Shmeasurement!"

Taking Things to be More Quantitative than They Truly Are

This subsection and the next one describe two ways that measurement can be misapplied. The first is to apply quantitative tools to variables that are not quantitative, or not as quantitative as we take them to be. The second consists in confounding qualitative enquiries with quantitative ones—answering yes/no questions with measurements, or reducing matters of magnitude to binary issues. The "shmeasurement" in the section's title is meant to be tongue-in-cheek (as shall be apparent in the next sections), but I think it reflects the intensity of some critiques of mainstream statistical methods. This section is based on their invaluable work.

(1) *Are preferences ordinal?* We are sometimes asked to rank things by order of preference (it can be activities, opinions, items we would want to buy, or candidates for an election). Most of the time these ordinal scales do a good job of representing our desires; but we do not know for sure that our preferences are structured in the format implied by ordinal scales. For this they should be (among other things) stable, other things being equal (no shifting preferences without a reason), and transitive (if X is preferred to Y and Y to Z, X is preferred to Z). There are good psychological reasons to doubt this (made famous by Tversky 1969, although see Regenwetter et al. 2011 for a rejoinder). Ordinal scales, thus, do not imply ordinal *variables*. Preferences, assuming they have a psychological reality, are arguably more complex than the linear orderings an ordinal scale takes them to be. Yet, preference rankings are routinely given a strongly quantitative interpretation in psychological or economic research.

(2) *Is happiness a quantity?* Some readers might be familiar with the claims of happiness research, a thriving branch of behavioral economics. One famous paper in the discipline is widely thought to show that money has a constant effect on well-being

[2] Thanks to Ann-Sophie Barwich for drawing my attention to it.

and happiness, more precisely that happiness is increased by wealth in a log linear relationship, with happiness increasing as a logarithmic function of wealth (Stevenson and Wolfers 2013).[3] This research is based on self-reports, with subjects rating their degree of happiness on a scale from 0 to 6. Like their colleagues and contradictors, the authors of that study assume happiness, so rated, to be an additive quantity—in other words, a move from 4 to 6 on the scale represents the same increase in happiness as a move from 2 to 4. Such an assumption would only make sense given another, independent measure of subjective happiness, but happiness researchers make little attempt at calibrating their measurements in this way. Instead, they simply assume that the happiness variable has an additive structure. The circularity is deepened, not lifted, when correlations such as that between happiness and financial standing are taken to validate the scales being used. To make matters worse, press reports tend to use turns of phrase, such as, "Americans are twice as rich as Brazilians but not twice as happy," that assume happiness to be a genuine quantity, with a nonarbitrary zero point, just like money—a problematic assumption, since no one claims to have found such a point.

(3) *Do causal factors add up?* Consider the following claims:

(a) Two thirds of man-made global warming is due to only 90 firms.
(b) Schizophrenia is 80 % genetic.

Both appear to measure the contribution of a given mechanism (human activities, genes) to a complex phenomenon that depends on at least a few other causes. As philosophers of science have noted (Keller 2010 summarizes the argument; see also Bookstein 2009), for such claims to make sense, the contribution of the mechanisms in question should be additive. If, say, a gene increases your chances of having schizophrenia if you are a vegetarian but decreases them if you are a meat-eater, we cannot quantify its overall contribution to the incidence of the disease in any simple way. The problem is that, strictly speaking, these assumptions of additivity are approximate at best, especially when applied to very complex systems like mental health or climate[4]: mechanisms interact in various and sundry ways, not always tractable. Causal impact, or explanatory power, are not genuinely additive variables: if so, the two claims above are difficult to make sense of, and both have indeed been criticized on this ground (see, e.g., Keller's criticisms of behavioral genetics as using meaningless measurements; Keller 2010, Chap. 1). There are, however, two reasons not to take that particular line of criticism too far. First, we do not know how wrong the assumptions of additivity are. Second, even if the assumption of additivity were wrong, a variable's explanatory power is a useful thing to know, even if we cannot really tell what kind of a quantity that explanatory power is—or whether it is a quantity at all. Taken in context, along with other indicators, it can usefully reorient scientific debates: the abandonment of the "schizogenic mother hypothesis," or the refutation of global warming skeptics, are not trivial achievements.

