Forthcoming in *Philosophy of Science*

Calibration, Coherence, and Consilience in Radiometric Measures of Geologic Time

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

In 2012 the Geological Time Scale, which sets the temporal framework for studying the timing and tempo of all major geological, biological, and climatic events in Earth's history, had one-quarter of its boundaries moved in a wide-spread revision of radiometric dates. The philosophy of metrology helps us understand this episode, and it, in turn, elucidates the notions of calibration, coherence, and consilience. I argue that coherence testing is a distinct activity preceding calibration and consilience, and highlight the value of discordant evidence and tradeoffs scientists face in calibration. The iterative nature of calibration, moreover, raises the problem of legacy data.

---

1. Introduction

Geochronology is the science of measuring geologic time through the dating of the materials of the Earth. One of the most important dating methods is radioisotope geochronology, which is the investigation of the age of these materials through the radioactive decay of the isotopes they contain. Radiometric methods, which have been around for a hundred years, provide one of most powerful tools for measuring absolute time. There are several different radioactive methods, based on different chemical elements, each with a different half-life, which makes it more or less suitable for measuring particular substances and particular periods of geologic time. The most well-known radiometric method is radiocarbon dating, which is based on the decay of ^{14}C. However, because radiocarbon (^{14}C) has a relatively short half-life, it can only reliably measure dates going back approximately 60,000 years, hardly making a dent in the measurement of geologic time. The two radiometric clocks most relevant to geological time are uranium-lead ($^{206}Pb/^{238}U$) dating and argon-argon ($^{40}Ar/^{39}Ar$) dating, both of which can be used to date materials going back to the formation of the Earth around 4.5 billion years ago.

[†] This paper was written while a visiting researcher in the Earth and Ocean Sciences Division at Duke University. I would like to express my deep gratitude to Brad Murray and the other researchers there for providing such a stimulating and welcoming environment in which to explore these questions. I am also grateful to Blair Schoene for reading over the penultimate version of the paper and to the anonymous referees, whose probing questions led to a much improved paper.

The history of the Earth, from its formation up until the present day, is measured by the official Geological Time Scale (GTS), which provides a hierarchical set of divisions for describing the units of geologic time. The GTS is divided at the largest scale into the Phanerozoic and the Precambrian; the Phanerozoic, for example, is then further divided into the Cenozoic, Mesozoic, and Paleozoic. Each of these is again divided into further subunits; for example, the Mesozoic is divided into the Cretaceous, Jurassic, and Triassic, and so on (the full official GTS can be viewed at www.stratigraphy.org). Although many different "geologic clocks" go into the construction of the GTS, the uranium-lead and argon-argon radiometric clocks provide a crucial backbone for the GTS in that they anchor the various stratigraphic clocks to absolute time, which is measured in (typically millions of) years before the present. The GTS is crucial to the geosciences, in that it sets the temporal framework for studying the timing and tempo of all major geological, biological, and climatic events in Earth's history.

Judgments about time—temporal ordering, coincidence, and the rates of various processes—are central to many (if not most) scientific investigations. Usually we take the conceptualizations and measurements of time, on which these scientific judgments are based, for granted. There are many scientific contexts, however, when such unreflective habits are ill-advised, and a concrete understanding of how time is measured in a domain—and its associated uncertainties—is essential for drawing appropriate scientific or philosophical conclusions. Radiometric measures of geologic time are a case in point. To motivate this claim, I focus on a recent episode in radioisotope geochronology that is prima facie puzzling given our standard, pre-philosophical understanding of radiometric methods: In 2012 one-quarter of all the Geological Time Scale boundaries were revised, and almost half of these changes had the boundary move by more than 4 million years. Such a dramatic change is puzzling to those who thought that the boundaries of geological time periods, such as the Jurassic, were either purely a matter of convention or were settled by the relevant radiometric dates. Why would so many radiometric dates, which are firmly grounded in a fundamental physics-based method for measuring time, need to be revised? In order to understand this widespread revision of radiometric dates—and whether they are likely to be revised again—one must take a closer look at exactly how these radiometric methods work.

My aim in this paper is to show how recent work in the philosophy of metrology (the scientific study of measurement) can help us understand episodes such as this, and in turn how radioisotope geochronology provides a particularly rich case study in which to further advance philosophical work in the epistemology of measurement. In particular, I argue that the current philosophy literature fails to adequately distinguish coherence testing from both calibration and consilience. For example, in his seminal work Eran Tal (2017a,b) defines calibration as a kind of coherence testing; this identification obscures the substantive decisions that scientists must make in light of a failed coherence test. As we will see in the radiometric case, calibration is a further step beyond the information-gathering stage of coherence testing, and there are different possible ways to calibrate a measurement method in light of a coherence failure, each with its own epistemic costs and benefits. The consilience literature has also failed to appreciate the full epistemic role of coherence testing, by focusing only on the concordant outcomes of a coherence test. Although the concordance that grounds consilience arguments is certainly important,

it misses the epistemic value of discordant lines of evidence. One of the few philosophers to recognize the insights that a failure of coherence provides is George Smith (2014), who in the context of Newtonian gravitational research, discusses how the discrepancies that arise between observations (or measurements) and theory are productive of new scientific knowledge in an iterative process he calls 'closing the loop'. Here I want to explicitly extend these insights regarding the scientific value of discordance to the case of a failure of coherence between two measurements, which can similarly be epistemically fruitful.

I begin in section 2 by reviewing some important concepts and distinctions from metrology and the philosophy of metrology, such as the distinction between precision and accuracy, the distinction between a measurement indication and a measurement outcome, and the notion of a measurement standard. I argue that the radiometric case requires that we extend the distinction between indications and outcomes beyond the simple case of direct measurements to the more complicated case of what Wendy Parker (2017) has called derived measurements. In addition, the radiometric case reveals a more complex role for measurement standards than has yet been examined in the philosophy of metrology literature.

Sections 3, 4, and 5 are devoted to the radiometric case studies. In section 3, I briefly review the current philosophical literature on radiometric dating, specifically the work of Alison Wylie (e.g., 2016, 2017) and colleagues, which has been restricted to radiocarbon dating in the context of archaeology. The radiocarbon case illustrates the central notion of calibration and introduces the concept of a calibration curve. This section advances that literature by providing deeper insight into where such calibration curves come from and the epistemological issues that arise in their construction. After drawing four lessons from this work, I turn, in sections 4 and 5, to the two radiometric methods most relevant to the measurement of geologic time: uranium-lead and argon-argon dating.

Section 4 illustrates the process of coherence testing, which precedes either calibration or consilience arguments, and highlights how scientists learn from discordant lines of evidence. As we will see, there was a failure of coherence between uranium-lead dates and argon-argon dates for key events such as the Permian mass extinction. Geochronologists were then faced with the choice between reestablishing coherence by intercalibrating the argon-argon method with the uranium-lead method, or by making changes elsewhere, in order to keep these two radiometric methods independent so that they could be used for consilience. Rather than intercalibrating the two methods, scientists opted to use the discordance between these measurements as a resource for revising background scientific knowledge (or auxiliary hypotheses) that then could be used to refine (or independently recalibrate) these two measurement methods.

Although recalibrations of both these radiometric methods were involved in the puzzling 2012 revision of the Geological Time Scale, the most substantial revisions were due to changes made to the standard used in argon-argon dating. Thus, section 5 turns to an examination of the role that standards play in radiometric measurements, and how a standard can come to be revised. Traditionally in the philosophy of metrology literature, standards are thought to provide accurate values of the quantity being measured (the "measurand"). However, in the case of argon-argon dating, we see a more subtle role for measurement standards: rather than supplying a value of the measurand for calibration,

the standard provides a value for a quantity that goes into the calculation of the measurand for this derived measurement. Understanding this more complex role for standards also advances work in the philosophy of metrology.

In the concluding section, I show how the detailed exploration of this case reveals that the calibration of radiometric dates is importantly an iterative process, and that far from being a rare or unique event, we should expect many further revisions to the Geological Time Scale as geochronologists learn to better identify, manage, and reduce the various sources of uncertainty involved in radiometric measures of geologic time. I argue that it is crucial to recognize the provisional and iteratively-improved status of the GTS because these on-going recalibrations of radiometric dates create the problem of 'legacy data'. Legacy data, very briefly, are data whose method of collection or storage, inhibits their continued use. In the present case, the problem of legacy data arises because relevant radiometric dates obtained prior to 2012 cannot be meaningfully compared to those obtained after 2012 without significant recalibration. A failure to appreciate this point can, in some instances, lead to spurious conclusions for the geological, paleobiological, and paleoclimate studies that rely on these radiometric dates, again underscoring the scientific and philosophical importance of a more detailed understanding how geologic time is measured.

2. Calibration, Coherence, and Consilience

One of the most important distinctions in the philosophy of metrology is that between an instrument *indication* and a measurement *outcome*. A measurement indication is a property of a measuring instrument in its final state after a measurement process. As Eran Tal explains, an indication "does not presuppose reliability or success in indicating anything, but only an *intention* to use such outputs for reliable indication of some property of the object or event being measured" (Tal 2017a, p. 34). A measurement outcome, by contrast, is a knowledge claim that attributes a particular value of a variable or a property to the object or event being measured. This distinction has often been overlooked, as Tal notes, because many instruments are designed to conceal their difference. Familiar instruments, like the thermometer, have "black boxed" the complex inferential process that went into the design and calibration of the instrument, allowing the user to take the indication as the outcome (Chang 2004). A calibration function allows scientists to infer a measurement outcome from one or more measurement indications, along with other auxiliary assumptions and background knowledge.

The term 'indication' is potentially misleading in that it suggests measurement results are always obtained directly from an instrument. As Wendy Parker has noted, however, this is only the case in what she calls direct measurements: "In 'direct measurement' an instrument indication is produced via a process that involves no explicit symbolic calculation, and the raw instrument reading assigns a preliminary value to the parameter that is ultimately of interest" (Parker 2017, p. 280). She contrasts this with a second type of measurement that she terms 'derived measurement', in which the parameters that are directly measured are not the quantity of interest, and the latter must instead be calculated (or derived) from the former via relevant scientific laws, principles, or definitions (p. 281). Radiometric measures of geologic time are precisely such an example of a derived measurement, where the quantities that are directly measured are

not the quantities of interest that motivate the measurement. I argue that the distinction between indications and outcomes is no less relevant for derived measurements than it is for direct measurements, as I show next in the radiometric case.

