Forthcoming in *Studies in History and Philosophy of Science*

Understanding Data Uncertainty[†]

Alisa Bokulich
Department of Philosophy
Boston University
abokulic@bu.edu

Wendy Parker
Department of Philosophy
Virginia Tech
wendyparker@vt.edu

Abstract: Scientific data without uncertainty estimates are increasingly seen as incomplete. Recent discussions in the philosophy of data, however, have given little attention to the nature of uncertainty estimation. We begin to redress this gap by, first, discussing the concepts and practices of uncertainty estimation in metrology and showing how they can be adapted for scientific data more broadly; and second, advancing five philosophical theses about uncertainty estimates for data: they are substantive epistemic products; they are fallible; they can be iteratively improved; they should be judged in terms of their adequacy-for-purpose; and these estimates, in turn, are essential for judging data adequacy. We illustrate these five theses using the example of the GISTEMP global temperature dataset. Our discussion introduces a novel adequacy-for-purpose view of uncertainty estimation, addresses a weakness in a recent philosophical account of data, and provides a new perspective on the "safety" versus "precision" debate in metrology.

1. Introduction

Uncertainty is an inescapable part of science. Yet while much has been written in recent years on uncertainty in computational modeling contexts (especially related to climate modeling; see, e.g., Parker 2010; Frigg et al. 2013; Knutti et al. 2019), comparatively little philosophical attention has been given to uncertainty associated with scientific data collected via observation, measurement, and experiment. Important preliminary work on this topic examines the evaluation of systematic uncertainty in measurement (Staley 2020), changing conceptions of uncertainty in metrology (de Courtenay and Grégis 2017; Grégis 2019b), and the representation of uncertainty when measurements are discordant (Grégis 2019a). Overall, however, when it comes to uncertainty associated with scientific data obtained via observation, measurement, and experiment—what we will call "data uncertainty"—the territory remains largely unexplored from a philosophical point of view.

Conceptions of data uncertainty and practices of data uncertainty estimation have changed in recent decades. Within the field of metrology (the scientific study of measurement), the conceptualization of uncertainty has evolved and bifurcated: in addition to the traditional "error" view, which conceptualizes uncertainty as the magnitude of possible error in an estimate of a quantity, there has emerged an alternative "epistemic" view, which takes uncertainty to reflect the inexactness of our knowledge of a quantity value.[1] In scientific practice, community standards for uncertainty estimation have become more demanding. While uncertainty estimates for data were once seen as an optional "fudge factor," an increasingly rigorous and thorough evaluation of data uncertainty is now expected in many domains.[2]

[†]We dedicate this paper to the memory of Margie Morrison (1954-2021) whom we lost too soon and who was a great inspiration and mentor to us both.

[1] As we will reiterate below, although the epistemic view of uncertainty is increasing in popularity, both conceptualizations are still widely used in science.

[2] By an *uncertainty estimate* we will mean an estimate of the total uncertainty associated with a datum, along with information about any estimated contributions to that total from individual sources of uncertainty and how those contributions were combined to arrive at the total.

In this paper, we call attention to these conceptual and practical changes in data uncertainty estimation and begin to redress the lack of philosophical attention to data uncertainty. We take metrology to be a rich resource in this regard. This is not only because many scientific data and data products are – like measurements – intended to be estimates of quantity values, but also because the metrology community has detailed vocabularies and guides for uncertainty estimation, articulated by internationally coordinated institutions. Thus, our discussion will engage in substantive ways with work in contemporary metrology.[3] At the same time, we see opportunities for augmenting the metrologists' contributions. In particular, we will propose that an adequacy-for-purpose perspective, originally developed in philosophy of science in the context of model evaluation, be employed when evaluating uncertainty estimates for data.

We proceed as follows. In Section 2, we discuss the recent "epistemic turn" in metrology and articulate two conceptions of data uncertainty. Section 3 outlines a process for uncertainty estimation advocated within metrology, which involves the production of a detailed uncertainty budget. While this approach has much to offer, we argue that some of the requirements are ill-suited for scientific practice, where data may be produced for purposes other than precision measurement. With this background in place, in Section 4 we introduce five important but underappreciated philosophical theses about uncertainty estimates for data: first, they are substantive epistemic products; second, they are fallible; third, they can be iteratively improved; fourth, they should be developed and judged in terms of their adequacy-for-purpose; and, finally, they are essential for judging the adequacy of data. Section 5 illustrates all five of these theses with the example of the GISTEMP global temperature dataset. Section 6 applies our adequacy-for-purpose view to offer a new perspective on the safety versus precision debate in metrology, which first arose in connection with concerns about how to represent uncertainty when measurements are discordant. In Section 7 we draw together the findings of the preceding sections and suggest some future lines of research.

2. The Error and Epistemic Conceptions of Uncertainty

The most comprehensive and detailed discussions of uncertainty and its estimation today are found in the field of metrology. The basic concepts and community standards for expressing uncertainty are determined by an international metrology organization, the International Bureau of Weights & Measures (BIPM[4]), and its Joint Committee for Guides in Metrology (JCGM). These concepts and standards are codified and distributed via published documents, such as the "Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement" known colloquially as "GUM" (JCGM 2008). The JCGM is also responsible for standardizing the terminology used to describe measurements, which is communicated in the periodically updated "International Vocabulary of Metrology" (VIM). These shared terminologies and coordinated community standards help data to travel and to be meaningfully compared across different places, times, and contexts (e.g., JCGM 2008, p. *viii*; Leonelli & Tempini 2020).

A core tenet of metrology today is that measurement results are incomplete without a statement of their associated uncertainty. As the JCGM explains:

> The objective of a measurement is to determine the value of the measurand, that is, the value of the particular quantity to be measured. . . . In general, the result of a measurement is only an approximation or estimate of the value of the measurand and thus is complete only when accompanied by a statement of the uncertainty of that estimate. (JCGM 2008, p. 4)

As they emphasize here, the data that result from measurement are only *estimates* of the quantities being measured (the "measurands"), and they have some associated *uncertainty*. How this uncertainty is

[3] It is unfortunate that work in metrology and in the philosophy of data are not in more substantive conversation; we aim to begin to change this with the present paper.

[4] The traditionally used acronyms are typically for the French versions of the names and titles, e.g., BIPM for Bureau International des Poids & Mesures and VIM for Vocabulaire International de Métrologie.

conceptualized, however, has shifted in recent years, from an "Error Approach" to an "Epistemic Approach". The shift is a subtle one worth unpacking.

Assessments of the quality of measurements were traditionally analyzed following the Error Approach. As explained in the third edition of the Vocabulary of Metrology (VIM3):

> "[t]he objective of measurement in the Error Approach is to determine an estimate of the true value that is as close as possible to that single true value. The deviation from the true value is composed of random and systematic errors. . . . [T]he total error is estimated, sometimes loosely named 'uncertainty'." (VIM3 2012, p. *viii*)

As the quote suggests, the Error Approach involves three key concepts: true value, random error, and systematic error. The *true value* of a measurand is the ideal target of a measurement—what the measurement *aims* to obtain. *Random error* is defined [2.19] as that "component of measurement error that in replicate measurements varies in an unpredictable manner" (VIM3, p. 23). Typically, there are some influences on the measurement process that vary spatially and temporally in a stochastic or random way such that they will cancel out in the limit of "an infinite number of replicate measurements of the same measurand" (*ibid.*). Not all errors have this character, however; some systematically skew a measurement result in particular direction and hence cannot be eliminated or reduced through replication. *Systematic erro*r is defined as that which "in replicate measurements remains constant or varies in a predictable manner" (VIM3, p. 22). If a source of systematic error is well-understood, then a correction factor might be applied to the measurement data. In many cases, however, the precise magnitude of a systematic error is unknown and hence cannot be corrected for. The *total error* associated with a measurement result, then, is taken to be a sum of these two kinds of possible error: random error plus the error stemming from imprecisely quantifiable systematic sources. In principle, the total error indicates how far a measured value could be from the true value.

A number of concerns have been expressed about the Error Approach and the concepts upon which it depends. One is that its key quantities (true value, error) are, strictly speaking, unknowable: "Although these two traditional concepts are valid as ideals, they focus on *unknowable* quantities: the 'error' of the result of a measurement and the 'true value' of the measurand" (JCGM 2008, p. 3).[5] This was one motivation for reframing discussions of measurement uncertainty in terms of *inexactness of knowledge*. Uncertainty on this alternative Epistemic Approach "reflects the lack of exact knowledge of the value of the measurand" (JCGM 2008, p. 3). From this perspective, uncertainty is an epistemic concept and characterizes the state of our knowledge, whereas on the Error Approach it describes a characteristic of the measuring process itself, namely, the total possible error that might result from its limitations. Nadine de Courtenay and Fabien Grégis (2017) refer to this shift as the "epistemic turn" in metrology.

As the GUM authors further emphasize, the differences between these two approaches are substantive, not merely terminological. Perhaps most notably, small uncertainties -- understood epistemically—do *not* necessarily mean small errors:

> But, even if the evaluated uncertainties are small, there is still no guarantee that the error in the measurement result is small; for . . . a systematic effect may have been overlooked because it was unrecognized. Thus the uncertainty of a result is not necessarily an indication of the likelihood that the measurement result is near the value of the measurand; it is simply an estimate of the likelihood of nearness to the best value that is consistent with presently available knowledge. (JCGM 2008, p. 51)[6]

---

[5] A possible exception is the special case where the true value is stipulated to be that provided by some measurement standard.

