How Accurate Is the Standard Second?

Eran Tal[†‡]

Contrary to the claim that measurement standards are absolutely accurate by definition, I argue that unit definitions do not completely fix the referents of unit terms. Instead, idealized models play a crucial semantic role in coordinating the theoretical definition of a unit with its multiple concrete realizations. The accuracy of realizations is evaluated by comparing them to each other in light of their respective models. The epistemic credentials of this method are examined and illustrated through an analysis of the contemporary standardization of time. I distinguish among five senses of "measurement accuracy" and clarify how idealizations enable the assessment of accuracy in each sense.

1. Introduction. A common philosophical myth states that the meter bar in Paris is exactly 1 meter long—that is, if any determinate length can be ascribed to it in the metric system. One variant of the myth comes from Wittgenstein (1953, sec. 50), who tells us that the meter bar is the one thing "of which one can say neither that it is one metre long, nor that it is not one metre long." Kripke (1980, 56) famously disagrees but develops a variant of the same myth by stating that the length of the bar at a specified time is rigidly designated by the phrase "one meter." Neither of these pronouncements is easily reconciled with the 1960 declaration of the General Conference on Weights and Measures, according to which "the international Prototype does not define the metre with an accuracy

†To contact the author, please write to: Department of Philosophy, University of Toronto, 170 St. George, Toronto, ON M5R2M8, Canada; e-mail: eran.tal@utoronto.ca.

‡I am indebted to Judah Levine, David Wineland, Till Rosenband, and their colleagues at the Time and Frequency Division at the U.S. National Institute of Standards and Technology (NIST) for valuable discussions. My visit to NIST in summer 2009 was made possible with kind help from Ian Hacking and Allan Franklin as well as a travel grant from the University of Toronto. I am also grateful to Margaret Morrison, Ian Hacking, Anjan Chakravartty, Marcel Boumans, Stephan Hartmann, Carlo Martini, Isaac Record, Paul Teller, Jonathan Weisberg, and several members of the audience at the Philosophy of Science Association meeting for helpful feedback on earlier versions of this article.

Philosophy of Science, 78 (December 2011) pp. 1082–1096. 0031-8248/2011/7805-0030$10.00
Copyright 2011 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/662268 Published online by Cambridge University Press

adequate for the present needs of metrology" and is therefore replaced by an atomic standard (CGPM 1961). There is, of course, nothing problematic with replacing one definition with another. But how can the accuracy of the meter bar be evaluated against anything other than itself, let alone be found lacking?

Wittgenstein and Kripke almost certainly did not subscribe to the myth they helped disseminate. There are good reasons to believe that their examples were meant merely as hypothetical illustrations of their views on meaning and reference.[1] This article does not take issue with their accounts of language but with the myth of the absolute accuracy of measurement standards, which has remained unchallenged by philosophers of science. The meter is not the only case where the myth clashes with scientific practice. The second and the kilogram, which are currently used to define all other units in the International System (i.e., the "metric" system), are associated with primary standards that undergo routine accuracy evaluations and are occasionally improved or replaced with more accurate ones. In the case of the second, for example, the accuracy of primary standards has increased more than a thousandfold over the past four decades (Lombardi et al. 2007).

This article will analyze the methodology of these evaluations and argue that they indeed provide estimates of accuracy in the same senses of "accuracy" normally presupposed by scientific and philosophical discussions of measurement. My main examples will come from the standardization of time. I will focus on the methods by which time and frequency standards are evaluated and improved at NIST. These tasks are carried out by *metrologists*, experts in highly reliable measurement. The methods and tools of metrology—a live discipline with its own journals and controversies—have received little attention from philosophers.[2] Recent philosophical literature on measurement has mostly been concerned either with the metaphysics of quantity and number (Swoyer 1987; Michell 1994) or with the mathematical structures underlying measurement scales (Krantz et al. 1971). These "abstract" approaches treat the topics of uncertainty, accuracy, and error as extrinsic to the theory of measurement and as arising merely from imperfections in its application.

By contrast, the current study is meant as a step toward a practice-oriented epistemology of physical measurement. The management of un-

1. Wittgenstein mentions the meter bar only in passing as an analogy to color language-games. Kripke carefully notes that the uniqueness of the meter bar's role in standardizing length is no more than a hypothetical supposition (1980, 55 n. 20).

