Making measurement useful: Integrating measurement, uncertainty, and sensitivity

Pierre-Hugues Beauchemin* Kent W. Staley†

Keywords: measurement, pragmatism, inquiry, experiment, uncertainty, evidence

Abstract: We employ a pragmatist model of inquiry to explain how measurement in physics can solve the problem of usefulness. In spite of the fact that a variety of resources, including theory, simulation, heuristics, rules of thumb, and practical considerations contribute to the context of a specific measurement inquiry, the measurement inquiry process partially decontextualizes its results, making them useful for other inquiries. This measurement inquiry process involves a process of transformation of data we call "entheorization," which happens in conjunction with the evaluation of uncertainty of measurement results. These uncertainty estimates then serve to define the sensitivity of the result to the aims of subsequent inquiries. On this approach, the epistemology of measurement requires treating measurement procedure, uncertainty estimation, and sensitivity to targets of inquiry as equally fundamental to understanding how measurement yields knowledge. To help understand how the abstract

*Department of Physics and Astronomy, Tufts University
†Corresponding author, Department of Philosophy, Saint Louis University: kent.staley@slu.edu

elements of our epistemological model of experimental inquiries are applicable to concrete episodes of measurement, we use the example of the W-boson mass measurement at the Large Hadron Collider to illustrate our arguments.

1 Introduction

Central to the turn against "theory-dominated" modes of philosophical inquiry into science has been a shift: from trying to understand the relationship of evidence to theory to seeking to understand how experimental results are obtained and used in the production of scientific knowledge. In this perspective, how experimental results *become* evidence for theoretical claims takes center stage. In Ian Hacking's influential works (Hacking, 1981, 1983), such a turn centers on the validation of evidence: how is trust in experimental results established? For him, answering this question requires paying attention to the details of scientific practice, and identifying the arguments supporting trust that the results of experiments can be rationally and objectively used as evidence in support of some theory. This perspective has allowed the pursuit of arguments to address problems posed by post-positivists and social constructivists that have been understood as calling into quesion the objectivity of science. Examples include the experimenters' regress (Collins, 1985, 1994) and the theory-ladeness of experimental evidence (Hanson, 1958). For example, Allan Franklin proposes the use of calibration for establishing trust in the results of an experiment, thus coping with the fact that experiment is fallible and sometimes produces discordant results (Franklin, 1994, 2002). He argues that experimenters replicate known results obtained by other experiments to generate trust that the results they obtained are not artifacts of the instrument and do fulfill their evidential function. The fact that this strategy is effectively articulated in many reports of experimental inquiries, however, does not on its own answer an important question: Does this strategy contribute epistemically to

establishing trust in experimental results, or does it succeed on simply a rhetorical level?

If, with Bogen and Woodward (1988), we recognize the idiosyncrasies of experimental data, we must ask how the fact that an apparatus produces an expected result in one context serves to show that a different result, obtained in a different context, is to be trusted. Bogen and Woodward argue that it is via the inference from data to experimental phenomena that science generates trustable and stable evidence. That doesn't fully address the problem: how *could* phenomena applicable to a broad variety of experimental contexts be distilled from raw data bearing the idiosyncratic marks of the context of their production? This debate endures today. For example, to explain how calibration offers a strategy for validating a measurement process, Eran Tal proposes that measurement relies on an idealized model of the measuring process, enabling measurement outcomes to become objective in attributing a property to the measured object rather than to the idiosyncrasies of the concrete measurement process used in the production of that outcome (Tal, 2017). From the perspective of Nora Boyd's account of enriched evidence (Boyd, 2018), such an idealization away of the details of data production poses the risk of preventing an adequate response to the problem of the experimenters' regress: Details omitted from an idealized model may in fact be relevant metadata, necessary for trustworthy use of past experimental results in later inquiries.

From Boyd's position, we can see that what really lies behind the problem of the experimenters' regress is not so much the logical structure of evidential argumentation in experimental science, nor how agents, individually or collectively, gain trust that a given experimental result can serve to warrant theoretical beliefs. Instead, we can ask: how can the idiosyncrasies of the results of an experiment be acknowledged and maintained as relevant, as argued by Boyd, while being decontextualized for use elsewhere, as discussed by Tal? How do experimental results become useful resources for new and

different inquiries? This is *the problem of usefulness*. This problem has epistemological advantages over the question of trust. It does not prioritize the question of trust over the one of evidence, but allows framing both questions in the same context of securing evidence (Staley, 2012, 2020). Normative judgments about what constitutes a trustable experimental result are not decoupled from the question of what those results are meant to be evidence for. This change of perspective also avoids artificially isolating experiments from one another. The question is no longer "how can one trust this result?" but "how can this result be used, jointly with others, in the pursuit of other inquiries?" Finally, the problem of usefulness invites a more open-ended solution: Unlike trustworthiness, usefulness is not an intrinsic or isolated feature of a particular experimental result; it is a judgment bearing on a result's use in a given context. A resource can be useful for one inquiry, and not for another one. In brief, focusing on the problem of usefulness encourages making sense of scientific practice, integrating epistemological observations about experimentation that have been highlighted by both the "New Experimentalist" and the "Social Constructivist" traditions, while opening up a breadth of new interesting philosophical interrogations in a coherent way.

In the following, we propose a solution to the challenge posed by the problem of usefulness, using a pragmatist account of science, as outlined in Sect. 4 below (see also Beauchemin and Staley, 2024). However, we will restrict our reflection to a special class of experimental inquiries – measurement – for which the problem of usefulness is most crucial. Applying the pragmatic method, we ask what practical difference is made by practices of measurement: What is done in producing a measurement, and what is achieved by doing that, that makes a difference in how the results of such practices are used? Showing that a measurement result is useful as a resource for inquiry goes beyond showing that it can be trusted. We will show that the problem of usefulness (as detailed in Sect. 2) and our pragmatist framing, promote understanding of experimental practices related to the production and use of measurement results,

by capturing their epistemological significance. We decline to begin our analysis by treating measurement itself and its relation to models. That approach lends itself to treating measurement uncertainty as a secondary issue of how to characterize the quality of a measurement process or its product, while leaving the role of sensitivity largely unarticulated (Joint Committee for Guides in Metrology Working Group II, 2012; Joint Committee for Guides in Metrology Working Group I, 2008). Instead, we argue that outlining the centrality and complete entanglement of these aspects of measurement is key to a solution to the problem of usefulness.