[6] An anonymous referee correctly notes that there is no guarantee under the Error Approach either. Still, the *meaning* of a report of small uncertainty differs under the two approaches. Under the Error Approach, to report small uncertainty is to *claim* or *assert* that the possible error in a result is small; on the Epistemic Approach, to report small uncertainty is to claim only that current information strongly contrains the values that can be reasonably assigned to the measurand.

As the reference to "the best value" suggests, another notable difference is that, while the Error Approach typically assumes that there is a single true value for the measurand and seeks to estimate it, the Epistemic Approach makes no such assumption and instead delivers "an infinite number of values dispersed about the result that are consistent with all of the observations and data and one's knowledge of the physical world, and that with varying degrees of credibility can be attributed to measurand" (JCGM 2008, p. 51). The best value is that which is most credible, from the perspective of current knowledge.[7]

A second concern about the Error Approach pertains to the distinction between random and systematic error. The concern is that this distinction is context dependent—or perhaps more pointedly that the notion of types of error has ontological aspirations that are inconsistent with the context-dependent nature of these types. As explained in the GUM, "a 'random' component of uncertainty in one measurement may become a 'systematic' component of uncertainty in another measurement in which the result of the first measurement is used as an input datum" (JCGM 2008, p. 6). The scenario they are envisioning here occurs in what Wendy Parker calls a *derived measurement*, in which measurement outcomes are calculated or derived from directly measured values for other parameters, using reliable scientific principles or definitions (Parker 2017, p. 9).

GUM thus recommends abandoning the random-systematic classification and instead using the designations Type A and Type B measurement uncertainty, emphasizing again that this is not merely a terminological shift. The Type A and B classification is meant to move away from a focus on the nature or source of the components to instead focus on how the sources of uncertainty are evaluated or handled. In particular, Type A uncertainty is defined as those components of uncertainty that are evaluated through statistical analysis from a series of repeated observations. Type B uncertainty is, rather unsatisfyingly, defined only negatively as everything else: "evaluation of a component of measurement uncertainty determined by means other than a Type A evaluation" (JCGM 2021, p. 21). What are these other methods of evaluation? The GUM notes that Type B evaluation involves "scientific judgement based on all of the available information on the possible variability" (JCGM 2008, p. 11). This approach to characterizing uncertainty in terms of method of evaluation (as Type A and B) goes beyond the epistemic turn to what might be called the *operationalist turn*—uncertainty components are classified by the operations used to evaluate them.

Although one cannot (on the Epistemic Approach) continue to speak in terms of systematic and random *error*, it is notable that a version of this distinction is maintained even after the operationalist turn, in discussions of systematic versus random *effects*. In the section on uncertainty, for example, the GUM authors note that even after one has attempted to correct a measurement for known systematic effects, there is still uncertainty arising from imperfect corrections for those systematic effects, as well as uncertainty from random effects. They add, however, that the "uncertainty of a correction for a known systematic effect may in some cases be obtained by a Type A evaluation while in other cases by a Type B evaluation, as may the uncertainty characterizing a random effect" (JCGM 2008, p. 6).

Drawing on this work in metrology, we can formulate two different conceptions of *data uncertainty*. On an Error Approach, data uncertainty refers to the possible error in a datum, i.e., how far off the datum value might be from the true value of the target quantity (or, more generally, the target variable, which might be categorical or qualitative). For an example of the latter, consider a datum indicating that an observed insect belongs to a particular species; an uncertainty report might indicate that the insect might instead be a member of one of two other species. The uncertainty report here is in effect a claim about the data collection process, in particular, that the species-discriminating power of the observing method was limited in particular ways. On an Epistemic Approach, data uncertainty reflects the inexactness of our knowledge; there is a set of values that, from the perspective of current knowledge, can reasonably be attributed to the target variable. In the insect example, the same uncertainty report would be interpreted as follows: from the perspective of current knowledge, it is also credible that the insect is a member of one of these other two species.

[7] As an anonymous referee noted, one could adopt the Epistemic Approach and still assume that a true value exists for a given measurand.

We wish to emphasize that both of these ways of understanding data uncertainty are employed in current scientific practice. Indeed, despite the shift in metrology, an Error Approach remains standard in many scientific domains. That is, data uncertainty is often conceptualized as a measure of possible error in the data. The Epistemic Approach is also sometimes employed – even in the scientific domains where an Error Approach tends to be more common—and has obvious affinity with a Bayesian perspective. Indeed, sometimes the same scientist will use an error conceptualization of uncertainty in one scientific project and an epistemic conceptualization in a different project; each can be useful. It is thus helpful to keep both approaches in mind and to consider which is being employed—whether explicitly or implicitly—in a given case.

3. Budgeting Sources of Uncertainty

The field of metrology also provides comprehensive guidance on how to assess measurement uncertainty. It begins with the specification of a detailed model of the measurement process, or what is termed a *measurement model*. As laid out by the GUM, constructing a measurement model involves the following five steps (JCGM 2020, pp. 3-4):

a) Select and specify the *measurand* (quantity to be measured).

b) Model the *measurement principle* that connects the input quantities actually measured to the desired output quantity (estimate of the measurand) in an indirect or derived measurement. For example, to measure the volume of a cylinder, one typically first measures the length, L, and diameter, d, of the cylinder and then uses these input quantities and the measurement principle $V = \pi L d^2/4$ to derive the volume. The measurement principle can be based on theory, empirically modeled, or some combination.

c) Identify all the relevant *effects* contributing to the measurement result. This goes beyond the factors specified in the simple, ideal relationship identified in (b) to include influences that arise in the practical implementation of the measurement process, including interfering environmental factors, instrument drift, and many more (as discussed in detail below).

d) Extend the basic model specified in (b) to account for the effects identified in step (c); this involves first dividing the effects into two categories:

1. "*Well-understood effects*" whose influence can be measured or estimated with sufficiently small uncertainty to be added to the measurement model as a *correction* (JCGM 2020, p. 29; for philosophical discussion see Bokulich 2020b).
2. "*Poorly understood effects*" which, though known to exist, have a magnitude or perhaps even direction of influence that cannot be estimated exactly enough to be added as a correction, but nonetheless contribute appreciably to the measurement *uncertainty* and thus should be part of the measurement model (JCGM 2020, p. 33). These can be either systematic or random effects.

A complete measurement model will indicate how both types of effects should be taken into account to arrive at a measurement outcome, including an uncertainty estimate.

e) Assess the adequacy of the measurement model. According to GUM, a "measurement model is adequate if the estimate of the measurand obtained by using it is corrected for all known effects, and the associated uncertainty reflects all factors that could reasonably affect the estimate" (JCGM 2020, p. 57).

As the GUM notes, these five steps need not be taken in this particular order, and multiple iterations of them may be needed before an acceptable measurement model is obtained.

Steps (c)-(e) in the process can be particularly demanding and difficult. To aid step (c), the GUM provides a long (yet non-exhaustive) list of "effects arising in the practical implementation of the measurement" (JCGM 2020, p. 28) as well as a taxonomy of "possible sources of uncertainty in a measurement" (JCGM 2008, p.6). These include the following:

1. *Incomplete definition of the measurand*. If the measurand is only vaguely specified, this can be a source of uncertainty for the measurement result. For example, if one seeks a high-precision measurement of the length of a metal bar, then the definition of the measurand should also include a specification of the temperature at which the length of the bar is to be measured, since temperature differences will cause the metal bar to expand or contract.

2. *Imperfect realization of the definition of the measurand*. Even if the definition of the measurand is complete, the measuring conditions may not be realizing that definition perfectly. For example, even if the definition of the measurand indicates that the length of the metal bar should be measured at a temperature of 20.0 degrees Celsius, it may not be possible to keep the bar uniformly at precisely that temperature.

3. *Nonrepresentative sampling*. Related to (2) above but applied to a population; the sample that is measured may also fail to adequately represent the defined measurand.

4. *Imperfect knowledge of (or measurement of) environmental influences*. This common source of uncertainty in measurement encompasses a wide range of possible influences, arising, for example, when one has not adequately shielded from, or been able to precisely correct for, an interfering factor.

5. *Personal bias in reading analogue instruments*. A classic example is reading a fluid level in a graduated cylinder—different observers might not bring it to same eye level or interpret the level with the curved surface tension (meniscus) in exactly the same way.

6. *Finite instrument resolution or discrimination threshold*. All instruments have a finite resolution: no analog display can be infinitely discriminated, and no digital display has an infinite number of decimal places. A kilogram scale, for example, might be designed to detect weight differences of a gram but unable to detect differences of a milligram.

7. *Inexact values of measurement standards and reference materials*. Many complex measurements make use of measurement standards or references materials. Any uncertainties associated with these standards will thus contribute to the uncertainty of the measurement.[8]

8. *Inexact values of fundamental constants or other parameters*. In the case of indirect or derived measurements, the measurement principle which is used to calculate an estimate of the measurand from various input quantities may make use of a fundamental constant or other parameter in the calculation. Any uncertainties associated with these constants or parameters will thus also contribute to the uncertainty of the measurement outcome.[9]

9. *Approximations and assumptions of the measurement method*. This broad category encompasses the many different ways that the actually implemented measurement process

[8] For a discussion of the uncertainty associated with the age of a mineral standard used in the production of argon-argon radiometric dates, and its ultimate revision, see Bokulich 2020c.
[9] For a discussion of uncertainties in the empirical determination of the values of fundamental constants, see Bokulich & Bocchi 2024.

deviates from the theoretical measurement process, as well as ways in which that theoretical process is idealized, incomplete, etc.[10]

10. *Drift of the measuring system.* Drift is a phenomenon that can happen to any physical component of the measuring system and occurs for a variety of reasons, such as normal "wear and tear" of an instrument or a buildup of debris (a problem that plagued even the prototype kilogram standard; see Bokulich 2020a). A measuring system may have multiple components whose drift may be difficult to specify precisely, contributing to the uncertainty associated with the measurement result.