2. Notable exceptions are Chang (2004) and Boumans (2007). Metrology has been studied by historians and sociologists of science (see, e.g., Latour 1987, chap. 6; Schaffer 1992; Galison 2003; and Gooday 2004).

https://doi.org/10.1086/662268 Published online by Cambridge University Press

certainty and error will be viewed as intrinsic to measurement. Specifically, I will use insights into metrological practices to outline a novel account of the underexplored relationship between uncertainty and accuracy. Scientists often include uncertainty estimates in their reports of measurement results, but whether such estimates warrant claims about the accuracy of results is an epistemological question that philosophers have overlooked. Using an analysis of time standardization, I argue that inferences from uncertainty to accuracy are justified when a doubly robust fit—among instruments as well as among idealized models of these instruments—is demonstrated. My account will shed light on metrologists' claims that the accuracy of standards is being continually improved and on the role played by idealized models in these improvements.

2. Five Notions of Measurement Accuracy. Accuracy is often ascribed to scientific theories, instruments, models, calculations, and data, although the meaning of the term varies greatly with context. Even within the limited context of physical measurement, the term carries multiple senses. For the sake of the current discussion, I offer a preliminary distinction among five notions of measurement accuracy. These are intended to capture different senses of the term as it is used by physicists and engineers as well as by philosophers of science. The five notions are neither coextensive nor mutually exclusive but instead partially overlap in their extensions. As I argue below, the sort of robustness test metrologists employ to evaluate the uncertainty of measurement standards provides sufficient evidence for the accuracy of those standards under all five senses of "measurement accuracy."

1. **Metaphysical measurement accuracy:** Closeness of agreement between a measured value of a quantity and its true value (correlate concept: truth).[3]
2. **Epistemic measurement accuracy:** Closeness of agreement among values reasonably attributed to a quantity based on its measurement (correlate concept: uncertainty).[4]
3. **Operational measurement accuracy:** Closeness of agreement between a measured value of a quantity and a value of that quantity obtained by reference to a measurement standard (correlate concept: standardization).
4. **Comparative measurement accuracy:** Closeness of agreement among

3. Compare "Measurement Accuracy" in the *International Vocabulary of Metrology* (*VIM*) (JCGM 2008b, 2.13). My own definitions for measurement-related terms are inspired by, but in some cases diverge from, those of the *VIM*.

4. Compare JCGM (2008b), 2.13 n. 3.

https://doi.org/10.1086/662268 Published online by Cambridge University Press

measured values of a quantity obtained by using different measuring systems or by varying extraneous conditions in a controlled manner (correlate concept: reproducibility).

5. **Pragmatic measurement accuracy ("accuracy for"):** Measurement accuracy (in any of the above four senses) sufficient for meeting the requirements of a specified application.

Let us briefly clarify each of these five notions. First, the metaphysical notion takes *truth* to be the standard of accuracy. For example, a thermometer is metaphysically accurate if its outcomes are close to true ratios between measured temperature intervals and the chosen unit interval. If one assumes a traditional understanding of truth as correspondence with a mind-independent reality, the notion of metaphysical accuracy presupposes some form of realism about quantities. The argument advanced in this article is nevertheless independent of such realist assumptions, as it neither endorses nor rejects metaphysical conceptions of measurement accuracy.

Second, a thermometer is epistemically accurate if its design and use warrant the attribution of a narrow range of temperature values to objects. The dispersion of reasonably attributed values is called "measurement uncertainty" and is commonly expressed as a value range.[5] Epistemic accuracy should not be confused with measurement precision, which constitutes only one aspect of epistemic accuracy. Measurement precision is the closeness of agreement among measured values obtained by repeated measurements of the same (or relevantly similar) objects using the same measuring system.[6] Imprecision is therefore caused by uncontrolled variations to the equipment, operation, or environment when measurements are repeated. This sort of variation is a ubiquitous but not exclusive source of measurement uncertainty. As will be explained below, some measurement uncertainty stems from other sources, including imperfect corrections to systematic errors. The notion of epistemic accuracy is therefore broader than that of precision.

Third, operational measurement accuracy is determined relative to an established measurement standard. For example, a thermometer is operationally accurate if its outcomes are close to those of a standard thermometer when the two measure relevantly similar samples. The most common way of evaluating operational accuracy is by calibration, that

5. Compare "Measurement Uncertainty" (JCGM 2008b, 2.26). Note that this term does not refer to a degree of confidence or belief but to a dispersion of values whose attribution to a quantity reasonably satisfies a specified degree of confidence or belief.