In a way, the making of questionable quantitative assumptions looks like nothing more than an expression of the healthy scientific habit of theorizing ahead of the facts. In a way, many "shmeasurers" are in the position that Kuhn described regarding early work on temperature:

> Many of the early experiments involving thermometers read like investigations *of* that instrument rather than like investigations *with* it. How could anything else have been the case during a period when it was totally unclear what the thermometer measured? Its readings obviously depended on the "degree of heat", but apparently in immensely complex ways. (Kuhn 1961, p. 189)

Today, the progression of temperature measurement, from ordinal to interval and finally to ratio scales, seems to us a foregone conclusion; but Kuhn rightly insisted that the decision to focus research on those aspects of temperature that were most amenable to quantification was a risky bet. That bet could have been lost: there was no way of *knowing* that a genuine magnitude lay underneath their measurements. The researchers who have put themselves in this uncomfortable (but exciting) position deserve to be given some slack. Even cross-national studies of happiness, for all their premature conclusions, could be on the right track in construing contentment as a magnitude.

These researchers differ from early students of temperature in one key respect, though: they are not actively investigating the structure of the variable they study. They are not busy proving its quantitative nature. In part, this is because of the nature of subjective happiness—a somewhat solipsistic phenomenon. Yet, in our examples (1) and (3), where independent calibration would be more practicable, metrological issues are not high on the agenda. Instead, the view is sometimes mooted that elegant and robust results can somehow validate questionable quantitative assumptions.

[3] I must apologize for singling out this one paper to raise a much more general problem. The methodological issue at stake here reaches much beyond the work of these two scientists, and beyond the field of happiness research.

[4] My example, not Keller's or Bookstein's.

Using Measurement Tools to Answer Qualitative Questions

This second form of misapplication of quantitative tools is the mirror image of the first. While type-1 "shmeasurement" pretends to measure things that cannot be measured, type-2 "shmeasurement" fails to measure things, preferring to use them to ask qualitative questions instead.

This is my interpretation of the criticisms (some of them, at least) that have been leveled at the current practices of statistical testing, their most notorious target being null hypothesis significance testing (NHST). NHST has been attacked on various grounds (for instance, it does not test the hypothesis of interest, but assumes a null hypothesis that is often meaningless). Nevertheless, most accusations revolve around NHST's detachment from quantification: p-values are notoriously hard to interpret and have little to do with effect sizes (Gigerenzer 2004; Ziliak and McCloskey 2008; Cumming 2013). The threshold of 0.05 is an arbitrary convention: meeting it or not tells us nothing about the observed magnitudes. Confidence intervals are routinely neglected (Cummings 2013, along with the critics already cited). In Deirdre McCloskey's felicitous phrasing, the sciences that rely on NHST fail to "ask How Much" (McCloskey 2002). They have lost track of the quantitative.

In other words, NHST is what we might call a yes/no test, a statistical tool that presents itself as a quantitative measure, but actually serves to answer qualitative questions in a binary way. NHST is simply the most prominent yes/no test in current use, not the only one. A similar criticism could be addressed to any test built to determine whether an effect meets some prespecified conventional threshold. That includes critical ratios—the predecessors of p values—and possibly some Bayesian tests (as Cummings 2013 notes).

The two forms of "shmeasurement" we just reviewed have one thing in common: they blur the line separating the quantitative from the nonquantitative. Making unwarranted quantitative assumptions (type-1 "shmeasurement") means trying to measure things that cannot be measured, because they are not as quantitative as we think they are. To use yes/no tests (type-2 "shmeasurement") is to use quantitative tools to answer qualitative questions. Yet both practices have proven extremely popular in the empirical sciences, from biology to sociology, and precisely among those researchers who strive to make their research as "quantitative" as possible. Why is that?

What "Physics Envy" Does Not Explain

Critics of mainstream quantitative science have a simple—too simple—explanation for the misuse of quantitative methods in the sciences. It boils down to a sad tale of arrogance, laziness, and hypocrisy. Everything starts (to summarize) with an inferiority complex. The burgeoning social and biological sciences (economics and psychology in particular) modeled their methods and their jargon on physics in an attempt to gain professional legitimacy. Auguste Comte's positivism is often presented as the main foil in this story. Various other brands of scientism are also pointed at. "Physics envy" generates meaningless parodies of true measurement, what some authors call "cargo cult science" (Feynman 1974; McCloskey 2002).[5] Since psychologists, economists, and other social scientists cannot reproduce the achievements of the harder sciences, they are constantly blurring the line between true quantitative inquiries and their inferior imitation of it.