In the context of radiometric measures of geologic time, the distinction between indications and outcomes is reflected in the geochronologist's distinction between *dates* and *ages*. A date is the number one obtains by measuring both the remaining amount of a radioactive "parent" isotope (aka nuclide) and the amount of the radiogenic daughter that has been produced by decay, and then solving the following radiometric date equation:

$$t = \frac{1}{\lambda} \ln\left[1 + \frac{N_D(t_1)}{N_P(t_1)}\right], \quad (1)$$

where t is time elapsed, λ is the decay constant related to the half-life of the radioactive nuclide, $N_D(t_1)$ is the current amount of the radiogenic daughter, and $N_P(t_1)$ is the remaining amount of the radioactive parent isotope. An age is then inferred from a date, along with a host of auxiliary assumptions, including, as we will see, nontrivial assumptions about the geologic history of the object being measured.

The process of turning a radiometric date into a geologic age, or more generally of establishing a reliable relation between a measurement indication and the relevant feature of the object being measured (a measurement outcome) is known as calibration. The term calibration is used in a variety of ways in different disciplines; the relevant notion here derives from metrology.[1] In the *International Vocabulary of Metrology*, calibration is defined as follows:

> Calibration: operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result ["outcome"] from an indication. (JCGM 2012, p. 28)

In this definition we see the aforementioned distinction between an indication and a measurement result (or outcome), and that calibration is about obtaining the latter from the former. The Joint Committee for Guides in Metrology (JCGM) goes on to note that a calibration may be expressed by a statement, a calibration table, or a calibration curve. An example of a calibration curve, and a discussion of where it comes from, will be given in the next section on radiocarbon dating.

In the context of the philosophy of metrology, Eran Tal has more broadly defined calibration as follows: "In its full generality, *calibration is the activity of modeling different processes and testing the consequences of such models for mutual compatibility* (Tal 2017b, p.12; emphasis original). Although one could define calibration in this way, I argue that it is more helpful to clearly separate out coherence testing, which is an information-gathering activity, from calibration, which is a decision to revise a

[1] As Tal (2017a) notes this is distinct from the sense of calibration one finds, for example in climate modeling, which involves tuning the free parameters of a model to fit the data (pp. 33-34).

measurement procedure in light of that (and/or other) information. Drawing this distinction is helpful because not all calibrations involve coherence testing, not all coherence testing is for the purpose of measurement calibration, and even when coherence testing is used for calibration there can be different ways to calibrate a measurement in light of a failed coherence test. When a coherence test yields discordant measurement results, scientists can decide to modify one (or the other) measurement procedure, revise both measurement procedures, or revise neither.[2] On the approach I am urging here, it should not count as a calibration until this decision is made. This does not of course mean that a calibration is fixed once and for all; Tal is absolutely right to note that calibration is typically an iterative process: when scientists make a significant advance in their understanding of a measurement process, the various sources of error, and how to control for those errors, they will usually decide to recalibrate (i.e., revise) the measurement procedure in light of that information. These recalibrations can involve changing the concrete measurement procedure, changing the way a measurement outcome is inferred from a measurement indication, or changing both. Indeed we will see examples of all three when we get to the radiometric case studies.

Because coherence testing has not been adequately understood apart from calibration or consilience arguments, let me very briefly discuss a context for coherence testing that involves neither. In his landmark work "Closing the Loop: Testing Newtonian Gravity Then and Now", George Smith (2014) highlights the distinct form of testing arising from Isaac Newton's *Principia*, which he argues is not simply a testing of the theory against observations (such that the theory is verified if it agrees and falsified if it does not). Rather, the test is

> whether robust physical sources can be found for each systematic discrepancy between those calculations and observations—with the further demand of achieving closer and closer agreement with observation in a sequence of successive approximations. (Smith 2010, Preface)[3]

The program Smith identifies is one of iterative coherence testing, where the coherence test in this case is between theory and observation. Smith's (and arguably Newton's) key insights are that, first, the discrepancies themselves can be a source of evidence for new facts about the world, which can then be taken into account; and, second, the trajectory over time of the success of this method in iteratively decreasing the discrepancies becomes itself evidence for the reliability of the theory, as the loop is closed.[4]

Although Smith's focus is on coherence testing between theory and observations, I argue that we can extend many of these same insights to coherence testing between two measurement methods. For example, Smith enumerates six possible sources of discrepancies that arise on the observation/measurement side: 1. Simple error—'bad data',

---

[2] The decision to revise neither might arise if scientists discover that two measurements that were thought to be measuring the same quantity, process, or stage of a process in fact are measuring different things.

[3] This concise summary appears only in the 2010 preprint of this article, not in the final published version (Smith 2014), though the rest of the manuscript is essentially identical.

[4] There are many subtleties involved that Smith (2014) discusses, such as that the requirements that the sources of discrepancies be physically robust and not ad hoc.

2. Limits of precision, 3. Systematic bias in instruments, 4. Imprecise fundamental constants, 5. Inadequate corrections for known sources of systematic error, and 6. Not yet identified sources of systematic error (Smith 2014, p. 297). These are the same six sources of discrepancies geochronologists look for when a coherence test between two dating methods yields a discordance. When, in light of the information provided by a coherence test, scientists decide on a particular way to revise the measurement procedure, then it is a measurement calibration. As in the Newtonian case, the history of radiometric geochronology is a history of successfully identifying these sources of discrepancy and resolving them in a way that reduces the discrepancies in each successive iteration.

The aim of calibration is to improve the reliability of a measurement process, where reliability can be understood as a function of two distinct notions: precision and accuracy. Precision reflects the reproducibility of experiments, namely, how close together the values of a sequence of measurements are. Precision is increased by reducing random errors. Over the last couple decades, radiometric methods have entered the era of what is known as high-precision geochronology, meaning that the instruments, techniques, and laboratory protocols for measuring isotope ratios have been improved and refined to the point that random errors are quite small, and both intra- and inter-laboratory reproducibility is quite high (indeed the results are often reproducible to within a remarkable .1% of the age of the object being dated). Although high precision is an important component of the reliability of a measurement, it is important to remember that precision is not the same thing as accuracy.

Accuracy, as it is used in the geochronology literature, is how close the measured values are to the "true" value.[5] Accuracy is increased by reducing any systematic errors. Systematic errors skew the measurement value away from the true value by a given amount. Examples of systematic errors include errors in the values of the decay constants or errors in the natural relative abundances of the potassium isotopes (discussed in section 5). As long as one is only interested in a relative comparison of dates obtained within a single dating method (e.g., comparing uranium-lead dates with each other), then systematic errors can typically be ignored. If, however, one is interested in the absolute value of a date, or one is comparing dates across different dating methods (e.g., comparing radiocarbon dates with dendrochronology (tree-ring) dates, or comparing uranium-lead dates with argon-argon dates), as is typical in the construction of the Geological Time Scale, then systematic errors must be addressed and their associated uncertainties included.

A depiction of the difference between precision and accuracy within the context of radiometric dating is given in Figure 1.

[5] Although this is typically how the notion of accuracy is explicated, recently the International Organization of Standards has proposed calling this "trueness", while using "accuracy" to describe a combination of high precision and high trueness. We will return to the issue of how to understand this "true value" below.

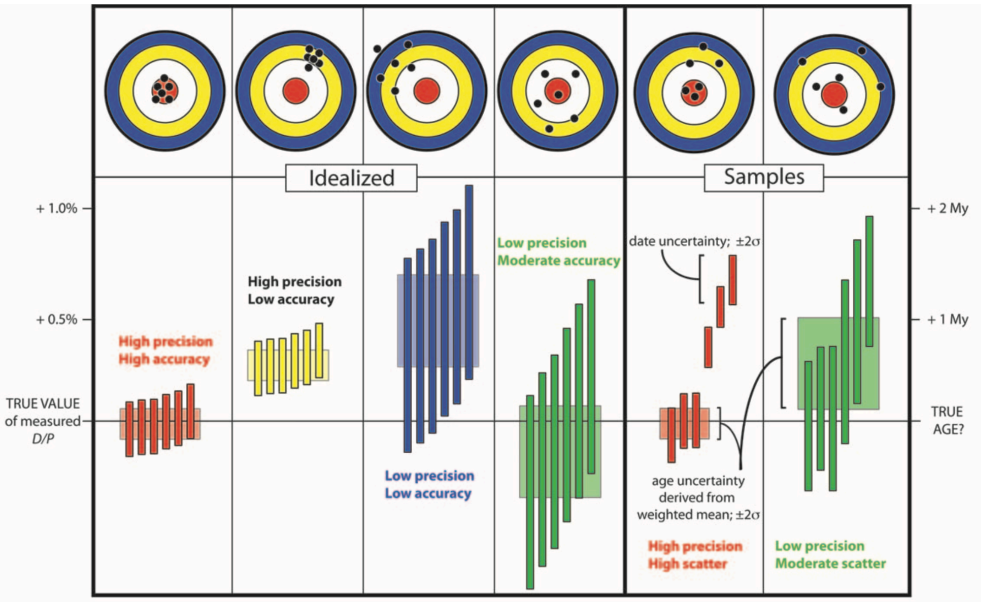

Figure 1: Precision vs. accuracy in radiometric dating. Top row is a typical bulls-eye depiction of the difference between precision and accuracy. The bottom row is an idealized depiction of radiometric data, involving the ratio of radiogenic daughter isotope (D) to radioactive parent isotope (P) (on left), and the associated ages (time elapsed in millions of years) inferred from these measurements (on right). The height of the bar indicates the uncertainties associated with the measurement values. (From Schoene et al. 2013, Figure 1, with permission from the Mineralogical Society of America).

Precision and accuracy are independent from one another, and ideally one aims for both high precision and high accuracy. Indeed, it was the fact that the precision of radiometric methods had begun to outstrip their accuracy that precipitated the large-scale recalibration of radiometric dates in the 2012 GTS.