6. Compare "Measurement Precision" (JCGM 2008b, 2.15) and "Measurement Repeatability" (2.21). My concept of precision is narrower than that of the *VIM* (see also n. 8).

https://doi.org/10.1086/662268 Published online by Cambridge University Press

is, by modeling an instrument in a manner that establishes a relation between its indications and standard quantity values.[7]

Fourth, comparative accuracy is the closeness of agreement among measurement outcomes when the same quantity is measured in different ways. The notion of comparative accuracy is closely linked with that of reproducibility. To say that a measurement outcome is comparatively accurate is to say that it is closely reproducible under controlled variations to measurement conditions and methods.[8] For example, thermometers in a given set are comparatively accurate if their outcomes closely agree when applied to relevantly similar samples.

Finally, pragmatic measurement accuracy is accuracy sufficient for a specific use, such as a solution to an engineering problem. There are four subsenses of pragmatic accuracy, corresponding to the first four senses of measurement accuracy. For example, a thermometer is pragmatically accurate in an epistemic sense if the overall uncertainty of its outcomes is low enough to reliably achieve a specified goal, for example, keeping an engine from overheating. Of course, whether a measuring system (or a measured value) is pragmatically accurate depends on its intended use.[9]

In the physical sciences quantitative expressions of measurement accuracy are typically cast in epistemic terms, namely, in terms of uncertainty. This does not mean that scientific estimates of accuracy are always and only estimates of epistemic accuracy. What matters to the classification of accuracy is not the form of its expression but the kind of evidence on which estimates of accuracy are based. As I argue below, metrological evaluations provide evidence of the right sort for estimating accuracy under all five notions. Before delving into the argument, the next section will provide some background on the concepts, methods, and problems involved in the standardization of time.

3. The Multiple Realizability of Unit Definitions. A key distinction in the standardization of physical units is that between definition and realization. Since 1967, the second has been defined as the duration of exactly 9,192,631,770 periods of the radiation corresponding to a hyperfine transition of cesium-133 in the ground state (BIPM 2006). This definition

7. Compare "Calibration" (JCGM 2008b, 2.39).

8. Unlike precision, reproducibility concerns controlled variations to measurement conditions. I deviate slightly from the *VIM* on this point to reflect general scientific usage of these terms (cf. "Measurement Reproducibility," JCGM [2008b], 2.25).

9. Pragmatic accuracy may be understood as a threshold (pass/fail) concept. Alternatively, pragmatic accuracy may be represented continuously, for example, as the likelihood of achieving the specified goal. Both analyses of the concept are compatible with the argument presented here.

https://doi.org/10.1086/662268 Published online by Cambridge University Press

pertains to an unperturbed cesium atom at a temperature of absolute zero. Being an idealized description of a kind of atomic system, no actual cesium atom ever satisfies this definition. Hence, a question arises as to how the reference of "second" is fixed. The traditional philosophical approach would be to propose some "semantic machinery" through which the definition succeeds in picking out a definite duration, for example, a possible-world semantics of counterfactuals. However, this sort of approach is hard pressed to explain how metrologists are able to experimentally access the extension of "second" given the fact that it is physically impossible to instantiate the conditions specified by the definition. Consequently, it becomes unclear how metrologists are able to tell whether the actual durations they label "second" satisfy the definition. By contrast, the approach adopted in this article takes the definition to fix a reference only indirectly and approximately by virtue of its role in guiding the construction of atomic clocks. Rather than picking out any definite duration on its own, the definition functions as an ideal specification for a class of atomic clocks. These clocks approximately satisfy—or in the metrological jargon, "realize"—the conditions specified by the definition. The activities of constructing and modeling cesium clocks are therefore taken to fulfill a semantic function, that is, that of approximately fixing the reference of "second," rather than simply measuring an already linguistically fixed time interval.

The construction of an accurate primary realization of the second—a "meter stick" of time—must make highly sophisticated use of theory, apparatus, and data analysis in order to approximate as much as possible the ideal conditions specified by the definition. But multiple kinds of physical processes can be constructed that would realize the second, each departing from the ideal definition in different respects and degrees. In other words, different clock designs and environments correspond to different ways of de-idealizing the definition. As of 2009, 13 atomic clocks around the globe are used as primary realizations of the second. There are also hundreds of official secondary realizations of the second, that is, atomic clocks that are traced to primary realizations. Like any collection of physical instruments, different realizations of the second disagree with one another, that is, "tick" at slightly different rates. The definition of the second is thus *multiply realizable* in the sense that multiple real durations approximately satisfy the definition, and no method can completely rid us of the approximations.