Our pragmatist model of experimental inquiry analyzes scientific practice in terms of a "use mode" and a "critical mode" of inquiry performed in view of attaining certain aims. Such a general account of scientific inquiry allows understanding how the practices of attributing uncertainties to measurement results, and using them to assess the sensitivity of an inquiry to its targeted objectives, contribute to scientific knowledge. We show how, in spite of the fact that a variety of resources – including theory, simulation, heuristics, rules of thumb, and practical considerations – contribute to the context of a specific measurement inquiry, the measurement inquiry process partially decontextualizes its results, making them useful for other inquiries. This measurement inquiry process involves a process of transformation of data we call "entheorization," which happens in conjunction with the evaluation of uncertainty of measurement results. Such uncertainty estimates then serve to define the sensitivity of the result to the aims of subsequent inquiries. It is essential that reports of measurement results contain enough information about the uncertainty estimate, and about the dependence of entheorization on the context in which it has been performed, to allow using the result in other inquiries. That is the core of the solution we propose to the problem of usefulness, as explained in Sect. 6. To help understand how the abstract elements of our epistemological model of experimental inquiries are applicable to concrete episodes

of measurement, we use the example of the W-boson mass measurement at the Large Hadron Collider to illustrate our arguments. This is discussed in Sect. 5.

2 The problem of usefulness

Two prominent features of scientific measurement appear to be in tension with one another, arising from two central features: (1) In measuring, inquirers aim to achieve a result that enjoys evidential support. (2) The results of measurement are meant to be useful as evidential support in subsequent inquiries.

Achieving (1) is necessary for (2), but in ways that can limit the extent to which (2) is achieved. On the one hand, the content of measurement results and their evidential support depend upon details of the particular conditions of their production. Both the content of the result and the nature of the evidence supporting that result are relevant to the question of the use that can be made of the result. Yet, measurement results are produced for the purpose of being used in inquiries and other activities that will be conducted in conditions distinct from those in which they were produced. This tension between dependence upon the conditions of production and the aspiration for usefulness beyond those conditions gives rise to what we call "the problem of usefulness": How do specific and concrete measurement procedures executed in one context produce results that can be used in the conduct of scientific inquiry in a broader range of contexts?

Like the Roman god Janus, the problem of usefulness faces both forward and backward, in a way that can be expressed by two further questions: (A) How do investigators warrant that they appropriately used evidence from previous or ancillary inquiries to inform their choices about how to produce measurement results? (B) How do investigators warrant that the results they produce can be used as evidence in the context of distinct inquiries with their own varied aims?

3 Background

An adequate treatment of the problem of usefulness for measurement requires an account of measurement that takes seriously the importance of the aims of measurement activities, both in the context of the production of measurement results and in the context of the use of measurement results (which often involves the production of another measurement result). We regard such an emphasis on the centrality of aims to be a defining feature of an *inquiry* approach to measurement. In this section we note some recent contributions to the literature on data, evidence, and measurement that take helpful steps toward a satisfactory account of useful measurement, while clarifying what our approach will provide to go beyond these previous efforts.

The need for an account that deals explicitly with this issue is evident from the guidance documents of metrology, the GUM (Guide for the Expression of Uncertainty in Measurement) and VIM (International Vocabulary of Metrology). Pervasive through both documents are references to constraints imposed by the demands of usefulness. For example, in a section on "Practical considerations," in the GUM, we read that "implicit in this *Guide* is the assumption that a measurement can be modelled mathematically to the degree imposed by the required accuracy of the measurement" (Joint Committee for Guides in Metrology Working Group I, 2008, 7). By whatever means it is decided what accuracy is *required* in a given case, the decision must involve some consideration of the use to which the given measurement, or given type of measurement, will be put. In other words, in any given case one cannot make sense of what is 'required' of a measurement without being able to answer the question 'required for what?'

That how a measurement is to be used constrains (without determining) how it is to be performed is thus implicit in the GUM and VIM. These documents also invoke untheorized normative concepts. Some characterize what is needed for competent execution of the measurement process, as when the GUM states that the framework it

provides for assessing uncertainty "cannot substitute for ciritical thinking, intellectual honesty and professional skill" (ibid., 8). In other instances, the normative concepts play a more constitutive role in characterizing measurement, perhaps most notably in the VIM's definition of measurement itself as a "process of experimentally obtaining one or more quantity values that can *reasonably* be attributed to a quantity" (Joint Committee for Guides in Metrology Working Group II, 2012, 16, emphasis added). One contention of the approach to measurement defended here is that the practical considerations imposed by the context of use that the GUM implicates in the passage quoted in the previous paragraph are inseparable from the required judgments of reasonability implied in the VIM's definition of measurement.

Eran Tal's model-based epistemology of measurement treats measurement as "a set of procedures whose aim is to coherently assign values to model parameters based on instrument indications" (Tal, 2020), where an indication is a "property of a measuring instrument in its final state after the measurement process is complete" (Tal, 2017, 235). The parameters of interest in a measurement activity (the measurands) are parameters in a model of the measuring process. The instrument and its indications are part of the concrete measurement process, while a measurand is a parameter of the abstract, idealized model of that process.

In Tal's account, the accomplishment of the aim of measurement results in a meaurement *outcome*, which is a knowledge claim, inferred from instrument indications and background knowledge, "associating one or more parameter values with the object or event being measured." This achievement requires that the claim in question "be abstracted away from its concrete method of production and pertain to some quantity objectively," meaning that it is "attributable to the measured *object* rather than to the idiosyncrasies of the measuring instrument environment and human operators" (ibid., emphasis in original).

Without articulating explicitly the problem of usefulness, Tal in this way describes a criterion for successful measurement that might be invoked as a solution to that problem. More precisely, the satisfaction of this requirement would render a measurement outcome (in Tal's sense) useful to the extent that the object to which the claim is attributed can be taken to exist and bear its attributed properties in contexts distinct from that in which the measurement outcome was produced.

What is missing from Tal's account, however, is the consideration of the aims of measurement related specifically to the *use* of measurement results and their implications for the aims of the measurement process itself. Tal singles out the aim of coherently assigning values to parameters of the measurement model, but does not incorporate into his epistemological account the bearing that downstream aims have on decisions that human operators make about how to employ the idiosyncrasies of the measuring instrument, decisions that matter not only to the content of the measurement result but to the evidence that supports that result and the contexts in which the measurement result can and cannot be warrantedly employed. The ideal of attributing a result to the measured object full stop does not allow for consideration of such complications and thus stands in the way of an adequate treatment of the problem of usefulness.

Nora Boyd's account of "enriched evidence" helpfully draws attention to ways in which the usefulness of evidence depends on recording details about the process through which inquiry was conducted as an integral part of the evidence such inquiry yields (Boyd, 2018). Although Boyd's account aims at a broader target than just measurement, it includes features well worth retaining in an inquiry-based approach to measurement. Boyd provides an interpretation of the empiricist requirement that theories should be consistent with empirical evidence in terms of requiring consistency with "enriched lines of evidence." A line of evidence consists of "a sequence

of empirical results including the records of data collection and all subsequent products of data processing generated on the way to some final empirical constraint," and it becomes enriched when it is accompanied by auxiliary information in the form of "metadata regarding the provenance of the data records and the processing workflow that transforms them" (ibid., 406-407).