11. *Instability or inhomogeneity of the artefact measured*. This category focuses on sources of uncertainty that arise not from the measuring apparatus, but rather from variations in the object being measured. An unstable artefact changes or varies over time during the measurement process, contributing to a dispersion of measured values and uncertainty about its current state. Inhomogeneity involves variation in space; an inhomogeneous sample can yield different values for a given measurand, depending on where exactly the measurement is taken.

12. *Variations in repeated observations*. This final category is a catch-all for the many other ways in which no two replicate measurements are ever executed in an exactly identical fashion.

In a given measurement context, effects related to many of these categories may be present at once. Some may be identified as "well-understood" and corrected for, but the remaining will be considered sources of uncertainty; moreover, corrections for well-understood effects may themselves have non-negligible associated uncertainty.

In step (d), the basic measurement model specified in (b) is extended to take account of the various effects identified in (c). Researchers must estimate their magnitudes and determine how to combine them to arrive at a corrected estimate of the measurand and an estimate of its associated uncertainty. For the latter, researchers can construct an *uncertainty budget*: an itemized list or table in which each source of uncertainty is entered in its own row, with a column for providing a quantified estimate of the uncertainty arising from each source. Additional columns can be added to provide metadata about how the quantitative estimate for each source was derived. One can also have columns indicating which sources are evaluated as Type A and which sources as Type B, as well as information about the nature of an associated probability distribution for that source (e.g., normal or rectangular). A complete uncertainty budget should also specify how these different sources are to be combined to produce the total uncertainty estimate for the measurement result. Table 1 shows a very simple example of an uncertainty budget for a measurement of the length of a string. In the first column of the table, each source of uncertainty is listed, with every row being a particular instance of one of the types of sources of uncertainty listed in the taxonomy above. In the second column, a quantitative estimate of the magnitude of the uncertainty associated with each source is given.

[10] Pierre Duhem famously called attention to this source of uncertainty in highlighting the difference between the schematic instrument and the concrete instrument: "When a physicist does an experiment, two very distinct representations of the instrument . . . fill his mind: one is the image of the concrete instrument that he manipulates in reality; the other is a schematic model of the same instrument, constructed with the aid of symbols supplied by theories" (Duhem ([1914] 1954), pp. 155-156).

SOURCES OF UNCERTAINTY IN LENGTH MEASUREMENT	VALUE OF UNCERTAINTY
Uncertainty due to imperfect manufacture/calibration of metal tape measure.	0.2 mm
Bending of measuring tape during measurement	0.3 mm
Thermal expansion of measuring tape	0.1 mm
Reduction in string length due to string not lying straight	3.0 mm
Variation due to stretching or shrinking of string	2.0 mm
Uncertainty due to aligning tape with frayed ends of string	2.0 mm
Length deviation due to tape & string not being parallel	0.5 mm
Resolution limits reading numerical value from tape	0.5 mm
Combined uncertainty:	**4.2 mm**

Table 1: Example of an uncertainty budget for a measurement of the length of a string (inspired by Bell 1999, p. 21 & JCGM 2020, p. 23). Taking the square root of the sum of the squares yields the combined uncertainty. The measured value plus/minus the combined uncertainty yields a 1σ uncertainty interval.

Finally, step (e) calls for evaluating the adequacy of the measurement model produced in (d), which includes both the corrections and the uncertainty budget. As noted above, according to the GUM, "A measurement model is adequate if the estimate of the measurand obtained by using it is corrected for all known effects, and the associated uncertainty reflects all factors that could reasonably affect the estimate" (JCGM 2020, p. 57). Assessing adequacy thus also can be a very challenging step, insofar as it involves not only assessing whether all known effects have been corrected for and deciding on some criterion for when a factor could "reasonably" affect the measurement estimate, but also engaging in critical second-order reflection on the extent to which one is in a position to make such an assessment.[11]

From the perspective of scientific practice, the GUM's account of an adequate measurement model seems both too demanding and not demanding enough. It seems too demanding because, in many scientific contexts, correcting for all known (well-understood) effects is not needed for measurement data to be used successfully for their intended purposes, and attempting to do so could slow scientific progress considerably. It is not demanding enough, because it says nothing about the accuracy or rigor with which corrections and uncertainty contributions should be estimated; for some purposes, highly accurate corrections and relatively comprehensive uncertainty budgets are required, while for other purposes, rough and ready corrections and uncertainty estimates will suffice. Similar concerns are expressed by Graham White in a paper on measurement uncertainty in biomedical contexts: "The GUM bottom-up approach can quickly become unwieldy and mathematically complex…. The effort and cost of estimating MU [measurement uncertainty] should be commensurate with the clinical quality of measurement required" (White 2008, p. S55). In a clinical context, speed of result and knowledge of the measurand as simply being above or below a certain threshold may be what is required—a detailed uncertainty estimate in such a context is not only unnecessary, but potentially harmful in the context of urgent patient care.

In the next section, we will advocate for an alternative perspective on the evaluation of data uncertainty estimates that better fits with the varied goals and practical realities of scientific investigation. We believe that the GUM's basic framework (a)-(e) discussed above has the potential to be highly useful when applied to scientific data more broadly—not just in the context of precision measurement—once some important adjustments are made.

[11] We are grateful to an anonymous referee for encouraging us to underscore this point.

4. Uncertainty Estimates for Data: Five Theses

With the foundation provided by the preceding discussion, we are now in a position to articulate five important but underappreciated philosophical theses about uncertainty estimates for data. These theses will build on each other, culminating in the final two theses, which involve a novel view of what constitutes an adequate uncertainty estimate for data.

4.1. *Data uncertainty estimates are substantive epistemic products.*

Our first thesis follows naturally from the discussion of uncertainty budgets in the last section. Uncertainty budgets—and, more generally, uncertainty estimates—are not just "fudge factors" expressing ignorance, nor are they a simple, algorithmic product of statistical analysis. Instead, producing them can require substantial expertise and can involve an extended, careful, and detailed investigation. As the authors of GUM emphasize, "The evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement" (JCGM 2008, p. 8). Indeed, the activity of uncertainty estimation can have a rich and complex epistemic structure. Sometimes new experiments are conducted, or computer simulations are run, in order to arrive at estimates of the contributions of different sources of uncertainty. In many cases, providing an uncertainty estimate, such as a detailed uncertainty budget, should be recognized as a *significant epistemic achievement* in its own right. That data uncertainty estimation is such an achievement has hitherto been inadequately appreciated in the philosophy of science.

4.2 *Uncertainty estimates are fallible.*

Because data uncertainty estimates are substantive epistemic products, it is no surprise that they are also *fallible*. This is true under both the Error Approach and the Epistemic Approach.

Under the Error Approach, uncertainty estimates are claims about the possible error associated with a datum. Such claims can be mistaken for various reasons. Recognized sources of uncertainty might be omitted from the uncertainty budget, or the contributions of recognized sources might be inaccurately estimated. But they can also fail when there are *un*recognized effects on the data collection process, i.e., what might be termed "unknown unknowns." Although not original to him, Donald Rumsfeld infamously popularized this category as part of the following typology:

> "[A]s we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know...it is the latter category that tend to be the difficult ones" (Donald Rumsfeld, U.S. Secretary of Defense, 12 February 2002).[12]

When such unrecognized effects are present in a data production process, it can easily happen that the magnitude of possible error associated with the datum is underestimated and that, consequently, the true value of the data target is not within the uncertainty bounds specified. In fact, the history of science reveals that unknown unknowns are quite common in frontier and experimental science. Even in the case of the high-precision measurement of fundamental physical constants, values regularly get revised outside of the uncertainty bounds associated with the previous measured value (e.g., see Tiesinga et al. 2021, and for a philosophical discussion Bokulich & Bocchi 2024).

While, by definition, researchers can't specify what the unknown unknowns are, they can sometimes recognize that the situation is one in which unknown unknowns *of some sort or other* are probably present (Parker and Risbey 2015). That is, sometimes researchers can recognize the situation as one in which unknown factors are probably significantly affecting the data. This might be because the data target (and what can affect it) is recognized to be relatively poorly understood, or because the data are found to contradict long-established or foundational results in the domain. In some cases, researchers may even be able to roughly estimate what the magnitude of such unknown unknowns could plausibly be,

[12] Retrieved from https://archive.ph/20180320091111/http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636" \l "selection-1053.96-1053.343

given failed applications of data (e.g., when an arrow fails to hit its target after measuring windspeed and calculating trajectory) or discrepancies among data obtained using substantially different measurement techniques (see Bokulich & Bocchi 2024 for examples). Having uncertainty estimates and interpretations of data that acknowledge the possibility of—and attempt to account for—such unknown unknowns will be important for some uses of data; it may make the difference between accurate and inaccurate conclusions being drawn from the data.