That the definition of the second is multiply realizable does not mean that there are as many "seconds" as there are clocks. What it does mean is that metrologists are faced with the task of continually evaluating the accuracy of each realization relative to the ideal cesium transition frequency and correcting its results accordingly. But the ideal frequency is

https://doi.org/10.1086/662268 Published online by Cambridge University Press

experimentally inaccessible, and primary standards have no higher standard against which they can be compared. The challenge, then, is to forge a unified second out of disparately ticking clocks. This is an instance of a general problem that I will call the "problem of multiple realizability of unit definitions."[10] This problem is semantic, epistemological, and methodological all at once. To solve it is to specify experimentally testable satisfaction criteria for the idealized definition of "second," a task that is equivalent to that of specifying grounds for making accuracy claims about cesium atomic clocks, which is in turn equivalent to the task of specifying a method for reconciling discrepancies among such clocks. The conceptual distinction among three axes of the problem should not obstruct its pragmatic unity, for as we see below, metrologists are able to resolve all three aspects of the problem simultaneously.

Prima facie, the problem can be solved by arbitrarily choosing one realization as the ultimate standard. Yet this solution would bind all measurement to the idiosyncrasies of a specific artifact, thereby causing measurement outcomes to diverge unnecessarily. Imagine that all clocks were calibrated against the ticks of a single apparatus: the instabilities of that apparatus would cause clocks to run faster or slower relative to each other depending on the time of their calibration, and the discrepancy would be revealed when these clocks were compared to each other. A similar scenario has recently unfolded with respect to the International Prototype of the kilogram when its mass was discovered to systematically "drift" from those of its official copies (Girard 1994). Hence, a stipulative approach to unit definitions exacerbates rather than removes the challenge of reconciling discrepancies among multiple standards.

The latter point is helpful in elucidating the misunderstanding behind the myth of absolute accuracy. Once it is acknowledged that unit definitions are multiply realizable, it becomes clear that no single physical object can be used in practice to completely fix the reference of a unit term. Rather, this reference must be determined by an ongoing comparison among multiple realizations. Because these comparisons involve some uncertainty, the references of unit terms remain vague to some extent. Nevertheless, as the next sections make clear, comparisons among standards allow metrologists to minimize this vagueness, thereby providing an optimal solution to the problem of multiple realizability.

4. Uncertainty and De-idealization. The clock design currently implemented in most primary realizations of the second is known as the cesium "fountain," so called because cesium atoms are tossed up in a vacuum

10. Chang's (2004) "problem of nomic measurement" is a closely related, albeit distinct, problem concerning the standardization of instruments.

https://doi.org/10.1086/662268 Published online by Cambridge University Press

cylinder and fall back due to gravity. The best cesium fountains are said to measure the relevant cesium transition frequency with a fractional uncertainty of less than 5 parts in 10^{16} (Jefferts et al. 2007).[11] It is worthwhile to examine how this number is determined. To start off, it is tempting to interpret this number naively as the standard deviation of clock outcomes from the ideally defined duration of the second. However, because the definition pertains to atoms in physically unattainable conditions, the aforementioned uncertainty could not have been evaluated by direct reference to the ideal second. Nor is this number the standard deviation of a sample of readings taken from multiple cesium fountain clocks. If a purely statistical approach of this sort were taken, metrologists would have little insight into the causes of the distribution and would be unable to tell which clocks "tick" closer to the defined frequency.

The accepted metrological solution to these difficulties is to de-idealize the theoretical definition of the second in discrete "steps" and estimate the uncertainty that each step contributes to the outcomes of a given clock. The uncertainty associated with a specific primary frequency standard is then taken to be the total uncertainty contributed to its outcomes by a sufficient de-idealization of the definition of the second as it applies to that particular clock. The rest of this section will describe how this uncertainty is evaluated, and the next section will describe what kind of evidence is taken to establish the "sufficiency" of de-idealization.

Two kinds of de-idealization of the definition are involved in evaluating the uncertainty of frequency standards. These correspond to two different methods of evaluating measurement uncertainty that metrologists label type A and type B.[12] First, the definition of the second is idealized in the sense that it presupposes that the relevant frequency of cesium is a single-valued number. By contrast, the frequency of any real oscillation converges to a single value only if averaged over an infinite duration, due to so-called random fluctuations. De-idealizing the definition in this respect means specifying a set of finite run times and evaluating the width of the distribution of frequencies for each run time. Uncertainties evaluated in this manner, that is, by the statistical analysis of a series of observations, fall under type A.