A key philosophical question, however, is how one makes sense of the notion of accuracy—and indeed how does one go about improving the accuracy of a method—without independent access to the true ages, which are the measurement target? Although the metaphysical notion of accuracy as "closeness to true value" is an important regulative ideal, guiding a community of practitioners in the activity of measurement correction (as argued by de Courtenay & Grégis 2017, p. 22), what is needed for scientific practice is a way to get a handle on accuracy given only the resources that are empirically available to scientists (Tal 2016; 2011). The way geochronologists solve this problem in practice is by assessing accuracy through iterative coherence tests.[6]

In the radiometric context, coherence tests involve picking a key object or event in Earth's history (such as the Permian mass extinction), using multiple dating methods to assign an age to that event, and then assessing the convergence or failure of convergence of these various independent methods. The point of these coherence tests is not (in the

[6] Hasok Chang has similarly talked about how through a process of epistemic iteration progress was made in the development of thermometers (2004). More recently, Fabien Grégis (2019), in his historical discussion of approaches to the adjustment of physical constants, introduces a dynamical notion of accuracy, similar to what is being defended here.

first instance) to arrive at a specific age for the event, but rather to pinpoint potential sources of error in the individual methods and assess the magnitude of their effects. The importance of these sorts of tests has also been emphasized by Alessandra Basso, who describes it as a kind of measurement robustness. Unlike most discussions of robustness, however, she too argues that it is not about corroborating individual results, but rather about evaluating and improving the reliability of the measurement procedure. She writes,

> Based on the recognized sources of uncertainty, scientists formulate expectations (predictions) about the way in which the procedures should converge. . . .This comparison [provides] information about whether and to what extent, the sources of uncertainty actually affect the result in the expected way. (Basso 2017, p. 64)

The information from these pairwise comparisons can be used to help identify potential weaknesses or errors in each of the methods. Geochronologists can then go back and try to fix the hypothesized problematic elements in that dating method (such as by experimentally re-measuring the value of the decay constant), then re-perform the coherence test, and assess the new level of convergence.

It should be emphasized that in a coherence test, one is not forcing a convergence, just assessing its degree. The next key question is how one goes about resolving a failure of convergence. One option, as noted above, is to find independent avenues for correcting a given method. A second option is to intercalibrate the two methods, thereby forcing their agreement. This option can be desirable when one either does not have good independent means for correcting a method or one has good reasons to believe the method that is being treated as a standard in the intercalibration is highly reliable. Taking this second option of intermethod calibration, however, has a price: the agreement between these two dating methods no longer carries the same epistemic weight in arguments of consilience.

Consilience arguments, like calibrations, can begin with a coherence test, though they are again importantly different. Consilience refers to the convergence or concordance of multiple independent lines of evidence on a particular hypothesis or result. The term was coined by William Whewell in 1840 with his phrase 'consilience of inductions,' where the word means literally a 'jumping together' of facts. According to Whewell, a "*Consilience of Inductions* takes place when an Induction obtained from one class of facts, coincides with an Induction, obtained from another different class" (Whewell 1840, p. xxxix). More broadly, consilience has recently been described as

> a mode of reasoning that involves assigning a high degree of plausibility to a given hypothesis (H) when it is supported by a diverse set of independent lines of evidence, which would be unlikely to converge unless H were correct. (Vézer 2015, pp. 3-4)

As emphasized by this definition, consilience is not just any agreement of evidence, but rather an unlikely convergence of independent lines of evidence. The greater the degree of independence (in respects that are relevant) of these evidential lines, the more unlikely it is that they would converge apart from H being true, then the more inductive support such a consilience is said to lend to a hypothesis.

The importance of consilience arguments for drawing inferences about the deep past is discussed in detail by Patrick Forber and Eric Griffith (2011) and Adrian Currie

(2018). They too emphasize the central role played by the degree of independence between the different lines of evidence, where independence increases as the number of shared auxiliary hypotheses decreases. Drawing on a distinction from Alison Wylie (2011), Currie argues that the relevant notion of independence in consilience is horizontal independence:

> Vertical independence concerns cases where different lines of evidence play different roles in an inference, while horizontal independence concerns lines of evidence whose results support the same hypothesis but rely on different auxiliary premises (i.e., consilience). (Currie 2018, p. 151)

The example Forber and Griffith discuss is the consilience of evidence for the impact (Chicxulub asteroid) hypothesis as the cause of the end-Cretaceous mass extinction that wiped out the nonavian dinosaurs. Forber and Griffith distinguish between a strong and weak version of this impact hypothesis. They argue that while the consilience of evidence supports the weak claim that the impact caused some of the extinction events, it does not support the stronger claim that all, or nearly all, the extinctions at the end of the Cretaceous can be attributed to this impact. This is because many of the extinctions appear to have occurred prior to the time of the impact,[7] and there are other potential causes of these extinctions, such as the massive Deccan trap volcanism. Particularly relevant for our discussion here, resolving this long-standing debate requires developing a more fine-grained chronology capable of discriminating the time ordering of these key events, which in turn depends on increasing both the precision and accuracy of our radiometric methods.

3. Lessons in Calibration: Radiocarbon Dating in Archaeology

So far we have defined calibration in very abstract terms as the process of turning an instrument indication into a reliable measurement outcome, or in the context of geochronology more specifically, as the process of turning a radiometric date into a geologic age. In order to gain a deeper philosophical understanding of how calibration works, and the sort of epistemological issues that arise, it is helpful to turn to concrete examples. Before discussing the more complex cases of age calibration involved in the 2012 revision of the Geological Time Scale, we begin with the more well-worked-out example of radiocarbon calibration. Philosophical work on radiocarbon dating in archaeology has revealed four lessons about radiometric methods, which provide a foundation for the uranium-lead and argon-argon cases that follow.

Radiometric methods, like radiocarbon dating, have revolutionized the historical sciences, leading to an unprecedented ability to reconstruct both human and geologic time. Nonetheless, as work in the philosophy of archaeology has shown, the way that revolution has unfolded in the case of radiocarbon (^{14}C) dating was not as straightforward as one might have hoped. Building on Sturt Manning's (2015) history of the three

[7] Even establishing this can be difficult, given artefacts such as the Signor-Lipps effect, which makes biodiversity appear to decline prior to an extinction event; sorting this out requires reconstructing more accurate paelodiversity curves (see Bokulich 2018 for a discussion).

radiocarbon (^{14}C) dating revolutions in archaeology, Robert Chapman and Alison Wylie note that it was soon realized that radiocarbon could not be treated as a "silver bullet" for resolving time, and that it would be another 40 years before the various sources of random error (e.g., effects of electromagnetic impurities, ambient radiation, radon contamination and fractionation) were addressed and protocols ensuring inter- and intra-laboratory reliability had been instituted (Chapman and Wylie 2016, p. 148). This first lesson can be understood as the realization that radiocarbon dates could not be read straightforwardly as the true ages as initially hoped, and that a long process of calibration would be required.

The second lesson that emerged in the history of the radiocarbon revolution is that radiometric methods aren't simple in a geologically complex world. When you look at the decay equation that relates radiocarbon decay to time it is strikingly simple:

$$t = \lambda \, ln\left(\frac{N(t)}{N_0}\right), \qquad (2)$$

where t is time elapsed, λ is the decay constant related to the half-life of radiocarbon (^{14}C), N_0 is the initial amount of radiocarbon in the organism, which was in equilibrium with atmospheric concentrations of ^{14}C up till time of death, and $N(t)$ is the residual amount of radiocarbon remaining. Once the laboratory measurement protocols have been ironed out, you might think that radiocarbon dating is then relatively straightforward; this, however, turns out not to be the case.

As Manning (2015, p.129) and Chapman and Wylie (2016, p. 148) recount, one of the assumptions that initially went in to radiocarbon dating was that atmospheric concentrations of ^{14}C are constant throughout space and time; this turns out, in our geologically complex world, not to be the case. First, concentrations of ^{14}C vary in time, due to variations in the rate of radiocarbon production in the atmosphere, caused by changes in the Earth's magnetic field, solar activity, and the carbon cycle (Hua 2009, p. 379). Second, concentrations vary in space: there is, for example, a variation in atmospheric ^{14}C concentrations between the Northern Hemisphere and the Southern Hemisphere. This is because of the larger ocean surface area of the Southern Hemisphere (~60%, rather than ~40%) and the greater average wind speeds, resulting in more ^{14}C being transferred to the oceans through air-sea exchange of CO_2 (Hua 2009, p. 381). Third, the initial ^{14}C concentrations vary depending on whether it is a terrestrial or marine organism being dated. This is because of the ocean reservoir effect: oceans have much lower ^{14}C concentrations, especially in deep sea environments, because they are not exchanging radiocarbon as frequently with the atmosphere, and hence they become depleted through decay. Planktonic foraminifera, for example, which lived in the surface ocean (and, for example, play a crucial role in biostratigraphy), will appear around 400 years older than contemporaneous terrestrial samples when radiocarbon dated—and for deep-sea organisms, the discrepancy with their terrestrial cousins will be even greater (Hua 2009, p. 380). This ocean-reservoir offset varies by location (e.g., is greater near sites of upwelling) and has been shown to historically vary up to a couple of thousand years during times like the Late-glacial period, due to climatic changes (Hua 2009, p. 380).

So how do scientists deal with all this geological complexity and variation throughout space, time, and type of organism? The answer is that they must go outside of the radiocarbon dating and use other dating methods to develop various calibration curves

and age offsets (e.g., for reservoir effects) to bring the radiometric dates into alignment with ages.

There are different calibration curves synthesized by the international calibration (InCal) working group and approved by the oversight committees for the Northern Hemisphere (IntCal13), Southern Hemisphere (SHCal13), and marine context (Marine13). These calibration curves are understood to be a "work in progress" and are updated every few years as significant new data become available; the most recent is IntCal13 in 2013, shown in Figure 2, which is an update of previous calibration curves IntCal09 and IntCal04.

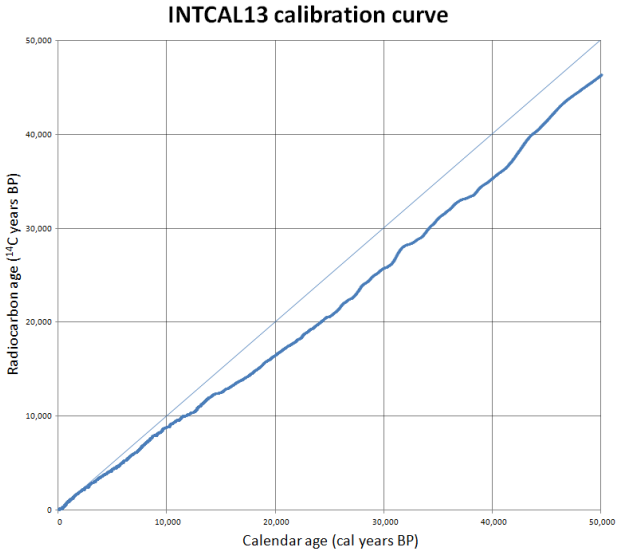

Figure 2: IntCal13: The 2013 version of the calibration curve (darker, wiggly line) to be used in correcting radiocarbon dates (obtained from the decay equation) for terrestrial (non-marine) samples obtained in the Northern (not Southern) Hemisphere. (From Wikimedia Commons, "Graph of INTCAL13" based on data from Reimer et al. 2013a.)