My summary is rather superficial, but the polemics against mainstream quantitative methods have not presented a much less sketchy account. As far as I know, there has been no equivalent for "shmeasurement" of Friedric Hayek's masterful genealogy of scientism, *The Counter-Revolution of Science* (1955). Even Steven Ziliak and Deirdre McCloskey's excellent *Myth of Statistical Significance* (2008), in spite of the authors' genius for dramatizing the driest technical issues, only manages to expose a rather banal dispute between an agronomer and a biologist. Michell brilliantly retraces the history of measurement theory, but his contention that psychology is "haunted by the ghost of Pythagoras" is polemics, more than explanation; it is hard to tell whether Gigerenzer (2004) takes his own Freudian speculation very seriously; Gould's "physics envy" (1996), though a powerful slogan, is little more than sarcasm.

The "cargo cult" narrative fails to address key issues. Mainstream quantitative methods simply do not look like a bad imitation of physics. The statistical methods of social science do not come from physics; if anything, statistical physics borrowed tools that had been perfected by state administrations (which had borrowed them, in turn, from gamblers and insurers) to deal with social phenomena (Hacking 1990). Influential as they might have been, neither scientism nor positivism (on which see Hayek 1955) started as quantitative inquiries. Auguste Comte is a case in point (Comte 1864): his positive sociology was not quantitative at all and had no place for statistics, economics, or psychology (three disciplines he strongly despised). It was brought to maturity by followers (like Émile Durkheim) who took their quantitative tools from state statistics, not from the hard sciences. Turning to the present, one is struck by how little there is in common between the quantitative

[5] "Cargo cult science," an expression made popular by Feynman (1974)—refers to Melanesian messianic movements that famously involved imitating the trappings of Western technology, but not its substance: headphones made from wood, airplanes made of straw, etc.

toolkit shared by many biomedical, social, or psychological sciences, and that of physics.

The critics nonetheless strike a chord when they point at the gap between the quantitative pretensions of most of our statistical measures, and the reality of scientific practice. Still, condemnation takes us only so far. We can agree that the toolkit of much empirical science is not quantitative in the way it purports to be, and still try to figure out what it is and what purposes it serves.

Collations and Tallies Versus Measurements

Collations and tallies have already been introduced. They share some properties with measurement:

A *collation* is the act of putting a number of observations together, in a commensurable format, in order to get a general picture. Collations take their name from philology: one collates several manuscripts to establish the authoritative version of a text. Words and sentences are included in the canon if they figure in a sufficient number of versions, or in sufficiently authoritative versions (with the occasional minority report included in the footnotes). The observations being collated will often not be measurements, but simple qualitative reports. Collations, thus, are simply aggregate empirical descriptions. They may or may not aggregate measurements. Collations and measurements go together so often in scientific practices that the distinction is easily missed; outside science, though, many measurements apply to one case only. Collations may aggregate quantitative observations (measurements), but qualitative indicators may also be collated. The World Color Survey, for instance, collated the colors associated with the words of a variety of languages, in a way that was systematic but not quantitative (Berlin and Kay 1969).

