How Does Measuring Generate Evidence? The Problem of Observational Grounding

Eran Tal

1Assistant Professor, Department of Philosophy, McGill University, Montreal, Canada
E-mail: gran.tal@mcgill.ca

Abstract. The epistemology of measurement is an area of philosophy that studies the relationships between measurement and knowledge. One of its central aims is to explain how measurement can function as a reliable source of scientific evidence. Key to such explanation is a clear characterization of the dependence of measurement on observation, but such characterization has remained elusive. This article traces the recent historical trajectory of views on the observational grounding of measurement, clarifies the current state of the problem, and proposes new directions for progress. Specifically, I argue in favour of viewing measurement outcomes as the best predictors of observed instrument indications under a given theoretical-statistical model of the measurement process. The evidential efficacy of measurement outcomes is explained by their relatively high epistemic security, rather than by their inferential or structural closeness to observation.

1. Introduction
Measurement outcomes are commonly reported as unconditional factual claims. Examples are statements such as “the electric current in this wire at time t\textsubscript{1} is 1.0±0.3 ampere” and “the CO\textsubscript{2} concentration in this air sample at time t\textsubscript{2} is 400±10 ppm”. Claims of this form are habitually used as evidence for testing scientific hypotheses, for designing and testing instruments, and for making decisions in areas such as healthcare, public policy and everyday life.

And yet, as is well-known to anyone who designs measuring systems, no measurement outcome is unconditional. Assumptions are always involved in designing instruments, in executing measurement procedures, in interpreting and analyzing data, in evaluating bias and uncertainty, and in calibrating measuring devices against standards. Despite this, scientists do not usually report measurement outcomes in the form: “if assumptions \{A\textsubscript{1}, A\textsubscript{2} ... A\textsubscript{n}\} about the measurement process hold, then the electric current in this wire at time t\textsubscript{1} is 1.0±0.3 ampere”. What warrants the neglect of background assumptions in the final report, and is the epistemic status of measurement outcomes as unconditional evidence justified?

Explaining the evidential efficacy of measurement is a central task of the epistemology of measurement, an area of philosophy that studies the relationships between measurement and knowledge [1][2][3]. This task becomes particularly interesting when one compares measurement to other kinds of estimation, such as the derivation of quantity values through theoretical prediction and computer simulation. Much like measurement, these other forms of estimation involve a combination

---

1 To whom any correspondence should be addressed.
of observed data, inference and background assumptions. Unlike measurement, scientists rarely treat the results of these other kinds of estimation methods as unconditional evidence. Instead, the results of theoretical predictions and computer simulations are usually viewed as conditional on the assumptions that informed the estimation – at least until such results are confirmed through measurement.

What warrants this seemingly privileged evidential status of measurement? Traditionally, answers have appealed to a special sort of connection between measurement and qualitative observation. In what follows I will discuss and reject the traditional view, formulate a problem that arises from the rejection of the traditional view, and outline a solution to that problem. Ultimately, I will argue that the warrant for treating measurement outcomes as unconditional evidence is their high epistemic security, rather than their inferential or structural closeness to observation.

2. Is measurement a kind of observation?

2.1. Foundationalism about measurement

Measurement has been traditionally viewed as quantitative observation, namely as an activity in which numbers are assigned to objects or events based on observation [4]. The claim that measurement is a kind of observation is trivially true if the term ‘observation’ is used so loosely as to encompass all measurement by default. In some contexts such loose usage may be warranted [5]. However, the classification of measurement as observation only becomes philosophically interesting when ‘observation’ is understood more narrowly, for example, as knowledge one obtains through immediate and unaided sense perception. Although the precise explication of such a concept of observation is riddled with philosophical difficulties, for current purposes we can suppose that scientists observe the indications of their measuring instruments, such as the positions of pointers and numerals appearing on displays. There is no doubt that measurement involves observation in this sense. Measurement also typically involves inference, theory, statistics, abstraction, idealization, prediction, calculation and instrument-building, but one is not usually drawn to claim that measurement is a species of any of these. Why, then, is it tempting to believe that measurement is a kind of observation, and is this belief justified?

A central advantage of classifying measurement as observation is that it immediately becomes clear why scientists are justified in using measurement outcomes as evidence for testing hypotheses and designing new instruments. If measurement outcomes are no more than observational reports cast in mathematical language, the empirical content of any measurement outcome is reducible in principle to some set of observations. The most comprehensive attempt to analyze measurement along these foundationalist lines is due to Norman Campbell [6]. Campbell’s ‘fundamental measurement’ is a type of procedure that exhibits the additive structure of a quantity, such as length, weight or time duration, based on qualitative observations alone and without requiring antecedent measurements. Other, ‘derived’ procedures can then extend the discovery of structure to quantities such as density and temperature, for which no fundamental measurement procedure is known. Viewed in this way, the entire edifice of physical measurement is ultimately grounded in nothing but qualitative observation.