A striking illustration of the way that unknown unknowns can affect data, with the consequence that the uncertainty bounds of results fail to include the true value of the data target, can be found in the 2011 OPERA experiment in high-energy physics.[13] OPERA results initially seemed to show that neutrinos were traveling faster than the speed of light, which is prohibited by special relativity. The experimental data indicated that neutrinos sent from CERN observatory in Geneva Switzerland arrived at the OPERA detector in Gran Sasso National Laboratory near L'Aquila Italy 60.7 nanoseconds faster than light would if traveling the same distance in a vacuum (OPERA 2011, p. 22; Reich 2011, p. 520). The OPERA scientists constructed an uncertainty budget with line contributions for all the known sources of uncertainty both "statistical" (random), which they calculated to be ±6.9 nanoseconds and "systematic", which they calculated to be ±7.4 nanoseconds. On the basis of their analysis, they concluded that their result—indicating that neutrinos were traveling faster than the speed of light—had an extremely high significance level of 6.0 σ (OPERA 2011, p. 22). As David Chandler explains,

> Technically, the results of that experiment had a very high level of confidence: six sigma [6σ]. In most cases, a five-sigma result is considered the gold standard for significance, corresponding to a one-in-a-million chance that the findings are just a result of random variations; six sigma translates to a one chance in a half-billion that the result is a random fluke. (Chandler 2012)

Given the small uncertainties assessed by the OPERA team, and the carefulness of the experiment, there seemed to be extremely strong evidence for the striking result that neutrinos can travel faster than the speed of light.

As a matter of fact, however, almost no one in the physics community believed this result to be correct, and—6σ notwithstanding—most were sure that some systematic effect or other had not been properly accounted for. And they were right. In this case there were two unrecognized sources of systematic error. First, the main factor contributing to the apparent faster-than-light result was that an optical fiber cable involved in the measurement process was not screwed in properly all the way (see Strassler 2012 for an explanation). Second, there was a drift in the clock used to time the neutrinos' arrival. Once researchers were able to determine the effects of these two sources of systematic error, they were able to apply corrections to the measurement data and show that the neutrinos were in fact traveling slower than the speed of light, as everyone had expected. Although there was no way for the OPERA scientists to specify and correct for these sources of error *before* they were discovered, it was possible to recognize—as many other physicists did—hat it was highly likely that s*ome unknown unknown or other* was affecting the experimental result, rather than that a foundational principle of modern physics was incorrect. At the very least, when situations arise in which the presence of significant unknown unknowns seems plausible, this can be acknowledged when discussing what should be concluded from data. We will return to this point below in connection with our fourth thesis.

Turning now to the Epistemic Approach, an uncertainty estimate under this approach is a claim about the limitations of current knowledge. By identifying a range of values that can be reasonably attributed to the data target, the uncertainty estimate *reports the extent to which, in light of current information, our knowledge of that target remains inexact*. Such a report is inaccurate to the extent that it misidentifies the values that can be reasonably attributed to the measurand in light of current information. This misidentification can happen because sources of uncertainty known to be present are nevertheless omitted from an uncertainty budget (what are sometimes called "known neglecteds") or because mistakes,

[13] The OPERA acronym stands for Oscillation Project with Emulsion-tRacking Apparatus. The OPERA results were announced on 22 September 2011 on the High Energy Physics Preprint ArXiv (see OPERA 2011). For an accessible discussion of the errors later discovered, see Strassler 2012.

simplifications, or idealizations were made when applying available information to estimate the contributions of particular sources of uncertainty. As noted before, however, an uncertainty estimate that *is* accurate under the Epistemic Approach need not accurately characterize the extent of error that is possible, given the data collection process that was actually undertaken. This is because, insofar as the current understanding of the data target (or of how it can be estimated) is itself mistaken in some way, an uncertainty estimate might be highly accurate according to the standard of the Epistemic Approach – it might accurately specify the values that*, in light of current information*, can be reasonably attributed to the data target – yet substantially underestimate how far off a datum could be from the truth, *given the actual data collection procedure employed*.

4.3 ***Uncertainty estimates can be iteratively improved.***

The flipside of the thesis that uncertainty estimates are fallible is that they can be iteratively improved. Under the Error Approach, they are improved as they more accurately characterize the possible error associated with a measured value. Under the Epistemic Approach, they are improved when they come to more accurately characterize the dispersion of values that, given available information, can be reasonably attributed to the measurand.[14] Under both approaches, making an uncertainty budget *more complete* by adding lines for recognized sources of uncertainty that previously were omitted (perhaps for reasons of cost or lack of expertise) and *refining* estimates of the contributions of individual sources of uncertainty already in the budget, can improve an uncertainty estimate. Examples of efforts to iteratively improve uncertainty estimates in these ways are readily found in scientific practice, sometimes extending through numerous iterations over long periods of time; this can occur, in part, in response to evolving community standards, as the case study in the next section will illustrate. Uncertainty estimation can then be seen as an historically extended process, and tracking its progression over time, rather than just examining it at a single moment, may yield better philosophical insight into the practice of knowledge production in that context.

It is worth noting that improving an uncertainty estimate does not necessarily result in reduced uncertainty about the data target. On the contrary, making an uncertainty budget more complete can, in general, be expected to *increase* the estimated total uncertainty. Likewise, correcting the estimated contribution of a particular source of uncertainty might mean increasing the estimate of that contribution. Somewhat paradoxically, this means that gains in scientific knowledge – as "unknown unknowns" become "known unknowns" and "known unknowns" are better understood and more accurately quantified – can sometimes lead to *larger* estimates of total uncertainty.

4.4 ***Uncertainty estimates should be judged as adequate or inadequate for purpose.***

While uncertainty estimates *qua descriptions of possible error* associated with a datum (or *qua descriptions of limitations of current knowledge* of a target quantity) can be improved in the ways just described, in practice this sort of improvement is often unnecessary. This is because, in practice, data are often produced not for the sake of learning as precisely as possible what the value of a quantity is – for what might be called "precision measurement" – but for some other scientific purpose for which coarser information will be sufficient, such as discriminating among competing hypotheses, or informing a practical decision about a course of medical treatment, or identifying a fruitful pathway for further research. Thus, while data uncertainty estimates qua description of possible error (or qua description of the inexactness of current knowledge) will be better the more complete and refined they are, we contend that, in general, data uncertainty estimates should be evaluated relative to the purpose(s) for which the data are being produced or used. For many such purposes, a complete and highly accurate uncertainty estimate will not be needed. Indeed, attempting to produce one may be counterproductive, as we explain below.

We are inspired here by a broader adequacy-for-purpose perspective that has grown in prominence recently in the philosophy of science, according to which the quality of a scientific resource

[14] A complication here is that "reasonably" will require some definition or specification.

is to be evaluated relative to the purposes for which and context in which it will be used; quality is a matter of adequacy-for-purpose (see, e.g., Parker 2020; Bokulich and Parker 2021; Lusk and Elliott 2022). Adequacy-for-purpose, in turn, is tied to successful use: a scientific resource (model, datum, instrument, method, etc.) is *adequate-for-purpose* when its use in the context of interest will result in users' achieving their epistemic purpose(s).[15] These purposes are often narrow and specific: predicting tomorrow's high temperature in London within a specified level of accuracy, explaining a puzzling aspect of the fossil record, identifying flu hotspots in a region, etc. Because adequacy is tied to successful use, whether a resource is adequate can depend not just on intrinsic features of the resource, but also on how features of the resource fit with broader features of the context in which it will be used, such as the skills and knowledge that users have, what methodologies they will use, and what the background circumstances of use will be (Parker 2020). On this view, a resource that is adequate-for-purpose can be understood as a solution to a problem that is defined by the various epistemic and pragmatic constraints arising in a particular context – a solution in a given "problem space" (ibid.). A generalization of the notion of adequacy-for-purpose is fitness-for-purpose, which is applicable when the purpose(s) of interest can be achieved to a greater or lesser extent, rather than just in a binary yes/no manner; the *fitness-for-purpose* of a scientific resource in a given context is greater to the extent that its use by researchers will result in their achieving their purpose(s) to a greater extent (ibid.).

An adequacy-for-purpose view for uncertainty estimates, then, holds that the quality of an uncertainty estimate is a matter of its suitability for the specific purpose(s) for which it will be used.[16] As with all scientific resources, this purpose will vary from case to case. Nevertheless, to flesh out the adequacy-for-purpose view, we briefly consider three common *types* of purpose for which uncertainty estimates are produced, the first two are associated with the Error Approach and the last is associated with the Epistemic Approach:

(i) to accurately characterize the possible error associated with a datum;
(ii) to learn whether the possible error associated with a datum is so large as to render the datum inadequate for some further scientific or practical purpose, P_s;
(iii) to learn, with sufficient accuracy, the largest (or smallest) value of the target quantity that is plausibly consistent with current knowledge.

Each is associated with a different type of purpose for which data are produced, as we explain.

(i). When the purpose of data production is to measure a quantity, independent of any further application, then an uncertainty estimate produced under the Error Approach will be intended to serve a purpose of type (i): to accurately characterize the possible error associated with a datum (where the datum is the estimate of the target variable that is produced via the data collection and correction process). This purpose can be achieved to a greater or lesser extent, as discussed above in connection with fallibility and iterative improvement. The uncertainty estimate in this situation would be assessed for its *fitness-for-purpose*, which will be greater *the more accurately it characterizes the possible error* associated with the datum. In general, an uncertainty estimate will have greater fitness for (i) the more complete it is with regard to recognized sources of uncertainty and the more accurate its estimates of the contributions of individual sources of uncertainty are. To the extent that it is plausible that significant unknown unknowns

[15] See Parker 2020 and Bokulich and Parker 2021 on different varieties of adequacy-for-purpose, covering cases where use of the resource "is likely to" result in success or "could" result in success.