Boyd's conception of enriched evidence helpfully draws attention to the ways in which the usefulness of evidence (including evidence in the form of measurement results) depends on recording details about the process through which inquiry was conducted, details that in fact must be regarded as integral to the evidence itself. The question of how to distinguish relevant from irrelevant metadata leaves the concept of enriched evidence open to vagueness, however. We propose that a crucial step toward resolving this vagueness requires placing that question within the context of an account of inquiry that treats inquiries as aim-directed activities. Usefulness is not, after all, an intrinsic property of evidence but exists in relationship to some end or set of ends.

Other authors have incorporated pragmatic perspectives into our understanding of data and data models in ways that deserve inclusion in the epistemology of measurement. Alisa Bokulich and Wendy Parker defend what they call a *pragmatic representationalist* view of data. Data are, according to them, "records of the results of a process of inquiry that involves interacting with the world" (Bokulich and Parker, 2021, 6). What constitutes a process of inquiry is left unspecified, but a pragmatic treatment of this question coheres with their explicitly pragmatic commitment to the idea that data and data models are to be evaluated "in terms of their adequacy for particular purposes" (ibid., 2). On this view, the pragmatic dimension of data (and, implicitly, of measurement) emerges explicitly when considering how data are evaluated for use. Yet data are defined in terms of their production, the pragmatic character of which is left unspecified apart from their provenance in an untheorized process of inquiry.

Sabina Leonelli, in an extended series of studies of the production, transformation, and use of data in a variety of data-centric life science case studies, presents a view of data and data models in which data are "tools for communication" that are "defined in terms of their function within specific processes of inquiry" (Leonelli, 2016, 69). On Leonelli's view, neither the representational value of data nor how they are distinguished from data models is fixed independently of use, but depends on "ascriptions of evidential value" or what is "*usable as evidence*" (Leonelli, 2019, 16-17, emphasis in original). She contrasts her view with the formal approach taken by (Suppes, 1962), which neglects "*the protential of the objects produced by researchers in a given situation to function as data or data models in other situations of inquiry*" (Leonelli, 2019, 19, emphasis in original).

By treating the evidential usefulness of data – in "situations of inquiry" different from that in which they are produced – not merely as desirable features but as defining characteristics of data, Leonelli's account constitutes significant progress toward a philosophical account of measurement that is adequate for understanding how investigators solve the problem of usefulness in measurement. Indeed, Leonelli even adopts and builds upon a view about inquiry that is rooted in the pragmatist tradition in philosophy, which is that proposed by John Dewey (Brown, 2012; Dewey, 1938). According to Dewey, inquiry is "the controlled or directed transformation of an indeterminate situation into one that is so determinate in its constituent distinctions and relations as to convert the elements of the original situation into a unified whole" (Dewey, 1938, 104-105, emphasis in original).[1] Leonelli notes several significant ways in which Dewey's account of inquiry supports insights that her studies have highlighted, such as "by stressing how the value of any dataset, and indeed its very identification as 'data,' depends on the ways in which it is presented and related to other data and to the situation at hand" (Leonelli, 2016, 183).

[1] A full explication of this formulation, including Dewey's concept of a *situation* would digress from our purposes here. See (Brown, 2012) for a careful and informative discussion.

As much as Dewey's account of inquiry has informed our own thinking about inquiry (in science and in general), we propose a different model that we think better supports the project of putting scientific practices into a pragmatic perspective. For example, our purposes are not served by identifying a determinate set of functional phases of inquiry, as Dewey does (Brown, 2012, 280-297). Having different philosophical aims than Dewey, we propose a view of inquiry that is tailored to our purpose of making sense of the usefulness of specific experimental-scientific practices, which in the present case center on measurement in experimental physics. Thus, we turn next to the articulation of our own pragmatic model of inquiry. Our stance with respect to such models is pluralist. Our own model is meant to provide a framework for elucidating how scientific practices of inquiry contribute to the production of knowledge. Other accounts might carve up the conceptual space differently. To the extent that we present any argument in favor of our own model in this paper, it will take the form of demonstrating its capacity to elucidate the problem of usefulness in measurement.

4 A pragmatic model of inquiry

We approach experimental inquiries from a pragmatist perspective in which knowledge is a product of successfully executed processes of inquiry by a community of inquirers. The pragmatic significance of such knowledge rests on its forward-looking stability and suitability for use as a resource in future episodes of inquiry. The outcome of an inquiry must be sufficiently informative to enable epistemic ends not previously achievable, or to open up means of achieving such ends not previously available. We have previously used this approach to argue for the epistemological value of paying attention to the details of scientific practices in general, and to understand the epistemic value of the exploratory character of some experimental inquiries, in particular (Beauchemin and Staley, 2024). In this perspective, the problem of usefulness is the problem of how inquiry produces knowledge, i..e how an inquiry produces results that become

resources to be used in future inquiries. Answering this question requires an epistemological general modeling of the process of inquiry, suitable to account for scientific experimental practices. Articulating the epistemology of scientific knowledge in terms of the problem of usefulness illustrates the power of applying the pragmatic method to scientific inquiry, rendering a philosophical problem tractable.

To this end, we analyze experimental inquiries as a sequence of tasks, i.e. as actions carried out in order to accomplish some aims or objectives. These objectives include producing knowledge in the form of judgments that exhibit stability and suitability as resources for future inquiry, but are not limited to a propositional conception of knowledge. Objectives could also involve developing a technique, a set of skills, or a material resource to be used to address other scientific challenges than the ones immediately targeted by the inquiry. Aims can also be non-epistemic, such as simplifying a protocol, speeding up a process, etc.

Tasks can be analyzed at different levels of precision and scrutiny. Upper level general tasks include analyzing a data sample, manipulating an instrument, writing a report, arguing for an assumption, etc. At a lower, more fine-grained level of scrutiny, tasks could consist in estimating the signal gain of a photomultiplier tube, calibrating a thermocouple thermometer at the triple point of water, or estimating the systematic uncertainty on an estimate of the red-shift of a specific galaxy like M82. There is not a unique way to define, at any level, the tasks performed in an inquiry process; they can be identified and delineated differently by different epistemological investigators. What is essential is that the analysis is sufficient to account for the practice under study and enables support for the epistemological claims obtained from it.

Performing a task requires using resources. Resources include data, hardware, background knowledge, know-how, etc. Resources also can be epistemologically analyzed at different levels of resolution, offering different perspectives for making sense of

scientific practices. We could, for example, refer to a Python script written to continuously store data images taken from a CCD camera in a database as a "computational resource." This multilevel perspective on tasks and resources, but also on objectives pursued by these tasks, is a key element toward a solution to the usefulness problem: it allows mapping context-specific elements of an experiment to big-picture aspects of scientific inquiries, linking proximate to distant aims.

Tasks are performed within an inquiry, using various resources, in order to produce evidential claims constituting the main epistemic outcome of the experiment. We refer to this aspect of an inquiry as the *use mode of the inquiry*. For an inquiry's outcome to take the form of a "warrantably assertible" judgment (Dewey, 1938), that claim, along with the process that leads to it, needs to be critically assessed. That involves answering questions such as: Are the tasks adequate for their use? Are the tasks appropriate for the resources available and the aims sought? Are the aims achievable given the tasks and resources available? How compatible are the aims with one another?