Despite their appeal, foundationalist views of measurement suffer from several interconnected problems, of which two are especially relevant to the current discussion. The first problem concerns idealization. Campbell and his followers assumed that qualitative structures were given in observation, but in practice such structures are rarely observed in a reproducible manner. For example, consider any two clocks that are meant to instantiate the same frequency, that is, to ‘tick’ at the same rate. Even the most stable atomic clocks exhibit a systematic frequency drift when compared accurately enough. It is easy to show that systematically drifting clocks provide inconsistent estimates of duration, and in some cases even assign the same time intervals a completely reverse order of duration. Despite this, practicing scientists do not infer the existence of multiple, fundamentally measureable kinds of quantity, ‘time-A’, ‘time-B’ etc. Rather, they suppose an ideal ordering of time intervals and assume that concrete clocks merely approximate this ideal. This is by no means special to clocks. Indeed, nearly every comparison between the indications of measuring instruments reveals nonlinear
systematic discrepancies if carried out accurately and for a sufficiently large sample. In practice, then, Campbell’s qualitative structures are not given by observation, but are idealizations of the measurement process in the hypothetical absence of error.

An additional difficulty is that the choice of idealized structure for the underlying quantity is not uniquely determined by observation. The underdetermination of measurement outcomes by observation is the second major problem facing foundationalist views of measurement. Going back to the clock example, no amount of data obtained through clock comparisons can by itself reveal which clock has the more stable frequency. Multiple ways of distributing errors among different clocks exist that are all compatible with observation, but yield inconsistent estimates of time duration. The inferential gap between instrument indications and measurement outcomes must be closed by other means. Theoretical considerations, along with considerations of convenience, fill this gap [7].

Taken together, the problems of idealization and underdetermination undermine the foundationalist attempt to reduce the content of measurement outcomes to raw, qualitative observations. This point must have been partially recognized by the authors of the Representational Theory of Measurement [8]. RTM provides an axiomatic treatment of qualitative structures, but calls such qualitative structures ‘abstractions’ and admits that not all of the axioms are empirically testable [9]. This is a step in the right direction, because it implicitly recognizes that extra-observational (such as theoretical, statistical, and pragmatic) assumptions inform choices of measurement scale and systematic error distribution. On the other hand, RTM says little about how its axioms may be justified, or about what sorts of observations would provide evidential support to those axioms that are empirically testable.

2.2. Observational grounding

The ambiguity surrounding the relationship between measurement and observation in post-foundationalist accounts gives rise to the problem of observational grounding [10][11]. The problem may be construed as a logical inconsistency among three compelling claims:

1. Measurement is a more reliable source of evidence than other forms of quantitative estimation, such as those based on theoretical prediction and computer simulation
2. If measurement is a more reliable source of evidence than other forms of quantitative estimation, it is because observation plays a special role in justifying measurement outcomes
3. Observation does not play a special role in justifying measurement outcomes compared to other forms of quantitative estimation, such as those based on theoretical prediction and computer simulation.

At least one of these claims must be rejected to avoid contradiction. Foundationalists like Campbell would reject (3), and argue that the content of measurement outcomes mirrors the structure of qualitative observation, whereas the content of quantitative estimates derived through theoretical prediction and simulation does not. The challenges of idealization and underdetermination block this move: at least some of the content of measurement outcomes is imposed from the ‘top down’, by adjusting inconsistent observations in light of idealized background assumptions. Observations partially constrain measurement outcomes but do not fully determine them. The same may be said of other forms of quantitative estimation. After all, observation plays important roles in theoretical and statistical modelling, prediction and simulation, such as providing raw data and determining initial and boundary conditions. In all of these activities, much like in measurement, observation is combined with inference and background assumptions to generate knowledge claims. What, then, is the special role observation plays in justifying measurement outcomes, if such unique role exists? And if such unique role does not exist, how does one explain the elevated epistemic status measurement is commonly thought to enjoy as a reliable source of evidence?

---

2 This problem should not be confused with the more familiar underdetermination of theory by observation.
A good first step is to consider the theoretical context in which measurements are performed. Historical examples show that the measurement of a certain kind of quantity and the theory behind that kind of quantity are co-dependent and progress through a process of mutual refinement. Hasok Chang has provided an illuminating account of how this process unfolded in the case of temperature [12]. Bas van Fraassen similarly describes the ‘empirical grounding’ of scientific theories as a “process of simultaneously, harmoniously extending both the theory and the range of relevant evidence” [13]. According to van Fraassen, a theory is empirically grounded when its significant parameters can be uniquely determined through measurement. Such determination must itself depend on the theory for modelling the measurement apparatus and analyzing instrument indications. Moreover, such determination must be refutable, in the sense that the theory does not simply fix measurement outcomes in advance but allows the production of outcomes that are inconsistent with the theory’s own predictions.