The challenges and reasoning behind radiocarbon dating calibration can be put roughly as follows: In order to obtain correct radiocarbon dates (e.g., to find out how old a terrestrial fossil bone really is), one needs to know the precise concentration of ^{14}C in the atmosphere at the time the fossil organism died; however, one cannot just look up a table for what the atmospheric concentrations were, for example, back in 31,250 years before present, because that number (31,250 yrs. BP) is the unknown you are looking for (the age of the fossil). So what you do instead is assume, for the purpose of the decay equation, some arbitrary, conventionally decided upon, constant-through-time concentration of ^{14}C, and just measure the residual amount of ^{14}C left (in your sample). What you need for calibration is some other absolute time scale (i.e., measured in years) that was constructed by some other, highly reliable, non-radiocarbon (i.e., independent) dating means. Now, absolute dating methods are already hard to come by, but even more challengingly one needs an absolute time scale built on objects that *can be* radiocarbon dated, but *were not* for the purposes of constructing that absolute time scale. If you can find such a reliable, independent absolute time scale (and one that is, moreover, on land and in the Northern Hemisphere, for example), then one can apply radiocarbon dating to

these other objects associated with each of the predetermined years, and find out which amounts of residual ^{14}C correspond to which absolute ages. The residual amount of ^{14}C is the "bridge" that takes you from the not-quite-correct radiocarbon date to the correct (or at least better) date on the other absolute time scale.

Paula Reimer, who is chair of the IntCal Working Group, along with over twenty other colleagues (including, for example, Manning), detail the various different data sources used in constructing the calibration curve and how they are integrated into the resulting curve (Reimer et al. 2013a, 2013b). Although all the gory details are outside the scope of this work, they provide the following summary:

> The Northern Hemisphere calibration is well defined by tree-ring measurements from 0 to 13,900 cal BP and supplemented by the addition of the Lake Suigetsu macrofossil data . . . from 13,900 cal BP to the end of the range of the dating method [~50,000 yrs. BP]. (Reimer et al. 2013a, p.1870)

Rather than just one alternative, highly reliable, non-radiocarbon absolute time scale to cover the entire range of years for which radiometric methods can be used, the calibration curve uses a patchwork of (predominantly) two different absolute time scales: tree-ring counting for the more recent time periods and Lake Suigetsu macrofossil data for the older time periods. In the tree-ring method, one has an *absolute* timescale only if one can correlate the different tree-ring widths, and patch together from different trees a continuous succession of tree rings for every year from the present back to 13,900 years ago. This is in fact what dendrochronologists successfully do. Moreover, tree rings are "ideal recorders of atmospheric ^{14}C" and can be radiocarbon dated, providing the requisite bridge.

The other primary source of data for calibrating the radiocarbon dates comes from annually-deposited, clearly-marked layers of lake sediments, known as 'varves' (or laminae) on Honshu Island in Japan. This is essentially a "relative-time" stratigraphic clock using the principle of superposition (older layers on the bottom), but with the additional property that each layer can be correlated 1:1 with a year. What makes these varves at Lake Suigetsu an absolute (rather than relative) clock is that, like the tree-rings, the layers can be traced all the way to the present. In order to be useful for radiocarbon calibration, the absolute clock needs to be able to have its "ticks" (here the annual deposits) radiocarbon dated: this is done by radiocarbon dating leaf fossils trapped within the varve layers. In this way, each (or at least enough) annual varve sediment layer is correlated with a residual ^{14}C amount, which can then be matched with the residual ^{14}C in your radiocarbon-dated-but-only-roughly-known-age fossil bone, giving it a more precise and accurate age.

Hence, radiocarbon dating, rather than replacing other dating methods such as dendrochronology (counting tree rings) and stratigraphy (counting sedimentary layers), actually *depends on* these other methods for its proper calibration. As Wylie puts it in the context of archaeology, "a third radiocarbon revolution has taken shape that centers on contextualizing radiocarbon dates in relation to a wide range of other sources of chronological evidence, including the archaeological chronologies they were meant to render obsolete" (Wylie 2017, p. 214). Using one dating method to calibrate another, however, means that those two methods are no longer independent, hence a consilience of dates no longer provides the same evidential weight for the accuracy of the age. This,

of course, depends on how fully the alternative dating method is implicated. For example, if it is just the Lake Suigetsu varve data used for calibration, then it is only that data implicated, not necessarily all varve-dating methods; hence, a consilience between calibrated radiocarbon dates and, for example, the Swedish varve chronology (which, say, was not used in the calibration curve) would still carry additional evidential weight. This is why it is essential that adequate metadata be included with radiometric dates, so users of radiocarbon dating can know exactly where the data for calibration came from and can adequately assess its evidential impact.[8]

A related set of issues has to do with error propagation. Radiocarbon dates that have been calibrated have an additional set of potential error uncertainties that must be included: in addition to any errors or uncertainties arising in the radiocarbon methodology itself, there are now also uncertainties or potential errors arising from the calibration data. There are two sorts of errors one might be concerned with: First, there are "local" errors due to the quality of sampling and measurement (e.g., at that particular varve data collection site). This is the sort of error that can typically be quarantined and corrected (in which case, for example, varve chronologies in general would still be useful for calibration). Second, there can be more "global" or systematic errors in the underlying assumptions of a particular dating methodology, which might travel farther (e.g., a problem implicating *all* varve chronologies).

In their discussion of how to choose the material (data) to be used in constructing the official radiocarbon calibration curve, Reimer and colleagues point to both these sorts of problems and allude to two cases:

> Several factors challenge the exercise of deriving a robust calibration curve from a suite of calibration data sets. Key amongst these factors are the integrity of the samples used in the $^{14}C/^{12}C$ data set (e.g. closed systems, stratigraphically consistent) and secondly, the quality of the independent time-scale. Much work has been expended in the past to generate high-quality ^{14}C measurements on records that were later shown to deviate from other calibration records well beyond statistical uncertainties. The Swedish varved clays . . . and the initial single Lake Suigetsu core . . . are 2 examples where hiatuses in sediments caused deviations in the timescales. (Reimer et al. 2013b, pp. 1925-6)

Although they do not elaborate on these two examples further, a few more details are instructive for our discussion. The problem in both cases was ascribed to 'hiatuses' or gaps in the stratigraphic (varve) data; but there are distinct kinds of hiatuses, which are importantly different. In the Suigetsu varve case, the hiatus was due to sampling and measurement errors. A reanalysis of the initial Suigetsu varve data revealed that the main source of error was "caused by both varve counting uncertainties and gaps in the sediment column of unknown duration between successively-drilled core sections" (Staff et al. 2010, p. 960). The "hiatus" here was just human error in sampling, which was subsequently fixed by more careful core drilling and with four parallel sets of cores to

[8] This point reinforces Nora Boyd's broader insights about the importance of "enriched evidence," by which she means "evidence enriched by auxiliary information about how those lines were generated . . . [including] metadata regarding the provenance of the data records and the processing workflow that transforms them" (2018a, pp. 406-407).

cross check for gaps (see, e.g., Bronk Ramsey 2012 for discussion). The Swedish varve case, however, is more problematic. Here the problem did not seem to be human measurement error, but rather actual gaps in the sedimentary record (rock-time) itself (Wohlfarth and Possnert 2000, p. 323). These latter kinds of "hiatuses" can be caused either by no sedimentation (stasis) during certain years, or by deposited varve layers being subsequently eroded. Although researchers still seem to be sorting out the causes of the gaps in the Swedish case, these kinds of hiatuses potentially implicate all varve chronologies in a way that the Suigetsu hiatuses do not, in that stasis may be more common than is typically thought (Tipper 2015).

Antecedently one might have thought that radiometric (absolute) clocks, with their high precision and grounding in fundamental physics, would be all that we need to measure time; and, moreover, that they would have rendered obsolete all other (non-radiometric) chronologic methods, such as dendrochronology and chronostratigraphy. In the case of radiocarbon (^{14}C) dating, however, we have seen that this is clearly not the case. Turning a radiocarbon date into an accurate age depends on calibration curve data, which, in turn, comes primarily from non-radiometric methods, such as tree-ring and varve chronologies. As I have emphasized here, it is epistemically important to know where the data for such calibration curves comes from, and the epistemic tradeoffs that arise in their construction. As is evident from the preceding discussion, calibration curves are also subject to scientific uncertainty, and their iterative improvement is a strength, not weakness, only if they are adequately understood. This discussion thus advances work in the philosophy of metrology by highlighting the epistemic challenges and tradeoffs scientists face in the construction of calibration curves, and why it is important that these issues not be black-boxed.

Building on discussions in the philosophy of archaeology literature about radiocarbon calibration, I have here advanced four general lessons regarding radiometric methods and their calibration: First, there is no silver bullet, which I argue can be understood as the insight that radiometric dates cannot be taken as true ages without significant calibration; second, simple methods aren't simple in a geologically complex world, and much of the progress in a measurement method comes from the refinement of the relevant auxiliary knowledge; third, radiometric dating methods are not autonomous: they depend on both coherence testing and intercalibration with other dating methods; and, fourth, inter-method calibration has a price: as I will make clearer in the subsequent sections, this price (though often one worth paying) is both the loss of the epistemic resource provided by discordances and the opportunity for arguments of consilience. In the next two sections I turn to the radiometric methods that are most relevant for reconstructing geologic time (i.e., the uranium-lead and argon-argon clocks) and show how these same four lessons apply.

4. Coherence Tests with Uranium-Lead Dating

The most useful radioactive elements for reconstructing geologic time are those with long half-lives, such as the two uranium-lead decay chains: ^{238}U to ^{206}Pb, with a half-life of around 4.5 billion years (roughly the age of the Earth), and ^{235}U to ^{207}Pb, with a half-life around 710 million years. Decay constants for the various radioactive elements are not known a priori and must be determined empirically; attempts to

determine these constants through direct counting, however, are typically imprecise for the long-half-life elements, unless they have energetic decay modes. Both ^{238}U and ^{235}U have energetic alpha decay modes, leading to the measurements of their decay constants being called the "gold standard" of geochronology. Nonetheless, even here some systematic errors and the need for further refinements in the values of the decay constant have come to light (Schoene et al. 2006).