The second kind of de-idealization of the definition has to do with systematic effects. For example, one way in which the definition of the second is idealized is that it presupposes that the cesium atom resides in

11. "Fractional" or "relative" uncertainty is the ratio between measurement uncertainty and the best estimate of the value being measured (usually the mean).

12. See JCGM (2008a) for a comprehensive discussion. The distinction between type A and type B uncertainty is unrelated to that of type I vs. type II error. Nor should it be confused with the distinction between random and systematic error.

https://doi.org/10.1086/662268 Published online by Cambridge University Press

a completely flat space-time, that is, a gravitational potential of zero. When measured in real conditions on earth, general relativity predicts that the cesium frequency would be redshifted depending on the altitude of the laboratory housing the clock. The magnitude of this "bias" is calculated on the basis of a theoretical model of the earth's gravitational field and an altitude measurement.[13] The measurement of altitude itself involves some uncertainty, which propagates to the estimate of the shift and therefore to the corrected outcomes of the clock. This sort of uncertainty, that is, uncertainty associated with corrections to systematic errors, falls under type B.

In addition to gravitational effects, numerous other effects must be estimated and corrected for a cesium fountain. With every such de-idealization and correction, some type B uncertainty is added to the final outcome, that is, to the number of ticks the clock is said to have generated in a given period. The overall type B uncertainty associated with the clock is then taken to be equal to the root sum of squares of these individual uncertainties.[14] In other words, the type B uncertainty of a primary standard is determined by the accumulated uncertainty associated with corrections applied to its readings. The general method of evaluating the overall accuracy of measuring systems in this way is known as "uncertainty budgeting." Metrologists draw up tables with the contribution of each correction and a "bottom line" that expresses the total type B uncertainty (an example will be given in sec. 6). Such tables make explicit the fact that "raw ticks" generated by a clock are by themselves insufficient to determine the uncertainty associated with that clock. Uncertainties crucially depend not only on the apparatus but also on how the apparatus is modeled and on the level of detail with which such models capture the idiosyncrasies of a particular apparatus.

5. A Robustness Condition for Accuracy. We saw that metrologists successively de-idealize the definition of the second until it describes the specific apparatus at hand. The type A and type B uncertainties accumulated in this process are combined to produce an overall uncertainty estimate for a given clock. This is how, for example, metrologists arrived at the estimate of fractional frequency uncertainty cited in section 4.

A question nevertheless remains as to how metrologists determine the

13. The calculation of this shift involves the postulation of an imaginary rotating sphere of equal local gravitational potential called a geoid, which roughly corresponds to the earth's sea level. Normalization to the geoid is intended to transform the proper time of each clock to the *coordinate time* on the geoid. See, e.g., Jefferts et al. (2002, 328).

14. This method of adding uncertainties is only allowed when it is safe to assume that uncertainties are uncorrelated.

https://doi.org/10.1086/662268 Published online by Cambridge University Press

point at which de-idealization is "sufficient." After all, a complete de-idealization of any physical system is itself an unattainable ideal. Indeed, the most difficult challenges that metrologists face involve building confidence in descriptions of their apparatus. Such confidence is achieved by pursuing two interlocking lines of inquiry: on the one hand, metrologists work to increase the level of detail with which they model clocks. On the other hand, clocks are continually compared to each other in light of their most recent theoretical and statistical models. The uncertainty budget associated with a standard is then considered sufficiently detailed if and only if these two lines of inquiry yield consistent results. The upshot of this method is that the uncertainty ascribed to a standard clock is deemed adequate if and only if the outcomes of that clock converge to those of other clocks within the uncertainties ascribed to each clock by appropriate models, where appropriateness is determined by the best currently available theoretical knowledge and data-analysis methods. This kind of convergence is routinely tested for all active cesium fountains (Parker et al. 2001; Li et al. 2004; Gerginov 2010) as well as for candidate future standards, as is shown below.

The requirement for convergence under appropriate models embeds a double robustness condition, which may be generalized in the following way:

> **RC**.—Given multiple, sufficiently diverse realizations of the same unit, the uncertainties ascribed to these realizations are adequate if and only if
>
> 1. discrepancies among realizations fall within their ascribed uncertainties; and
> 2. the ascribed uncertainties are derived from appropriate models of each realization.