Answering these questions is the goal of the *critical mode of the inquiry*. This mode of inquiry is also accomplished through tasks using resources aiming at some objectives. Debugging a piece of software, cross-checking a calibration, testing model assumptions used in a statistical analysis of the significance of a quantitative claim from a dataset, and estimating the systematic uncertainties on a correction factor are examples of tasks accomplished in the critical mode of an inquiry that require resources. Tasks, resources, and aims are all critically evaluated in relation to one another because criticism can result in revision of tasks, resources, and/or aims of both the use mode and the critical mode of inquiry. The use and critical modes are therefore not separate activities but are entangled, providing different ways of understanding inquiries. This intertwinement – using resources to perform tasks in order to achieve some goals, which are then claimed to be attained thanks to a critical assessment – forms an indissociable and irreducible triad, illustrated in Fig. 1.

Each of these elements affects the others. For example, the critical assessment of the inquiry process might lead to a revision of the tasks done and/or the resources used to get the sought experimental results, but they can also lead to a revision of the main objectives of that inquiry. This triad of feedback relationships offers a visual representation of one aspect of our model of inquiry briefly sketched here, but it will also prove helpful in understanding how this epistemological model can be used to provide a solution to the problem of usefulness. To illustrate this, we propose to use this analysis matrix to study an example of a complex measurement done in High Energy Physics, keeping in mind our main objective of addressing the problem of usefulness.

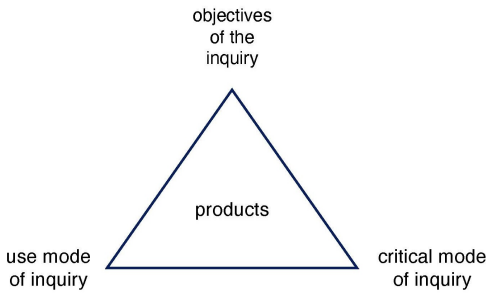

Fig. 1 Triad illustrating the feedback relationships between key aspects of a scientific inquiry. In the use mode, tasks are performed using resources in order to attain some objectives. This leads to a product or outcome of the inquiry. For this product to become a useful resource for future inquiry, it however needs to be critically assessed. This evaluation might lead to a change in the tasks performed or resources used in the inquiry, resulting in changes in the product of the inquiry. The critical assessment might also lead to a change of objectives, resulting in a change of products. These three aspects of an inquiry are related, feeding back into one another, and all contributing to the final outcome of the inquiry, the products that will become the resources readily available for new inquiries when the objectives are met.

5 The case of measuring the mass of the W boson

Measurement constitutes an important class of experimental inquiries in science. In general terms, measurement aims to learn what value(s) one may attribute to something treated as evaluable by a community of scientists and subject to normative constraints. The normative constraints on measurement might feature some

field-specific variants, but in all cases of the sort that concern us here[2] a successful measurement will rely on some measuring instrument and produce a result that includes a claim about values attributed to a targeted quantity called the measurand. There are epistemic, in addition to pragmatic reasons, for conducting a measurement process leading to such an evaluation of a measurand; something will be done with such a measured value. In a broadly Deweyan sense, a measurement hence leads to a change in the inquirers' relation to their environment and their ability to navigate (in a broad pragmatic sense) in that environment. Hence, measurements constitute inquiries. Measurements therefore lend themselves to an epistemological analysis using the pragmatist framework discussed in Sec. 4.

To illustrate this, consider the case of a measurement of the mass of the W boson (M_W) at the Large Hadron Collider (LHC). The W boson is one of the particles of the Standard Model (SM) of particle physics, and its mass is a parameter of the theory. According to the SM, that mass parameter shares very specific relationships to other parameters of the theory such as the Z-boson mass, the top-quark mass, and the Higgs-boson mass. Precise measurements of M_W (the measurand), in coordination with equivalently precise measurements of other parameters of the SM, offer powerful consistency tests of that theory (for example through global fits of the so-called electroweak parameters). Better yet, a quantification of a discrepancy between the measurement results of M_W and the expectations from the SM would provide crucial evidence on which to ground the theoretical development of potentially empirically successful Beyond the Standard Model (BSM) theories. There are therefore clear epistemic goals justifying the pursuit of M_W measurements, but there are other objectives that could be simultaneously pursued. A M_W measurement can be performed to provide a more precise calibration reference for other measurements, it could be used to demonstrate the usefulness of a new analysis procedure, to take advantage of improved "background knowledge," or to outperform the competition, to name a few. These

[2]Broadly speaking, this includes measurement in experimental physics.

examples illustrate the idea that aims and objectives can be classified in various ways, for example as overarching or contributory, primary or secondary, external or internal, immediate or long-term, proximate or distal, addressing type-I or type-II kinds of errors (Beauchemin, 2020), etc. A successful inquiry might not result in the satisfaction of all these objectives, but will need to satisfy the core objectives of the team of inquirers involved in the measurement process (including peer review of what the inquiry produces). The goals of a measurement might even be subject to revision in the course of an inquiry. Specification of the resources to be used (instruments, data, etc.), the tasks to be performed (calibration, statistical analysis, etc.), and the standards of critical assessment to be applied (resolution, consistency, robustness, etc.) during the inquiry will depend on all these aims. As a consequence, there will be many different ways in which a measurement could be performed.

In addition of the plurality of objectives, different inquirers have the option of using different resources to perform their measurement, even if they pursue the exact same objectives. The choice of resources could be dictated by the resources that are available to the inquirers, or by those that constitute a better fit to their expertise. Different choices will involve performing different tasks and different critical assessments of the tasks and resources. Any measurement of a quantity like the W-boson mass will therefore involve a complex sequence of operations and transformations that are specific to the inquiry being conducted, while aiming to produce a measurement result that attributes values to the measurand that transcend the idiosyncrasies of the context of the measurement.

For example, the procedures put in place by the ATLAS Collaboration to produce a measurement of the W-boson mass will involve different detectors, different background subtraction techniques, different detector effects and efficiency corrections, and even different strategies of measurement (different relationships to the measurand) than what their colleagues in the CMS Collaboration will do. It is even generally the

case that within the same group of scientists, such as the ATLAS team performing the M_W measurement, successive measurements of the same quantity will be conducted differently, making different assumptions, and choosing resources most adequate to their objectives.[3] Nonetheless, the quantitative values attributed to the measurand by each experimental inquiry must be such that the results can be compared and considered as measurements of the same quantity.