Unlike the radiocarbon case, when it comes to uranium-lead and argon-argon dating, there is no well-worked out calibration curve one can look up to turn a radiometric date into a geologic age. Instead geochronologists must grope their way towards an adequate calibration by performing various coherence tests. As noted in the introduction, coherence tests involve comparing different dating methods for the same key object or event in Earth's history and then using the resulting concordance or discordance of dates to probe potential sources of weakness in the respective methods or associated background assumptions. A particularly prominent example of such a coherence test is the comparison of uranium-lead and argon-argon dates for the Permian mass extinction.

The Permian mass extinction (which was earlier and even worse than the end-Cretaceous mass extinction that wiped out the nonavian dinosaurs) is the most devastating known extinction in the Earth's history, with approximately 96% of all marine species, 70% of terrestrial vertebrate species, and 83% of all genera wiped out (see, e.g., Benton 2003). Having a precise and accurate dating of this event, as well as the timing of the various other geological, biological, and climatic events around it, is critical for establishing cause and effect relationships and for a deeper scientific understanding of how such ecological devastation is possible. Determining the age of the end-Permian extinction is also important for the geological time scale, in that it marks both the end of the Permian Period/System (and beginning of the Triassic) and the end of the end of the Paleozoic Era/Erathem (and beginning of the Mesozoic).

The stratigraphic (rock) boundary layer marking this event, known as the Permian-Triassic boundary, is particularly well preserved in the marine strata in the Meishan section of South China, which "are unique because they host both a rich pre- and postextinction biota and abundant volcanic ash beds . . . [that] can be used to constrain the timing and rates of change across the boundary" (Schmitz and Kuiper 2013, p. 26). These features make it an ideal candidate for coherence tests to probe the precision and accuracy of multiple radiometric methods. One of the prime suspects in the killings is the massive flood volcanism from a large igneous province known as the Siberian Traps, hence substantial efforts were undertaken in the mid 1990s to radiometrically date both the extinctions and the inception of the massive volcanism. The two events were dated using both the uranium-lead radiometric method and the argon-argon method (which will be discussed more below). In a seminal paper titled "Absolute Ages Aren't Exactly," Paul Renne and co-workers described the results of this coherence test:

> The end of the Paleozoic era, marked by the most extensive mass extinction in the geologic record, has been dated very precisely by $^{40}Ar/^{39}Ar$ methods at 250.0 ±0.2 million years ago, enabling a comparison with the inception of massive volcanism in the Siberian Traps–also dated by $^{40}Ar/^{39}Ar$ methods, at 250.0 ±0.3 million years ago. . . . These same two events have been dated by ^{206}Pb -^{238}U method at

> 251.4 ±0.3 and 251.3 ±0.2 million years ago. Both radioisotopic systems indicate synchrony of the two events, but comparison between the two systems would suggest discrepancies. (Renne et al. 1998, p. 1841)

Although both radiometric methods had a high precision, the coherence test revealed a problem of accuracy. Up against the "gold standard" of $^{206}Pb/^{238}U$ decay constants, the large discrepancies between the two radiometric clocks was largely attributed to errors in the ^{40}K decay constant, on which the $^{40}Ar/^{39}Ar$ is based, which subsequently was revised, leading to substantial improvements of the argon-argon method (as will be discussed in section 5).

Although the uranium-lead method was thought to be more accurate than the argon-argon method, it too was recognized to have some sources of inaccuracy. Even restricting the coherence tests to within the $^{206}Pb/^{238}U$ radiometric method revealed some discordances. As Blair Schoene notes in a recent review, "the Permian-Triassic boundary . . . has yielded four different nonoverlapping $^{206}Pb/^{238}U$ ages in the last 14 years from the same stratigraphic section in Meishan, China" (Schoene 2014, p. 364). Despite substantial technological and methodological improvements in $^{206}Pb/^{238}U$ dating over this same period, a consistent obstacle to better time resolution and accuracy has been geological complexity.

A fundamental assumption of radiometric methods is what is called "closed-system" behavior. In the case of the $^{206}Pb/^{238}U$ method, one examines zircon crystals (typically a tenth of a millimeter in size) which form at extremely high temperatures when magma from a volcanic eruption begins to cool. The crystal structure of zircon excludes lead, but is able to take up uranium, sealing it within the crystal as it forms. Once the crystal is formed it is highly durable and able to maintain its integrity through geological time. As the radioactive uranium in the crystal begins to decay, a determination of the ratio of lead (Pb) to uranium atoms sealed in the crystal provides a measure of the time elapsed since the crystal was formed. In the case of magmatic zircon, for example, the age provides a good estimate of when the volcanic eruption took place.

While this textbook account applies to some zircon crystals, in the real world the story is often more complex. One problem that arises is *Pb loss*, where some of the accumulated radiogenic lead has "leaked out", which can occur for a variety of reasons. There are often ways of detecting such loss, however, such as through the use of concordia diagrams (involving a comparison of the two independent decay chains of U-Pb), or eliminating such damaged crystals through a mechanical or the newer, improved chemical abrasion method. Nonetheless, subtle Pb loss can be difficult to detect and can limit the accuracy of this radiometric method. Pb loss will result in the rock (or formation event, such as an eruption) appearing *younger* (or more recent) than it actually is. Other forms of open system behavior include Pb gain, U loss, and U gain (see Schoene 2014 for a review).

Another geological complexity that arises is known as *inheritance*, which is when older zircon crystals are able to survive the volcanic eruption (intact) and get incorporated into younger magma, making the event appear *older* than it actually is. Different zircon growth patterns can also be problematic for dating, such as when new zircon crystal growth happens around an older zircon core, or prolonged zircon growth occurs. As researchers at the Berkeley Geochronology Center have shown, "Zircons in

silicic magmas begin to crystallize 10's to 100's of thousands of years (ka) prior to their eruption" (Simon et al. 2008, p. 182).[9] There are thus all sorts of geological complexities that can bias the radiometrically calculated dates, making them appear both younger (e.g., through Pb loss) or older (e.g., through inheritance or prolonged growth), limiting the accuracy of this absolute time clock. As we saw in the case of radiocarbon, geological complexity represents an additional source of error in these radiometric methods. In many ordinary cases of measurement, one is just interested in the state of the system at the time of the measurement. In the case of radiometric methods like uranium-lead dating, however, an accurate measurement depends not just on the current state of the zircon crystal, but also on correctly inferring the geological history of that particular crystal—from its formation up through the millennia till today. Any error in assessing that path-dependence will thwart inferring an accurate age from the radiometric date.

Returning to the example of dating the Permian mass extinction, one can see why establishing the relevant causal mechanisms and explanation is so difficult. Because of geological complexity, the data set, resulting from the collection and U-Pb dating of the zircon samples, is a complicated one, with a considerable scatter for the dates. Geochronologists have to take that scatter of dates and interpret from it an age for the events, and depending on the perceived sources of uncertainty, different statistical and data processing methods can be used. Schoene explains how, even in this era of improved analytical techniques and community-wide standards, different researchers can arrive at such different ages for the Permian-Triassic boundary:

> Those who think a combination of Pb loss, inheritance, and analytical scatter are the most important sources of error, extract the most statistically equivalent population of zircons and apply weighted means. . . . Those who consider pre-eruptive growth of zircon [or a significant reworking of ash material after eruption and deposition] as the source of the spread in dates focus on the youngest grain or subset of younger grains from an ash bed as the best estimate of the eruption age. (Schoene 2014, p. 360)

In the face of all of this geological complexity, these are legitimate differences in scientific practice, given the current state of scientific understanding.

Coherence testing—both within a dating method and between different methods, as we have seen here—reveals discordances arising from a wide variety of sources, and many different substantive decisions have to be made regarding how exactly the loop is to be iteratively closed. In response to the problem of Pb loss (one of the identified physical sources of the discrepancy), the radiometric method was recalibrated by incorporating a new chemical abrasion method as part of the proper measurement procedure. However, not all sources of discrepancy can be addressed through this sort of measurement recalibration. Some improvements required modifications to broader

[9] Although a complication, U-Pb geochronologists have developed various methods to deal with this geological complexity: for example, while Simon et al. (2008) argue that increasing uncertainty in U-Pb dates may mitigate the problem, other approaches such as focusing on the youngest zircon grains rather than a mean date, or using Bayesian statistical approaches to explore probable eruption dates (e.g., Keller et al. 2018) maybe more robust ways to address the problem (Blair Schoene, personal communication).

background knowledge, such as making a more accurate determination of the relevant decay constants. Addressing other sources of discordance afforded the opportunity to make new discoveries about how zircon crystals grow, or discoveries about the dynamics of silicic magma in volcanoes. As two prominent geochronologists note, "these potential differences between chronometers [have] brokered a rich field of research into silicic-magma dynamics—one seeking to disentangle the operative petrologic and volcanic processes via the geochemical, isotopic, and age archives in single zircon crystals" (Schmitz and Kuiper 2013, p. 28). And finally other improvements have to be handled through statistical and other down-stream, data-processing methods that are also not strictly a part of the measurement process.[10] As our examination reveals, even the gold standard radiometric dating method is not simple in a geologically complex world. Nonetheless, this iterative process of coherence testing and then closing the loop by a variety of means illustrates the relevance of Smith's (2014) two key insights, now extended to the context of coherence testing between measurements: First, the discordances themselves can be a source of evidence for new facts about the world, which can then be taken into account; and, second, the trajectory over time of the success of this method in iteratively decreasing the discrepancies becomes itself evidence for the reliability of the radiometric dating method, as the loop is closed.

5. Revising a Measurement Standard: Argon-Argon Dating

The failure of concordance in coherence tests between the two premier high-precision radiometric methods for key events such as the Permian mass extinction revealed the need for changes not only in the uranium-lead method, but also in the argon-argon method. In response, geochronologists reexamined the potential sources of systematic error that could be limiting the accuracy of this method. In order to understand the subsequent process of recalibration, one must first get a clearer picture of how the argon-argon ($^{40}Ar/^{39}Ar$) method works.