These two conditions loosely correspond to what Woodward (2006) calls "measurement robustness" and "derivational robustness." The first kind of robustness concerns the stability of a measured value under varying measurement procedures, while the second concerns the stability of a prediction under varying modeling assumptions. Note, however, that in the present case we are not dealing with two independently satisfiable conditions but with two sides of a single, composite robustness condition. Recall that the discrepancies mentioned in subcondition 1 already incorporate corrections to the values being compared, corrections that were calculated in light of detailed models of the relevant apparatuses. Conversely, the "appropriateness" of models in 2 is considered sufficiently established only once it is shown that these models correctly predict the range of discrepancies among realizations.

https://doi.org/10.1086/662268 Published online by Cambridge University Press

Metrology teaches us that RC is indeed satisfied in many cases, sometimes with stunningly small uncertainties. However, the question remains as to why one should take uncertainties that satisfy this condition to be measures of the *accuracy* of standards. This question can be answered by considering each of the five variants of accuracy outlined above.

To start with the most straightforward case, the comparative accuracy of realizations is simply the closeness of agreement among them, for example, the relative closeness of the frequencies of different cesium fountains. Clearly, uncertainties that fulfill subcondition 1 are (inverse) estimates of accuracy in this sense.

Second, from an operational point of view, the accuracy of a standard is the closeness of its agreement to other standards of the same quantity. This is again explicitly guaranteed by the fulfillment of RC under subcondition 1. That subcondition 1 guarantees two types of accuracy is hardly surprising because in the special case of comparisons among standards the notions of comparative and operational accuracy are coextensive.

Third, the epistemic conception of accuracy identifies the accuracy of a standard with the narrowness of spread of values reasonably attributed to the quantity realized by that standard. The evaluation of type A and type B uncertainties in light of current theories, models, and data analysis tools is plausibly the most rigorous way of estimating the range of durations that reasonably satisfy the definition of "second." The appropriateness requirement in subcondition 2 guarantees that uncertainties are evaluated in this way whenever possible.

Fourth, according to the metaphysical conception of accuracy, the accuracy of a standard is the degree of closeness between the estimated and true values of the realized quantity. Here one may adopt a skeptical position and claim that the true values of physical quantities are generally unknowable. The skeptic is, in principle, correct: it may be the case that despite their diversity, all the measurement standards that metrologists have compared are plagued by a common systematic effect that equally influences the realized quantity and thus remains undetected. But for a nonskeptical realist who believes (for whichever reason) that current theories are true or approximately true, condition RC provides a powerful test for metaphysical accuracy because it relies on the successive de-idealization of the theoretical definition of the relevant unit. Estimating the metaphysical accuracy of a cesium clock, for example, amounts to determining the conceptual "distance" of that clock from the ideal conditions specified by the definition of the second. As mentioned, the uncertainties that go into RC are consequences of precisely those respects in which the realization of a unit falls short of the definition. It is therefore plausible to consider cross-checked uncertainty budgets of multiple primary standards as supplying good estimates of metaphysical accuracy. Nevertheless,

https://doi.org/10.1086/662268 Published online by Cambridge University Press

it is important to note that condition RC and the method of uncertainty budgeting do not presuppose anything about the truth of our current theories or the reality of quantities. That is, RC is compatible with a nonskeptical realist notion of accuracy without requiring commitment to its underlying metaphysics.

Finally, from a pragmatic point of view the accuracy of a standard is its capacity to meet the accuracy needs of a certain application. Here the notion of "accuracy needs" is cashed out in terms of one of the first four notions of accuracy. As the uncertainties vindicated by RC have already been shown to be adequate estimates of accuracy under the first four notions, they are ipso facto adequate for the estimation of pragmatic accuracy.

6. Future Definitions of the Second. The methodological requirement to maximize robustness to the limit of what is practically possible is one of the main reasons why unit definitions are not chosen arbitrarily. If the definitions of units were determined arbitrarily, their replacement would be arbitrary as well. But, as metrologists know only too well, changes to unit definitions involve a complex web of theoretical, technological, and economic considerations. Before the metrological community accepts a new definition, it must be convinced that the relevant unit can be realized more accurately with the new definition than with the old one. Here, again, "accuracy" is cashed in terms of robustness. In the case of the second, for example, a new generation of "optical" atomic clocks is already claimed to have achieved "an accuracy that exceeds current realizations of the SI unit of time" (Rosenband et al. 2008, 1809). To demonstrate accuracy that surpasses the current cesium standard, optical clocks are compared to each other in light of their most detailed models available. Table 1 presents a comparative uncertainty budget for aluminum and mercury optical clocks recently evaluated at NIST. The theoretical description of each atomic system is de-idealized successively, and the uncertainties contributed by each component add up to the "bottom line" type B uncertainty for each clock. These uncertainties are roughly an order of magnitude lower than those ascribed to cesium fountain clocks.