In addition to the theory-based transformations applied to the data to enable the measurement result to remain relevant beyond its context of production, another essential aspect that allows these apparently singular outcomes of different measurement inquiries to be directly compared is the estimate of the uncertainty that is provided for each measurement. Every operation performed to collect the measurement data, every experimental and instrumental condition in which such operations are conducted, every transformation applied to the resulting data involve variations, fluctuations, and dependence on assumptions that the estimate of uncertainty aims to capture. Such an estimate of uncertainty is an integral part of the results of the measurement process. The measurement result does not assign only one value, but an interval of values to the measurand, typically quoted in reports as $x \pm \Delta x$. For example, the latest W-boson mass measurement by the ATLAS Collaboration quoted a result of $M_W = (80360 \pm 16)MeV$. Estimating this uncertainty requires performing an in-depth critical assessment of the measurement process, including performing ancillary experiments to study the dependence of the results on particular effects. In this estimate, all the details of a specific instance of a measurement process matter. This is illustrated in Fig. 2 featuring the breakdown of the various contributions to the uncertainty on M_W.[4] In the language of the pragmatist framework outlined above,

[3] Beauchemin 2017 provides a detailed description of the complexity of the measurements performed at the LHC, and an analysis of the assumptions, resources, and type of modeling involved in each measurement.

[4] Each of the reported uncertainties corresponds to a combination of many more sources of uncertainty discussed in the report, gathered into larger categories to ease the presentation of the results. The structure of this table reflects a widespread practice in HEP of disinguishing *statistical* from *systematic* uncertainty. The validity and usefulness of this distinction is disputed, particularly within metrology. The GUM includes a recommendation that the term 'systematic uncertainty' be avoided because it can be "misleading," and proposes distinguishing uncertainties type A ("evaluated by statistical methods") and type B ("evaluated

the estimate of the uncertainty corresponds to performing the measurement in the critical mode of inquiry, to yield a spectrum of values that are representative of what was done in the measurement process. Without such an uncertainty estimate, it would be impossible to make any consistency claims between different experimental results, or to compare measurement results with theoretical expectations because numbers would almost always simply be different, with no means of evaluating the significance of those differences (Staley, 2020).

Unc. [MeV]	Total	Stat.	Syst.	PDF	A_i	Backg.	EW	e	μ	u_{T}	Lumi	Γ_W	PS
p_{T}^{ℓ}	16.2	11.1	11.8	4.9	3.5	1.7	5.6	5.9	5.4	0.9	1.1	0.1	1.5
m_{T}	24.4	11.4	21.6	11.7	4.7	4.1	4.9	6.7	6.0	11.4	2.5	0.2	7.0
Combined	15.9	9.8	12.5	5.7	3.7	2.0	5.4	6.0	5.4	2.3	1.3	0.1	2.3

Fig. 2 Table presenting the impact of the different uncertainty categories on the total uncertainty of the W-boson mass measurement as reported in (ATLAS, 2024).

There are significant epistemic advantages to developing a measurement process that increases the precision relative to previous measurements of the same measurand. For example, smaller uncertainties mean smaller intervals of values assigned to a measurand. These enable more acute tests of a theory and impose stricter constraints on developing theories capable of accommodating the data. Greater precision also helps identify measurements that are problematic and doubtful. Such statements are quantified by physicists involved in a given inquiry in terms of sensitivity to a target effect. Suppose, for example, that one seeks to test a specific BSM theory that predicts a shift of 25 MeV in the mass of the W boson compared to the SM prediction. A measurement result that would have an uncertainty of 50 MeV on the M_W parameter could include both the SM and the BSM predictions in the interval of values provided by the inquiry, and would therefore not have any sensitivity to the theoretical assumption predicting such a mass difference. Alternatively, a measurement result with a 5 MeV uncertainty would provide a means to test that theoretical assumption. This is why so much effort

by other means") (Joint Committee for Guides in Metrology Working Group I, 2008, ix). We do not attempt to resolve these debates here and our argument does not rest on how or if they are resolved.

is put into improving the precision of measurements. For example, after the publication of a very precise measurement of M_W in 2018 (Aaboud et al., 2018), the ATLAS Collaboration used the exact same data, but refined the statistical analysis of these data, the method to estimate one of the backgrounds (the multi-jet background), and new, more precise, theoretical calculations about one of the theoretical inputs to the measurement (the parton distribution functions used in simulation). These changes allowed ATLAS to publish a new more precise result for the measurement of M_W in 2024 (ATLAS, 2024). A reason for having spent half a decade re-analyzing the same dataset but changing part of the measurement process to improve precision is that the CDF Collaboration had done the same exercise and produced results that featured tensions with other measurements and the theory that needed to be clarified (Aaltonen et al., 2022). The comparison of the CDF and ATLAS M_W measurement results before and after their respective updates is presented in Fig. 3 below. We can clearly see how the increase in sensitivity of a measurement to a target effect is intimately related to the reduction of its estimated uncertainty, and how completely different epistemic pictures about a parameter like the mass of W boson emerge from such inquiries.

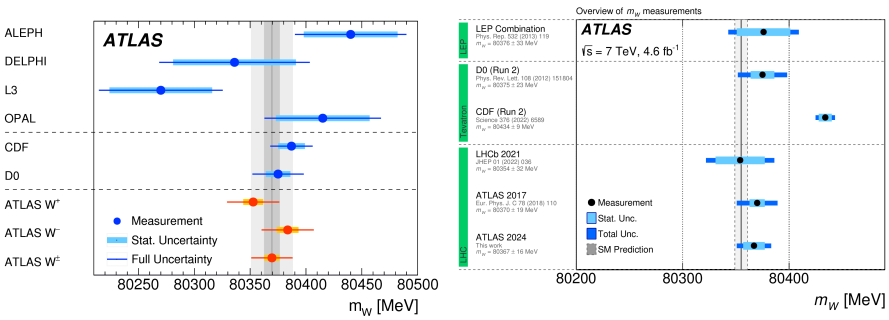

Fig. 3 Comparison between the M_W measurement results of different experiments and their comparison with the SM theoretical predictions before (left) (Aaboud et al., 2018) and after (right) (ATLAS, 2024) the reanalysis of the latest published data by the ATLAS and CDF collaborations respectively. Both experiments improved their precision, but also shifted the center of their quoted interval of values so that from a good agreement between the two, severe tensions have been featured.

The feedback relationships between the use of resources to yield a measurement of a given quantity (like the mass of the W-boson), the uncertainty estimated to quantify the dependence of the measured value on the idiosyncrasies of the measurement process, and the sensitivity resulting from that measurement to targeted effects can be illustrated by the triad measurement–uncertainty–sensitivity presented in Fig. 4. Such a triad offers a visual representation to facilitate understanding our treatment of the problem of usefulness in measurement. More importantly, this measurement–uncertainty–sensitivity triad maps on to the objective–use mode–critical mode triad of inquiry sketched in Sec. 4, helping to understand how our pragmatist framework can be used to provide a general solution to the problem of usefulness.