Many rocks and minerals contain potassium including the radioactive isotope ^{40}K, which decays via electron capture into ^{40}Ar, which is a gas. When the rock or mineral is melted, such as in a volcanic eruption, the argon gas escapes, essentially resetting the clock; once the rock starts to cool and crystalize, the radiogenic $^{40}Ar^{*}$ subsequently produced is trapped in the rock and starts to accumulate again. Hence the ratio of the amount of $^{40}Ar^{*}$ present in the sample to the amount of ^{40}K remaining can be used to calculate the time elapsed since the eruption event. The equation to calculate time is again strikingly simple:

$$t = \frac{1}{\lambda} ln \left[\frac{^{40}\mathrm{Ar}^{*}}{^{40}\mathrm{K}} \left(\frac{\lambda}{\lambda_e} \right) + 1 \right], \qquad (3)$$

[10] There is, of course, some ambiguity about where to draw the line between where a measurement ends and where the more complex data processing begins, especially when different scientists can process the same raw data in different ways, and even sometimes thousands of years after the initial observation or measurement was made (e.g., Boyd 2018b).

where t is time elapsed, λ is the total decay constant of ^{40}K, λ_e is decay constant of ^{40}K to $^{40}Ar^*$ (the in situ radiogenically produced argon). Like in the U-Pb radiometric method, the assumption of closed system behavior is essential: no radiogenic $^{40}Ar^*$ has leaked out, no atmospheric ^{40}Ar has been incorporated, and no further ^{40}K has been incorporated. Part of what makes a radiometric dating method a good one is the ability to ensure—or more often detect the failure of—this closure assumption. It is precisely the improved ability to assure closed-system behavior that explains why the direct radiometric method of K-Ar dating has largely been replaced with the higher-precision, indirect radiometric method of $^{40}Ar/^{39}Ar$ dating (McDougall and Harrison 1999, p. 12).

Unlike most radiometric method notation, the ratio $^{40}Ar/^{39}Ar$ does not represent a parent-daughter decay relationship (i.e., ^{40}Ar does not decay into ^{39}Ar). Instead, ^{39}Ar, which is created from ^{39}K by bombarding it with neutrons in a nuclear reactor, is used as a proxy for the potassium. This proxy relationship depends on the fact that the naturally occurring relative abundances of the three potassium isotopes (^{39}K, ^{40}K, and ^{41}K) is constant in nature. With $^{39}K/^{40}K$ a constant, the measured ratio $^{40}Ar/^{39}Ar$ is known to be proportional to the desired $^{40}Ar^*/^{40}K$. The naturally produced $^{40}Ar^*$ measured in the rock sample, along with the reactor-produced ^{39}Ar (created from the ^{39}K in the sample), can be used to calculate the time elapsed, t, (e.g., since the rock cooled from the erupted magma), using the following equation:

$$t = \frac{1}{\lambda} ln \left[1 + J \frac{^{40}\mathrm{Ar}^*}{^{39}\mathrm{Ar}} \right], \qquad (4)$$

where λ is total decay constant of ^{40}K and J is known as the neutron flux parameter. The dimensionless parameter J represents the other relevant factors controlling the amount of ^{39}Ar produced in the reactor, such as the length of irradiation, the neutron flux density, and the neutron capture cross section for ^{39}K. Because these factors that go into J are difficult to measure independently, one infers J by irradiating along with the unknown sample of interest, a mineral standard or "neutron flux monitor," whose age is already known, and hence can be used to calculate J for that experiment, according to the following equation:

$$J = \frac{e^{\lambda \tau_m} - 1}{^{40}\mathrm{Ar}^*/^{39}\mathrm{Ar}}, \qquad (5)$$

where $^{40}Ar^*/^{39}Ar$ is the measured ratio, τ_m is the age of the mineral standard (flux monitor), and λ is total decay constant of ^{40}K. Note that here we have an example of a measurement standard that is used, not to provide an accurate reference value for the measurand (t), but rather to provide an accurate value of another quantity (τ_m) that must go into the calculation of the measurand in this derived measurement.

Although the precision of $^{40}Ar/^{39}Ar$ dating has steadily increased with improved instruments (e.g., new generation multi-collector mass spectrometers) and with the refinement and standardization of laboratory protocols, there are two primary sources of systematic error that reduce the accuracy of this radiometric method: first, uncertainties in the decay constant of potassium (λ_K), and, second, uncertainties in the age of the mineral standard (τ_m). Both of these sources of error are places where intercalibration with another dating method could be used to try to reduce the associated uncertainty, and hence potentially improve the accuracy of the $^{40}Ar/^{39}Ar$ method. In both cases, however,

it is important to remember the lesson that inter-method calibration has a price, and decide whether that price is worth paying.

Let us begin with decay constant uncertainties, and return for a moment to our discussion of dating the Permian mass extinction and the Siberian Traps volcanism. A coherence test of the two radiometric methods revealed that the argon-argon dates for these events were systematically younger than the uranium-lead dates by about 1% (roughly 2 million years). As one possible explanation of this systematic discrepancy, Renne et al. 1998 suggested that there were likely errors in the stated values of the decay constants of potassium. Recall that fundamental constants were one of the six sources of observational discrepancies that Smith identified as an epistemically fruitful discordance in his discussion of coherence testing in the Newtonian context. A comprehensive review of systematic errors in $^{40}Ar/^{39}Ar$, including a reassessment of the ^{40}K decay constant, was undertaken by Kyoungwon Min and colleagues in 2000. As noted earlier, the two primary methods for determining decay constants involve either (1) counting the number of disintegration products per unit time emitted from the radioactive material, which can be difficult for the long-half-life elements, or (2) measuring the ratio of radiogenic daughter isotope to parent isotope for a rock or mineral whose age is already known by some independent dating method, which is an intercalibration approach.

The idea behind the intercalibration approach to refining decay constants is that one can potentially export the accuracy of one dating method, such as U-Pb with its more precisely known decay constant, to the $^{40}Ar/^{39}Ar$ radiometric method. If the minerals or rocks on which the two dating methods are used can be related to a set of processes that occurred at the same time, then "one can compare dates from different techniques (e.g., U-Pb on zircon and $^{40}Ar/^{39}Ar$ on sanidine) with uncertainties that are smaller than those in the decay constant experiments" (Schoene et al. 2013, p. 22). This improvement comes at a price, however, as Schoene et al. go on to emphasize:

> Though decay constants determined by intercalibration of different decay schemes provide a means to enhance the relative accuracy of dates, we must recognize that such systems are no longer independent measurements. . . . The resulting covariance between dates means that systematic uncertainties in the U-Pb system propagate through every other system. (Schoene et al. 2013, p. 22)

In other words, two chronometers intercalibrated in this way can no longer be used as independent checks on each other. Because U-Pb and $^{40}Ar/^{39}Ar$ are the two most widely used high-precision radiometric methods, it is useful to keep them separate for the independent information they can provide in arguments from consilience.

A second worry about the intercalibration of these two methods is that it is not clear that the assumption that they relate to two geological processes that occurred *at the same time* always holds. As mentioned earlier, zircons can begin to crystalize in magmas tens-to-hundreds of thousands of years prior to their eruption. Thus, even setting aside the option to intercalibrate these two methods, this portion of the discordance revealed by the coherence test between these two radiometric methods would not require a recalibration of either method. Again we see that coherence testing precedes—but is not identical to—calibration. Even when coherence tests reveal a discordance, that does not mean that either measurement methods must be recalibrated; and even when they do need

to be recalibrated, the coherence test alone does not determine how or when the recalibration will take place.

In the case of reducing decay-constant uncertainties, Min and colleagues decided to focus on improving the calculations based on the already existing data from the direct method for determining decay constants. In particular, some of the other "constants" that had been used in previous calculations of the decay constant had since been revised. Recall, from the discussion above, that the use of ^{39}Ar as a proxy for K depends on the relative abundances of the three potassium isotopes being a constant. While the ratios are still believed to be (mostly) constant in nature (unlike the case of ^{14}C discussed in section 3) , the value of the isotopic ratio $^{40}K/K$ has been revised from 1.18×10^{-4} to $1.17 \pm 0.02 \times 10^{-4}$. Similarly, other relevant constants for calculating λ_K have been revised, including the atomic weight of potassium and Avogadro's number. Min et al. conclude that the previous value of the decay constant, 5.543×10^{-10} traditionally used by geochronologists (e.g., Steiger and Jäger 1977) should be changed to $5.463 \pm 0.107 \times 10^{-10}$. Although this new value (of Min et al. 2000) is the one used for the 2012 Geological Time Scale, ongoing research has proposed other possible values, reminding us that fundamental physical constants are not handed down a priori, and, moreover, their empirically determined values can be—despite their name—inconstant.[11]

The other main source of systematic error limiting the accuracy of the high-precision $^{40}Ar/^{39}Ar$ method stems from uncertainties in the age of the mineral standards (τ_m). In order to be an effective neutron flux monitor, the standard must be of a comparable age to the sample being dated. Here too the issue of intercalibration arises. Recall that the age of the mineral standard (which is co-irradiated with the material to be dated as a neutron flux monitor) goes into the $^{40}Ar/^{39}Ar$ date calculation in this derived measurement; hence any uncertainty in the mineral standard age will contribute to the uncertainty of the material being dated. Geochronologist have thus expended considerable effort to determine accurate and precise ages for the standards, though this process, not surprisingly, runs into many of the same challenges.

One of the most widely used standards in $^{40}Ar/^{39}Ar$ dating, valued for its high reproducibility, is sanidine taken from the Fish Canyon Tuff, which is the remains of an eruption in the San Juan Mountains of southern Colorado that occurred circa 28 million years ago. [12] However, recent attempts to determine the absolute age of the Fish Canyon Tuff sanidine (FCTs) have themselves had a number of twists and turns, which again illustrate the lessons from geological complexity and the price of intercalibration. First, it turns out that FCTs cannot, as initially hoped, serve as a primary standard dated directly by the K-Ar method, because of "the difficulty in quantitatively extracting all the $^{40}Ar^*$ [gas] from the highly viscous K-feldspar melts" (Morgan and Cosca 2017). The second option is to date the FCT by applying the "gold standard" U-Pb radiometric dating method to the zircon crystals at Fish Canyon, and using that age as the same age for the sanidine. As emphasized above, such an intercalibration must pay the high price of making the two most widely used radiometric methods for measuring geologic time

[11] For an excellent historical discussion of epistemological issues in the adjustment of physical constants, see Grégis 2019.

[12] Sanidine is a high-temperature form of potassium feldspar that crystalizes in volcanic rocks, such as obsidian.

interdependent, and it is arguably more useful to keep these two radiometric methods separate, as an independent check on—or comparison for—each other. As discussed in section 2, consilience, or the convergence of multiple independent lines of evidence, is a critical tool when trying to draw inferences about the deep past.