A *tally* similarly gathers and synthetizes observations according to a rule, but it does so in order to produce one single and unambiguous answer to a question, a question specified in advance along with the rules of the tally. Unlike collations, tallies do not usually have solely descriptive aims. The clearest example of a tally is a vote: a vote is a codified way of aggregating preferences that yields a decision that every voter commits to respecting, whatever the result turns out to be. The voting rule may not involve numerical calculations (think of voting by acclamation) but it often does (e.g., in majority voting). Sports competitions are tallies as well: their goal is to select a winner according to a rule; the rule may often mention quantitative measurements (as in swimming races), but it does not need to (think of acrobatic diving or ice skating). Here again the rule may often involve numerical calculations but need not do so. Tallies may use various kinds of indicators, some measurements among them, but this does not turn tallies into measurements. They are not designed to measure (or simply to represent) the variables they take into account. Voting is a case in point. Mathematical theories of voting have shown that the outcomes of voting procedures necessarily misrepresent at least some of the voters' preferences (Pacuit 2012). There are always several ways of tallying peoples' choices, which often could lead to dramatically different results (list voting, majority voting, voting by ranking, etc.). A political psychologist who would want to assess the voters' inclinations would obviously need to take several such indicators into account. Likewise, the results of sports competition do not necessarily reflect the overall quality of the participants' performance. This remark is true, but it clearly misses the point of votes and competitions: their aim is not to measure or represent anything, but to simplify reality in a way that produces a clear choice. Getting such a clear outcome can be desirable because some controversy must be terminated, because a political problem needs to be solved, or simply for the fun of picking a winner and a loser for a game. Tallies are supposed to manufacture unambiguous choices from a controversial material: unlike collations, their purpose is decision, not description.

Collations and tallies only bear an indirect relation to measurement as previously defined. Collations may collate measurements—in fact, the measurements treated by statistical tools are often collations of measurements, not individual data points—but they can also use other indicators. Tallies may or may not use measurements. Thus, collations and tallies do not necessarily deal with magnitudes; besides, they necessarily apply to a plurality of observations, while a measurement can be unique.

On the other hand, the indicators used in collations and tallies fit the mainstream view of measurement quite well. Stanley Stevens' definition (anathema to purists like Michell) sees measurement as "the assignment of numerals to objects according to rules" (Stevens 1946, p. 677). If we add "events" and "choices" to "objects," we get a passable definition of the use of numerical indicators in tallies and conventions. The rule mapping events and observations to numerical indicators is left deliberately unspecified, in the place where the classic definition of measurement would mention the ratio of a magnitude upon a unit of measure.

The point has sometimes been made that what I call collations and tallies should fall into the category of measurement, broadly construed. This is one of the main contentions in O. D. Duncan's *Notes on Social Measurement* (1984; I thank Fred Bookstein for attracting my attention to it). Duncan's ambition is to bring under one

overarching concept ("social measurement") a variety of scales and classifying tools, ranging from votes, social ranks, and prices, to weights and measures, psychophysical scales, and demographic tools like census. (See Duncan 1984, Chap. 3, for a complete list.) It is hard to think of a more inclusive notion of measurement, but Duncan defends the view that all these things share at least two properties. First, they are all more or less based on social conventions and responding to social needs (1984, Chap. 2). Second, they are all modes of measurement: even voting should be studied in the light of historical metrology (1984, Chap. 3).

Both contentions are debatable. I agree that measurements, collations, and tallies are all social in some way; but this is not saying much. Measurements are social in the minimal sense that measurement units are arbitrary conventions, shaped by societies (they may conceivably be purely individual, as happens when a scientist devises a measurement tool for his or her own individual use). Most measurement units are social conventions, but the measurements themselves (the ratio of the unit over a magnitude) are not, and the quantities being measured are often nonsocial. Voting is social in a much stronger sense: the variables that it takes into account usually concern social issues, and its outcome is a collective decision.

As for collations and tallies being a form of measurement, I am not sure that Duncan takes his own proposal very seriously. Indeed, he insists on treating measurement of physical magnitudes as a distinct category (1984, Chap. 5). He even faults Stevens for "implying that all classifications are attempts at 'measurement', inferior attempts at that" (Duncan 1984, p. 137). I agree with these remarks: the variety of classifications, collations, and tallies surveyed by Duncan should not be put in the same category as quantitative measurement, or judged by the same standards. They have their own, specific standards, to which I now turn (Table 1).

The Inverted Values of Collations and Tallies

A good tally or collation may not make for good measurement, and vice versa. This section will suggest that the misuses of measurement considered above ("shmeasurements"), actually fulfill the functions of collations, or tallies. They are more than failed or fake measurements. Once we understand their true function many of their drawbacks may appear to be (as programmers say) features, not bugs.

There is a logic of inversion to collations and tallies, which turns the precepts of sound measurement on their heads. With them, in Orwellian fashion, *ignorance is strength* (one should sometimes choose crude and partial indicators over more complete and refined ones) and *freedom is slavery* (submitting to arbitrary conventions is useful).