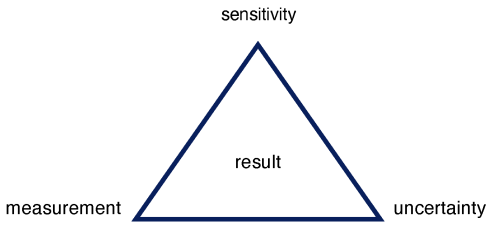

Fig. 4 Triad illustrating the feedback relationships between key aspects of a scientific measurement. During the measurement, tasks are performed using resources in order to attain some objectives. This process is critically assessed resulting in an uncertainty estimate. They both constitute the measurement result, but also specify how sensitive the measurement is to a target quantity. A lack of sensitivity would lead to a need to improve the measurement process, or the estimate of the uncertainty, or both. Failing that, a change in objectives, entailing a shift in sensitivity expectations, may be appropriate. Mutual adjustment of each element of the triad during the measurement yields an outcome reflecting that mutual adjustment.

6 Solving the problem of usefulness

To understand how investigators performing measurement activities solve the problem of usefulness, we start from three claims: (1) *Measurement* is inquiry directed at the aim (among others) of learning what value(s) one may attribute[5] to something treated

[5]We understand attribution here to be a communicative action. It is not merely assigning quantity values to a quantity, but producing an expression of such an assignment and making that expression part of a publicly available resource. Attribution is thus a necessary step in transforming the act of learning about something into a contribution to *scientific* inquiry (Joint Committee for Guides in Metrology Working Group II, 2012, 16).

as evaluable by a community of scientists and subject to normative constraints.[6] (2) Every measurement result includes, implicitly or explicitly, an estimation of *uncertainty*. (3) The uses of a result in subsequent inquiries are guided by its *sensitivity* to the targets of such inquiries. Achieving the central aim of a measurement inquiry (as specified in claim 1) requires decisions specifying resources to be used, tasks to be performed, and standards of critical assessment to be applied, in light of the aims of the inquiry. In this way, the features of the measurement process are chosen with reference to its capacity to produce a result with particular uncertainty characteristics (claim 2), enabling it to have sensitivity with respect to specific anticipated uses or objectives (claim 3). The triad in Fig. 4 can then be understood as a visual representation of this way of viewing measurement. We now proceed to defend these claims by showing how the resulting view of measurement illuminates the solution to the problem of usefulness.

Performing a measurement requires answering a number of questions. These include: What are we trying to measure? What means do we have access to for measuring it? Why do we want to measure it? What procedure can we use to achieve a result? How careful do we need to be for our result to be usable for our aims? How will we know whether our result meets those needs? The answers to these questions connect the measurement triad of Fig. 4 to the inquiry triad of Fig. 1.

For example, in the measurement of the W-boson mass, the ATLAS group produced a new measurement in 2024 using the same data they had used in their 2018 measurement. Why? The answer to the question 'What are we trying to measure?' does not explain why ATLAS produced the 2024 result. They had already measured

[6] We note longstanding disputes over the types of evaluations that should count as measurement. Some insist that measurands must satisfy such criteria as additivity and independence Campbell (1938) or some other criteria distinguishing *quantities* from other ordered attributes Michell (2008). Others, drawing upon the idea that measurement consists simply in "the assignment of numerals to objects or events according to rules" (Stevens, 1946, 677) permit a much broader scope to measurement. We adopt a non-committal stance here on these disputes. We seek not to define the scope of measurement in terms of what may be taken as its objects (what is measured), but rather to understand the relationships, in the pursuit of measurement, between what is measured, how it is measured, and why it is measured. We situate our present discussion within experimental physics, where such relationships are rendered most explicit, but we expect our model to be applicable to any type of measurement.

the same quantity in 2018. But answering questions about the means and procedures for measuring and the reasons for measuring do point to facts relevant to explaining why they produced a new result from the same data, and in ways that connect the measurement and inquiry triads. Let's start with means and procedures: The 2024 result used improved *resources* (relative to what was available in 2018) in the form of new parton distribution functions needed for the simulations that are integral to the method used for measuring M_W. The 2024 result also involved the use of a more advanced statistical method (the profile likelihood technique) for extracting an estimate that had not been used in 2018 and an improved method for estimating an important background. These new measurement *tasks* were performed in the service of arriving at a result that, when subjected to the critical mode assessment that produces an uncertainty evaluation, would yield a smaller uncertainty interval than had been achieved in 2018.

That brings us to the question "Why are we trying to measure it?" Achieving a reduction in uncertainty promoted multiple objectives answering to different dimensions of this question: A smaller uncertainty would constitute a display of physics prowess in the 2024 ATLAS group's competition with other collaborations (including the 2018 version of ATLAS itself, demonstrating an ability to innovate and improve). A smaller uncertainty would provide an opportunity to exploit the resources and techniques that made this result possible, such as those just described. A smaller uncertainty would increase the sensitivity of the measurement with respect to potential BSM theoretical predictions. Finally, independently of the improvement in precision, a new result based on improved measurement resources and tasks could be useful for understanding the previously mentioned tensions with theory and existing measurement results introduced by the CDF result of 2022. The fact that the ATLAS 2024 result came out to be even less consistent with the CDF result than the ATLAS 2018

result contributed to the evidence that the CDF result would be difficult to reconcile with data collected at the LHC, a question already of significant interest to HEP researchers (Amoroso et al., 2024).

We can see clearly in this example how objectives are important for making decisions about resources and tasks in measurement inquiries, but we also see how the resources and tasks that are accessible to investigators may motivate or facilitate the objectives that may be warranted to pursue. An account of inquiry that is suitable for the activity of measurement must therefore be capable of representing how the objectives of a measurement may embed the objective of learning what value(s) one may attribute to something treated as evaluable within a complex structure of other epistemic and non-epistemic objectives. It must also allow those objectives to be criticizable and revisable in relation to tasks and resources.

This brings us to our second claim, which states that measurement results include uncertainty outcomes. This is a consequence of the fact that measurement objectives do not determine the resources and tasks to be performed in a given measurement. Because resources, tasks, and objectives are related to one another through relations of interdependence and feedback, deciding how to conduct a given measurement constitutes a coordination problem involving choices at each vertex of the inquiry triad. Including uncertainties when reporting experimental results in physics is required in order to arrive at the objective of warranted measurement results in light of the choices that are made with regard to tasks and resources.

When ATLAS uses a new method for estimating the multi-jet background in the 2024 measurement of the W-boson mass, they commit in a practical sense to the adequacy of this method for the purpose of contributing to their measurement procedure Parker (2010). But this commitment is not a matter of faith. A critical mode assessment of that resource is a prerequisite for the warranting of the result that relies on that method, and that critical assessment reveals that the background estimate

has an uncertainty of its own. Consequently, a warranted measurement result must include that uncertainty. Similar comments apply to every other choice regarding the measurement procedure that is judged to have the potential to made a difference to the measurement result, and that can be subjected to a critical inquiry yielding an estimate of its contribution to the uncertainty. The uncertainty estimate, which is a quantitative summary of a complex inquiry into the consequences of variations in resources employed and tasks performed in generating a measurement result, provides an account of how that result depends upon resources in the completion of tasks that may behave in ways that vary from expectations, or that might justifiably have been chosen or performed differently.