Van Fraassen’s criterion of empirical grounding is meant to apply to scientific theories rather than to measurement procedures. Nonetheless, the co-dependence of measurement and theory suggests that the observational grounding of measurement and the empirical grounding of theories are but two sides of the same epistemic coin. Measurement and theory have the right sort of relationship with observation when the production and analysis of observations supports a fruitful, progressive coherence between measurement and theory. A plausible criterion of observational grounding for measurement may therefore be obtained by inverting Van Fraassen’s criterion of empirical grounding for theories:

**Preliminary criterion of observational grounding.** A measurement outcome is grounded in a set of observations if and only if those observations were obtained and analyzed in a manner that determines the values of relevant theoretical parameters in a unique, coherent and refutable way.

### 2.3. A model-based perspective

The preliminary criterion of observational grounding is still very general. Before one can discern whether this criterion can resolve the problem of observational grounding, one needs to answer the question: what sort of relation between measurement outcomes and observations satisfies the criteria of uniqueness, coherence and refutability?

An answer may be gleaned from studying the methods metrologists use to calibrate measuring instruments and to trace uncertainties to measurement standards. Calibration is a kind of modelling activity, that is, the activity of constructing and testing a theoretical and / or statistical model of a measurement process [14][15]. The kinds of models constructed during calibration vary widely, from simple input-output (‘black-box’) representations of a measuring instrument to detailed and modular (‘white-box’) representations. Despite their diversity, all calibration procedures have a common aim, which is to reliably extrapolate backwards from the observed indications (or ‘end-states’) of the measuring procedure to values of the quantity intended to be measured.

Due to uncertainties, there is rarely a single value that would exclusively predict the observed indications. Calibration therefore aims at selecting the best range of predictors of each possible set of observed indications. The ‘best’ set of predictors is coherent with background theoretical and statistical assumptions, maximally accurate, and invariant (‘unique’) under changes to the measurement procedure and environment [15]. In addition, the best predictors need to be selective (or ‘refutable’) in the sense that they only predict a proper, and preferably small, subset of the possible indications of the measurement apparatus. In other words, the criteria for the selection of best predictors during calibration coincide with the criteria for the observational grounding of measurement, which in turn mirror the criteria for the empirical grounding of theories.

We are finally in a position to offer an answer to the opening question: is measurement a kind of observation? The answer is that measurement is not a kind of observation if by that one means that the empirical content of measurement outcomes is epistemically or semantically reducible to relations among qualitative observations. At the same time, measurement is observationally grounded, in the
sense that measurement outcomes have a specific sort of relationship with qualitative observations. Measurement outcomes constitute the best predictors of the observed end-states of a measurement process relative to a particular theoretical and statistical model of that process. The best predictors are those that maximize the predictive power of the model as well as that of underlying theories, and thereby enable the measurement and theory of the relevant quantity to cohere in the most accurate and universal manner practically possible.

3. How does measuring generate evidence?

The analysis of measurement outcomes as model-based predictors strengthens claim (3). There is no reason to think that observations constrain measurement results in a fundamentally different way than they do other sorts of model-based, quantitative scientific estimates. This does not mean that theoretical prediction and computer simulation are in general equally reliable sources of evidence compared to measurement. Although observation does not play a special role in justifying measurement outcomes, there are good reasons for thinking that measurement outcomes are usually more secure. The notion of secure evidence has been proposed by Kent Staley [16][17]. Roughly speaking, the security of a knowledge claim is higher for an epistemic agent when there are fewer possible scenarios in which that agent would deem the claim incorrect. Security is closely (and inversely) tied to the likelihood that, for all that a given scientific community knows, a knowledge claim would require revision in the foreseeable future. Note that security is not simply the inverse of uncertainty: a knowledge claim can be reported as having a low uncertainty but rest on assumptions that are likely to be revised in the foreseeable future. Such a claim would have a low security.

Measurement outcomes, especially those reported with low uncertainties, are never fully secure. At least some of the sources of information that contribute to the production of high-quality measurement outcomes are fallible. Theories are occasionally revised, statistical assumptions could be proven wrong, simplifying assumptions may be refined, and so on. To counter these sources of insecurity, scientists and engineers who design measuring systems try to rely only on uncontroversial and well-tested portions of theory and statistics, and to err on the side of caution when evaluating uncertainties. Moreover, scientists habitually compare measurement outcomes formed under different circumstances and assumptions to each other and test them for consistency within the limits of their respective uncertainties. Indeed, such comparison is implicit in the process of calibrating a measuring instrument against a standard. Consistency with a standard strengthens the likelihood of consistency with other measuring instruments calibrated against that standard, and recursively with an entire web of instruments linked to the standard through mutual comparisons [15]. These various precautions and robustness tests do not render measurement outcomes completely immune to revision, but they make it less likely that measurement outcomes will require revision in the foreseeable future given the current state of knowledge in the scientific community.