[16] In some places, the GUM seems to adopt an adequacy-for-purpose perspective on measurement. It is pointed out, for example, that the choice of measurement model can be shaped by the purposes of measurement and by practical considerations, such as the cost of implementation; what matters is that the measurement model is "adequate" or "fit for purpose" given the intended application of the measurement result (see, e.g., JCGM 2020, p.1, p.4, p.5, p.16, p.33). Nevertheless, little or no explication of adequacy or fitness accompanies such remarks and this purpose-relative approach seems inconsistent with other key statements in the GUM about what constitutes an adequate measurement model, such as that the estimate of the measurand must be corrected for all known (well-understood) effects (see JGCM 2020, Section 12.1, p.57).

are present the context of the measurement, it *may* increase an uncertainty estimate's fitness-for-(i) to try to factor this in when arriving at an estimate of total uncertainty; whether it actually does so depends on whether there actually are unknown unknowns and how well they are accounted for. Here we see one opportunity for considerations of inductive risk to come into play in uncertainty estimation: the investigator risks underestimating the uncertainty if she chooses to simply ignore the possibility of error due to unknown unknowns, and she risks overestimating uncertainty if unknown unknowns are factored in when in reality they are absent or have effects smaller than estimated.[17]

(ii). When data are produced for the sake of some further scientific or practical purpose, P_s, an uncertainty estimate will often be intended to serve a purpose of type (ii): to learn whether the possible error associated with the datum is so large as to render that datum inadequate for P_s. That is, an uncertainty estimate is produced to probe a particular way in which a datum can turn out to be inadequate for a scientific purpose of interest, namely, by having too much associated uncertainty. An uncertainty estimate in this case is *adequate-for-purpose* when it allows researchers *to successfully discern* whether the uncertainty is too large. Returning to the simple string example in Section 3, if the measurement of the length of the string is made for the purpose of (P_s) learning whether the string is longer than 16 mm, then the uncertainty estimate will be adequate-for-purpose if it allows researchers to discern whether the measured value of the string's length *cannot* be used to reach a confident conclusion about whether the string is longer than 16 mm, because its associated uncertainty is too large.

Note that an uncertainty estimate that is adequate for a purpose of type (ii) does not necessarily need to be complete; it need only be *complete enough*, in the sense that accounting for additional sources of uncertainty will not make a difference to whether researchers can successfully discern whether the possible error associated with the datum is too large for it to be used successfully for P_s (see also Parker and Risbey 2015). In some cases, just first-order recognized sources of uncertainty will need to be roughly accounted for, while in others a detailed uncertainty budget that, moreover, includes a line that estimates the potential contribution of unknown unknowns, may be required. For a simple illustration, suppose that researchers are collecting data using a rain gauge and want to use the data to (P_s) determine whether more than 9 inches of rain fell in a given month. They might know that uncertainty associated with a correction they applied for wind loss is by far the largest source of uncertainty for their data, swamping all others; if they estimate the uncertainty associated with the wind loss correction and find that its magnitude doesn't come remotely close to rendering their data inadequate for determining whether more than 9 inches of rain fell, then there will be no need for them to proceed with further lines in their uncertainty budget. In fact, in many cases, it will be counterproductive for researchers to try to produce a complete and highly accurate uncertainty estimate. Doing so can be time consuming and expensive. If P_s is time-sensitive, and if a simpler and rougher uncertainty estimate would do the job, then pursuing the simpler uncertainty estimate may even be required if the data are to be used successfully for P_s.[18]

(iii). In some practical contexts, decision makers aim to demonstrate due diligence in avoiding courses of action that risk very harmful outcomes. In this situation, an uncertainty estimate might be produced for a purpose of type (iii): to learn, with sufficient accuracy, the largest (or smallest) value of a target quantity that is plausibly consistent with current knowledge. Such information may be sought in order to help decision makers choose a course of action with a desirable margin of safety. In this case, how complete the uncertainty estimate needs to be and how accurate/precise the estimated contributions of individual sources of uncertainty need to be in order for the uncertainty estimate to be adequate-for-purpose will depend on what counts as a "sufficiently accurate" estimate of the extreme value of the target quantity; this in turn will be a function of the particular decision problem faced. When uncertainty estimates are being produced for a purpose of type (iii), considerations of inductive risk may also come into play: when two reasonable ways of estimating the contribution of a particular source of uncertainty

[17] Inductive risk is the risk of erring in one way or another in one's conclusions, with attendant consequences. See Douglas 2000.

[18] This is not to say that a complete and highly accurate uncertainty estimate, were it already in hand, wouldn't be adequate-for-purpose; it might well be. The issue is one of pursuit.

are available but one risks overestimating the uncertainty while the other risks underestimating it, then the former might be chosen, resulting in a broader range of plausible values of the target quantity.

4.5 ***Uncertainty estimates are essential for judging data adequacy-for-purpose.***

This fifth thesis was already implicit in the discussion of the fourth but is worth making explicit, since it is sometimes overlooked in philosophical discussions of data. Uncertainty estimates allow investigators to probe a particular way in which data can be inadequate for a particular scientific or practical purpose, P_s, namely, by being too uncertain. When the aim is to use data for P_s, an adequate uncertainty estimate is one that allows investigators to *learn whether* the uncertainty associated with the data is so large as to render those data inadequate for P_s. Ruling out this source of inadequacy is required if data are to be judged adequate-for-purpose. Of course, there are other reasons that data can fail to be adequate for a given scientific purpose: they might contain large systematic errors that haven't been corrected for, or they might be in the wrong format, or they might be lacking needed metadata, etc. So, having an adequate uncertainty estimate, and learning from it that data are not inadequate on the grounds of the magnitude of their associated uncertainty, is necessary but not sufficient for establishing that data are adequate-for-purpose.

Another way of arriving at this fifth thesis is simply to recognize that, without information about the uncertainty associated with data, it is impossible to determine what conclusions can and cannot be drawn from those data. Data are, in this sense, relatively uninformative when they are not accompanied by an uncertainty estimate. It is thus unsurprising that metrologists consider a measurement to be *incomplete* unless it includes a statement of the uncertainty associated with a measured value (JCGM 2008, p.4).

The fact that uncertainty estimates are essential for judging data adequacy-for-purpose was overlooked in our recent discussion advocating an adequacy-for-purpose perspective on data. In Bokulich and Parker (2021), we introduced what we call the pragmatic-representational (PR) view of data, according to which data are representations that are the product of a process of inquiry and should be evaluated in terms of their adequacy or fitness for particular purposes. While we emphasized the relevance of accuracy, precision, and resolution of data – as well as their portability, manipulability, and metadata, etc. – to the assessment of data adequacy-for-purpose, we did not discuss in that paper uncertainty estimation at all. The foregoing helps to remedy this omission: an adequate uncertainty estimate is an essential ingredient to the assessment of data fitness-for-purpose; researchers cannot determine that *data* are fit-for-purpose without an adequate estimate of their associated *uncertainty*.[19]

In fact, we propose that the GUM process (a)-(e) discussed in Section 3 can be usefully adapted to enrich the PR view of data, once adjustments are made to reflect an adequacy-for-purpose perspective. The main adjustments are indicated below in bold:

> a) Select and specify the *data target(s)*, i.e., the variable(s), object(s), or phenomena about which information is sought, and the **purposes** for which data are to be used.[20]
>
> b) Model the basic *data-collection principle*, i.e., give an account of how the data-collection process can deliver information about the data target(s).
>
> c) Try to identify any other **relevant** *effects* arising in the practical implementation of the data-collection process, i.e., effects that need to be accounted for in order for the collected data to be used successfully for the intended **purposes**.

[19] An exception might be the rare case in which a purpose is achieved just by producing *some* reasonable value for the target quantity.

[20] In some cases, whole objects and phenomena (rather than variables representing particular properties of them) are data targets; think of data in the form of photographs, imprints, etc.

d) Extend the basic model articulated in (b) to account for the **relevant** effects identified in (c). This requires trying to estimate the magnitudes of those effects with **sufficient** accuracy and precision, where what is sufficient will depend on the **purposes** for which data will be used. The result will be a complete datum or dataset, i.e., a (possibly corrected) datum and an estimate of its associated uncertainty, for each data target.

e) Assess the adequacy- or fitness-for-purpose of the complete datum/dataset. This will require consideration of the extent to which (b), (c) and (d) were successfully carried out and, in some cases, whether **pragmatic features** of data—its format, portability, etc.—are suitable for the context at hand, including its users and their circumstances.

In this way, the production and evaluation of an adequate uncertainty estimate are seen to be built into the process of data evaluation.

5. Illustrating the Five Theses: Uncertainty Estimation for GISTEMP

In this section, we illustrate our theses with the real-world scientific example of uncertainty estimation for a prominent dataset within climate science known as GISTEMP, developed at NASA's Goddard Institute for Space Studies (GISS). This illustration serves both to further establish the plausibility of our five philosophical theses individually and to show how they are jointly instantiated in a single, coherent case study from current scientific practice.