The experimenters showed that successive comparisons of the frequencies of these clocks indeed yield outcomes that fall within the ascribed bounds of uncertainty, thereby applying the robustness condition above. The fact that these clocks involve two different kinds of atoms was taken to strengthen the robustness of the results. Nevertheless, it is unlikely that the second will be redefined in the near future in terms of an optical transition. More optical clocks must be built and compared before metrologists are convinced that such clocks are modeled with sufficient detail. Meanwhile, the accuracy of current cesium standards is still being improved by employing new methods of controlling and correcting for errors. In the long

https://doi.org/10.1086/662268 Published online by Cambridge University Press

TABLE 1. Comparison of Uncertainty Budgets of Aluminum (Al) and Mercury (Hg) Optical Atomic Clocks.

Shift	$\Delta\nu_{Al}$	σ_{Al}	$\Delta\nu_{Hg}$	σ_{Hg}	Limitation
Micromotion	−20	20	−4	4	Static electric fields
Secular motion	−16	8	−3	3	Doppler cooling
Blackbody radiation	−12	5	0	0	DC polarizability
313-nm Stark	−7	2	—	—	Polarizability, intensity
DC quadratic Zeeman	−453	0.5	−1,130	5	B-field calibration
AC quadratic Zeeman	0	1	0	10	Trap RF B-fields
Electric quadrupole	0	0.5	0	10	B-field orientation
First-order Doppler	0	1	0	7	Statistical imbalance
Background gas collisions	0	0.5	0	4	Collision model
AOM phase chirp	0	0.1	0	6	RF power
Gravitational redshift	−5	1	—	—	Clock height
Total	−513	23	−1,137	19	

Source.—Rosenband et al. (2008), 1809. Reprinted with permission from the American Association for the Advancement of Science.

Note.—This table was used by the group at NIST to support the claim that both clocks are more accurate than the current cesium standard; $\Delta\nu$ stands for fractional frequency bias and σ stands for uncertainty, both expressed in 10^{-18}. Uncertainties (σ) are added in quadrature. AC = alternating current; AOM = acousto-optic modulator; DC = direct current; RF = radio frequency.

run, however, increasing technological challenges involved in improving the accuracy of cesium fountains are expected to lead to the adoption of new sorts of primary realizations of the second such as optical clocks.

7. Implications and Conclusions. As the foregoing discussion has made clear, measurement standards are not absolutely accurate, nor are they chosen arbitrarily. Moreover, unit definitions do not completely fix the reference of unit terms, unless "fixing" is understood in a manner that is utterly divorced from practice. Instead, choices of unit definition, as well as choices of realization for a given unit definition, are informed by intricate considerations from theory, technology, and data analysis.

The study of these considerations reveals the ongoing nature of standardization projects. Theoretically, quantities such as mass, length, and time are represented by real numbers on continuous scales. The mathematical treatment of these quantities is indifferent to the accuracy with which they are measured. But in practice, we saw that the procedures required to measure duration in seconds change with the degree of accuracy demanded. Consequently, a necessary condition for the possibility of increasing measurement accuracy is that unit-concepts are continually recoordinated with new measuring procedures.[15] Metrologists are responsible for performing such acts of recoordination in the most seamless

15. See van Fraassen's (2008, chap. 5) discussion of the problem of coordination. I take my own robustness condition RC to be a methodological explication of van Fraassen's "coherence constraint" on acceptable solutions to this problem.

https://doi.org/10.1086/662268 Published online by Cambridge University Press

manner possible, so that for all practical purposes the second, meter, and kilogram appear to remain unchanged. This is achieved by constructing and improving primary and secondary realizations and (less frequently) by redefinition. The dynamic coordination of quantity concepts with increasingly robust networks of instruments allows measurement results to retain their validity even when standards are improved or replaced. Moreover, increasing robustness minimizes vagueness surrounding the reference of unit terms, thereby providing an optimal solution to the problem of multiple realizability of unit definitions.