Finally, a third strategy for dating the FCTs involves calibrating against a *non-radiometric* method for measuring absolute time, such as astrochronology. Astrochronology involves two components: cyclostratigraphy, which is the study of cyclic variations in the stratigraphic record, and the correlation of those variations in sedimentation with variations in the Earth's orbit (e.g., variations in orbital eccentricity and the obliquity of the ecliptic). The theoretical basis for how these various orbital variations combine into dominant periods at roughly 100,000 years and 405,000 years, leading to differences in solar radiation (insolation), climate, and hence sedimentation (which can then be read off as a series of bands in the stratigraphic record) was first determined in the 1920s by the mathematician Milutin Milanković. Astrochronology is an absolute dating method in that using classical (celestial) mechanics and the current "initial" conditions one can calculate back in time (currently up to around 50 million years ago) to determine what the orbital conditions, and hence insolation conditions, were at that time.[13]

Astrochronology has come to the aid of the $^{40}Ar/^{39}Ar$ radiometric method by providing an alternative way to date the FCTs standards. This was done, first, by finding a geographical location with a clear cyclostratigraphic sequence that has been astronomically tuned (i.e., had its strata dated by correlation with the orbital solutions), and which, furthermore, has a layer (or layers) of tephra (volcanic ejecta) with sanidine minerals interspersed. Such a location is found, for example, in the Melilla-Nador Basin of Morocco. The location of the tephra horizon within the cyclostratigraphic layers provides an astronomically determined absolute age for those Melilla sanidines contained therein. These astronomically dated Melilla sanidines can then be used as the $^{40}Ar/^{39}Ar$ standards (neutron flux monitors) when co-irradiated with the FCTs, using the latter as the unknown to be dated. This study was carried out by Klaudia Kuiper and colleagues, who conclude

> We compared astronomical and $^{40}Ar/^{39}Ar$ ages of tephras in marine deposits in Morocco to calibrate the age of Fish Canyon sanidine, the most widely used standard in $^{40}Ar/^{39}Ar$ geochronology. This calibration results in a more precise older age of 28.201 ± 0.046 million years ago (Ma) and reduces the $^{40}Ar/^{39}Ar$ method's absolute uncertainty from ~2.5 to 0.25%. (Kuiper et al. 2008, p. 500)

This approach intercalibrates the radiometric $^{40}Ar/^{39}Ar$ clock with the astrochronologic clock (as recorded in the stratigraphic layers), perhaps ironically making the radiometric clock dependent on a stratigraphic clock. This is reminiscent of the radiocarbon case where we saw that the ^{14}C dates similarly needed to be corrected using stratigraphic data (i.e., the Lake Suigetsu varve record). Hence we again see the lesson that radiometric

[13] Astrochronology can still be useful further back in time, but it no longer functions as an absolute dating method, and instead becomes a relative-time clock that must be anchored by radiometric ages.

methods are not autonomous, and that they in fact depend on some of the very dating methods they were initially thought to replace (e.g., Wylie 2017, p. 214).

Another key lesson to take away from all this is that standards can play a much more subtle and complicated role in the calibration of measurements than just providing an accurate reference value of the measurand. Indeed in cases of so-called derived measurements (Parker 2017), there can be a complicated nesting of measurements and standards, as we saw in the FCTs case. Even more importantly, perhaps, is the lesson that, despite their subtle role in measurement, when such standards are revised, they can have a big impact on measurement outcomes. Indeed as we will see next, it was largely this decision to revise the FCTs standard on the basis of an intercalibration with the astrochronology dating method, that led to the dramatic 2012 revision of the Geological Time Scale.

6. Conclusion: Revising the GTS and the Problem of Legacy Data

With this more detailed picture of how radiometric methods work, we are now in a position to solve the puzzle of why there was such a widespread revision to the Geological Time Scale in 2012. For the purposes of the integrated GTS and for meaningfully comparing dates obtained by different researchers, it is essential that the same decay constant values and FCTs standard ages be consistently used in all $^{40}Ar/^{39}Ar$ ages. At the meeting of the EARTHTIME IV workshop, a quorum of $^{40}Ar/^{39}Ar$ researchers met and decided to use the Kuiper et al. 2008 astrochronology-calibrated age for the Fish Canyon Tuff sanidine standard and the Min et al. 2000 value of the total potassium decay constant in the construction of the new official 2012 Geological Time Scale. As a result of this decision, however, all of the $^{40}Ar/^{39}Ar$ dates, which were used as time scale calibration points in the previous 2004 GTS, had to be recalculated using the newly accepted value of the potassium decay constant and new age of the FCTs standard. This resulted in all of the $^{40}Ar/^{39}Ar$ dates being pushed back 0.64% older than they were thought to be in the 2004 GTS. A key factor in this decision was that this combination provided the best available convergence or coherence between the $^{40}Ar/^{39}Ar$ and $^{206}Pb/^{238}U$ clocks, while still keeping them independent.

Although these recalibrations significantly reduce the previously mentioned 1% gap between the $^{40}Ar/^{39}Ar$ and $^{206}Pb/^{238}U$ radiometric clocks, they do not entirely eliminate it. It is thus important to realize that the calibration of radiometric dates—that is, the process of identifying the various sources of uncertainty, and reducing the random and systematic errors that limit the precision and accuracy of these methods—is an ongoing scientific research program. The iterative nature of calibration has two broader consequences: first, it reveals that the GTS is not fixed, but rather a dynamic work in progress, and second, it creates the problem of legacy data.

Defined quite broadly, legacy data refers to data whose method of collection or storage inhibits their continued use. It can arise because the data are stored on floppy disks that no one still has the machines to read, or because the measurement protocols or standards used in the data collection have changed. Legacy data are data that are typically not usable without significant further data processing or curation. Since data can be difficult and expensive to obtain, and some data sources are ephemeral, there is a growing interest in how one ought to collect and store data so that they are reusable by

future generations and for different scientific projects. The topic of legacy data has not yet received the full philosophical attention it deserves, though Sabina Leonelli's recent work on the timescales of data use (2018) and the notion of data journeys (e.g., 2016) is relevant, as is Wylie's (2017) work on putting old archaeological data to reuse. While a full discussion of legacy data is outside the scope of this work, this issue is relevant to our discussion of the recalibration of radiometric dates.

Radiometric dates published prior to the 2012 recalibration cannot be meaningfully compared to those obtained after, without themselves undergoing further recalibration. For example, older argon-argon radiometric dates used what is now taken to be the incorrect value for the decay constant of potassium and incorrect age of the mineral standard used as the neutron flux monitor. Without recalibration, two argon-argon dates (one obtained prior to 2012 and one obtained after) that should have identical ages in the Cambrian period, for example, will appear more than 3 million years apart. An inadequate appreciation of this problem of legacy dates has recently been noted in connection with a number of paleobiological studies of dinosaur evolution by Fowler (2017), who provides a recalibration table using the new Min et al. decay constants and Kuiper et al. mineral standard age for 200 key (pre-2012 measured) argon-argon dates during the Cretaceous period. As he notes, accounting for these changes in calibration can completely reverse some paleobiological conclusions, yet this fact has not been adequately appreciated by many researchers (Fowler 2017, p. 2). Cases such as this again underscore the broader scientific and philosophical importance of a more detailed understanding of how radiometric measures of geologic time work.

As we have seen, the recalibration of radiometric dates at the heart of the 2012 revision of the GTS provides a rich case study with which to deepen our philosophical understanding of calibration, coherence, and consilience. I have argued that coherence testing should be recognized as an epistemically important activity that is distinct from both calibration and consilience. Coherence testing plays a central role in many contexts, including comparing theory predictions with observations, as well as comparing two different measurement outcomes. When a coherence test reveals discrepancies, there are substantive decisions scientists must make regarding how to go about resolving these discordances.

There are at least four broad options available to scientists when a coherence test between two measurements fails: First, they can choose to intercalibrate the two methods. In the radiometric case examined here, that would mean tying the argon-argon method to the uranium-lead one, with the consequence that they would no longer be independent dating methods.[14] A second option would be to independently recalibrate one (or both) of the measurement methods. This is in fact what the geochronologists decided to do: changes were made both to the uranium-lead method (e.g., incorporating the new chemical abrasion procedure) and to the argon-argon method (e.g., revising the decay constant and adjusting the age of the FCTs standard). Third, discordances can be addressed by changing some of the relevant auxiliary hypotheses or background knowledge. This was seen in the radiocarbon case, where scientists had to revise background assumptions about the constancy of atmospheric radiocarbon levels at different time periods and places (e.g., Northern vs. Southern Hemisphere, marine vs.

[14] This approach has been defended by Renne et al. (2010).

terrestrial contexts). A fourth option in light of a failed coherence test is to do nothing at all. Recall, that some of the discrepancy between the uranium-lead and argon-argon methods was due to the fact that these two methods are not actually measuring geological processes that happened at the same time (i.e., uranium-lead is dating zircons crystalizing up to a hundred thousand years prior to an eruption, while the argon method is dating potassium minerals cooling just after that eruption). In such cases, a discrepancy between the two measurement outcomes is to be expected, and does not indicate a problem with either method. In light of all these substantive decisions, it makes more sense to identify coherence testing as an activity that often precedes calibration, but is not identical to it. A calibration is a decision to resolve a discrepancy or revise a measurement process in a particular sort of way.

This case study also reveals the epistemic fertility of coherence tests even in the absence of the concordance required for arguments of consilience. In particular, the discordances revealed by coherence tests can themselves become evidence used in the discovery of new facts about the world. Indeed, this was one of the reasons that geochronologists decided not to intercalibrate the argon-argon and uranium-lead methods: in addition to precluding arguments of consilience, it would have preemptively shut down this potential source of new knowledge, rendering the price of intercalibration too high.[15] Finally, although this pattern of iteratively revising radiometric dates creates the problem of legacy data, it is precisely the trajectory over time of geochronologists' success in finding and resolving these many sources of discordance that itself becomes evidence for the reliability of the radiometric methods they employ.

[15] Whether or not one ultimately agrees that the price is too high, as Schoene (personal communication) notes, arguably the most important things to keep in mind are the particular limitations that any given approach to calibration introduces, and the necessity of making one's calibration decisions explicit to potential users of those radiometric ages, again raising the issue of legacy data.