Producing estimates of global temperature variation from local measurements and observations is a complicated process. A very simplified overview is as follows: first, researchers bring together millions of thermometer readings that have been made at various locations on land and at sea since the mid/late-19th century; these data undergo quality control, conversion from absolute temperatures to temperature anomalies[21], correction for recognized sources of systematic error, and spatial and temporal averaging. In the case of GISTEMPv4, much of this processing is performed external to GISS, by the U.S. National Oceanic and Atmospheric Administration (NOAA), resulting in the ERSSTv5 gridded monthly sea surface temperature dataset and the GHCNm monthly land station temperature dataset. These two datasets are inputs to the GISTEMPv4 production process, which involves some further processing and spatial interpolation to produce monthly temperature anomalies on a global latitude/longitude grid from 1880 onward. GISTEMPv4 was preceded by three other GISTEMP versions, the first produced in the 1980s. Because discussions of the GISTEMP uncertainty analysis are strongly suggestive of an Error Approach, we assume that uncertainty estimates here are estimates of possible error.[22]

Uncertainty estimation for GISTEMPv4 clearly supports our first thesis, that ***data uncertainty estimates are substantive epistemic products.*** NOAA produces estimates of uncertainty associated with ERSSTv5 and GHCNm using ensemble (Monte Carlo) techniques. This involves sampling the uncertainty associated with the various inputs to the dataset production process – the thermometer readings and assumptions made in processing them -- and propagating this through to results. It is a brute force way of investigating the joint effects of component sources of uncertainty. The result is a large set of plausible alternatives to ERSSTv5 and GHCNm; 100 of each are selected for use in the GISTEMP uncertainty analysis. The procedure for producing GISTEMPv4 is applied to each of these ERSSTvt-GHCNm pairs to produce 100 plausible alternatives to GISTEMPv4. To each of these, GISTEMP researchers add two different estimates of the "sampling uncertainty" associated with the spatial interpolation step of the GISTEMPv4 production process, which transforms GHCNm station data into data on a latitude/longitude grid. This sampling uncertainty is estimated with the help of hybrid

[21] An anomaly is just a deviation from a reference value, such as the average temperature at a station over the period 1951-1980.

[22] While an Error Approach to data uncertainty seems more common in meteorology and climate science, in some cases researchers in these fields understand uncertainty in epistemic terms (see, e.g., the Bayesian assessment of climate sensitivity in Sherwood et al. 2020).

simulation-observation products known as *reanalyses,* which have complete global coverage (see Lenssen et al. 2024).

The end result is 200 plausible alternatives to GISTEMPv4, each of which reflects one realization of possible errors associated with input thermometer data and its processing. From these 200 datasets, estimates of total uncertainty for individual grid points or for regional averages can be straightforwardly calculated in terms of the spread of values at relevant grid points. Because some grid points are missing values in GISTEMPv4 for some time periods (where station data in an area were very limited or absent entirely), GISTEMP researchers produced an additional 200-member uncertainty ensemble for use in large-scale analyses (including global analyses), which takes account of additional uncertainty stemming from these missing data values, again with the help of reanalyses. Figure 1 shows how this ensemble-based estimate of total uncertainty for GISTEMPv4 global annual mean temperature anomalies (*GISTEMP Ensemble*) compares to a previous attempt (*GISTEMP*; see Lenssen et al. 2019) and to estimates of uncertainty that other research groups produced for their own global temperature analyses using various methods.

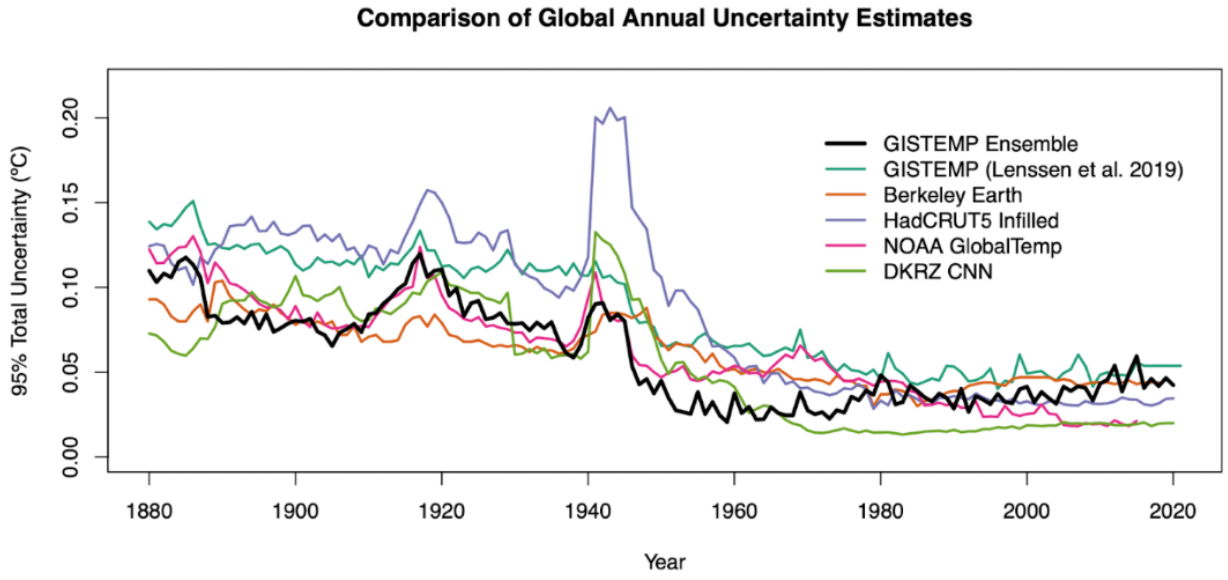

Figure 1. The latest estimate of total uncertainty for GISTEMPv4 annual global temperatures (*GISTEMP Ensemble*), the previous uncertainty estimate for GISTEMPv4 (Lenssen et al. 2019), and estimates that other research groups have produced for their own global temperature analyses. Figure available at: https://data.giss.nasa.gov/gistemp/uncertainty/.

Given the nontrivial assumptions about individual sources of uncertainty that underlie *GISTEMP Ensemble*, it readily illustrates our second thesis, that ***uncertainty estimates are fallible***. Indeed, in Figure 1, the *GISTEMP Ensemble* uncertainty estimate differs noticeably from the previous uncertainty estimate (*GISTEMP,* from Lenssen et al. 2019); if these are intended to characterize the possible error associated with the *same* data production procedure, they *cannot* both be correct. Lenssen et al. (2024) argue that the *GISTEMP Ensemble* analysis is superior to the previous analysis on multiple grounds, including that it has improved (i.e., more *refined*) estimates of some component sources of uncertainty. Similarly, when the *GISTEMP* uncertainty analysis (Lenssen et al. 2019) was produced several years earlier, it was argued to be an improvement on the uncertainty analysis that preceded *it*, which was described as a "rough estimate" of uncertainty that accounted for sampling uncertainty only (ibid., p.6307); the *GISTEMP*

analysis was more c*omplete,* since it included estimates of other important component sources of uncertainty. The history of uncertainty analysis for the GISTEMP datasets thus illustrates nicely our third thesis, that ***uncertainty estimates*** (qua descriptions of possible error) ***can be iteratively improved***.

These improvements were sought because datasets like GISTEMP are produced first and foremost for *measurement* purposes (Parker 2024).[23] The aim is to estimate gridded temperature anomalies with sufficient accuracy to allow the climate research community to correctly answer various further questions of interest about global and regional temperature changes. With each iteration of the uncertainty analysis, GISTEMP researchers note any significant remaining limitations and consider the implications for the range of scientific questions that might be addressed. For instance, for the *GISTEMP* analysis (Lenssen et al. 2019), they investigated whether a failure to account for temporal correlations among uncertainties would make much difference when estimating the probability that any given year was the warmest on record. They concluded that it would not make much difference for this application but cautioned that, for some other important purposes, the dataset could be unsuitable. Remedying this limitation was one of the motivations for producing the newer *GISTEMP Ensemble* analysis, which does account for such temporal correlations (see Lenssen et al. 2024).[24] Scientific practice here thus aligns well with our fourth and fifth theses, that ***uncertainty estimates should be judged as adequate or inadequate for purpose*** and ***are essential for judging data adequacy-for-purpose***.

Other aspects of the GISTEMP analyses also accord well with the adequacy-for-purpose perspective that we advocate. For instance, the most recent GISTEMP uncertainty analysis involved producing *two* uncertainty ensembles, which are explicitly identified as being suitable for different purposes (see Lenssen et al. 2024). In addition, GISTEMP researchers emphasize that a strength of these ensembles is their *ease of use*. They note that, while a "potential reason for the near ubiquitous omission of observational uncertainty in analyses involving historical climate data is the lack of accessible, interpretable, and easily implemented uncertainty products" (ibid., p.2), uncertainty ensembles like those produced for GISTEMPv4 can be implemented "nearly trivially" (ibid.) by users and are easy to interpret. Put differently, their uncertainty ensembles have greater fitness-for-purpose on *pragmatic* grounds too, in accordance with our view.

6. A New Perspective on the Safety vs. Precision Debate

The adequacy-for-purpose perspective that we advocate also offers new insight into the long-standing "safety" versus "precision" debate in metrology (e.g., Cohen & DuMond 1965; Taylor 1971; Grégis 2019a; and Staley 2020). This is a debate over how widely or narrowly to characterize uncertainty, especially in situations where multiple measurements have been performed and their uncertainty estimates are not in statistical agreement. "*Safety"* (also sometimes called "security") is the view that uncertainty

[23] An anonymous referee doubts that GISTEMP is aptly characterized as the outcome of a measuring procedure, suggesting that its production is more akin to what philosophers, since Suppes (1962), have sometimes called "data modeling". We agree that GISTEMP's production involves data modeling in the sense of cleaning, transforming, and regimenting scientific data (see also various forms of model-data symbiosis discussed in Bokulich 2020b), but we see this as entirely consistent with the GISTEMP dataset *also* being the outcome of a measuring procedure. Contemporary views of measurement (e.g., van Fraassen 2008; Tal 2012; Parker 2017) encompass much more than simple, "direct" measuring procedures; data modeling (in the philosophers' sense) can be *part of* a measuring procedure on these views. The *full* procedure underlying GISTEMP's production – starting with the taking of thermometer readings – seems to readily qualify as a measuring procedure on these views, albeit a highly complex and extended one, and GISTEMP's individual gridded temperature anomaly values, as well as the estimates of global temperature change derived from them, are appropriately characterized as measurement outcomes (see Parker 2024). Their uncertainties stem from *both* how data were collected (e.g., shifts in the type of buckets used to collect ocean water (Winsberg 2018, Ch.2), changes in the siting or exposure of station thermometers) and how those data were cleaned, corrected, and gridded.