This last point deserves emphasis: producing an uncertainty estimate constitutes on its own an *inquiry*, and as such may incorporate the results of other measurements. Indeed, this is typical of measurements in HEP and many other domains. Uncertainty estimates are consequently themselves the product of inquiries conducted in some context. Uncertainty estimation assesses potential variability or discrepancy from a variety of sources; useful measurements need to incorporate an adequate accounting of these. Rather than treating this as simply expressing a limitation on knowledge (we don't know the exact value of the measurand), this is a condition for producing scientific knowledge. Context-independence or decontextualization is always partial and subject to critical evaluation by a sufficiency standard.[7]

We turn finally to our third claim: The uses of a result in subsequent inquiries are guided by its sensitivity to the targets of such inquiries. A measurement result is useful for learning about possibilities in a given domain to the extent that it is sensitive to the differences among those possibilities. As illustrated in Section 5, a measurement result is sensitive to the difference between possibilities only if the uncertainty of the measurement result is smaller than the difference between predictions regarding the measured quantity based on the possibilities considered. The sensitivity is greater to

[7]We are grateful to an anonymous reviewer for encouraging us to emphasize this point.

the extent that the difference between quantity predictions exceed the uncertainty on the measured quantity.

This account emphasizes how the aims of an inquiry influence the results of that inquiry via the uncertainty estimate, enabling a judgment of the adequacy of the measurement process to its aims via its sensitivity. If the uncertainty is too large for the measurement to be sensitive to its target objectives, the process will be changed until the required sensitivity is achieved by reducing the uncertainty and possibly changing the central value. The objective, for ATLAS, of measuring M_W to better understand the differences between LHC and Tevatron results was not met with the 2018 result, because the uncertainty was too large. To achieve that objective, ATLAS developed a new method to estimate one of the largest sources of background, a more precise statistical analysis, and more precise theoretical calculations. In this way, uncertainty reporting is related strongly to objectives of the measurement that are assessed via sensitivity and call forth changes to the measurement process and the result reported.

Of course, some ways in which objectives might influence measurement results would be suspect. Suppose researchers simply *directly* reduced an uncertainty interval by 10% to achieve the objective of producing a smaller uncertainty than their competitors. Such an ad hoc adjustment would not survive a critical mode scrutiny. Indeed, such a procedure would not contribute to an *inquiry* into the measurement uncertainty at all.

Our treatment of measurement as inquiry treats sensitivity considerations as being concerned with the objectives of a measurement inquiry. But not all objectives of a given measurement will relate directly, or at all, to sensitivity. Some aims of an inquiry may be unrelated to sensitivity, such as training less-experienced physicists in a particular experimental technique. Others may only be related to sensitivity through a particular specification of an aim not explicitly stated in terms related to sensitivity. For example, the aim of "doing better than the competitors" could be reformulated as

"having tighter limits than the competitors," which could in turn be reformulated in terms of an increase in sensitivity to an expected BSM signal. Finally, some objectives in a measurement will intrinsically entail sensitivity considerations. If an objective of a measurement is to produce a result that can be used for testing between two alternative PDF models, then the result must be sensitive to the difference between them, meaning that it must have an uncertainty that is smaller than the difference between the values of the measurand predicted using those two models.[8]

The example of measuring the mass of the W boson illustrates the crucial role of *entheorization* of data in the production of useful measurement results. Entheorization refers to the process of transforming data with inputs from background knowledge, theory, simulation, and so on. Such a process should not be mistaken for the idealization of data or measurement processes, but is better understood as a kind of "dressing up" of data with theoretical resources to render them less idiosyncratic to their circumstances and more relevant for deployment in subsequent inquiry. For example, in measuring M_W, experimental data on kinematic variables in candidate W decay events are fitted to templates of simulated data derived from a range of theoretical hypotheses about the value of M_W. The latter form the basis for the simulation of processes of W-boson production and decay, which are in turn "passed through" a simulation of the ATLAS detector. In this and other ways, theory is built into the process of measuring M_W. Such entheorization is not, however, simply a matter of assuming theory and imposing it on the data, but is subject to its own process of critical scrutiny in terms of the resources serving as inputs and the ways those are used in the form of tasks performed with and on them. Particular entheorizations thus receive critical scrutiny that includes uncertainty estimation, which may serve to prevent biasing from circular forms of theory-dependence (Ritson and Staley, 2021). Such critical scrutiny is conducted with reference to the full range of aims of the measurement inquiry to which such entheorization contributes.

[8]We thank an anonymous reviewer for drawing our attention to the need to clarify this point.

The crucial role of uncertainty estimation for solving the problem of usefulness should now be clear. By assessing the variability in measurement reports resulting from choices made regarding measurement tasks and resources, investigators can address questions crucial for establishing the contexts and purposes for which a measurement result can be useful. In the case of a resource that has been taken from inquiries previously conducted (such as a parton distribution function), these questions include: "What is the impact on the current inquiry due to the quoted uncertainty on the resource used?" and "Are the epistemic standards and risk attitudes in the production of that resource acceptable for the objectives of the present measurement inquiry?" In the case of choices and tasks that are internal to the measurement inquiry presently undertaken (such as relying on a new method for estimating the mult-jet background), such questions include: "Does the result reported adequately account for the assumptions and choices relied upon in executing those choices?" Carrying out assessments such as these and documenting them is a requirement of the responsible conduct of inquiry and essential for producing useful measurement results.[9]

7 Conclusion

We have adopted a pragmatic model of inquiry to understand how measurement conducted in specific ways and in specific conditions can be useful for the purposes of inquiries conducted in different ways in and in different conditions. We treat measurement, which involves using resources in the execution of tasks, as inseparable from uncertainty estimation, which is a critical mode activity of inquiry aimed at accounting for potential variance due to choices of measurement resources and tasks, and as inseperable from sensitivity, a feature of measurement results directly related to the satisfaction of measurement objectives and dependent upon the results of uncertainty

[9] In emphasizing the importance of documentation of this sort our account resembles Boyd's as discussed in Section 3 (Boyd, 2018).

estimation. The usefulness of a measurement result is achieved through the successful coordination of tasks, resources, and objectives, as achieved under the scrutiny of a critical assessment. That critical assessment is the source of a crucial feature of a fully successful measurement result: it must be expressed by an uncertainty interval representing the sensitivity of the output of the measurement procedure to the choices made in the coordination process that led to the use of that procedure.

We have reached these conclusions as the outcome of applying the pragmatic method to measurement practices, guided by the framework of our pragmatic model of inquiry. We do not claim that our approach is the only way to arrive at these concusions, but we have shown that an understanding of the epistemological significance of considerations of uncertainty and sensitivity follows naturally from our approach.[10]

A caveat deserves emphasis, however. Our first claim is that measurement aims to *learn what may be attributed to the measurand*. This passive construction risks obscuring the active role of the investigator doing the attributing. Learning that *we may attribute* to the quantity "mass of the W boson" the interval $80360 \pm 16MeV$ is not the same thing as learning that the value of the quantity "mass of the W boson" lies in the interval $80360 \pm 16MeV$. The latter claim would constitute an inference from the former. The attribution that is expressed in terms of an uncertainty interval summarizes the investigator's understanding of what was done in the measuring procedure and what is warranted to report with regard to the achievement of the measurement's objectives.