These considerations suggest that the key to understanding the source of evidential efficacy of measurement outcomes may lie in their relative security. The epistemic credentials of measurement are not different in kind from those of other modes of quantitative estimation, such as theoretical prediction and computer simulation. However, the outcomes of measurement procedures tend to be more secure, because they are obtained by following the precautions and passing the robustness tests detailed above. This partially explains why scientists are warranted in neglecting background assumptions when reporting measurement outcomes. The assumptions underlying measurement outcomes are usually selected in a conservative manner, and calibration further guarantees that the outcomes are invariant under changes to as many background assumptions as is practically possible to test. This is not generally the case for theoretical predictions and simulation-based estimates, which are often meant to explore speculative theoretical scenarios and are often not calibrated with the same level of rigour.

At the same time, nothing in principle prevents theoretical predictions and computer simulations from matching or even surpassing the degree of security of the best measurements available for a given quantity. This already appears to be the case for some ab initio calculations of the properties of...
simple atoms and molecules, such as the static electric dipole polarizability of helium [18]. Such results are not only accurate, but also considered sufficiently secure to underwrite the official determination of the values of fundamental physical constants, such as the Boltzmann constant [19]. In such cases it may be useful to think of the computation as a kind of measurement process, inasmuch as it generates evidence that is as secure as the best equivalent ‘traditional’ measurement outcomes.

4. Conclusions
Measuring generates evidence, not by preserving the structure of ‘raw’ qualitative observations, but by constraining the region on parameter space where the best predictors of observed indications are likely to lie. Such constraining involves observation along with other sources of information, including theories, tools of data analysis, simplifying assumptions, and computational techniques. Measurement outcomes are model-based predictors that have attained a high degree of security through various strategies, such as robustness tests. It is their security, rather than inferential or structural closeness to observation, that makes measurement outcomes suitable for serving as unconditional evidence claims.

This way of thinking about measurement leads to a novel solution to the problem of observational grounding. Claim (3) is retained, and no unique role is attributed to observation in justifying the outcomes of measurement. Instead, claim (2) is rejected, and the grounds for a difference in evidential reliability shift to a difference in epistemic security. Additionally, claim (1) is qualified, so that the elevated evidential status of measurement outcomes depends on their security. This proposed solution is but a sketch for an eventual, comprehensive solution to the problem of observational grounding. A full solution would require careful explication of the relevant notions of measurement, observation, evidence and information, as well as a clear account of the ways in which the epistemic security of measurement outcomes may be assessed and enhanced.

Acknowledgments
Work on this article was supported by a Marie Curie Intra-European Fellowship within the 7th European Community Framework Programme, held at the University of Cambridge.

References
[1] Mari L 2003 Measurement 34 17-30
[2] Leplège A 2003 Social Science Information 42 451-462
[3] Tal E 2016 British Journal for the Philosophy of Science 67 297-335
[4] Bunge M 1998 Philosophy of Science Volume 2 (New Brunswick and London: Transaction)
[5] Shapere D 1982 Philosophy of Science 49 485-525
[6] Campbell N R 1920 Physics: The Elements (London: Cambridge University Press)
[7] Poincaré H 1898 The measure of time The Value of Science (NY: Dover, 1958) pp 26–36
[8] Krantz D H, Luce R D, Suppes P, and Tversky A 1971 Foundations of Measurement volume 1 (San Diego and London: Academic Press)
[9] Luce R D, Krantz D H, Suppes P, and Tversky A 1990, Foundations of Measurement volume 3 (San Diego and London: Academic Press), section 21.8
[10] Tal E 2013 Philosophy Compass 8 1159-1173
[11] Tal E 2015 Measurement in science Stanford Encyclopedia of Philosophy ed E N Zalta
[12] Chang H 2004 Inventing Temperature (Oxford: Oxford University Press)
[13] van Fraassen B C 2012 Philosophy of Science 79 773-784
[14] Mari L 2005 Models of the measurement Process Handbook of Measuring Systems Design vol 2 ed P Sydenham and R Thorn (Wiley) chapter 104
[15] Tal E Measurement, calibration and inference to the best prediction (under review)
[16] Staley K W 2004 Philosophy of Science 71 467-488.
[17] Staley K W 2012 European Journal for Philosophy of Science 2 21-43
[18] Lach G, Jeziorski B and Szalewicz K 2004 Phys. Rev. 92 233001
[19] Fischer J 2015 Metrologia 52 S364–S375