- *Ignorance is strength*: *crude and partial indicators may be desirable*. Collators face a trade-off between sophistication and robustness: they need to explore many data points with a few commensurate indicators. The most accurate indicators are context-sensitive, and require information that may not be available for every case. Because of this, cruder indicators may be preferred: they increase the range of cases over which collators can aggregate their observations. The cruder the indicators, the wider the range of cases covered, the more robust the collations. If we go to extremes, a collation that would attune its indicators to the idiosyncrasies of every case would not be a collation but a mere description. Tallies, insofar as they involve collating observations, are also subject to this constraint—even more so, since they are often asked to yield an unambiguous answer to one question.
- *Freedom is slavery*: *blindly following arbitrary conventions can be a good thing*. Tallies are meant to settle disputes, which is to say that their users can be expected to disagree on many things, up to and including the way of settling the dispute. If one participant devises a test, it can only be suspected of advantaging his own side of the controversy. Trying to agree upon a novel procedure to solve the conflict is one sure way of prolonging the dispute. Conventional and arbitrary tests (especially when imbued with some vaguely ritual authority) are especially useful here, since they preempt the need for explicit agreement.

Here we might find a way to make sense of the practice of yes/no testing. Besides noting that NHST does not measure anything (as it tries to reduce quantitative realities to simple, dichotomous answers) its critics are fond of noting its ritual, almost religious character (Gigerenzer 2004; Ziliak and McCloskey 2008; Bookstein 2014). If we consider yes/no tests as a tally, though, we can look on all these things as features, not bugs: NHST is not a

Table 1 Sketch of a typology of several, more or less "quantitative" modes of inquiry

	Measurements	Collations	Tallies
Dealing with magnitudes?	Yes	Depends	Depends
For description or decision?	Description	Description	Decision
Aggregating several observations?	Depends	Yes	Yes

measurement, because that is not its function. It serves to terminate certain disputes. Scholars (when they play the game honestly, which as we know is not always the case) agree in advance to give up on some hypotheses that fail to pass the test. Here again, the use of one crude and arbitrary standard of success (instead of many subtle ones) is instrumental: a tally should not have too many degrees of freedom, as any ambiguity in the tally may be used to save one's pet hypothesis and keep the dispute going on. The quasi-religious attachment to a meaningless ritual also makes sense in this context: stubbornly sticking to one arbitrary standard obviates the need to justify one's chosen method of dispute resolution.

Sociologists of science often remark that textbook methodologies do not provide scientists with means of terminating controversies (Collins and Pinch [1993]2012, Chap. 5). They cannot, because they do not yield definite certainties. One can always keep one's theories in an artificial coma, fed by ad hoc reasoning and methodological quibbles. Yes/no tests, however, are an (imperfect) way of bringing disputes to an end. For this, though, they necessarily have to contravene some of the prescriptions of sound scientific method. Once NHST is gone, another form of yes/no testing is likely to occupy its niche: yes/no testing is likely to stay in existence as long as there is a need to put scientific controversies to a premature end. Since our time and resources are limited, that need is not likely to disappear.

Conclusion

Most of the tools discussed here could be loosely described as "quantitative" in certain ways: they often use numbers and figures to describe the world or settle issues; they often measure quantities (and sometimes pretend to do so). Yet, I have argued, the surest way of misjudging all these ways of doing research is to assume that they form part of a coherent method, with the measurement of magnitudes at its core. Estimating or predicting magnitudes is not the goal of most inquiries, even in the most ostensibly mathematized fields of the empirical sciences. The efforts spent on promoting or criticizing "quantitative" research are mostly wasted: the various styles of quantitative inquiry distinguished here do not have enough in common that they could be defended or attacked as one block.