REFERENCES

Basso, A. (2017), “The Appeal to Robustness in Measurement Practice”. *Studies in History and Philosophy of Science* 65-66: 57-66.

Benton, M. (2003), *When Life Nearly Died: The Greatest Mass Extinction of All Time*. London: Thames and Hudson, Ltd.

Bokulich, A. (2018), "Using Models to Correct Data: Paleodiversity and the Fossil Record" *Synthese* https://doi.org/10.1007/s11229-018-1820-x.

Boyd, N. (2018a), "Evidence Enriched" *Philosophy of Science* 85: 403-421.

Boyd, N. (2018b), "Zombie Data from Babylon". (Talk presented at the Philosophy of Science Association Biennial Meeting on November 1st, 2018 in Seattle Washington). https://psa2018.philsci.org/en/74-program/program-schedule/program/110/methodologies-of-integration.

Chang, H. (2004), *Inventing Temperature: Measurement and Scientific Progress.* Oxford: Oxford University Press.

Chapman, R. and A. Wylie (2016), *Evidential Reasoning in Archaeology*. London: Bloomsbury Academic.

Condon, D. and M. Schmitz (2013), "One Hundred Years of Isotope Geochronology, and Counting" *Elements* 9: 15-17.

Currie, A. (2018), *Rock, Bone, and Ruin: An Optimist’s Guide to the Historical Sciences*. Cambridge, MA: The MIT Press.

de Courtenay, N. and F. Grégis (2017), “The Evaluation of Measurement Uncertainties and its Epistemological Ramifications” *Studies in History and Philosophy of Science* 65-66: 21-32.

Forber, P. and E. Griffith (2011), “Historical Reconstruction: Gaining Epistemic Access to the Deep Past” *Philosophy, Theory, and Practice in Biology* 3 (3): e203. : http://dx.doi.org/10.3998/ptb.6959004.0003.003.

Fowler, D. (2017), “Revised Geochronology, Correlation, and Dinosaur Stratigraphic Ranges of the Santonian-Maastrichtian (Late Cretaceous) formations of the Western Interior of North America” PLoS ONE 12 (11): e0188426. https://doi.org/10.1371/journal.pone.0188426

Gradstein, F., J. Ogg, and F. Hilgen (2012), “On the Geological Time Scale” *Newsletters on Stratigraphy* 45(2): 171-188,

Grégis, F. (2019), "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.

Hua, Q. (2009), "Radiocarbon: A Chronological Tool for the Recent Past" *Quaternary Geochronology* 4: 378-390.

Joint Committee for Guides in Metrology (2012), *International Vocabulary of Metrology: Basic and general concepts and associated terms (VIM)*, 3rd Edition. https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf.

Keller, C.B., B. Schoene, and K.M. Samperton 2018, "A Stochastic Sampling Approach to Zircon Eruption Age" *Geochemical Perspectives Letters* 8: 31-35.

Kuiper, K., A. Deino, F. Hilgen, W. Krijgsman, P. Renne, J. Wijbrans (2008), "Synchronizing Rock Clocks of Earth History" *Science* 320 (5875): 500-504.

Leonelli, S. (2018), "The Time of Data: Timescales of Data Use in the Life Sciences" *Philosophy of Science* 85 (December): 741-754.

Manning, S. (2015), "Radiocarbon Dating and Archaeology: History, Progress, and Present Status" in R. Chapman and A. Wylie (eds.) *Material Evidence: Learning from Archaeological Practice*. London: Routledge, pp. 128-158.

McDougall, I. and T. M. Harrison (1999), *Geochronology and Thermochronology by the $^{40}Ar/^{39}Ar$ Method*, Second Edition. Oxford: Oxford University Press.

Min, K. , R. Mundil, P. Renne, K. Ludwig (2000), "A Test for Systematic Errors in $^{40}Ar/^{39}Ar$ Geochronology through Comparison with U/Pb Analysis of a 1.1-Ga Rhyolite" *Geochimica Acta* 64 (1): 73-98.

Morgan, L. and M. Cosca (2017, June 23), "Sanidine from the Fish Canyon Tuff and its Use as a $^{40}Ar/^{39}Ar$ Geochronology Standard" [Earthtime Blog post]. Retrieved from http://www.earthtimetestsite.com/blog/.

Parker, W. (2017), "Computer Simulation, Measurement, and Data Assimilation" *British Journal for the Philosophy of Science* 68: 273-304.

Phillips, D., E. Matchan, M. Honda, and K. Kuiper (2017), "Astronomical Calibration of $^{40}Ar/^{39}Ar$ Reference Minerals Using High-Precision, Multi-Collector (ARGUSVI) Mass Spectrometry" *Geochimica et Cosmochimica Acta* 196: 351-369.

Reimer, P, E. Bard, A. Bayliss, J.W. Beck, P. Blackwell, C. Bronk Ramsey, C. Buck, H. Cheng, R. L. Edwards, M. Friedrich, P. Grootes, T. Guilderson, H. Haflidason, I. Hajdas, C. Hatté, T. Heaton, D. Hoffmann, A. Hogg, K. Hugen, K. F. Kaiser, B. Kromer, S. Manning, M. Niu, R. Reimer, D. Richards, E. M. Scott, J. Southon, R. Staff, C. Turney, J. van der Plicht (2013a), "INTCAL13 and MARINE13 Radiocarbon Age Calibration Curves 0-50,000 Years Cal BP" *Radiocarbon* 55 (4): 1869-1887.

Reimer, P, E. Bard, A. Bayliss, J.W. Beck, P. Blackwell, C. Bronk Ramsey, D. Brown, C. Buck, R. L. Edwards, M. Friedrich, P. Grootes, T. Guilderson, H. Haflidason, I. Hajdas, C. Hatté, T. Heaton, A. Hogg, K. Hugen, K. F. Kaiser, B. Kromer, S. Manning, R. Reimer, D. Richards, E. M. Scott, J. Southon, R. Staff, C. Turney, J.

van der Plicht (2013b), "Selection and Treatment of Data for Radiocarbon Calibration: An Update to the International Calibration (INTCAL) Criteria" *Radiocarbon* 55 (4): 1923-1945.

Renne, P., D. Karner, and K. Ludwig (1998), "Absolute Ages Aren't Exactly" *Science* 282 (5395): 1840-1841.

Renne, P., R. Mundil, G. Balco, K. Min, and K. Ludwig 2010, "Joint Determination of ^{40}K Decay Constants and $^{40}Ar^{*}/^{40}K$ for the Fish Canyon Sanidine Standard, and Improved Accuracy for $^{40}Ar/^{40}K$ Geochronology" *Geochimica et Cosmochimica Acta* 74: 5349-5367.

Schmitz, M. (2012), "Radiogenic Isotope Geochronology" in F. Gradstein, J. Ogg, M. Schmitz, and G. Ogg (eds.) *The Geological Timescale 2012*. Oxford: Elsevier BV.

Schmitz, M. and K. Kuiper (2013), "High-Precision Geochronology" *Elements* 9: 25-30.

Schoene, B. (2014), "U-Th-Pb Geochronology" in A. Davis (ed.) *Treatise on Geochemistry*, *2nd ed. Volume 1: Meteorites and Cosmochemical Processes*. Amsterdam: Elsevier Science, pp. 341-378.

Schoene, B., D. Condon, L. Morgan, N. McLean (2013), "Precision and Accuracy in Geochronology" *Elements* 9: 19-24.

Simon, J., P. Renne, and R. Mundil (2008), "Implications of Pre-Eruptive Magmatic Histories of Zircons for U-Pb Geochronology of Silicic Extrusions". *Earth and Planetary Science Letters* 266: 182-194.

Smith, D., R. Bailey, P. Burgess, and A. Fraser (2015), "Strata and Time: Probing the Gaps in our Understanding" in Smith, D., R. Bailey, P. Burgess, and A. Fraser (eds.) *Strata and Time: Probing the Gaps in our Understanding*. Geological Society, London, Special Publications, 404: 1-10.

Smith, G. (2014), "Closing the Loop: Testing Newtonian Gravity, Then and Now" in Z. Biener and E. Schliesser (eds.) *Newton and Empiricism*. Oxford: Oxford University Press, pp. 262 -351.

Staff, R., C. Bronk Ramsey, C. Bryant, F. Brock R. Payne, G. Schlolaut, M. Marshalll, A. Brauer, H. Lamb, P. Tarasaov, Y. Yokoyama, T. Haraguchi, K. Gotanda, H. Yonenobu, T. Nakagawa, and Suigetsu 2006 Project Members (2011), "New ^{14}C Determinations from Lake Suigetsu, Japan: 12,000 to 0 Cal BP". *Radiocarbon* 53 (3): 511-528.

Steiger, R. and Jäger, E. (1977), "Subcommission on Geochronology: Convention on the Use of Decay Constants in Geo- and Cosmochronology" *Earth and Planetary Science Letters* 36(6): 359-362.

Tal, E. (2011), "How Accurate is the Standard Second?" *Philosophy of Science* 78: 1082-1096.

Tal, E. (2016), "Making Time: A Study in the Epistemology of Measurement" *British Journal for the Philosophy of Science* 67: 297-335.

Tal, E. (2017a), "Calibration: Modeling the Measurement Process" *Studies in History and Philosophy of Science* 65-66: 33-45.

Tal, E. (2017b), "A Model-Based Epistemology of Measurement" in N. Mößner and A. Nordmann (eds.) *Reasoning in Measurement*. London: Routledge, 233-253.

Tipper, J. (2015), " The Importance of Doing Nothing: Stasis in Sedimentation Systems and its Stratigraphic Effects " in Smith, D., R. Bailey, P. Burgess, and A. Fraser (eds.) *Strata and Time: Probing the Gaps in our Understanding*. Geological Society, London, Special Publications, 404: 105-122.

Vézer, M. (2015), "Aggregating Evidence in Climate Science: Consilience, Robustness and the Wisdom of Multiple Models" (2015). Electronic Thesis and Dissertation Repository. 2837. https://ir.lib.uwo.ca/etd/2837.

Whewell, W. (1840), *The Philosophy of the Inductive Sciences, Founded Upon Their History,* Volume I. London: John W. Parker.

Wohlfarth, B. and G. Possnert (2000), "AMS Radiocarbon Measurements from the Swedish Varve Clays". *Radiocarbon* 42 (3): 323-333.

Wylie, A. (2017), "How Archaeological Evidence Bites Back: Strategies for Putting Old Data to Work in New Ways" *Science, Technology, & Human Values* 42 (2): 213-225.