The usefulness of a measurement result is a product of the care that the investigator takes in accounting for the conditions in which and the manner in which the measurement result was produced, as well as the process of entheorization that expands the range of contexts in which a result may be used. The user of that result, however, bears their own responsibility to decide whether their own context and objectives permit them to use that result for the particular inquiry the user intends to pursue. We

[10] We thank an anonymous reviewer for encouraging us to clarify our position.

can therefore see the Janus-faced nature of the problem of usefulness reflected in the responsibility for solving that problem. That responsibility lies with both those who measure and those who use measurement results.

Acknowledgments

We are grateful to Richard Dawid and Harald Wiltsche for organizing the "Philosophy of Experiment" conference in Stockholm in November 2023, thus providing us the opportunity to begin to formulate the ideas of this paper. Conversations with the audience in Stockholm helped us significantly in the formulation of our arguments. We are particularly grateful to two anonymous reviewers for this journal whose careful consideration of an earlier draft helped guide us to important improvements in this paper.

Declarations

- Funding: N/A. No external funding was relied on in this research.
- Conflict of Interest: N/A. The authors have no conflicts of interest.
- Ethical approval: N/A. No human subjects were used.
- Informed consent: N/A. No human subjects were used.
- Author contribution: Pierre-Hugues Beauchemin: Conceptualization, Investigation, Writing – Original Draft, Writing – Review and Editing; Kent Staley: Conceptualization, Investigation, Writing – Original Draft, Writing – Review and Editing
- Data Availability Statement: N/A.

References

Aaboud, M. et al. 2018. Measurement of the W-boson mass in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Eur. Phys. J. C 78*(2): 110. https://doi.org/10.

1140/epjc/s10052-017-5475-4. arXiv:1701.07240 [hep-ex].

Aaltonen, T. et al. 2022. High-precision measurement of the W boson mass with the CDF II detector. *Science 376*(6589): 170–176. https://doi.org/10.1126/science.abk1781 .

Amoroso, S. et al. 2024. Compatibility and combination of world W-boson mass measurements. *The European Physical Journal C* 84: 451. arXiv:2308.09417 [hep-ex].

ATLAS. 2024. Measurement of the W-boson mass and width with the ATLAS detector using proton-proton collisions at $\sqrt{s} = 7$ tev. *Accepted by Eur. Phys. J. C*. arXiv:2403.15085 [hep-ex].

Beauchemin, P.H. 2017. Autopsy of measurements with the ATLAS detector at the LHC. *Synthese* 194: 275–312. https://doi.org/10.1007/s11229-015-0944-5 .

Beauchemin, P.H. 2020. Signature-based model-independent searches at the Large Hadron Collider: An experimental strategy aiming at safeness in a theory-dependent way. *Philosophy of Science 87*(5): 1234–1245. https://doi.org/10.1086/710515 .

Beauchemin, P.H. and K. Staley. 2024. The epistemological significance of exploratory experimentation: A pragmatist model of how practices matter philosophically. *European Journal for Philosophy of Science* 14: 59 .

Bogen, J. and J. Woodward. 1988. Saving the phenomena. *Philosophical Review 97*(3): 303–352 .

Bokulich, A. and W. Parker. 2021. Data models, representation and adequacy-for-purpose. *Eur J Philos Sci.*: 11(1)–31. https://doi.org/0.1007/s13194-020-00345-2 .

Boyd, N. 2018, July. Evidence enriched. *Philosophy of Science* 85: 403–421 .

Brown, M.J. 2012. John Dewey's logic of science. *HOPOS: The Journal of the International Society for the History of Philosophy of Science 2*(2): 258–306 .

Campbell, N.R. 1938. Symposium: Measurement and its importance for philosophy. I. *Proceedings of the Aristotelian Society, Supplementary Volumes* 17: 121–142 .

Collins, H. 1985. *Changing Order: Replication and Induction in Scientific Practice.* Beverly Hills and London: Sage.

Collins, H. 1994. A strong confirmation of the experimenters' regress. *Studies in History and Philosophy of Modern Physics* 25(3): 493–503 .

Dewey, J. 1938. *Logic: The Theory of Inquiry.* New York: Henry Holt and Company.

Franklin, A. 1994. How to avoid the experimenters' regress. *Studies in the History and Philosophy of Science* 97–121: 25 .

Franklin, A. 2002. *Selectivity and Discord: Two Problems of Experiment.* Pittsburgh, PA: University of Pittsburgh Press.

Hacking, I. 1981. Do we see through a microscope? *Pacific Philosophical Quarterly* 62: 305–322 .

Hacking, I. 1983. *Representing and Intervening.* Cambridge: Cambridge University Press.

Hanson, N. 1958. *Patterns of Discovery: An Inquiry into the Conceptual Foundations of Science.* Cambridge University Press.

Joint Committee for Guides in Metrology Working Group I. 2008. *Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement.* Joint

Committee for Guides in Metrology.

Joint Committee for Guides in Metrology Working Group II. 2012. *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*. Joint Committee for Guides in Metrology.

Leonelli, S. 2016. *Data-centric biology: A philosophical study*. Chicago: University of Chicago Press.

Leonelli, S. 2019. What distinguishes data from models? *European Journal for Philosophy of Science 9*(2): 22. https://doi.org/10.1007/s13194-018-0246-0 .

Michell, J. 2008. Is psychometrics pathological science? *Measurement 6*(1-2): 7–24 .

Parker, W.S. 2010. Scientific models and adequacy-for-purpose. *The Modern Schoolman* 87: 285–293 .

Ritson, S. and K. Staley. 2021. How uncertainty can save measurement from circularity and holism. *Studies in History and Philosophy of Science* 85: 155–165 .

Staley, K.W. 2012. Strategies for securing evidence through model criticism. *European Journal for Philosophy of Science* 2: 21–43 .

Staley, K.W. 2020. Securing the empirical value of measurement results. *British Journal for the Philosophy of Science 71*(1): 87–113. https://doi.org/10.1093/bjps/axx036 .

Stevens, S.S. 1946. On the theory of scales of measurement. *Science 103*(2684): 677–680 .

Suppes, P. 1962. Models of data, In *Logic, Methodology and Philosophy of Science Proceedings of the 1960 International Congress*, eds. Nagel, E., P. Suppes, and A. Tarski.

Tal, E. 2017. A model-based epistemology of measurement, In *Reasoning in Measurement*, eds. Mößner, N. and A. Nordmann, 233–253. New York: Routledge.

Tal, E. 2020. Measurement in Science, In *The Stanford Encyclopedia of Philosophy* (Fall 2020 ed.)., ed. Zalta, E.N. Metaphysics Research Lab, Stanford University.