Should we, then, stop using that slippery word, "quantitative"? It is now used to point at two completely different things (at least). The first is the ideal of measurement defined in the classic, Euclidean fashion (as in the "Measurement: The Standard View" section). The second is the use of mathematical formalism, often thought to bring precision and clarity to our theories. In the latter acception, it has nothing to do with measurement. Some theoreticians and model builders create mathematical constructions that are never compared with any empirical quantity. On the other hand, some measurements can easily be visualized without going through any kind of mathematical formalism.[6] As for precision and exactness, they may be desirable qualities for measurement to have, but one can do with a modicum of fuzziness. Approximation cannot always be avoided, and is not always to be avoided at all costs (as Kuhn 1961 made clear). Qualitative inquiries can be more exact and precise than quantitative ones: compare climatology or econometrics with genealogy or philology. Any inquiry that "asks How Much" (to use McCloskey's phrase once more) is quantitative in the classic sense. This in itself does not make it more exact, clearer or more precise: the virtues of good measurement are not the only scientific virtues.

The word "quantitative," thus, is not just equivocal. It tempts us into judging scientific work by inadequate standards, the standards of quantitative inquiry. This, I have argued, prevents us from understanding many virtues (and vices) of our current way of working with data. When looked at through quantitative lenses, many tools in use in the empirical sciences appear distorted: at best, a clumsy attempt at quantifying what cannot be quantified, a parody of physics at worst. A less dismal picture may reveal itself if we change the lenses.

Acknowledgments I wish to thank Ann-Sophie Barwich, Fred Bookstein, Evelyn Fox Keller, and Isabella Sarto-Jackson for their valuable input during and after our 2014 workshop. Memories of this event bring us all back to the late Werner Callebaut—his warmth, his competence, his geniality. He is sorely missed.

References

Berlin B, Kay P (1969) Basic color terms: their universality and evolution. University of California Press, Berkeley

Bookstein FL (2009) Measurement, explanation, and biology: lessons from a long century. Biol Theory 4:6–20

Bookstein FL (2014) Measuring and reasoning: numerical inference in the sciences. Cambridge University Press, New York

Collins HM, Pinch T (1993) The golem: what you should know about science. Cambridge University Press, New York

Comte A (1864) Cours de philosophie positive. J. Baillère et fils, Paris

Cumming G (2013) The new statistics: why and how. Psychol Sci 25:7–29. doi:10.1177/0956797613504966

Duncan OD (1984) Notes on social measurement—historical and critical. Russell Sage Foundation, New York

Feynman R (1974) Cargo cult science. Eng Sci 37(7):10–13

Gigerenzer G (2004) Mindless statistics. J Socioecon 33(5):587–606

[6] Intricate debates surround the question of knowing whether all measured quantities can in principle be represented with numbers, but I won't get into those.

Girard C (2010) Acclamation voting in Sparta: an early use of approval voting. In: Laslier JF, Sanver MR (eds) Handbook of approval voting. Springer, Berlin, pp 15–17
Gould SJ (1996) The mismeasure of man. Norton, New York
Hacking I (1990) The taming of chance. Cambridge University Press, New York
Hayek F (1955) The counter-revolution of science: studies on the abuse of reason. Free Press, New York
Keller EF (2010) The mirage of a space between nature and nurture. Duke University Press, Durham
Kuhn TS (1961) The function of measurement in modern physical science. Isis 52:161–193
McCloskey D (2002) The secret sins of economics. Prickly Paradigm Press, Chicago
Michell J (2005) Measurement in psychology: a critical history of a methodological concept. Cambridge University Press, New York
Michell J (2011) Qualitative research meets the ghost of Pythagoras. Theory Psychol 21:241–259. doi:10.1177/0959354310391351
Pacuit E (2012) Voting methods. In: Zalta EN (ed) The Stanford encyclopedia of philosophy. Winter. http://plato.stanford.edu/archives/win2012/entries/voting-methods/. Accessed May 2015
Regenwetter M, Dana J, Davis-Stober CP (2011) Transitivity of preferences. Psychol Rev 118:42–56
Stevens SS (1946) On the theory of scales of measurement. Science 103(2684):677–680. doi:10.1126/science.103.2684.677
Stevenson B, Wolfers J (2013) Subjective well-being and income: is there any evidence of satiation? Am Econ Rev 103:598–604
Tversky A (1969) Intransitivity of preferences. Psychol Rev 76:31–48
Ziliak S, McCloskey D (2008) The cult of statistical significance: how the standard error costs us jobs, money, and lives. University of Michigan Press, Ann Arbor