[24] An additional impetus was the fact that ensemble approaches had become "current best practice" (ibid., p.2) for global temperature analyses. Evolving community standards can change what is expected of uncertainty estimates in measurement contexts and can be a driver for their iterative improvement.

ranges should be made large enough to (1) encompass discordant measurement results; (2) reduce the likelihood that future measurement values will be revised outside the uncertainty estimate of the previous value; and (3) likely contain the unknown true value. By contrast, the "*precision"* view (also known as "sensitivity") calls for excluding data that are judged to be less reliable, thereby keeping uncertainty estimates narrower and reflective of what are considered to be the highest quality measurements.

It is instructive to review the reasons given by parties involved in this debate for their positions. In a 1970 international conference organized by the U.S. National Bureau of Standards (NBS, renamed NIST in 1988) where this issue was debated, Peter Bender, who was a spokesperson for this safety view, argued that we must "find a way to avoid having the quoted uncertainty in the results be systematically too small because of throwing out data" (Bender in Langenberg and Taylor 1971, p. 494; also quoted in Grégis 2019a, p. 50). Bender was concerned about the practice of achieving concordance by excluding certain data sets from the analysis (i.e., "throwing out data"), both because of the subjectivity in judging which data set(s) to throw out, and because doing so could lead to an underestimate of the uncertainty and, hence, an overconfidence in the result.

By contrast, in their 1965 paper on the readjustment of physical constants, Richard Cohen and Jesse DuMond advocate for the precision approach, arguing that this concern for safety is misguided:

> for whom is such an overestimate 'safe'? Certainly not for the general scientific community who wishes to use the result. . . . [It is] an illusory safety for, because of the unwarranted exaggeration of his error estimate, a crucial discrepancy, which might otherwise reveal some basically important new fundamental fact, may have been buried and lost forever. (Cohen & DuMond 1965, p. 541)

Their concern is that expanded uncertainty estimates hide discordances that may reveal new facts. If one makes the uncertainty estimates wide enough, any two measurement results will "agree," prematurely resolving discordances that could be informative.[25,26]

A different justification is offered by another advocate of the precision view, NIST scientist Barry Taylor. Excluding data judged to be of low reliability is advantageous, he contends, because of how researchers will respond: "people will be more apt to make new calculations or to repeat a measurement (or devise new ways to do an experiment) . . . to prove the . . . [expurgator] incorrect" (Taylor 1971, p. 496). In this way, the current estimate will be put to the test, rather than just accepted. Taylor admits that this practice may lead to the more nearly correct data being expurgated and the erroneous data being kept, and indeed he points out this is what happened in the 1965 recommended value of the fine structure constant, where a later Josephson effect measurement showed the wrong choice had been made. Even in this case, however, Taylor argues that a more precise incorrect value "served a more useful purpose than would have the decision to use some kind of average value with an expanded error" (Taylor 1971, p. 496).

The views articulated in the quotations above suggest that *differing purposes or proximate aims*, rather than fundamental disagreements over the nature of measurement or uncertainty estimation, may explain at least some of the disagreement over safety versus precision. Working backward through the views presented, we can identify the following rationales for choosing one approach rather than the other:

1) A precision approach is more likely to spur further attempts at measurement and, in doing so, *to facilitate progress in homing in on the 'true' value*.
2) A precision approach is more likely *to prevent scientists from missing opportunities to uncover important new facts*.
3) A safety approach allows scientists *to avoid representing the measurand as more precisely known than it actually is*.

---

[25] For philosophical discussion of an historical example of the value of discordant measurements, see Ohnesorge 2021.

[26] Kent Staley emphasizes a similar trade-off when discussing his secure evidence framework: "as one weakens the conclusion to enhance the security of the evidence, one diminishes the sensitivity of the measurement result itself to the phenomena of interest" (Staley 2020, p. 89).

4) A safety approach will make it less likely that scientists will have *to revise their measurement values outside of the uncertainty bounds of the previous values*.

These rationales are not only distinct; they are somewhat different in kind. While the third seems purely epistemic, the first and second have a kind of hybrid pragmatic-epistemic character; the precision approach is preferred not on the grounds that it is more likely to deliver the truth immediately, but on the grounds that it will move inquiry forward in a more efficient process of error-correction (#1) or in a way that prevents missing out on substantial new knowledge (#2).

More recent work in the tradition of this debate has identified other subtle differences in contexts that might call for differing approaches.[27] For example, when there is a relatively unified body of data and the scientific community wants to arrive at a single measurement result, the question arises how best to treat outliers, which can motivate what Kent Staley (2012) calls the "robust" approach to statistics. Alternatively, there may be a meta-analytic context in which only the compatible measurements are combined, and other discordant results are simply reported as such, rather than being either combined or discarded. Such an example is discussed by deCourtenay and Grégis (2017), where the outlier then becomes a scientific problem to solve.

In the context of the adjustment of fundamental physical constants where the safety versus precision debate first arose, Taylor points out that few scientists need the last few decimal places of the values for fundamental constants for their purposes, hence a safety approach is not needed. He underscores this low-stakes context when he notes, "no cry of anguish was heard from the general scientific community when . . . the previously accepted values of the constant . . . [had] to be changed. . . . Indeed there was not even as much as a whimper!" (Taylor 1971, p. 496). Two key motivations for remeasuring the values of the constants, he argues, is to test the consistency of physics and detect new physical phenomena—purposes which are better served by the precision approach to uncertainty. By contrast, when Parker and Risbey (2015) argue for default requirements of "faithfulness and completeness" in uncertainty assessment, they have in mind contexts in which climate research is being used to inform practical decision making, where a safety approach (risking overestimating uncertainty) is often more appropriate.

Which of these approaches to uncertainty estimation is the best? Our answer is that there is no one right approach. If one adopts an adequacy-for-purpose perspective on uncertainty estimation, then which approach is more appropriate depends on the context and on the purposes for which the data will be used. A precision approach can be expected to deliver uncertainty estimates that have greater fitness for some purposes and contexts (e.g., some low-stakes contexts of basic research), while in other cases, a safety approach will deliver uncertainty estimates that have a greater fitness-for-purpose (e.g., some high-stakes contexts where the aim is to inform policy, with significant inductive risks). From this perspective, there is no general answer to whether a precision approach, a safety approach, or indeed some other more nuanced approach to handling discordant measurements is the "right one"; it can vary from case to case, depending on the context and purposes for which data will be used.

7. Conclusion

In this paper, we have started to redress the lack of philosophical attention to uncertainty estimation in the context of scientific data. We began by reviewing recent work in metrology that is useful for thinking about data uncertainty; we introduced the Error and Epistemic conceptualizations of data uncertainty and discussed the use of detailed uncertainty budgets when estimating data uncertainty. A key insight of our discussion, to loosely paraphrase Immanuel Kant, is that *data without uncertainty estimates are empty, uncertainty estimates without data are blind.* We went on to defend five philosophical theses about data uncertainty estimates: first, they are substantive epistemic products; second, they are fallible, even when understood as reports on the state of current knowledge; third, they

[27] We thank an anonymous referee for encouraging us to discuss these other proposals, which further underscore our general point about the context and purpose of the uncertainty estimation being important for its evaluation.

can be iteratively improved over time; fourth, they should be judged in terms of their adequacy or fitness for particular purposes; and fifth, they are essential for establishing data adequacy-for-purpose. We illustrated these five theses using the example of uncertainty estimation for the GISTEMP global temperature dataset, which offered further evidence from scientific practice in support of our philosophical views.

In the process, we made a number of further contributions. We offered a novel account of how the quality of data uncertainty estimates should be judged—the adequacy-for-purpose account—and argued that it is preferable to the completeness view implied by the GUM in metrology. Our account takes the quality of a data uncertainty estimate to be determined by its suitability for one or more purposes of interest, which in turn can depend not just on what those purposes are, but also on the context of use, including the users, their methodologies, and broader aims. In addition, we applied this adequacy-for-purpose view to offer a solution to the long-standing safety versus precision debate in metrology: both the context and purpose for which the data are being produced determine whether a "safe" (secure) or "precise" (sensitive)—or indeed some other type of uncertainty estimate—is most suitable. Finally, our discussion helps to build a bridge between two literatures that have unfolded largely independently from one another—work in contemporary metrology and the philosophy of data—demonstrating their mutual relevance.

We close by identifying a few promising avenues for future research. First, though there has been some excellent preliminary work discussing the "epistemic turn" in metrology (e.g., de Courtenay and Gregis 2017), there is room for further examination of the differences between the Error and Epistemic Approaches, the advantages and disadvantages of each, and what leads scientists to adopt different conceptualizations of uncertainty in different contexts. Second, detailed case studies of data uncertainty estimation in different scientific contexts would help to advance understanding of data uncertainty estimation practices and uncover further interesting methodological and epistemological questions for philosophers, such as the handling of outliers (e.g., Bailey 2018). Finally, the adequacy-for-purpose perspective on uncertainty estimation introduced here itself merits further development and critical examination. We hope our discussion provides a useful foundation and will spur further philosophical attention to the scientifically important topic of data uncertainty estimation.