The central role that comparisons among multiple procedures play in metrology suggests that robustness is likely to be a key concept in any comprehensive practice-oriented epistemology of measurement. The main strength of the approach to robustness outlined here is that it incorporates both the experimental and theoretical aspects of measurement in a single condition. Indeed, considerations that go beyond the scope of the current work suggest that a version of robustness condition RC may be taken as a general test for the accuracy of measurement procedures.

REFERENCES

BIPM (Bureau International des Poids et Measures). 2006. *The International System of Units* (*SI*). 8th ed. Sèvres: BIPM, http://www.bipm.org/en/si/si_brochure/.

Boumans, Marcel. 2007. "Introduction." In *Measurement in Economics: A Handbook*, ed. Marcel Boumans, 3–18. Bingley: Emerald Group.

CGPM (Conférence Général des Poids et Measures). 1961. *Proceedings of the 11th CGPM*. Sèvres: CGPM, http://www.bipm.org/en/CGPM/db/11/6/.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Galison, Peter. 2003. *Einstein's Clocks, Poincaré's Maps: Empires of Time*. New York: Norton.

Gerginov, Vladislav, N. Nemitz, S. Weyers, R. Schröder, D. Griebsch, and R. Wynands. 2010. "Uncertainty Evaluation of the Caesium Fountain Clock PTB-CSF2." *Metrologia* 47:65–79.

Girard, G. 1994. "The Third Periodic Verification of National Prototypes of the Kilogram (1988–1992)." *Metrologia* 31:317–36.

Gooday, Graeme J. N. 2004. *The Morals of Measurement: Accuracy, Irony, and Trust in Late Victorian Electrical Practice*. Cambridge: Cambridge University Press.

JCGM (Joint Committee for Guides in Metrology). 2008a. *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement*. Sèvres: JCGM, http://www.bipm.org/en/publications/guides/gum.html.

———. 2008b. *International Vocabulary of Metrology*. 3rd ed. Sèvres: JCGM, http://www.bipm.org/en/publications/guides/vim.html.

Jefferts, Steven R., et al. 2002. "Accuracy Evaluation of NIST-F1." *Metrologia* 39:321–36.

———. 2007. "NIST Cesium Fountains—Current Status and Future Prospects." *Acta Physica Polonica* A 112 (5): 759–67.

Krantz, David H., Duncan R. Luce, Patrick Suppes, and Amos Tversky. 1971. *Foundations of Measurement*. Vol. 1. New York: Academic Press.

Kripke, Saul A. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Latour, Bruno. 1987. *Science in Action*. Cambridge, MA: Harvard University Press.

Li, Tianchu, et al. 2004. "NIM4 Cesium Fountain Primary Frequency Standard: Performance and Evaluation." In *Proceedings of the 2004 IEEE International Ultrasonics,*

https://doi.org/10.1086/662268 Published online by Cambridge University Press

Ferroelectrics, and Frequency Control, 702–5. Piscatawny, NJ: Institute of Electrical and Electronics Engineers.
Lombardi, Michael A., Thomas P. Heavner, and Steven R. Jefferts. 2007. "NIST Primary Frequency Standards and the Realization of the SI Second." *Measure: Journal of Measurement Science* 2 (4):74–89.
Michell, Joel. 1994. "Numbers as Quantitative Relations and the Traditional Theory of Measurement." *British Journal for the Philosophy of Science* 45:389–406.
Parker, Thomas, et al. 2001. "First Comparison of Remote Cesium Fountains." In *Proceedings of the 2001 IEEE International Frequency Control Symposium*, 63–68. Piscatawny, NJ: Institute of Electrical and Electronics Engineers.
Rosenband, Till, et al. 2008. "Frequency Ratio of Al+ and Hg+ Single-Ion Optical Clocks: Metrology at the 17th Decimal Place." *Science* 319:1808–12.
Schaffer, Simon. 1992. "Late Victorian Metrology and Its Instrumentation: A Manufactory of Ohms." In *Invisible Connections: Instruments, Institutions and Science*, ed. Robert Bud and Susan Cozzens, 23–56. Cardiff: SPIE Optical Engineering.
Swoyer, Chris. 1987. "The Metaphysics of Measurement." In *Measurement, Realism and Objectivity: Essays on Measurement in the Social and Physical Sciences*, ed. John Forge, 235–90. Boston: Reidel.
van Fraassen, Bas. 2008. *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.
Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. London: Macmillan.
Woodward, Jim. 2006. "Some Varieties of Robustness." *Journal of Economic Methodology* 13 (2): 219–40.

https://doi.org/10.1086/662268 Published online by Cambridge University Press