Works Cited

Bailey, David (2018). "Why Outliers are Good For Science" *Significance* 15(1): 14–19. https://doi.org/10.1111/j.1740-9713.2018.01105.x

Bell, Stephanie (1999). "A Beginner's Guide to Uncertainty of Measurement" *Measurement Good Practice Guide*, Number 11, Issue 2. National Physical Laboratory.

Bender, Peter (1971). "Handling of Discrepant Data in Evaluations of the Fundamental Constants". In D. N. Langenberg, & B. N. Taylor (Eds.). *Precision Measurement and Fundamental Constants*: Proceedings of the International Conference held at the National Bureau of Standards, Gaithersburg, Maryland, August 3-7, 1970 National Bureau of Standards Special Publication 343.

Bokulich, Alisa (2020a). "Understanding Scientific Types: Holotypes, Stratotypes, and Measurement Prototypes" *Biology and Philosophy* 35 (5): 1 - 28. https://doi.org/10.1007/s10539-020-09771-1

Bokulich, A. (2020b). "Towards a Taxonomy of the Model-Ladenness of Data" *Philosophy of Science* 87 (5): https://doi.org/10.1086/710516

Bokulich, A. (2020c). "Calibration, Coherence, and Consilience in Radiometric Measures of Geologic Time" *Philosophy of Science* 87 (3): 425 -456. https://doi.org/10.1086/708690

Bokulich, Alisa and Federica Bocchi (2024). "Kuhn's '5th Law of Thermodynamics': Measurement, Data, and Anomalies" in *Kuhn's The Structure of Scientific Revolutions at 60*, ed. by K. Brad Wray. Cambridge University Press.

Bokulich, Alisa and Wendy S. Parker (2021). "Data Models, Representation, & Adequacy for Purpose" *European Journal for Philosophy of Science* 11: 31, pp. 1-26. https://doi.org/10.1007/s13194-020-00345-2

Chandler, David (February 9th, 2012). "Explained: Sigma" *MIT News: On Campus and Around the World.* https://news.mit.edu/2012/explained-sigma-0209.

Cohen, Richard and Jesse DuMond (1965). "Our Knowledge of the Fundamental Constants of Physics and Chemistry in 1965" *Reviews of Modern Physics* 37(4): 537-594.

de Courtenay, Nadine and Fabien Grégis (2017). "The Evaluation of Measurement Uncertainties and Its Epistemological Ramifications" *Studies in History and Philosophy of Science* 65-66: 21-32.

Douglas, Heather (2000). "Inductive risk and values in science." *Philosophy of Science* 67 (4):559-579.

Duhem, Pierre ([1914] 1954). *The Aim and Structure of Physical Theory*, 2nd edition. Translated from French by Marcel Rivière. Princeton, NJ: Princeton University Press.

Frigg, R., Thompson, E., and Werndl, C. (2015). Philosophy of climate science part II: modelling climate change. *Philosophy Compass* 10 (12):965-977.

Grégis, Fabien (2019a). "Accessing Accuracy in Measurement: The Dilemma of Safety versus Precision in the Adjustment of Fundamental Constants." *Studies in the History and Philosophy of Science* 74: 42-55.

Grégis, Fabien (2019b). "On the Meaning of Measurement Uncertainty." *Measurement* 133: 41-46.

JCGM (2008). "Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement [GUM]." BIPM, Joint Committee for Guides in Metrology, JCGM 100:2008. URL: https://www.bipm.org/documents/20126/2071204/ JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6

JCGM (2012). "International Vocabulary of Metrology — Basic and General Concepts and Associated Terms [VIM3]." BIPM, Joint Committee for Guides in Metrology, JCGM 200:2012. (3rd edition). URL: https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf/f0e1ad45-d337-bbeb-53a6-15fe649d0ff1.

JCGM (2020). "Guide to the Expression of Uncertainty in Measurement—Part 6: Developing and Using Measurement Models [GUM-6]." BIPM, Joint Committee for Guides in Metrology. https://www.bipm.org/documents/20126/2071204/JCGM_GUM_6_2020.pdf/d4e77d99-3870-0908-ff37-c1b6a230a337?version=1.9&t=1679905339843&download=true

JCGM (2023). "International Vocabulary of Metrology Fourth edition – Second Committee Draft [VIM4 2CD]. BIPM, Joint Committee for Guides in Metrology. https://www.bipm.org/documents/20126/115700832/VIM4_2CD_clean/c6d0dfb2-ddbf-059e-1f74-9b025c9c59d8

Knutti, Reto, Baumberger, Christoph & Hadorn, Gertrude Hirsch (2019). "Uncertainty quantification using multiple models - Prospects and challenges." In Claus Beisbart & Nicole J. Saam (eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*. Cham, Switzerland: pp. 835-855.

Lenssen, Nathan, Gavin Schmidt, James Hansen, et al. (2019). "Improvements in the GISTEMP Uncertainty Model" *Journal of Geophysical Research: Atmospheres*, 124: 6307–6326. https://doi.org/10. 1029/2018JD029522

Lenssen, Nathan, Gavin A Schmidt, Michael Hendrickson, et al. (2024). A NASA GISTEMPv4 Observational Uncertainty Ensemble. *Journal of Geophysical Research: Atmospheres*, 129 (17): e2023JD040179.

Leonelli, Sabina and Niccolò Tempini (Eds.) (2020). *Data Journeys in the Sciences*. Cham, Switzerland: Springer.

Lusk, Greg and Elliott, Kevin. C. (2022) "Non-epistemic values and scientific assessment: an adequacy-for-purpose view" *European Journal for Philosophy of Science* 12 (2): 35.

Ohnesorge, Miguel (2021) "How Incoherent Measurement Succeeds: Coordination and Success in the Measurement of the Earth's Ellipticity." *Studies in History and Philosophy of Science* 88: 45-62.

OPERA Collaboration (September 22, 2011). "Measurement of the neutrino velocity with the OPERA detector in the CNGS beam". arXiv:1109.4897v1 (https://arxiv.org/abs/1109. 4897v1) [hep-ex (https://arxiv.org/archive/hep -ex)].

Parker, Wendy S. (2010). Predicting weather and climate: Uncertainty, ensembles and probability. *Studies in History and Philosophy of Modern Physics* 41: 263-272.

Parker, Wendy S. (2017). "Computer Simulation, Measurement, and Data Assimilation" *British Journal for the Philosophy of Science* 68 (1): 273-304. https://doi.org/10.1093/bjps/axv037

Parker, Wendy S. (2020). "Model Evaluation: An Adequacy-for-Purpose View". *Philosophy of Science*, 87(3): 457-477. https://doi.org/10.1086/708691

Parker, Wendy S. (2024). *Climate Science*. Cambridge University Press.

Parker, Wendy S. and James S. Risbey (2015). "False Precision, Surprise, and Improved Uncertainty Assessment" *Philosophical Transactions of the Royal Society A* 373: 20140453-1-13. http://dx.doi.org/10.1098/rsta.2014.0453

Reich, Eugenie (2011). "Speedy Neutrinos Challenge Physicists," *Nature* 477: 520. https://doi.org/10.1038/477520a

Rumsfeld, Donald (2002). Department of Defense News Briefing - Secretary Rumsfeld and Gen. Myers. Presenter: Secretary of Defense Donald H. Rumsfeld, February 12, 2002 11:30 AM EDT. https://archive.ph/20180320091111/http:/archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636#selection-401.0-409.30

Sherwood, Steve C., Webb, Mark J., Annan, James D., et al. (2020). "An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence" *Reviews of Geophysics*, 58(4): e2019RG000678.

Staley, Kent (2012). "Strategies for Securing Evidence Through Model Criticism" *European Journal for the Philosophy of Science* 2: 21-43.

Staley, Kent (2020). "Securing the Empirical Value of Measurement Results" *British Journal for the Philosophy of Science* 71: 87-113.

Strassler, Matt (2012) "OPERA: What Went Wrong." http://profmattstrassler.com/articles-a nd-posts/particle-physics-basics/neutrinos/ne utrinos-faster-than-light/opera-what-went-wro ng/

Suppes, Patrick (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), Logic, methodology, and philosophy of science: proceedings of the 1960 international congress (pp. 252–261). Stanford: Stanford University Press.

Tal, Eran (2012). *The epistemology of measurement: a model-based approach.* Ph.D. Dissertation, University of Toronto.

Taylor, Barry (1971). "Comments on Least-Squares adjustments of the Constants." In D. N. Langenberg, & B. N. Taylor (Eds.). *Precision Measurement and Fundamental Constants*: Proceedings of the International Conference held at the National Bureau of Standards, Gaithersburg, Maryland, August 3-7, 1970 National Bureau of Standards Special Publication 343.

Tiesinga, Eite, Peter Mohr, David Newell, and Barry Taylor (2021). "CODATA Recommended Values of the Fundamental Physical Constants: 2018." *Reviews of Modern Physics* 93 (2): 025010. DOI: 10.1103/RevModPhys.93.025010.

van Fraassen, Bas C. (2008). *Scientific representation.* New York: Oxford University Press.

White, Graham (2008). “Basics of Estimating Measurement Uncertainty” *The Clinical Biochemist Reviews* 29, Supplement 1 (August): S53 – S60.

Winsberg, E. (2018). *Philosophy and Climate Science.* Cambridge: Cambridge University Press.